The *BCAR1* Locus in Carotid Intima-Media Thickness and Atherosclerosis

Freya Boardman-Pretty

University College London

Thesis submitted for the degree of
Doctor of Philosophy

Declaration of work

I, Freya Boardman-Pretty, confirm that the work presented in this thesis is my own and has been generated by me as the result of my own original research. Where information has been derived from other sources, this has been indicated in the text. My role in the work presented in each chapter is outlined below.

In **chapter 3** I carried out all bioinformatics analysis, including identification of SNPs in strong LD, analysis of regulatory marks associated with these SNPs using UCSC Genome Browser, HaploReg and ElDorado, and selection of candidate SNPs. I used the Gene-Tissue Expression (GTEx) browser and Gilad/Pritchard eQTL browser to investigate eQTLs.

In **chapter 4** I was responsible for genotyping the PLIC cohort for the SNP rs4888378. Genotype data and an analysis plan were sent to Danilo Norata and Andrea Baragetti (University of Milan), who carried out the statistical tests that I requested in PLIC and provided test results and the necessary data for the meta-analysis. I carried out the sex-stratified meta-analysis and additional event rate analysis.

In **chapter 5** I carried out electrophoretic mobility shift assays (EMSAs) on candidate SNPs, multiplex competitor EMSAs and supershift EMSAs. For the luciferase assay, I designed reporter fragments of interest and carried out the cloning experiments to create the four sets of reporter vectors, and carried out the luciferase assays. I also carried out transfection experiments.

In **chapter 6** I carried out all experiments involved in design of the 4C protocol. My understanding and design of the protocol was developed with the advice of contacts carrying out similar experiments: Johanna Fischer (Crick Institute), Marjon Verstegen (Hubrecht Institute, Utrecht) and Michal Mokry (UMC Utrecht).

In **chapter 7** I used exome sequencing data to identify the SNP rs1035539 for study, and carried out genotyping in the IMPROVE and PLIC cohorts. Genotyping data and analysis plan were sent to Danilo Norata and Andrea Baragetti, who carried out the analysis I requested and provided test results. I carried out statistical analysis in IMPROVE.

In **chapter 8**, the pEGFP-C2/BCAR1 and pENTR3C-BCAR1 plasmids had previously been created by the Cardiovascular Biology and Medicine group, and were provided to me for further use. I carried

out side-directed mutagenesis, cloning, virus production and assays on transfected and infected cells, with help and guidance from Ian Evans.

Acknowledgments

Many people gave invaluable help and advice throughout my PhD. I would first like to thank my supervisor Steve Humphries, who from the moment of offering me the PhD position to helping me with the final details of my thesis, has been hugely supportive. His guidance and encouragement have helped me to understand and enjoy new areas of genetics, and motivated me to persevere when things were less easy. Equal thanks go to my secondary supervisor Andrew Smith, whose expertise and knowledge have made a great difference to my PhD, helping me to shape my research and construct much of my thesis. He has always been willing to help me with long and difficult experiments, or to know when I needed to go and complain about them at the pub.

Thank you also to Philippa Talmud who has been there for support many times throughout my PhD, and to Ann Walker for the encouraging words during long writing sessions. I am also greatly appreciative towards Jackie Cooper and Fotios Drenos, who have given me a great deal of advice and insight into statistical methods, and were always happy to answer my questions.

Thank you to all past and present members of CVG who have made the lab such an enjoyable environment to work in; for being a welcoming and friendly group and for joining in with the bake-offs, plank-offs, cocktail parties, away days and more that make the group such a unique place to be. Thank you to Katherine for always being there to talk about topics both work-related and more inconsequential, Marta for being a constant friend and source of advice, and to Dauda for being my go-to for a chat or for advice on all things technology-related. I am also very grateful to Jutta for all the lab expertise she has shared with me over the years, and to Jay and Jacquie for their help and support. I would like to thank the British Heart Foundation for providing valuable funding to the group to support our research.

I would also like to thank the members of the Cardiovascular Biology and Medicine Group, particularly Ian Evans, who has taught me many new methods and has always been on hand to answer my many questions. His tireless support and assistance with both lab work and writing have helped shape a large portion of my thesis.

I am very grateful to my collaborators Andrea Baragetti and Danilo Norata, who worked with me on analysis of the PLIC cohort; big thanks to Andrea who was always willing to run new analyses for me and answer all my questions. Thank you too to Karl Gertow and the Karolinska group, who have been excellent collaborators and without whom the basis of this thesis would be very different.

I would also like to thank the organisers of the IMPROVE, WHII, EAS and MDC cohorts used in this thesis, and the individuals who participated in these studies.

Finally, thank you to my family and friends, who have given me a great deal of support throughout the last few years. I appreciate the patience and encouragement they have shown me and without their help I'm sure I would not be where I am today.

Abbreviations

AMD: age-related macular degeneration

ANOVA: analysis of variance

ASAP: Advanced Study of Aortic Pathology

BCAR1: breast cancer antiestrogen resistance 1

BiKE: Biobank of Karolinska Endarterectomies

BMI: body mass index

bp: base pairs

BSA: bovine serum albumin

CAD: coronary artery disease

CCA-IMT: common-carotid intima-media thickness

CFDP1: craniofacial development protein 1

CHARGE: Cohorts for Heart and Aging Research in Genomic Epidemiology

CHD: coronary heart disease

ChIP: chromatin immunoprecipitation

ChIP-seq: chromatin immunoprecipitation sequencing

CHST6: carbohydrate (N-acetylglucosamine 6-O) sulfotransferase 6

CIP: calf intestinal alkaline phosphatase

COS-7: CV-1 (simian) in origin, carrying SV40

CRISPR: clustered regularly-interspaced short palindromic repeats

CRP: C-reactive protein

CTRB1: chymotrypsinogen B1

CTRB2: chymotrypsinogen B2

DMEM: Dulbecco's modified eagle medium

DMSO: dimethyl sulphoxide

dNTPs: deoxynucleoside triphosphates

EAS: Edinburgh Artery Study

EGF: epidermal growth factor

EMEM: Eagle's minimal essential medium

ENCODE: Encyclopaedia of DNA Elements

ER: oestrogen receptor

EVS: exome variant server

FBS: foetal bovine serum

FDR: false discovery rate

FH: familial hypercholesterolaemia

GAPDH: Glyceraldehyde-3-Phosphate Dehydrogenase

GFP: green fluorescent protein

GTEx browser: Gene-Tissue Expression browser

GWAS: genome-wide association study

HDL: high density lipoprotein

HEK293: human embryonic kidney 293

HR: hazard ratio

HR: homologous recombination

HUVEC: human umbilical vein endothelial cell

IBC: ITMAT-Broad_CARe

iCOGS: iSelect Collaborative Oncological Gene-environment Study

IMPROVE: European Carotid IMT and IMT-Progression as Predictors of Vascular Events

IMT: intima-media thickness

KDR: kinase insert domain receptor (aka vascular endothelial growth factor receptor 2)

LD: linkage disequilibrium

LDHD: lactate dehydrogenase D

LDL: low density lipoprotein

MAF: minor allele frequency

MDC: Malmö Diet and Cancer Study

MDS: multi-dimensional scaling

NHEJ: non-homologous end joining

NRP1: neuropilin-1

nt: nucleotides

NTC: no-template control

OR: odds ratio

PBS: phosphate-buffered saline

PBST: phosphate-buffered saline with 0.1% Tween 20

PDGF: platelet-derived growth factor

PLIC: Progressione della Lesione Intimale Carotidea

SDS: sodium dodecyl sulphate

SDS: Sodium dodecyl sulphate

SH3: src homology 3

siRNA: short interfering RNA

SNAP: SNP Annotation and Proxy Search

SOC: super optimal broth with catabolite repression

TBE: Tris/Borate/EDTA

TE: tris-EDTA

TEMED: Tetramethylethylenediamine

TF: transcription factor

TMEM170A: transmembrane protein 170A

VEGF: vascular endothelial growth factor

VSMC: vascular smooth muscle cell

WHII: Whitehall II

ZFP1: zinc finger protein 1

ZNRF1: zinc and ring finger 1

Abstract

Carotid intima media thickness (IMT) is a marker of subclinical atherosclerosis that can predict cardiovascular events over traditional risk factors. This thesis focused on a chromosome 16 locus associated with IMT and coronary artery disease, in order to identify functional variation affecting protein structure or gene expression, and investigate the role of novel genes in atherosclerosis.

Bioinformatics tools were used to identify variants in strong LD with the lead SNP and to filter these to a shortlist of potential regulatory variants. Electrophoretic mobility shift assays on these 6 SNPs detected allele-specific protein binding to the lead SNP rs4888378 in *CFDP1,* and implicated the protein FOXA. Luciferase reporter assays showed a 35-92% decrease in gene expression with the A allele. Expression-QTL analysis confirmed associations of the protective allele of rs4888378 with higher expression of *BCAR1* in vascular tissues. Genotyping and analysis of the lead SNP in the PLIC cohort of 2144 healthy men and women suggested a sex-specific effect of the SNP on IMT progression. Meta-analysis of five cohort studies supported a protective effect of the A allele on common-carotid IMT in women only.

As illustrated by this locus, functional variation often lies in enhancers far from its target promoter. Circular chromosome conformation capture (4C) was explored to investigate interactions between enhancers and promoters at this locus, with the aim of capturing regions interacting with rs4888378 and the *BCAR1* promoter.

Analysis of exome sequencing data identified a SNP in *BCAR1* (coding amino acid change P76S), which was genotyped in two cohorts (IMPROVE and PLIC). Associations with lower plaque and IMT were found (wild-type proline form), but differed between the cohorts and require further validation. Wild-type and variant vectors were expressed in cells to assess the variant's effect on protein function and pathways involved in plaque, showing cells with wild-type protein to migrate more quickly.

These analyses on regulatory and protein-coding mechanisms implicate BCAR1 in atherosclerosis and provide new pathways for analysis in the understanding of cardiovascular disease.

Table of Contents

List of figures

List of tables

1 Introduction

1.1 Overview

It was said by Aldous Huxley that "the more we know, the more fantastic the world becomes and the profounder the surrounding darkness"[1]. Ninety years later, this statement still applies to many areas of science, with the genetics of complex disease no exception. Advances in genetics have uncovered novel genetic factors behind many conditions, including cardiovascular disease, a worldwide burden, but these results often raise more questions. Diseases linked to genes with unknown function, or areas of the genome previously thought to be non-functional, indicate roles for novel genes and pathways in disease, and for previously uncharacterised genetic regions in regulating gene expression. There is much yet to discover about the complexity of the human genome, and this thesis will attempt to explore a part of it in relation to cardiovascular disease.

This introduction will cover the background of cardiovascular disease, including its development, the burden on population health and strategies for prevention and treatment. Risk factors will be discussed with a focus on new genetic risk factors and a genetic locus associated with carotid intima-media thickness. BCAR1 and related proteins will also be described, which are involved in molecular processes that may be involved in the development of plaque, followed by the overall hypothesis and thesis aims.

1.2 Cardiovascular disease

Cardiovascular disease (CVD) is a general term for diseases of the heart and vascular system that involve narrowing or blocking of blood vessels; it comprises coronary heart disease (CHD), stroke, peripheral arterial disease and aortic disease. CVD is the most common cause of death in the developed world[2] and globally[3], having been responsible for 17.5 million deaths worldwide in 2012[3] (Figure 1). Incidence is rising worldwide as populations grow in size, age, and undergo epidemiological changes that influence cardiovascular risk[4].

Many environmental and behavioural factors for CVD are known to increase risk of the disease, with the most significant being poor diet, physical inactivity and smoking[3], and addressing these risk factors can reduce incidence. Cardiovascular disease is also known to have a significant genetic component, and many genetic factors conferring increased risk have been identified[5]. However, approximately 90% of the heritable factors of CVD remain unknown[6]. Identifying factors that contribute to CVD has aided the development of new treatments and prevention strategies[7], but to

further reduce the burden it will be necessary to identify additional factors that are involved in CVD risk. By understanding more of the molecular basis of the disease methods can be developed to reduce incidence. Part of this will involve discovering the missing heritability of CVD.

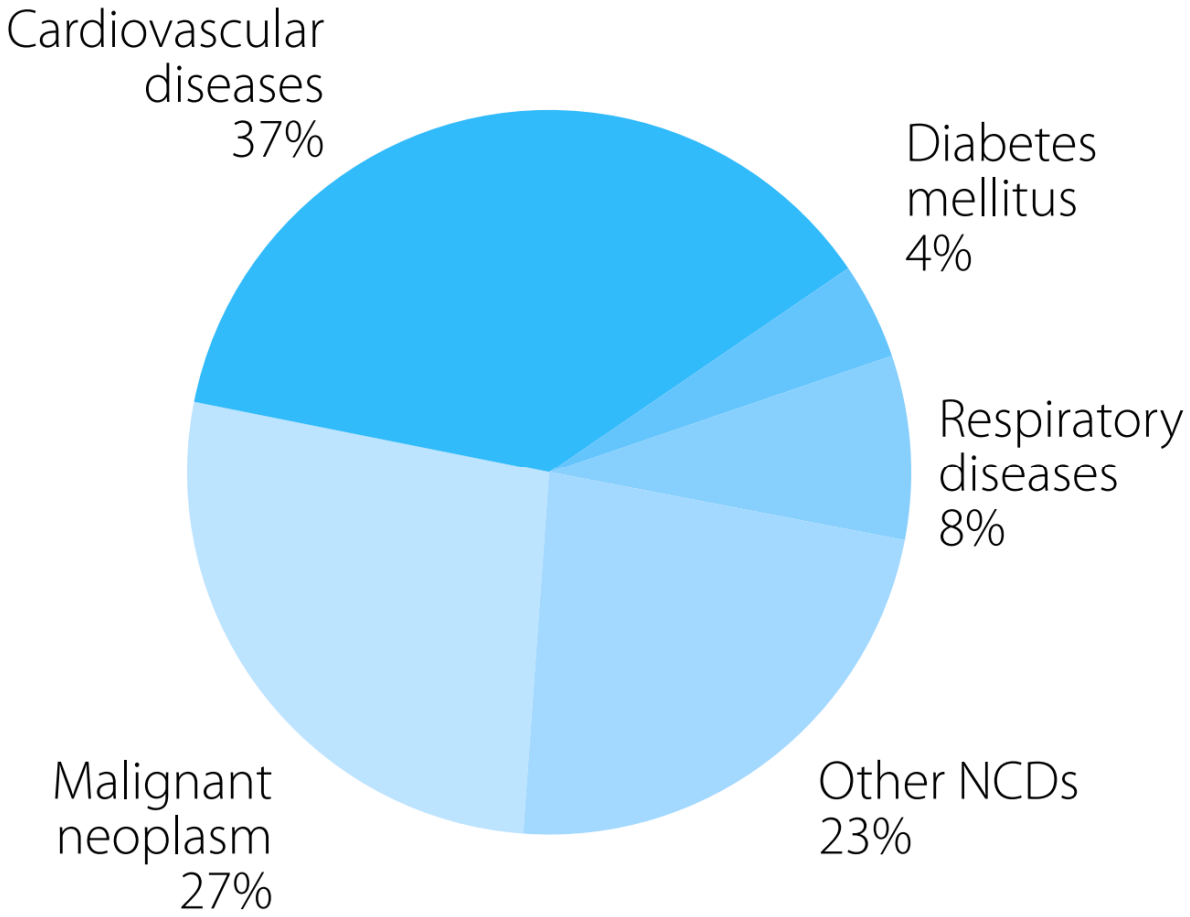

Figure 1: Proportion of global deaths caused by non-communicable disease, by cause of death, 2012. Figure from World Health Organisation (2014).

1.2.1 Coronary artery disease and atherosclerosis

Coronary artery disease (CAD), or coronary heart disease (CHD), is the most common subtype of cardiovascular disease, caused by the narrowing and hardening of the arteries supplying the heart (atherosclerosis). Atherosclerosis is caused by deposition of cholesterol and fats in the artery walls. As these deposits build up, blood flow through the arteries is restricted, with two main outcomes. Narrowing of the coronary arteries reduces the blood flow to the heart; here reduced oxygen supply forces cardiac muscle to respire anaerobically and lactic acid to build up, causing pain in the chest (angina). If atherosclerotic plaque ruptures, plaque matrix and foam cells are released into the plasma, and initiate thrombus formation on contact with platelets. If extensive enough to fully block the coronary artery, this leads to myocardial infarction, stopping blood flow to the cardiac muscle[8].

Over 40% of cardiovascular disease deaths were caused by CAD in 2012, with the second-largest cause being stroke[3].

Rupture of clots in other arteries also cause vascular disease: clots in the blood vessels supplying the brain block oxygen supply to the brain, causing stroke[8], and blockage in the blood supply to the limbs causes peripheral arterial disease and manifests as pain in the extremities[9].

1.2.1.1 Pathology of atherosclerosis

The structure of the artery walls comprises three discrete layers (Figure 2). The inner layer, the tunica intima, comprises a monolayer of endothelial cells, and a membrane of collagen and proteoglycans which bind it to the media layer. The media consists of smooth muscle cells (SMCs), while the outer adventitia consists of connective tissue, fibroblasts and smooth muscle cells[8].

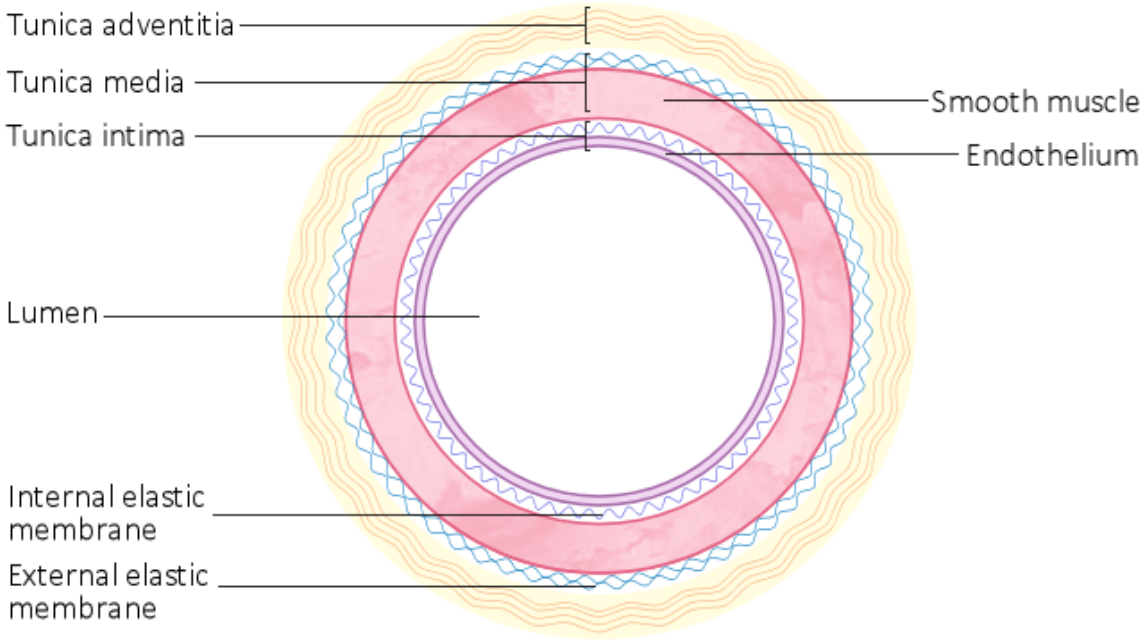

Figure 2: Structure of the artery wall. The artery wall consists of three layers: the intima, media and adventitia. The tunica intima consists of endothelial cells and the internal elastic membrane. The tunica media contains smooth muscle cells and connective matrix, and is connected to the adventitia by the external elastic membrane. The tunica adventitia consists of connective tissue, fibroblasts and smooth muscle cells.

Atherosclerosis is instigated with the movement of low-density lipoprotein (LDL) cholesterol from the plasma into the blood vessel wall; high circulating LDL levels promote this process[10]. The apolipoprotein B-100 (apoB-100) protein that is bound to LDL particles binds to proteoglycans in the extracellular matrix, trapping LDL particles in the intima[11]. Here reactive oxygen species cause the oxidation of LDL to oxidised LDL[10]. At the intima of the plaque site endothelial activation can occur,

in which endothelial cells enter an inflammatory state, activating a defence response[12]. In addition to high circulating LDL, a number of factors such as hypertension[13], free radicals increased by tobacco smoking[14] and diabetes[15] can cause endothelial activation. This encourages monocytes to adhere to the surface of the endothelium, then migrate into the intima, where they differentiate into macrophages[8]. These bind to epitopes on the oxidised LDL through scavenger receptors such as CD36, and subsequently internalise the particles[16]. The macrophages are unable to break down the internalised oxidised LDL, so it gradually accumulates. This causes the transformation of the macrophages into larger lipid-filled foam cells, which are too large to cross back into the lumen; this begins to form a fatty streak in the vessel wall. Foam cells that die release their lipid contents, which can combine to form a lipid-rich necrotic core.

Migration of vascular smooth muscle cells (VSMCs) into the fatty streak causes it to increase in size, which may be stimulated by the inflammatory response in the lesion[17], such as platelet-derived growth factor (PDGF) production by activated endothelial cells[18]. VSMCs begin to proliferate and form fibrous tissues, such that a fibrous cap forms over the lipid core[19]. Plaques that have a large necrotic core tend to contain more inflammatory cells and have a thinner fibrous cap, and are therefore regarded as unstable plaques. Stable plaques have smaller lipid cores and a thicker fibrous cap, making them less likely to rupture[20].

The physical growth of the plaque is first managed by remodelling of the vessel. The overall circumference of the vessel can initially enlarge to compensate for the increased intima thickness, allowing blood flow to continue unimpeded (Figure 3b). However, growth of the plaque beyond a certain size induces constrictive remodelling (Figure 3c), with the plaque encroaching upon the lumen and reducing the area available for blood flow[21].

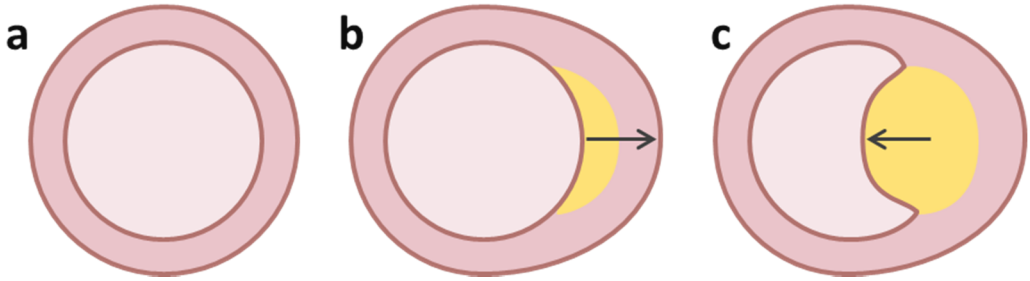

Figure 3: Atherosclerotic plaque growth and artery remodelling. (a) The artery in its healthy state has adequate space for blood flow. **(b)** The formation and growth of atherosclerotic plaque causes enlargement of the artery, at first without significantly reducing lumen size. **(c)** The plaque is sufficiently large to intrude into the lumen, restricting blood flow.

1.2.2 Risk factors for cardiovascular disease

Many variables have been identified which influence the likelihood of developing CVD. Some of these risk factors are unmodifiable; age, for example, increases risk and is the strongest predictor for cardiovascular disease[22]. Sex is another strong unmodifiable risk factor, with lifetime risk of incidence and mortality higher for men than women[23,24]. While men have a significantly higher risk of coronary heart disease, the risk of certain CVD subtypes such as stroke shows less of a difference between sexes[25]. Other strong non-modifiable risk factors are ethnicity (even after controlling for other risk factors)[26,27] and family history[28], indicating the significance of genetic factors in cardiovascular risk. The heritability of cardiovascular disease has indeed been found to be substantial, with premature atherosclerosis in a parent conferring a 3-fold increase in CVD risk[29].

While non-modifiable risk factors play a significant role in risk, behavioural risk factors are responsible for approximately 80% of CHD and cerebrovascular disease[30]. Cigarette smoking is a major behavioural cardiovascular risk factor, doubling the risk of stroke and CHD and increasing the risk of peripheral arterial disease and aortic aneurysm by over 300%[25,31]. Smoking promotes CVD through several mechanisms: it increases circulating LDL levels and makes the particles themselves more susceptible to oxidation, creating atherogenic oxLDL[32], increases the risk of clotting through elevating levels of clotting factors such as fibrinogen[33], and nicotine increases blood pressure, increasing the chance of endothelial damage and plaque rupture[34]. Physical activity is associated with lower risk of cardiovascular disease and mortality, through various processes, including weight regulation, insulin control, decreases in blood pressure and endothelial function[35].

Alcohol consumption is another behavioural risk factor, but an unusual one, in that many studies have shown a U- or J-shaped curve of association between alcohol and CVD risk, with moderate consumption of alcohol being atheroprotective[36,37]. This has been suggested to be a result of raised high-density lipoprotein (HDL) cholesterol (proposed to be atheroprotective[38]), lower levels of inflammatory markers and reduced aggregation of platelets[39]. However, a Mendelian randomisation study using genetic variants that predispose to lower alcohol consumption suggested that lower alcohol consumption always decreased risk of CHD, suggesting the suggested protective effect of alcohol may be due to confounding or selection bias[40].

One of the major modifiable risk factors is hypertension, which is highly prevalent, with 22% of people being hypertensive globally in 2014[2]. Hypertension is defined by the WHO World Health Organisation (WHO) as systolic blood pressure above 140 mmHg or diastolic blood pressure above

90 mmHg[41]. Hypertension increases the risk of cardiovascular events, by putting arteries under increased pressure, increasing chance of injury and promoting atherosclerosis[42]. It is responsible for at least 45% of deaths due to heart disease[43]. High blood pressure can be reduced with moderation of diet (particularly reducing salt intake) and increasing physical activity. A reduction in blood pressure of 10 mmHg is associated with a 41-46% reduction in cardiometabolic mortality[44]. Blood pressure-lowering medication such as ACE inhibitors have been shown to reduce cardiovascular events[45].

Obesity (defined as body mass index (BMI) > 30 kg/m^2) and being overweight (BMI > 25 kg/m^2) raise the risk of CVD[46], and rising obesity levels have increased levels of CVD, increasing the burden on health services and causing productivity losses[47,48]. Obesity increases cardiovascular risk when controlling for other risk factors, but also through increasing risk of hypertension[49] and type 2 diabetes[50]. Diabetes, characterised by hyperglycaemia and glucose intolerance, doubles the risk of cardiovascular disease[51] through hyperglycaemic effects on the vascular tissue[52].

Blood lipid levels are also strong modifiable risk factors: risk of CVD is increased by raised total cholesterol and low density lipoprotein (LDL) cholesterol[53]. Increased levels of circulating LDL particles raise the number of particles entering the blood vessel wall and becoming oxidised, beginning the atherosclerotic process. Increased circulating triglyceride levels also increase risk[54]. It has been suggested that triglyceride-rich lipoproteins may activate platelets and promote a pro-coagulant and pro-inflammatory phenotype[55]. HDL has been proposed as a protective blood lipid, with higher concentrations associated with decreased cardiovascular risk[56]; however, Mendelian randomisation analysis has suggested that this effect may be correlated with other factors rather than being causal[57]. These blood lipids are modifiable through changing behaviour such as improving diet and increasing exercise levels, which decreases risk as expected[58,59], but also have a strong heritable component[60].

As seen above, atherosclerosis is a process with a strong inflammatory influence, and inflammatory markers have been linked to CVD risk. For example, C-reactive protein (CRP) is an acute-phase reactant that acts as a general marker of systemic inflammation, which has been associated with higher risk of cardiovascular events[61,62]. For a time, CRP was therefore considered as a variable to target in reducing cardiovascular risk[63], but Mendelian randomisation studies failed to demonstrate a causal role of the protein, suggesting it is a marker rather than a causal factor[64]. The pro-inflammatory cytokine interleukin-6 (IL-6) has also been shown to be associated with increased risk

of CHD events[65]; in this case, Mendelian randomisation studies have suggested it is causal in the development of CHD[66].

1.2.2.1 Management of cardiovascular risk factors

As the influence of behavioural factors on CVD risk is so high, procedures to identify people at higher risk, and interventions to reduce this risk, are an important part of disease prevention. However, while the health benefits of improved diet, increased physical activity and smoking cessation are clear and well-established[67,68], persuading populations to make long-term lifestyle changes can be difficult.

Methods of reducing risk of CVD would ideally have a large effect and be easy to implement. Statins are a good example of such an intervention. These drugs inhibit the enzyme HMG-CoA reductase, required for the synthesis of endogenous cholesterol, thereby reducing the amount of cholesterol in circulation. Statins have been shown to reduce LDL concentration, and with it the incidence of cardiovascular events[69]. They are a cheap treatment with relatively few side effects[70], and are therefore widely prescribed to lower the likelihood of CVD for those in high-risk groups. However, efficacy is compromised by poor adherence to the treatment programme[71].

1.2.2.2 Carotid intima-media thickness

Markers of disease such as the above inflammatory factors allow potentially more accurate risk prediction than standard risk factors. A more direct marker of atherosclerosis is that of carotid intima-media thickness (IMT): a value comprising the thickness of the tunica intima and media layers in the carotid artery (Figure 4). It can thus be used as a marker for subclinical atherosclerosis[72]. It is a useful marker for large studies and screening, as it can be easily measured using standardised non-invasive ultrasound techniques[73].

IMT is useful for the study of CAD risk, having been shown by numerous studies to correlate with atherosclerosis in the coronary artery[74–76], and thus can be used as a surrogate of this measure. Its value as a marker and surrogate is shown by multiple studies that have shown it to be predictive of incident coronary and cerebrovascular events[77–79]. Specifically, a meta-analysis of 8 trials showed that a 0.1 mm increase in carotid IMT is associated with a 10-15% increase in risk of MI and a 13-18% increase in risk of stroke[80] (Figure 5). Carotid IMT therefore acts as both a marker of disease progression at the time and a risk factor for later events.

There is some debate over the utility of IMT in improving cardiovascular risk in individuals. IMT is associated with cardiovascular events even when controlling for traditional risk factors[79,81], but there is conflicting evidence about whether it improves individual risk prediction when added to existing risk prediction algorithms[82]. Results from the Atherosclerosis Risk in Communities study showed improvement in ten-year CHD risk prediction when adding IMT information to traditional risk factors[83]. The French Three-City Study found no value of adding IMT to prediction methods, although it used IMT as measured selectively in plaque-free areas and presence of plaque remained an independent risk predictor[84]. The Carotid Atherosclerosis Progression Study saw no improved risk classification when adding IMT to traditional models, although noted it remains predictive for cardiovascular endpoints[81].

As shown in Figure 4, carotid IMT can be measured at various segments within the carotid tree: in the common-carotid artery, at the bifurcation, and in the internal and external carotid arteries. There has been some debate about which IMT measurements are the most meaningful[85], and thickness at different segments may reflect different stresses and processes; for example, shear stresses have been shown to affect IMT, and are more likely to affect carotid walls at the bifurcation[86] . However, the different segment measurements do not show great differences in their ability to reflect risk[85].

Although its additive value in risk algorithms remains debated, IMT's use as an atherosclerotic marker gives it value for genetic studies. The underlying genetic mechanisms that contribute to increased or decreased IMT are not yet well known. Using association studies to search for loci associated with IMT, as a marker of subclinical disease, gives more specificity than looking at later cardiovascular events, and may allow the identification of novel genes involved in atherosclerosis.

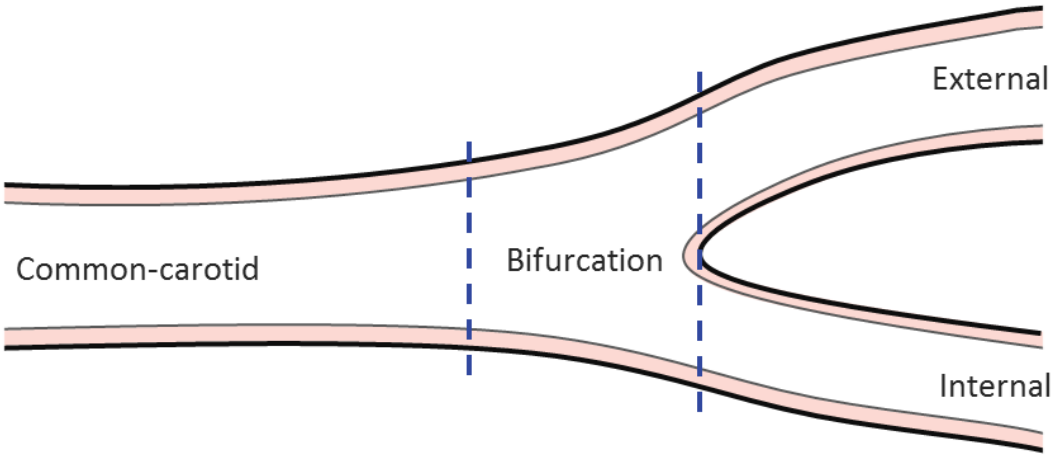

Figure 4: Segments of the carotid artery. IMT can be measured in the common-carotid artery, at the bifurcation itself, and in the internal and external carotid arteries. Measurements often comprise the mean value of all measurements at a segment, or the maximum value.

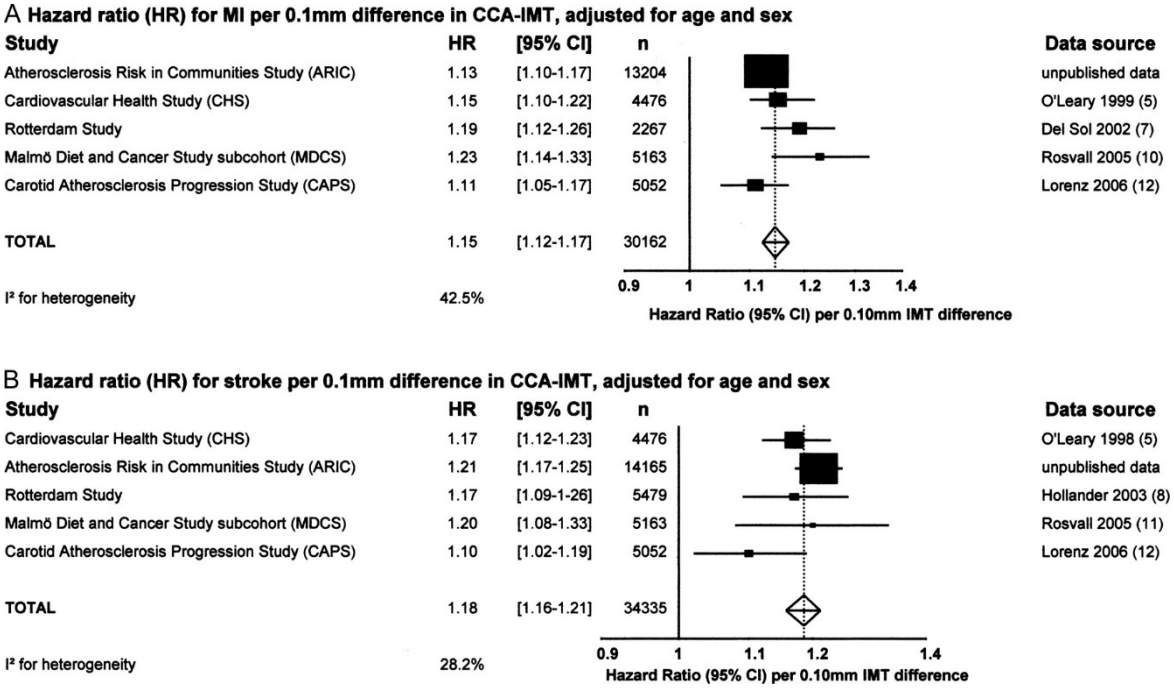

Figure 5: Forest plots for hazard ratios (HRs) per 0.1-mm difference in common-carotid IMT, adjusted from age and sex. Figure from Lorenz et al[80].An increase in 0.1 mm increases the hazard ratio for MI and stroke by 15% and 18% respectively.

1.2.4 Identifying new genetic risk factors for CVD

Approaches used to determine the genetic basis of disease are dependent on the pattern of inheritance the disease displays. Diseases inherited in a Mendelian fashion suggest a single causal gene with a strong effect on phenotype; for these conditions, family linkage studies and direct DNA sequencing have been successful in finding the causal genes. An early cardiovascular example is that of familial hypercholesterolaemia (FH), a genetic disease characterised by high circulating LDL cholesterol, and consequent early cardiovascular disease[87]. The majority of cases are caused by loss of function mutations in the low-density lipoprotein receptor (*LDLR*) gene, encoding the receptor that removes LDL from the blood circulation[88]. In 1985, direct sequencing of the *LDLR* gene in a patient with homozygous FH was used to detect a deletion of several exons, which left the LDLR protein without membrane-spanning or cytoplasmic domains, abolishing its function and causing high LDL in the plasma[89].

Family linkage analyses have also been used to uncover the basis of heart disease disorders. Subjects in a large family with familial hypertrophic cardiomyopathy were genotyped for numerous DNA markers throughout the genome, pinpointing a locus on chromosome 14 that was coinherited with the condition, afterwards identified as the beta cardiac myosin heavy chain[90,91].

The simple genetic basis of such Mendelian diseases allows these detection techniques to be carried out, as the disease state is generally easy to recognise and the inheritance pattern is simple. However, the genetics behind most cardiovascular diseases are much more complex, involving many genes of smaller effect and interplay with environmental factors. The common disease/common variant (CD/CV) hypothesis proposes that genetic variants which are present at relatively high frequency in the population, but with small effect on phenotype, are significant contributors to common diseases[92]. Rare variants of small effect, while they are also likely to be involved in disease, are difficult to detect using association studies due to lack of power[93] (Figure 6).

The CD/CV hypothesis may be particularly relevant for CAD, which is likely to occur later in life. It is therefore likely to have little effect on fitness until after reproductive age, and variants conferring greater risk of CAD less likely to have been reduced in frequency due to selection pressure. This is especially applicable for CVD, whose impact has grown substantially in recent decades, fuelled by worldwide demographic and lifestyle changes[2]. Mapping common variant loci involved in disease requires a different approach, made possible by the advent of genome-wide association studies.

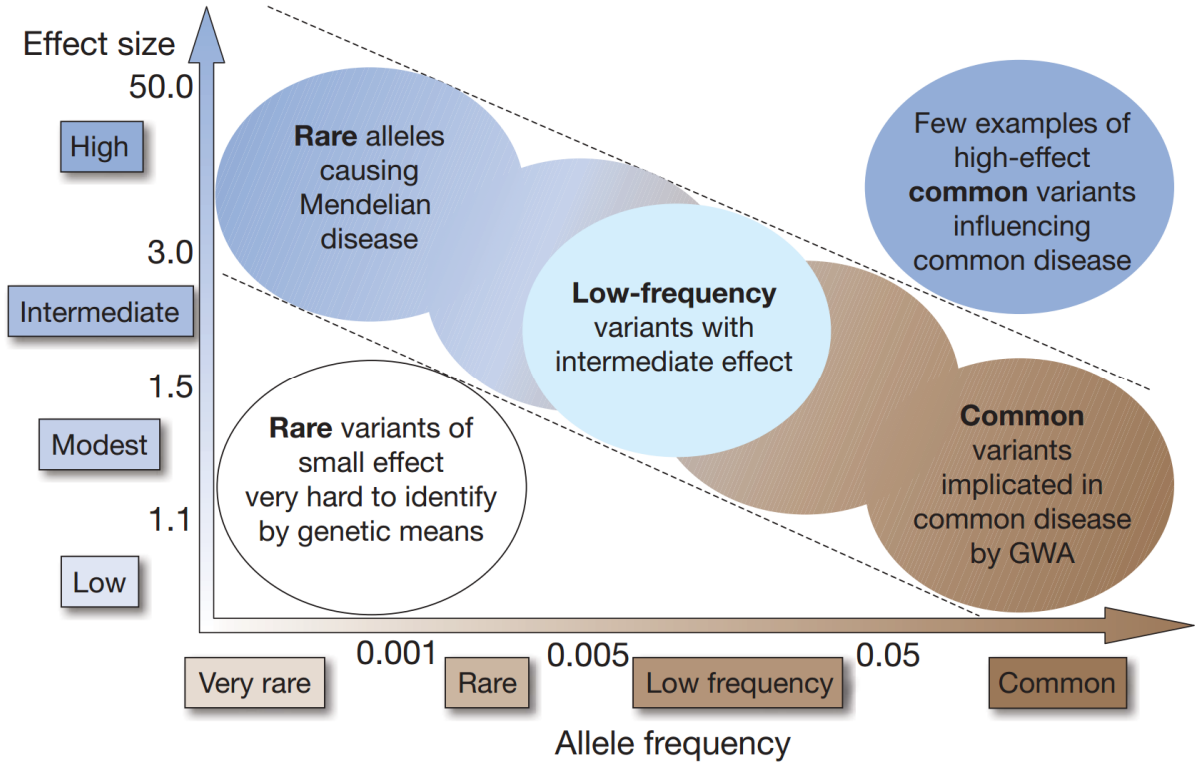

Figure 6: Genetic variants by risk allele frequency and strength of genetic effect. Figure from Manolio et al[94]. Rare alleles that cause Mendelian disease are easy to identify due to their strong phenotypic effect and simple inheritance pattern. Variants with small effect can be identified using genome-wide association studies if they are sufficiently common in the population; rare variants with small effect are very difficult to detect. Variants with a strong effect on disease are unlikely to also be common due to selection pressure.

1.2.5 Genome-wide association studies

Genome-wide association studies (GWAS) are a high-throughput method of scanning genetic markers across the genomes of many individuals, to look for genetic variants that associate with a particular disease or trait. They were made possible after the completion of the Human Genome Project in 2003[95] and the International HapMap Project in 2005[96], which provided a reference human genome and a map of over a million SNPs in the genome.

The development of high-throughput genotyping technologies, particularly genotyping platforms from Illumina and Affymetrix, allowed large numbers of SNPs to be genotyped at a low cost. These genotypes can be combined with phenotype data to carry out hundreds of thousands of SNP-phenotype association tests across the genome. SNPs that show a strong association are likely to indicate a genetic locus with an effect on the phenotype. The association p-values are often plotted against chromosomal location as a Manhattan plot (Figure 7), allowing associated loci to be identified. To detect variants of small effect, large cohorts are needed; the relatively low cost of the

high-throughput genotyping platforms allowed genotyping to be carried out in large numbers of samples.

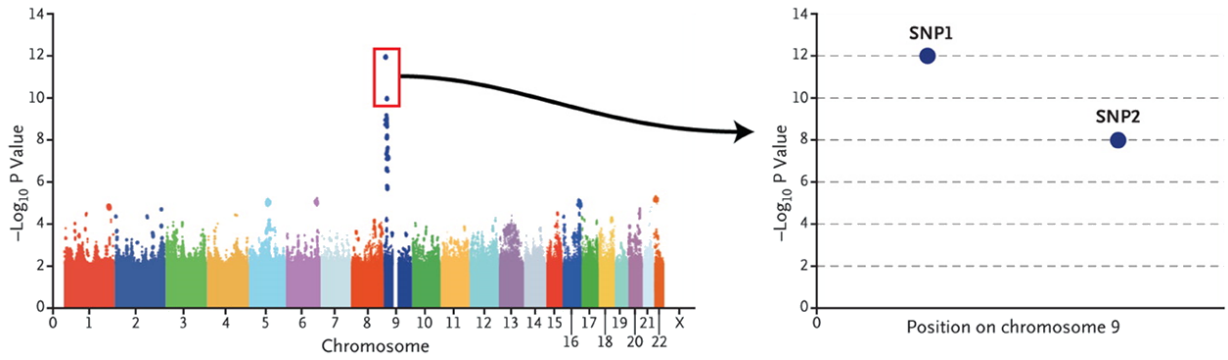

Figure 7: Example of a Manhattan plot for a genome-wide association study. Figure from Manolio[97]. The $-\log_{10}$ p-values of the association test for each SNP are plotted against their chromosomal location. A strong association can be seen by its small p-value. Here a locus on chromosome 9 is implicated in the phenotype. The second graph shows a smaller section of chromosome 9, showing the positions of the two most strongly associated SNPs.

GWAS is facilitated by linkage disequilibrium (LD), the non-random association of alleles at different loci in the genome. According to Mendel's second law, the alleles of different variants are inherited independently[98]. However, in some cases the genotypes of variants are correlated, with certain alleles more likely to be inherited together. This can happen when two variants are located in close proximity on a chromosome. Recombination events during meiosis progressively reduce contiguous chromosomal regions in a population, but variants that are closer together or have fewer recombination hotspots between them are more likely to remain as unbroken regions, with their alleles being inherited together. Recombination events shorten the length of these regions over time. Genotyping every variant in the genome would be unfeasible in terms of time and cost, but the LD blocks in the genome allow one genotyped variant to act as a tag or proxy to determine the allele of SNPs in LD[99]. In this way many more variants can be effectively genotyped than those that are directly assayed.

Clearly defined phenotypes are needed in GWAS to effectively characterise pathogenic variants. Unambiguous phenotypes that are easy to measure such as height are ideal, whereas heterogeneous phenotypes that are hard to accurately measure, or vary in onset, decrease the power to detect associations[100]. Where there is room for ambiguity in phenotype measurements, measurement protocols should be standardised and quality controlled to increase accuracy and reproducibility of the results.

One issue for GWAS interpretation is that association testing for each genotyped variant results in hundreds of thousands or millions of statistical tests being applied across the genome. To avoid the thousands of false positive associations that would thus be expected due to type I error, a statistical correction for multiple comparisons is used. This is often the conservative Bonferroni correction, in which the alpha value is divided by the number of tests, giving a much smaller p-value threshold for significance[93]. A threshold p-value of 5×10^{-8} is often used as a standard for GWAS[101]. An alternative to the Bonferroni correction is the false discovery rate (FDR) method, in which p-values are used to correct for the number of significant results that are false positives[102].

1.2.5.1 GWAS examples

One of the first GWAS identified the Complement Factor H gene as a major risk factor for age-related macular degeneration (AMD), locating the variants and finding the biological basis for the association[103]. The authors carried out a genome wide screen of over 100,000 SNPs in AMD cases and controls, locating an associated intronic SNP in the complement factor H (CFH) gene, a regulator of innate immunity. Exome resequencing identified the variant with the strongest association, a non-synonymous SNP coding for a tyrosine-histidine change in the gene. It was later shown that protein with the risk variant has reduced binding to C-reactive protein, heparin and retinal pigment endothelial cells[104].

After the first studies, the pace of GWAS publications increased rapidly as the cost of performing the genotyping decreased. One of the first large GWAS for cardiovascular disease was carried out by the Wellcome Trust Case Control Consortium in 2007, genotyping 17,000 subjects to look for associations with seven common diseases[105]. The study identified a locus on chromosome 9p21 as strongly associated with CAD (Figure 8), with a signal so strong that the finding was independently reported by three other studies within a short timeframe[106–108]. The 9p21 locus is the strongest and most well-replicated cardiovascular GWAS hit, yet still one that is still not well understood: the region contains no known protein-coding genes, and the functional variant has not been characterised[100]. The GWAS also looked for associations with other diseases, including the cardiovascular risk factors type 2 diabetes (T2D) and hypertension. Three loci were found for T2D, but none were found at genome-wide significance for hypertension. This may be due to inadequate tagging of functional variants by the genotyping array. Alternatively, it may be the case that the effects of individual blood pressure loci are smaller (compared to an OR of 1.47 per allele for the CAD locus), such that larger sample sizes are required to detect them, and that inappropriate

selection of cases and controls may have reduced power to detect associations. It has indeed been suggested by later studies that variants for hypertension are likely to have smaller effect sizes than those for CAD[109], and later studies using larger sample sizes and continuous blood pressure phenotypes have uncovered numerous new loci[110,111].

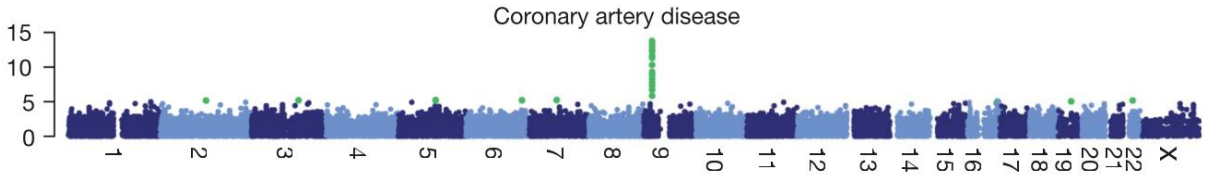

Figure 8: Manhattan plot showing association signal for coronary artery disease at the 9p21 locus. Figure from Wellcome Trust Case Control Consortium[105]). The chromosomal position of each SNP is plotted against $-\log_{10}$ of the association p-value, so that larger numbers indicate a stronger signal.

Another landmark in cardiovascular association studies was a 2010 meta-analysis of over 100,000 subjects, identifying several loci associated with blood lipids (total cholesterol, LDL-cholesterol, HDL-cholesterol and triglycerides)[60]. Some of the loci identified were novel, while others were located by genes previously known or proposed to be involved in lipid metabolism, supporting these previous findings. Many were also associated with CAD risk, demonstrating the significance of blood lipids – particularly LDL – as a cardiovascular risk factor.

Large GWAS, such as those described, use genotyping chips that cover as many variants as possible across the whole genome. On the other hand, once a locus has been identified, genotyping strategies that densely cover the locus of interest are valuable in order to fine-map the association signal to a smaller region. As the creation of locus-specific genotyping arrays is expensive, many fine-mapping strategies use targeting genotyping arrays for a certain disease or trait, such as the iSelect Collaborative Oncological Gene-environment Study (iCOGS) array for cancer[112] and the ITMAT-Broad_CARe (IBC)[113] and Illumina CardioMetabochip ("Metabochip")[114] for cardiovascular, metabolic and inflammatory traits.

The IBC array has been used to follow up on previous blood pressure and hypertension GWAS results to confirm previous blood pressure-associated loci, and identify two new loci at known candidate genes (*MDM4* and *HRH1*)[110]. Associated SNPs at the *MDM4* locus showed associations with expression of the *MDM4* gene. The array has also been used to identify new signals for central adiposity[115].

More recently, the Metabochip was designed for the study of CAD and type 2 diabetes risk loci and quantitative risk factors associated with these traits (including lipid levels and blood pressure) [114]. There are approximately 200,000 SNPs present on the array, these are concentrated at risk-associated loci to permit dense fine-mapping of GWAS association signals, with additional content based on nominal associations from GWAS studies to provide cost-effective replication studies. Willer and colleagues followed up on the 2010 blood lipid meta-analysis by conducting a meta-analysis using cohorts genotyped with both genomic arrays and the Metabochip[116]. This approach facilitated the discovery of many new lipid-associated loci and fine-mapped previously known loci to separate the strongest association signal from a larger region. Fine-mapping was aided by using samples from multiple ethnic groups, in which patterns of LD differ, leaving different SNPs in LD with the functional variant.

Further studies on cardiovascular events have used greater cohort sizes and improved genotype techniques to further refine known signals and identify new loci. The CARDIoGRAM (Coronary ARtery DIsease Genome wide Replication and Meta-analysis) consortium combines information from many large genetic studies to identify CAD and MI risk loci. A large meta-analysis in CARDIoGRAM identified 13 new loci associated with CAD and confirmed 10 others[117]. The majority of the new loci were not associated with traditional risk factors, nor were they located in previously implicated regions. Deloukas and colleagues used the Metabochip in the CARDIoGRAMplusC4D consortium in order to fine-map confirmed loci[6]. A further 15 CAD loci were identified, along with loci also associated with blood pressure and lipid traits. Network analysis identified lipid metabolism and inflammation as biological pathways important for CAD pathogenesis.

1.2.5.1.1 Carotid intima-media thickness association studies

Carotid intima-media thickness is another attractive phenotype for association analyses, being both a risk factor for cardiovascular events and a marker of atherosclerosis/subclinical disease[74,77]. Its ease of measurement using a standardised protocol allows unambiguous phenotyping of numerous subjects and therefore greater study power. A meta-analysis of GWAS in the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) consortium identified three loci associated with common-carotid IMT (*ZHX2*, *APOC1* and *PINX1*), and two with plaque (*PIK3CG* and *EDNRA*)[118]. The genes identified are involved in LDL metabolism, endothelial function and platelets. These distinct signals shown between the phenotypes indicate the interplay of numerous different pathways, although it should be noted that the definition of plaque differed between the included cohorts.

A scan was later performed using the Metabochip array. The scan was carried out in the European Carotid IMT and IMT-Progression as Predictors of Vascular Events (IMPROVE) cohort, a longitudinal multicentre observational study designed to look at IMT as a risk predictor[119]. The study identified a novel locus on chromosome 16 as robustly associated with IMT and CAD. The minor A allele of the lead SNP (rs4888378) was protective, being associated with lower carotid IMT and CAD risk (Figure 9).

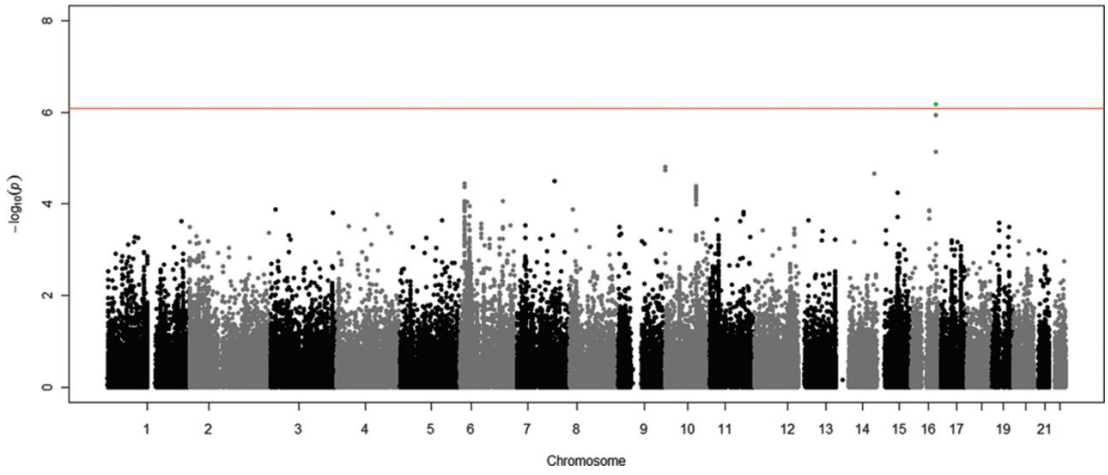

Figure 9: Manhattan plot of association p-values for IMT in IMPROVE. Figure from Gertow et al[119]. P-values are adjusted for age, sex and population substructure. The lead SNP at the association signal, rs4888378, is shown in green on chromosome 16. This is the only signal crossing the threshold for array-wide significance (indicated by the red line).

This signal was present at a novel locus for cardiovascular phenotypes, and hence was not a Metabochip candidate region covered densely by the chip. There was no clear candidate gene at the locus. The lead SNP is located within intron 6 of the *CFDP1* (craniofacial development protein 1) gene; located upstream are the genes *TMEM170A* (transmembrane protein 170A) and *CHST6* (carbohydrate (N-acetylglucosamine 6-O) sulfotransferase 6), and downstream are *BCAR1* (breast cancer antiestrogen resistance 1), *CTRB1* (chymotrypsinogen B1), *CTRB2* (chymotrypsinogen B2), *ZFP1* (zinc finger protein 1) and *LDHD* (lactate dehydrogenase D) (Figure 10).

Many of the genes at the locus are not yet well-characterised. *CFDP1* is known to be expressed widely in the embryo, but it is specifically involved in the development of teeth[120]. *TMEM170A* codes for a transmembrane protein that localises in the endoplasmic reticulum and nuclear envelope. Its expression promotes the formation of endoplasmic reticulum sheets and nuclear pore complexes[121].

BCAR1 codes for an adaptor protein with roles in signalling pathways for cell adhesion, migration and mechanical stress, and is described further in section 1.4. *CHST6* codes for an enzyme that

catalyses the transfer of sulphate to keratin in the cornea[122]. Keratan sulphate is necessary for maintenance of corneal transparency, and mutations in *CHST6* are associated with macular corneal dystrophy[123]. *CTRB1* and *CTRB2* code for chymotrypsinogens, serine proteases secreted into the GI tract as precursors of the digestive enzymes chymotrypsin[124]. *ZFP1* codes for a zinc finger protein which has been suggested to be involved in transcriptional regulation[125]. *LDHD* codes for an enzyme catalysing the conversion of lactate to pyruvate. It is widely expressed, with the highest levels in kidney and liver[126].

Expression data from the Advanced Study of Aortic Pathology and Biobank of Karolinska Endarterectomies studies revealed differences in expression of *TMEM170A* by rs4888378 allele, and nominally-significant differential expression in *BCAR1* and *LDHD* (Figure 11); it therefore appears that the functional variation may be affecting expression of one or more of these genes.

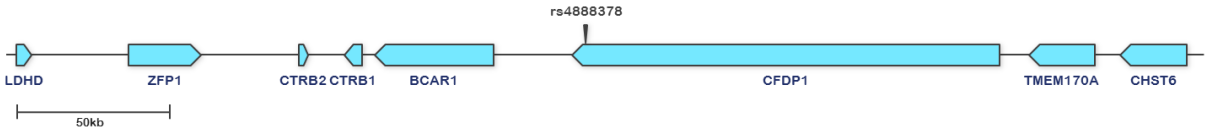

Figure 10: The position of the lead SNP rs4888378 in Gertow and colleagues' carotid IMT scan. The SNP is located in an intron at the 3' end of *CFDP1*. Gene coordinate data is from UCSC Genome Browser[127].

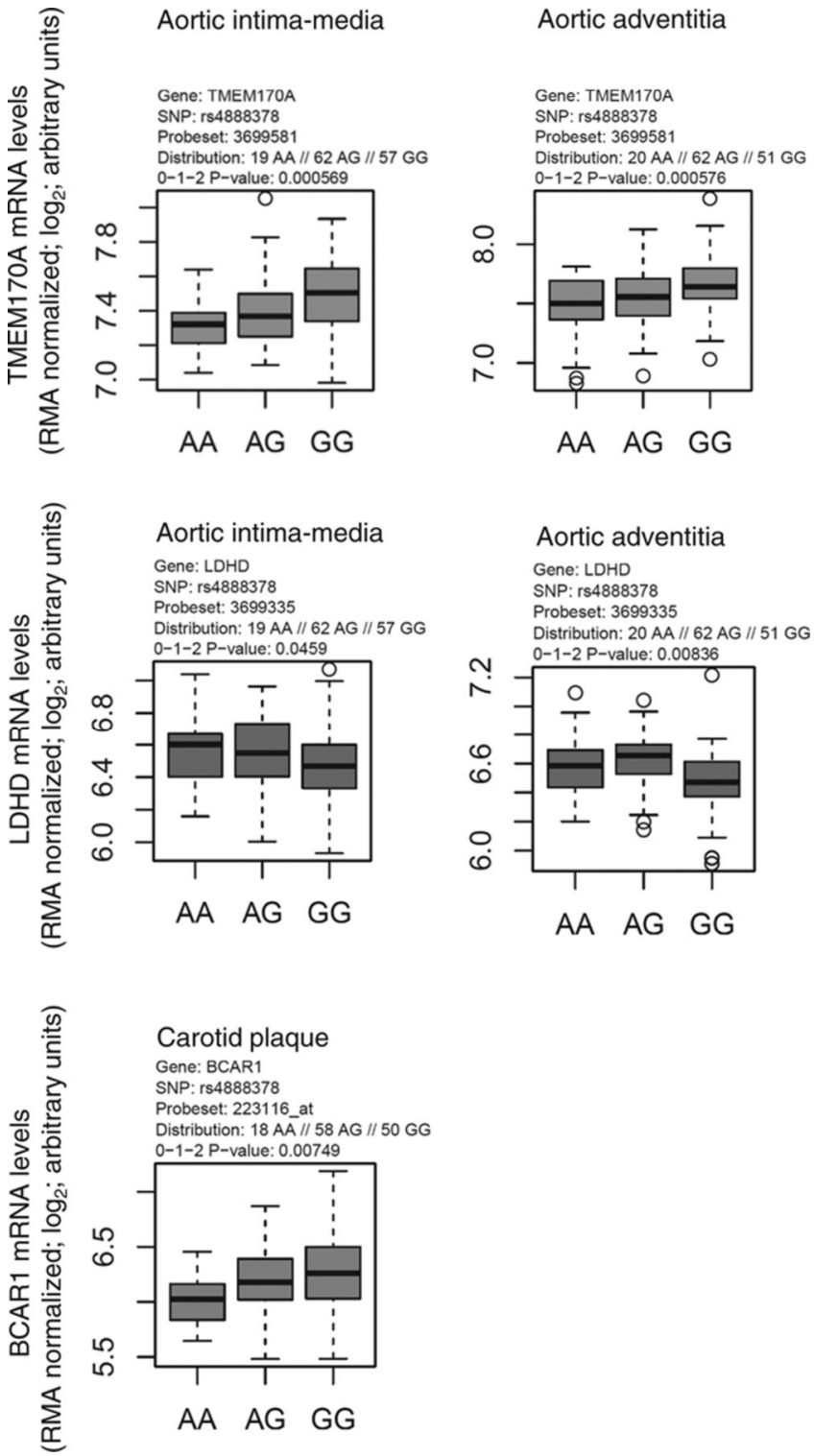

Figure 11: Association of rs4888378 with expression of *TMEM170A*, *LDHD* and *BCAR1* in target tissues. Figure from Gertow et al[119]. The SNP's G allele is associated with higher expression of *TMEM170A* and *BCAR1* and lower expression of *LDHD*. Only the association with *TMEM170A* is significant after correction for multiple testing.

1.3 Moving from GWAS loci to functional variation using genomic annotations

GWAS have identified numerous novel genetic associations with disease, but the majority of these have yet to be translated into clinically useful information[128]: aside from the incomplete knowledge we have regarding the full functions of human gene products, the majority of GWAS do not identify the causal variants. Identifying these variants is critical in understanding how a locus is involved in molecular pathways which contribute to disease pathology. The ideal analysis will determine how a variant exerts an effect at the molecular level; for example, disrupting the action of a promoter, enhancer or silencer. A nucleotide change may affect the binding of a transcription factor which drives expression of a gene, and finding out which transcription factor is affected may provide clues about which pathways are involved. Knowing this information – for example, that an inflammatory pathway is involved – may allow the discovery of novel factors that are involved and interact with risk of disease. Identification of regulatory elements may also implicate certain tissues in the disease process.

Part of the problem in the search for causal variants is the issue of linkage disequilibrium: while it allows for the tagging and analysis of hundreds of thousands of unprobed variants, it also presents a problem for finding the actual genetic change that causes the effect. When a variant identified by GWAS is in LD with many other SNPs, strength of association is not a useful instrument to distinguish between them: the correlation between genotypes means all variants will show a similar association. In order to assess the likelihood of functionality for these SNPs, other approaches must therefore be used.

Genomic annotation data can be used to study a genetic locus with a view to determining which variants are most likely to be causing an effect on phenotype. The ENCODE (Encyclopaedia of DNA Elements) Consortium aims to document the functional elements in the human genome[129], mapping information relevant to genetic regulation such as transcription, transcription factor binding, chromatin structure and histone modifications. These annotations allow the evaluation of likely functionality of regions of the genome; indeed, ENCODE claimed to have assigned biochemical function to 80% of the genome[129], although the magnitude of this claim has been met with controversy[130,131]. Nevertheless, variants that lie in regions undergoing regulatory activity, as assessed by these regulatory marks, are more likely to disrupt processes in gene expression. Such data can therefore be used to judge the likely functionality of variants identified by GWAS. The NIH Roadmap Epigenomics Mapping Consortium[132] is another publicly available resource of human

epigenomic data, focusing on stem cells and primary tissues to map epigenetic marks such as histone modifications and chromatin features.

The UCSC (University of California Santa Cruz) Genome Browser[127] facilitates visualisation of ENCODE and Roadmap annotations across the genome by displaying the data as tracks mapped to the genome sequence, allowing easy comparison of regulatory activity at loci of interest. The genome browser can be used to gain an overview of the regulatory and structural landscape at a locus, and make inferences about variants and their likelihood of affecting regulatory systems. To directly compare variants' regulatory potential, online tools such as HaploReg[133] summarise relevant annotations from these databases. At present there is no conclusive tool to rank non-coding variants by likelihood of functionality, but programs such as RegulomeDB[134] and CADD[135] provide their own systems of ranking using annotations from ENCODE and RoadMap.

Assessing likelihood of functionality for variants in protein-coding regions is more straightforward than for those who are intronic and intergenic. Variant effect is likely to depend on synonymous or non-synonymous nature of the base change, the properties of the original and new amino acid coded for, and the position of the amino acid in the protein. SIFT[136] and PolyPhen2[137] are examples of algorithms that compare these variables, assessing factors such as residue conservation, charge and hydrophobicity.

1.4 BCAR1

As discussed in Section 1.2.5.1.1, the chromosome 16 locus found to be associated with carotid IMT contains multiple genes, one of which is *breast cancer anti-oestrogen resistance 1* (*BCAR1*). As will be discussed, gene expression data and sequencing analysis carried out in this thesis identified *BCAR1* as the main gene of interest for IMT associations at the locus, and further work concentrated on identifying the role of this gene and variants within it in intima-media thickness and atherosclerosis. The BCAR1 protein has been well studied and is known to have roles in many molecular processes, (Figure 12). In this section the known functions of the gene and protein will be discussed, along with their possible involvement in atherosclerosis-related pathways.

BCAR1 is also known as Cas (Crk-associated substrate) or p130cas. It was identified in 1989 as a 130 kDa protein associating with the adapter protein Crk (CT10 (chicken tumour virus number 10) regulator of kinase) that was highly phosphorylated on tyrosine residues on transformation with v-Src and v-Crk oncogenes[138]. It is an essential protein in development, with mouse knockouts being embryonic lethal[139]. Mouse embryonic fibroblasts (MEFs) without the protein have disorganised actin stress fibres and defects in cell migration, implicating the protein in cell motility and actin fibre arrangement[139].

It does not have a kinase domain, but its many conserved sequence motifs and many post-translational modifications suggested that it was an adapter protein[140]. These proteins tend not to have enzymatic activity, as is the case with BCAR1, but instead contain interaction domains that link binding partners to create larger signalling complexes. Its structure is similar to that of other adapter proteins, such as DOK1 (downstream of tyrosine kinase 1) and Gab (GRB2-associated binding protein), with a well-defined domain in the N-terminal region and a large C-terminal region with less definition[141].

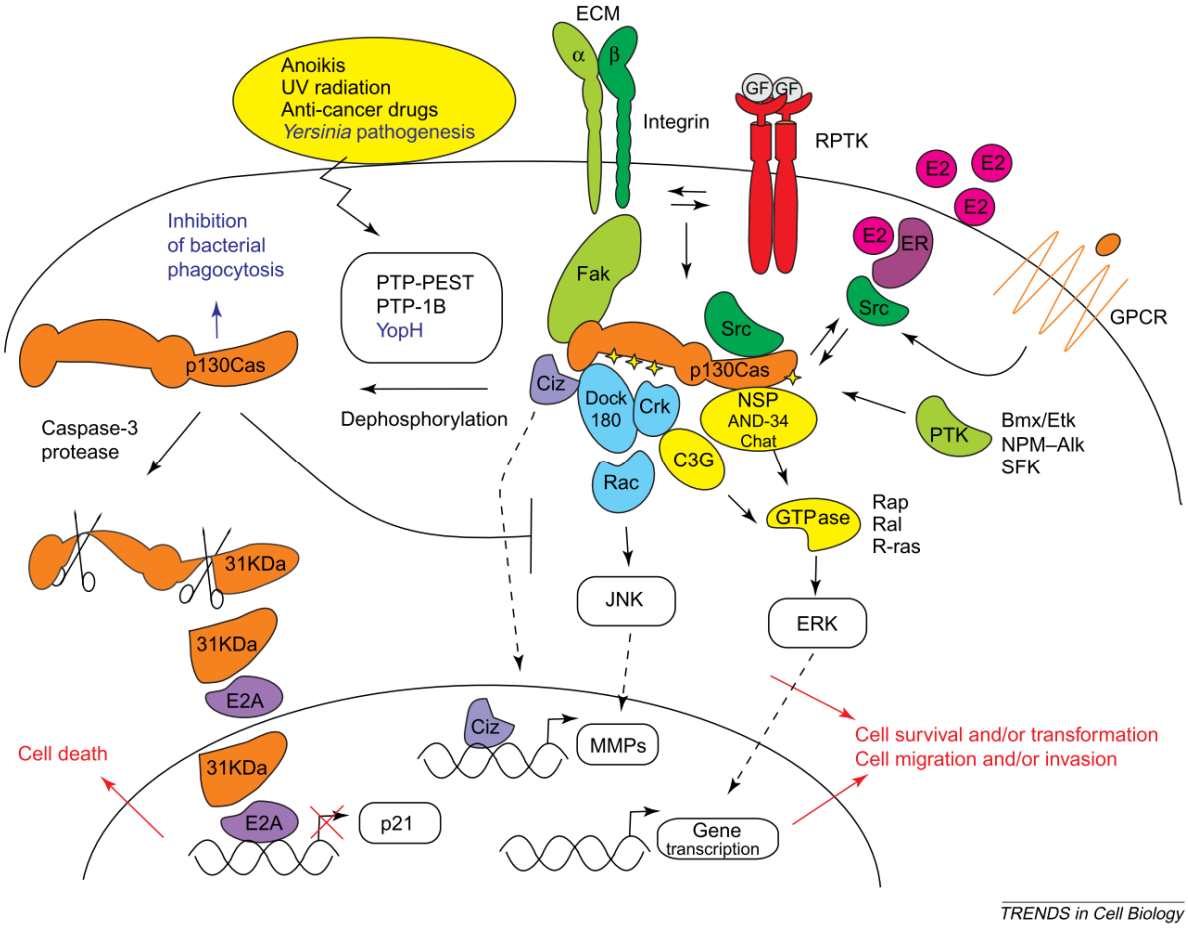

Figure 12: Signalling networks involving BCAR1 (p130cas). Figure from Defilippi[142]. Proteins such as integrins, receptor protein tyrosine kinases, oestrogen receptors and G-protein coupled receptors regulate BCAR1 through formation of a complex with Src and tyrosine phosphorylation. BCAR1 recruits proteins to activate downstream pathways, regulating cell survival and movement. Dephosphorylation of BCAR1 stimulates its cleavage and contributes to cell death.

1.4.1 BCAR1 structure

BCAR1 consists of SH3 (Src-homology 3), proline-rich, substrate, serine-rich and C-terminal domains (Figure 13). The SH3 domain binds FAK (focal adhesion kinase)[143], PYK2 (protein tyrosine kinase 2)[144] and C3G[145], and is fundamental for the localisation of BCAR1 to focal adhesions[146] and its ability to disassemble them[147]. Following the SH3 domain is a proline-rich region of about 40 aa, to which no known function has yet been attributed.

The substrate domain is approximately 300 amino acids in length and contains the major sites of tyrosine phosphorylation, which are 15 repeats of a YxxP amino acid sequence[140]. Src-homology 2 (SH2)-domain-containing proteins bind the phosphorylated tyrosine residues in this sequence. 9 of the repeats are a YDxP sequence, which best matches the binding motifs for the SH2/SH3 adapter proteins Crk and Nck, with which it binds[148]. The serine-rich domain follows the substrate domain,

forming a four-helix bundle which acts as a protein-interaction motif similar to that found in other adhesion-related proteins[149]. The large C-terminal domain contains little structural definition, but is known to bind to Src[150], which is responsible for YxxP phosphorylation in the substrate domain[141].

Figure 13: Protein domain structure of BCAR1. The protein is largely composed of an SH3, proline-rich, substrate, serine-rich and C-terminal domain. The major post-translational modifications occur in the form of phosphorylation of tyrosine residues in the substrate domain.

1.4.2 Cas family members

BCAR1 is part of the Cas protein family, whose members have high structural homology and conserved binding domains. However, their roles, tissue expression and tissue distribution vary. While BCAR1 is expressed ubiquitously, HEF1/Cas-L (human enhancer of filamentation/Crk-associated substrate in lymphocyte) is expressed in epithelial and nervous tissues and is involved in mitosis[151]. Efs/sin (embryonic fyn substrate) is expressed largely in the embryo and placenta but is also active in skeletal muscle and the brain[152], and has anti-inflammatory functions in T lymphocytes[153]. HEPL (HEF1-Efs-p130cas-like) was identified due to its structural similarity to the Cas family and is expressed in primary tissues[154].

1.4.3 BCAR1 localisation

BCAR1 is ubiquitously expressed throughout the body[140]. The gene-tissue expression browser (GTEx) shows it to be highly expressed in artery tissue, particularly the aorta[155] (Figure 14). The Human Protein Atlas also reports that it is expressed highly in telomerase-immortalised microvascular endothelial cells[156].

Its subcellular localisation is concentrated largely at the plasma membrane and in the cytoplasm, with some present at nucleoli, focal adhesions and stress fibres[157]. Unphosphorylated BCAR1 tends to be cytosolic[141], while ligand-induced tyrosine phosphorylation causes it to translocate to the cell membrane[140], where it localises in focal adhesions (FAs) with other FA proteins, such as the kinases Src, FAK and PYK2 and the adapter proteins Crk and Nck[143].

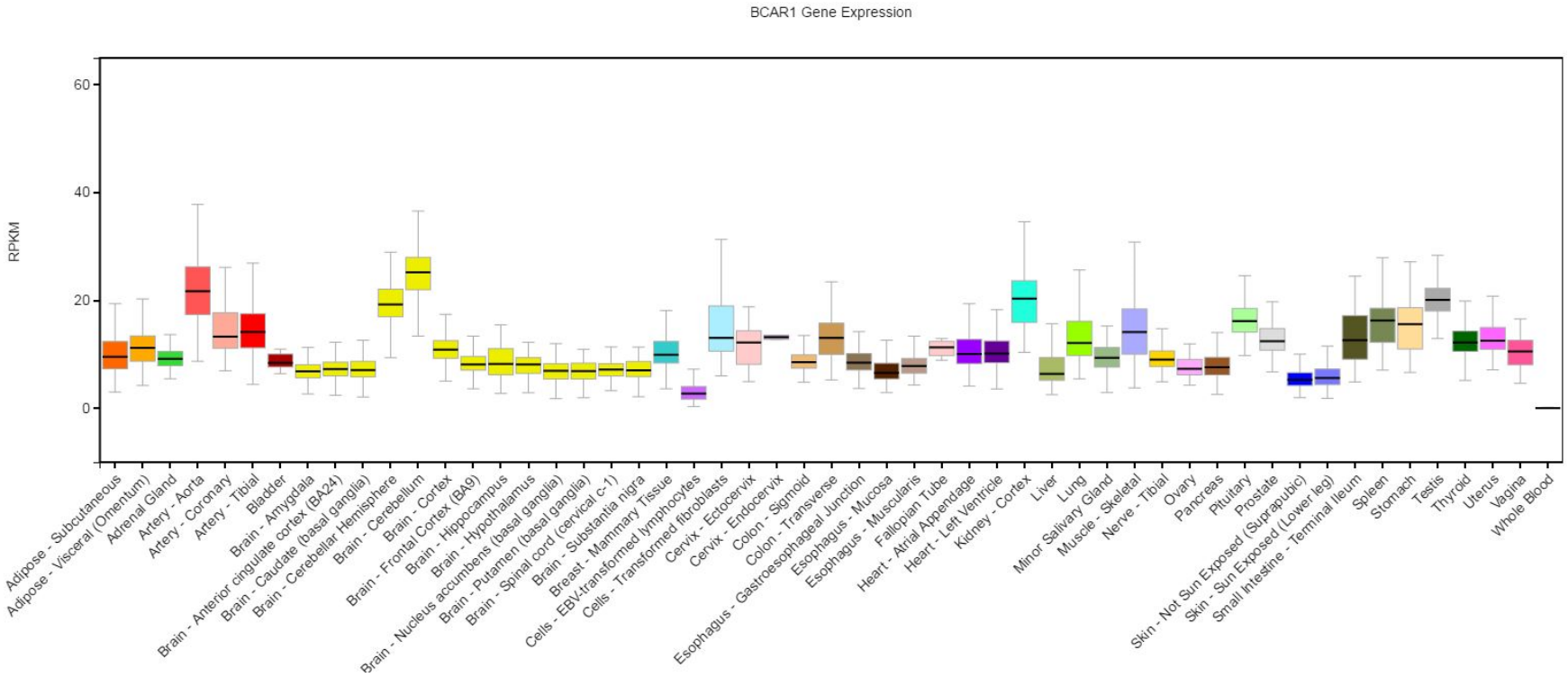

Figure 14: *BCAR1* expression in different human tissues. *BCAR1* is expressed ubiquitously, with the exception of whole blood, with particularly high expression in the cerebellum, aorta and kidney cortex. Data from GTEx (Gene-Tissue Expression Browser)[155].

1.4.4 BCAR1 at focal adhesions

BCAR1 localises at focal adhesions with the other focal adhesion proteins FAK, paxillin and tensin[157]. Focal adhesions are large protein complexes that link the extracellular matrix (ECM) to the cell's internal actin cytoskeleton. They act as both a physical anchor and a transmitter of signals from the ECM to internal signalling pathways, and thus can translate mechanical forces into signals affecting cell behaviour.

Focal adhesions are important for the maintenance and modification of cell shape. They frequently undergo assembly and disassembly, particularly in moving cells that are constantly making new connections. BCAR1 is important for focal adhesion assembly and disassembly: mouse embryonic fibroblasts (MEFs) without BCAR1 show slow FA disassembly which is restored with addition of the protein[147,158].

The physical link between the extracellular and intracellular spaces is created by integrins, which span the cell membrane as heterodimers of α and β proteins[159] (Figure 15). The non-catalytic cytoplasmic domains associate with adaptor proteins like BCAR1 to transmit signals from the ECM[160]. Integrin complexes link to the actin cytoskeleton through actin-regulating proteins such as vinculin, paxillin and talin[161]. It was recently demonstrated that phosphorylated BCAR1 is linked to the actin cytoskeleton through tensin 1[162], which drives cell migration.

Adhesion to proteins in the ECM such as fibronectin and vitronectin activate integrins, stimulating phosphorylation of BCAR1 and the other FA proteins FAK, paxillin and tensin[141]. Focal adhesions are also involved in growth factor signalling; both integrin- and growth-factor-mediated phosphorylation are discussed in Section 1.4.5.

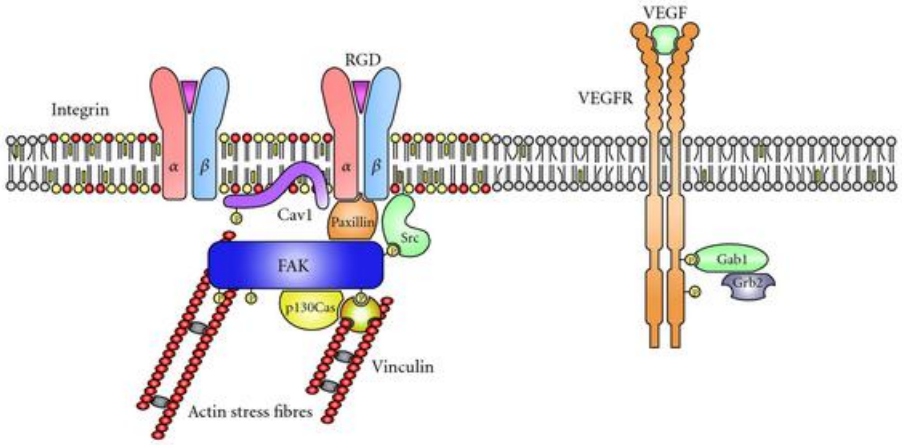

Figure 15: Integrins and VEGF-receptor signalling at focal adhesions. Figure from Ngalim et al[161]. α- and β-subunits of integrins bind to proteins in the ECM, activating cascades involving the phosphorylation of BCAR1 (p130cas).

1.4.5 Phosphorylation of BCAR1

BCAR1 is largely regulated by phosphorylation and dephosphorylation, at either tyrosine or serine/threonine amino acids. The main form is tyrosine phosphorylation, particularly the tyrosines of the 15 YxxP repeats contained in the substrate domain. BCAR1 is tyrosine-phosphorylated on stimulation with growth factors[141], activation by integrins[163] and peptide hormone ligands for G-protein coupled receptors (GPCRs)[164].

Growth factor-induced tyrosine phosphorylation of BCAR1 requires the presence of Src[165]. Growth factors that stimulate phosphorylation of BCAR1 do so through activation of their respective protein receptor kinase ligands. VEGF (vascular endothelial growth factor), the main chemotactic and angiogenic factor in endothelial cells, induces phosphorylation of BCAR1[166]; its stimulation of endothelial cells causes type 2 VEGF receptors to translocate from caveolae to focal adhesions and stimulate signalling cascades[167]. Other growth factors that stimulate BCAR1 phosphorylation are FGF2 (fibroblast growth factor 2)[168], PDGF (platelet-derived growth factor)[169], EGF (endothelial growth factor)[170] and IGF-1 (insulin-like growth factor 1)[171]. Src directly carries out growth-factor induced phosphorylation of BCAR1[165], and FAK is an important upstream activator involved in this process[172].

The other main initiator of tyrosine phosphorylation is integrin activation. As detailed in Section 1.4.4, BCAR1 forms a protein complex at focal adhesions with integrins, where it undergoes tyrosine phosphorylation in response to cell adhesion. When the cell adheres to ECM proteins such as fibronectin and vitronectin, integrins are activated to phosphorylate FAK and paxillin[173], tensin[174] and

BCAR1[143]. This stimulation of phosphorylation by different extracellular proteins allows the signalling cascades to be regulated according to the status of the ECM. As with growth-factor-induced phosphorylation, integrin-mediated BCAR1 phosphorylation is Src-dependent[175]. Tyrosine phosphorylation of FAK, Src and BCAR1 is required for the functioning of focal adhesions[161], and tyrosine phosphorylation is so abundant here that anti-phosphotyrosine antibodies are used to mark focal adhesions in immunostaining[176]. BCAR1 and FAK phosphorylation occur in association with the formation of actin stress fibres[163].

Serine/threonine phosphorylation of BCAR1 occurs to change the interactions of BCAR1 during mitosis. Focal adhesions disassemble in order to let cells change shape and lose attachments to the other cells and the ECM. Serine/threonine phosphorylation and tyrosine phosphorylation of BCAR1 allows the complex to disassemble, and the process is reversed after mitotic division[177].

BCAR1 tyrosine phosphorylation is key in controlling organisation of the cytoskeleton, which facilitates cell attachment, migration and invasion. Phosphorylated tyrosine residues in BCAR1 are recognised by SH2-domain-containing proteins such as Crk and Src[140]. Interaction with proteins in the Crk family (CrkI, CrkII and CrkL (Crk-like protein)) are essential to these downstream processes[178]. These interactions depend on phosphorylation of both proteins, with phosphorylation of the Crk proteins being mediated by the non-receptor tyrosine kinase Abl[179].

Src-mediated tyrosine phosphorylation causes the assembly of a complex of BCAR1, Crk, DOCK180 and ELMO at focal adhesion sites, which activates the GTPase Rac[180]. This causes actin polymerisation which is needed to form lamellipodia and membrane ruffles[165].

Phosphorylated BCAR1 also forms a scaffold for the non-catalytic region of the tyrosine kinase Nck in these membrane ruffles, forming a complex involved in linking growth factor signalling to the actin cytoskeleton[181].

1.4.6 BCAR1 in mechanosensing

BCAR1 is affected not only by chemical signals but mechanical force. Forces such as endothelial shear stress regulate expression of growth factors such as VEGF and PDGF, which stimulate BCAR1[182], but also directly influence tyrosine phosphorylation of BCAR1 itself without acting through Src family kinase activity[183]. Directly stretching of the cell and the BCAR1 protein increases tyrosine phosphorylation, suggesting this force exposes suitable tyrosine residues for phosphorylation[184]. Mechanical force can therefore be transduced into the BCAR1 signalling cascades.

1.4.7 BCAR1 in blood vessel tissues

BCAR1 and the focal adhesion complex are known to be involved in endothelial migration[141]. Growth-factor-dependent endothelial migration requires tyrosine phosphorylation of BCAR1 and the presence of the coreceptor neuropilin-1[185]. Growth factors that are important to the function of endothelial cells, such as VEGF, have been shown to stimulate BCAR1 phosphorylation in these cells[166]. As the cell types present proximal to the lumen of the blood vessel, it is endothelial cells that mainly experience fluid shear stress, with this stress increasing phosphorylation of BCAR1[183]. Phosphorylation of BCAR1 has been shown to differ depending on whether the protein is located at focal adhesions upstream or downstream of blood flow[141].

Vascular smooth muscle cells (VSMCs) are the other main constituent of the blood vessel wall, where they predominantly show the quiescent, or contractile, phenotype. BCAR1 is needed for the contraction of VSMCs through actin polymerisation[186]. After vascular injury, a healing response is activated and growth factors such as PDGF and VEGF released; this triggers VSMCs to change to the migratory/proliferative profile, allowing them to migrate to repair the damage[17]. This process is necessary for wound healing and vascular development, but can also be damaging in the context of atherogenesis. As an atherosclerotic lesion forms, oxidised LDL promotes an inflammatory response in the lesion, activating the migratory response of VSMCs and causing the lesion to thicken. These growth-factor-stimulated migratory responses are mediated through BCAR1, and deletion or dephosphorylation of BCAR1 from VSMCs decreases the migratory response[141]. Chen and colleagues showed that the expression and phosphorylation of BCAR1 directly promote formation of neointima after arterial injury, and that vascular injury modulates this phosphorylation. They conclude that BCAR1 may be a potential therapeutic target for vascular disease[187].

1.4.8 BCAR1 in disease

The phenotype of knockout mouse models (impaired cardiovascular development and impaired actin organisation) shows the importance of BCAR1 in cardiovascular processes[139]. In addition to the roles in blood vessel tissues described above, BCAR1 plays a role in myocytes in regulating organisation of the sarcomere[188]. It is also involved in pulmonary arterial hypertension (PAH): BCAR1 expression and phosphorylation was higher in endothelial and smooth muscle cells of people with PAH, causing more proliferation and migration of cells[189].

BCAR1 also has many known interactions with cancer. It was named *breast cancer anti-oestrogen resistance 1* for the fact that its overexpression confers resistance to anti-oestrogens such as

tamoxifen in ER-positive breast cancer cells[190]. In breast tumours, high levels of BCAR1 are associated with poorer tamoxifen response and decreased survival[191]. Silencing of BCAR1 in breast carcinoma cells *in vitro* decreased their invasive potential, suggesting it promotes breast cancer progression[192]. Furthermore, it is involved more generally in many processes important to cancer: it increases the invasive potential of Src-transformed cells, with associated increased tyrosine phosphorylation[193], it facilitates the repression of growth inhibition by TGFβ[194], and its breakdown is important for regulation of programmed cell death[195].

It can be seen that BCAR1 is involved in a wide variety of functions and cell types implicated in the development of atherosclerosis. Genetic variants affecting its expression or structure may have the capacity to disturb these processes and affect the development of atherosclerosis.

1.5 Aims and hypothesis

The hypothesis of this thesis is that genetic variants exist which contribute to the formation and progression of atherosclerosis, which can be measured by monitoring carotid intima-media thickness. By identifying these variants, greater understanding can be obtained about the mechanisms involved in the pathogenesis of CVD, and this can be used to assess and decrease people's cardiovascular risk.

To address the above hypothesis, the aims of this PhD project are as follows:

- Investigate the *CFDP1-BCAR1-TMEM170A* locus, with the aim of finding functional variation and identifying genes involved in atherosclerosis
- Investigate through what mechanisms functional variation affects atherosclerosis
- Explore methods of studying genetic regulation at loci such as *CFDP1-BCAR1-TMEM170A*

2 Methods

2.1 General methods

2.1.1 Agarose gel electrophoresis

Gel electrophoresis was used to analyse DNA sizes during various applications throughout the project. Agarose gels ranged from 1-2% depending on the size of DNA fragment to be resolved. For a 100ml 1% gel, 1g agarose was added to 90ml dH_2O, and microwaved on high power until dissolved. 10ml 10×TBE and 250ul 10 mg/ml Ethidium Bromide were added, and the mixture poured into a gel mould with comb. Gels were placed in the electrophoresis tank and submerged in 1×TBE. 6× loading buffer was added to 10-20 µl DNA product and the mixture loaded onto the gel. Electrophoresis was typically carried out at 120V for 45 minutes. Gels were visualised using the Syngene Gel Documentation and Genesnap v6.04 software, or the Syngene G:BOX Chemi XRQ and GeneSys v1.5.0.0 software.

2.1.2 PCR

Polymerase chain reaction (PCR) was used to amplify DNA templates for various applications throughout the thesis. PCR utilises the ability of DNA polymerase to synthesise copies of a specific region of DNA. A reaction consists of DNA template containing the region to be amplified, two primers of approximately 20 bp length that are complementary to the 3' ends of the sense and anti-sense strands of the DNA target, DNA polymerase enzyme, deoxynucleoside triphosphates (dNTPs) used by the polymerase to synthesise new DNA strands, and buffer solution.

The procedure relies on a thermocycling procedure consisting of multiple cycles of different temperature steps. In the denaturation step, the reaction is heated to a temperature of 95-98°C for 20-30 seconds, causing the double-stranded DNA to break apart into single-stranded DNA molecules. The annealing step, at approximately 50-65°C for 20-40 seconds, allows the primers to anneal to the single-stranded DNA template at the target sites. An extension step at approximately 72°C, for a length of time dependent on the target length to be amplified, allows the DNA polymerase to synthesise a new DNA strand complementary to the template strand, creating a double-stranded DNA molecule again. Repeated thermocycling allows the amount of target DNA to increase exponentially until sufficient amounts have been produced.

In this thesis, PCR was carried out using the NEB's Phusion DNA polymerase, Roche's Expand Long Template PCR System, and Acqua Science's PCR Mastermix. Reactions were set up according to the

protocols provided by the manufacturers. Thermocycling conditions were calculated on the basis of manufacturer-recommended conditions, DNA template length, primer Tm, and amount of PCR product required. PCR reactions were carried out on the BioRad C1000 Thermal Cycler.

2.1.3 TaqMan allelic discrimination

Life Technologies' TaqMan allelic discrimination genotyping system was used for genotyping SNPs in cohort studies. The system uses FRET technology: two target-specific probes, one for each allele of the SNP, are covalently linked to two different 5' reporter dyes (generally VIC and FAM), and a 3' quencher dye. When the probes are intact, the quencher dye represses fluorescence from the reporter dye due to their close proximity. At the annealing step of PCR, the probes anneal to the SNP; during extension, the 5' nuclease activity of the Taq polymerase releases the reporter and quencher dyes, leaving the reporter dye free to fluoresce. As a probe with a mismatched base is not recognised by the Taq polymerase, only the reporter matching the allele will be released. Reporter dye signals are detected and visualised in a plot, allowing each sample's genotype to be recognised by the relative signal of each dye.

To carry out the assay, 4 µl assay mix for the SNP under investigation was dispensed over 5 ng dry DNA in a 384-well plate format and thermocycled as per manufacturers' instructions. Fluorescence was detected with the ABI 7900HT Fast Real-Time PCR System.

2.1.4 KASP SNP genotyping system

KBioScience's KASP SNP genotyping was used for genotyping of a SNP in the PLIC cohort. This is a high-throughput genotyping system with a similar protocol to TaqMan, but different underlying chemistry. Each assay mix contains two allele-specific primers with a unique unlabelled 5' tail sequence, and a common reverse primer. The common reaction mix contains two secondary 5' fluor-labelled oligos (labelled with either FAM or HEX), which can bind to the allele-specific primer tails, with two complementary oligos with quenchers bound to the 3' ends. The complementary pairs bind to each other and fluorescent signal is quenched.

In the first round of PCR, the allele-specific primer binds upstream of the SNP, and the common reverse primer binds on the other strand. The unique 5' tail sequence is incorporated into the PCR product. In the second round, the common reverse oligo binds the template made in the first round and extends, producing a complement to the allele-specific 5' tail. In the third round, the secondary oligo with the attached fluorophore binds to the product, removing the fluorophore from its

quencher and incorporating it into the PCR product. In subsequent rounds of PCR, more fluorophores are incorporated into the products and allowed to fluoresce, producing more signal. Allele-specific signals are detected and visualised in a plot, allowing the identification of genotype by relative signal intensity.

3.6 µl assay mix was dispensed over 5 ng dry DNA in a 384-well plate format and thermocycled according to manufacturers' instructions. Fluorescence was detected with the ABI 7900HT Fast Real-Time PCR System.

2.1.5 Sanger sequencing

DNA that was sent for sequencing for verification of sequence and for genotyping used Source Bioscience's Sanger sequencing services. Plasmid DNA was sent at a concentration of 100 ng/µl, while PCR products were sent at 1 ng/µl per 100 bp. Sequencing primers were sent at a concentration of 3.2 pmol/µl.

2.1.6 Cell culture

Huh7 (human hepatoma) cells were obtained from the European Collection of Cell Cultures (ECACC) and cultured in DMEM (Dulbecco's Modified Eagle Medium) with 10% added foetal bovine serum (FBS). Human embryonic kidney 293 (HEK293) and CV-1 (simian) in origin, carrying SV40 (COS) cells were obtained from European Collection of Authenticated Cell Cultures and cultured in DMEM (Dulbecco's Modified Eagle Medium) with 10% FBS. HepG2 (human hepatoma) cells were obtained from European Collection of Authenticated Cell Cultures and cultured in Eagle's Minimal Essential Medium (EMEM) with 10% added FBS and non-essential amino acids. HUVECs were obtained from Promocell, or supplied by the Queen Mary Cardiovascular Genomics and Stratified Medicine group, and cultured in Promocell's Endothelial Cell Medium with FBS-containing Endothelial Cell Growth Supplement. All cells were cultured at 37°C, 5% CO_2.

When necessary, cell numbers and viability were measured using the Digital Bio ADAM (Advanced Detection and Accurate Measurement) series automatic cell counter (Digital Bio Technology)

2.1.7 Ethanol/isopropanol precipitation

Ethanol precipitation was carried out at various stages for purification and concentration of DNA. To DNA solution, 0.1 × volume sodium acetate (NaOAc) was added. 3 × volume (after addition of NaOAc) ethanol or 0.7 × volume isopropanol was then added. DNA was incubated on ice for 15 minutes or overnight at -20°C, then centrifuged at 15,000 g for 30-45 minutes at 4°C. Supernatant

was discarded and the pellet rinsed with 70% ethanol and centrifuged at 15,000 g for 15 minutes. Supernatant was discarded and the remainder left to evaporate, and the pellet dissolved in the desired buffer.

2.2 Chapter 3: bioinformatics methods

2.2.1 Bioinformatics

1000 Genomes project data[196] with the Broad Institute's SNAP (SNP Annotation and Proxy Search)[197] was used to identify variants in strong LD ($r^2 \geq 0.8$) with rs4888378, the lead SNP identified in Gertow and colleagues' paper[119]. These variants were examined for regulatory annotations using ENCODE Project and RoadMap Epigenomics data[129,132]. The UCSC Genome Browser was used to visualise the location of variants in strong LD[127]. UCSC Genome Browser and Haploreg[133] were used to find out whether SNPs were coding SNPs or located within splice junctions. Targetscan[198] was used to assess whether SNPs were in predicted miRNA binding sites.

Variants were selected for further analysis based on the following criteria: location within narrow peaks for transcription factor binding sites (defined from ChIP-seq data), location within narrow peaks for DNase-I hypersensitivity, and within either promoter or enhancer histone signatures (defined by signatures H3K4me1, H3K4me3 and H3K27Ac). Of these, the ElDorado genomic annotation tool (Genomatix) was used to select only variants with changes in strong transcription factor binding motifs (thresholds used were core similarity 1 and matrix similarity > 0.8).

The Gilad/Pritchard eQTL browser was used to look for any QTLs at the *CFDP1-BCAR1-TMEM170A* locus. Two DNase sensitivity QTLs (dsQTLs) were taken forward for further analysis: rs73605136 and rs247454. EMSA probes were ordered for the two dsQTL SNPS, and EMSAs were carried out on the two SNPs with Huh7 cell extract as in 2.4.1.

2.2.2 eQTL analysis

The Genotype-Tissue Expression (GTEx) portal[155] was used to calculate association between rs4888378 genotype and gene expression at the locus in the two blood-vessel-related tissues (aorta and tibial artery). All genes within 200 kb were analysed, following Gertow and colleagues' approach when looking at gene expression[119], giving a total of nine: *BCAR1*, *CFDP1*, *CHST6*, *CTRB1*, *CTRB2*, *LDHD*, *TMEM170A*, *ZFP1* and *ZNRF1*. SNP-expression associations were corrected for multiple testing using the Bonferroni correction.

2.2.3 Analysis of CardioMetabochip coverage

A short analysis was carried out on the ability of the Illumina CardioMetabochip to cover the *CFDP1-BCAR1-TMEM170A* locus. The area under study was chosen as the 305 kb region containing the lead SNP that was bordered on both sides by recombination hotspots, as identified using LDLink[199] and UCSC Genome Browser[127]. Variant data for all SNPs in the region was downloaded for all subjects in the CEU population from 1000 Genomes[196]. The 21 SNPs in this region that are on the Metabochip were identified; two were monoallelic in CEU population, leaving 19 Metabochip SNPs. Pairwise LD was calculated between each Metabochip SNP and every other SNP in the region using PLINK[200]. Minor allele frequency was calculated for every SNP. For every SNP at the locus, the maximum LD with any Metabochip SNP was chosen. The proportion of variants in strong LD with any Metabochip SNP was calculated for varying cut-offs of minor allele frequency. LD (r^2) values of the SNPs with MAF > 0.05 were plotted against their location.

2.3 Chapters 4 and 7: genotyping and association analyses

2.3.1 Study cohorts

IMPROVE (IMT and IMT-Progression as Predictors of Vascular Events) is a prospective multicentre longitudinal study set up to investigate carotid intima-media thickness in individuals at high risk of CVD[201]. 3711 participants (54-79 years) with at least three vascular risk factors were recruited in seven centres in Finland, France, Italy, the Netherlands and Sweden. Vascular risk factors were defined as: male sex, or female at least 5 years post-menopausal, hypercholesterolaemia, hypertriglyceridaemia, hypoalphalipoproteinaemia, hypertension, diabetes or impaired fasting glucose, smoking habits and family history of cardiovascular diseases. IMT variables were measured using ultrasonic scans; presence of plaque was defined as maximum IMT in the whole carotid tree greater than 1.5 mm[202]. The study was designed in accordance with the rules of Good Clinical Practice (GCP), and with the ethical principles established in the Declaration of Helsinki. Each participant provided two different informed consents; one for general participation in the study and one for genotyping.

PLIC (Progressione della Lesione Intimale Carotidea) comprises 2144 general population participants attending the Centre for the Study of Atherosclerosis, Bassini Hospital (Cinisello Balsamo, MI, Italy). The study was designed to study atherosclerotic lesions and IMT in the common carotid artery, and relationships with cardiovascular risk factors. Patient characteristics, blood biomarkers and lifestyle behaviours were recorded and cIMT measured using ultrasound, as published previously[203].

Informed consent was obtained from each patient and the study protocol conforms to the ethical guidelines of the 1975 Declaration of Helsinki as reflected in a priori approval by the institution's human research committee.

Characteristics of IMPROVE and PLIC are shown in Table 25 and Table 26 (chapter 3). Three additional cohorts were used for replication of rs4888378-IMT association in men and women: Whitehall II (WHII), Edinburgh Artery Study (EAS) and the cardiovascular arm of the Malmö Diet and Cancer study (MDC), the characteristics of which have been described previously[204–206].

2.3.2 Genotyping of rs4888378 in PLIC

The SNP rs4888378 was genotyped in PLIC using Life Technologies' TaqMan SNP genotyping assays as per section 2.1.3.

2.3.3 Statistical analysis of PLIC, IMPROVE and meta-analysis

Statistical analysis of PLIC data was performed using SPSS® v.19.0 for Windows® (IBM Corporation®, Chicago Illinois, USA) by Andrea Baragetti. The Shapiro-Wilk test was used to verify normal distribution of linear variables. Statistical analyses were carried out on log-transformed cIMT variables due to skewed distributions. Association between the rs4888378 SNP and continuous variables, adjusting for age and smoking, was performed by linear regression and ANCOVA (SPSS, General Linear Model) with Bonferroni post-hoc analysis. Analyses were controlled for age and sex (where applicable).

Statistical analysis of IMPROVE data and meta-analysis was performed using R version 3.0.2[207] and PLINK version 1.9[200]. Statistical analyses were carried out on log-transformed cIMT variables due to skewed distributions. Linear regression was used to assess genotypic association with continuous variables, and logistic regression was used to assess binary traits, such as presence of plaque. The Cox proportional hazard model was used to assess time-to event data after confirming that the proportional hazard assumption was met. Genotype-phenotype analyses used an additive genetic model. Analyses in IMPROVE controlled for age, sex (where appropriate) and multidimensional scaling (MDS) coordinates.

Meta-analysis was performed to analyse genotypic association of rs4888378 with common-carotid IMT separately in men and women using a random effects model. Pooled regression coefficients with 95% confidence intervals were calculated from the five cohorts under study: PLIC, IMPROVE, Whitehall II, Malmö Diet and Cancer Study (MDC) and Edinburgh Artery Study (EAS).

LD information was not available using online tools at the time of analysis for rs1035539 as it was not present in the 1000 Genomes Pilot 1 panel. Combined genotypes for rs4888378 and rs1035539 in IMPROVE were used to calculate the correlation coefficient, r^2, as a measure of LD.

2.3.4 Exome sequencing data analysis

Exome sequencing data from the chromosome 16 locus was obtained from the NHLBI GO Exome Sequencing Project (ESP) using the Exome Variant Server[208]. The data was analysed to identify any common variants that may affect protein structure. Variants were obtained for the nine genes within 200 kb of the lead SNP; these were filtered to select only those causing a missense mutation (change in coded amino acid) or change in splice site. In order to look for only common variants, only those with a MAF ≥ 0.05 were taken forward.

2.3.5 rs1035539 genotyping in IMPROVE

The SNP rs1035539, identified using Exome Variant Server, was genotyped in the IMPROVE cohort using KBioscience's KASP SNP genotyping system (as section 2.1.4). 3.6 µl assay mix was dispensed over 5 ng dry DNA in a 384-well plate format and thermocycled according to Table 1.

Table 1: KASP thermocycling conditions.

Cycle step	Temperature (°C)	Time (s)	Number of cycles
Initial denaturation	94	960	1
Denaturation	94	20	10
Annealing/extension	65-57 (decrease of 0.8°C per cycle)	60	
Denaturation	94	20	26
Annealing/extension	57	60	

2.3.6 rs1035539 genotyping in PLIC

KBioscience's KASP SNP genotyping system was attempted to genotype rs1053359 in the PLIC cohort; however, genotype clustering was poor, giving an inadequate assay call rate. DNA was sent to KBioscience who carried out genotyping of the SNP in this cohort.

2.4 Chapter 5: functional assays

2.4.1 Electrophoretic mobility shift assay

2.4.1.1 Nuclear extract- Huh7 and HUVEC

Nuclear extract was obtained from Huh7 and HUVEC cells for EMSA. One T175 flask of cells was grown to 100% confluence. Culture medium was removed and cells washed with 15 ml 1×PBS. 0.25% trypsin-EDTA was added to the flask; 5ml for Huh7, 3ml for HUVEC. Cells were left for 1-2 minutes until they dissociated with tapping of the flask. 15 ml of culture medium with FBS was added to the flask and the cell-medium solution centrifuged at 4°C, 1500rpm, for 5 minutes. The pellet was resuspended in 5ml ice-cold buffer A (1 ml 10 M HEPES, pH 7.9; 150 ml 1.5 mM $MgCl_2$; 500 ml 10 mM KCl; and 50ml 100× protease inhibitor (Thermo Scientific Protease Inhibitor Cocktail Kit)) and left on ice for 10 minutes. Cells were again centrifuged at 4°C, 1500rpm, for 5 minutes, supernatant discarded and the pellet resuspended in 2ml buffer A (10mM HEPES (pH 7.9), 1.5mM $MgCl_2$, 10mM KCl).

Cells were vortexed and centrifuged at 4°C, 13,000rpm, for 2 minutes. Supernatant was discarded and the pellet resuspended in 800µl buffer C (2 ml 20 mM HEPES, pH 7.9; 50 mls 25% v/v glycerol; 10.5ml 0.42M NaCl; 150 ml 1.5 mM MgCl2; 40 ml 0.2 mM EDTA) with 8µl protease inhibitor. The solution was vortexed for 1 minute and left on ice for 10 minutes; this step was repeated for a total of four times. The mix was then centrifuged at 4°C and 13,000rpm for 50 minutes, and the supernatant stored at -80°C.

2.4.1.2 Production of labelled probes

EMSA probes were designed to comprise the 25 bp flanking each candidate SNP, for both reference and alternative alleles. The flanking sequence and reverse complement were ordered so as to produce double-stranded probes on annealing. Probe sequences are shown in Table 2.

Table 2: Primer sequences for EMSA probes.

SNP no	SNP name	Allele	Orientation	Primer sequence (5' – 3')
1	rs4888378	A	For	ATAAAACAAATGAATTCTGTGTTTG
		A	Rev	CAAACACAGAATTCATTTGTTTTAT
		G	For	ATAAAACAAATGGATTCTGTGTTTG
		G	Rev	CAAACACAGAATCCATTTGTTTTAT
3	rs4888379	A	For	AGTATCACCCTAACCCATTCTGAAA
		A	Rev	TTTCAGAATGGGTTAGGGTGATACT
		T	For	AGTATCACCCTATCCCATTCTGAAA
		T	Rev	TTTCAGAATGGGATAGGGTGATACT
8	rs4888392	T	For	TACAGTCTGTTTTTGGGGTCAGCTA
		T	Rev	TAGCTGACCCCAAAAACAGACTGTA
		C	For	TACAGTCTGTTTCTGGGGTCAGCTA
		C	Rev	TAGCTGACCCCAGAAACAGACTGTA
9	rs2865530	G	For	AAAAATGGTCTTGAGCTTACAGGCA
		G	Rev	TGCCTGTAAGCTCAAGACCATTTTT
		T	For	AAAAATGGTCTTTAGCTTACAGGCA
		T	Rev	TGCCTGTAAGCTAAAGACCATTTTT
14	rs3743609	G	For	CGCACACGCCTCGGGCCGGAGCGGC
		G	Rev	GCCGCTCCGGCCCGAGGCGTGTGCG
		C	For	CGCACACGCCTCCGGCCGGAGCGGC
		C	Rev	GCCGCTCCGGCCGGAGGCGTGTGCG
15	rs11643207	C	For	CTCTGCTCTGCCCTTCTCTGCAGCG
		C	Rev	CGCTGCAGAGAAGGGCAGAGCAGAG
		T	For	CTCTGCTCTGCCTTTCTCTGCAGCG
		T	Rev	CGCTGCAGAGAAAGGCAGAGCAGAG

For 30 µl of labelled probe, the reaction contained 17.7µl dH_2O, 6µl 5×TdT buffer, 3µl biotin-11-dUTP and 0.3µl probe (5 pmol/µl). Tubes were incubated at 37°C for 90 minutes. 30µl chloroform:isoamyl alcohol (24:1) was added to extract TdT, the tubes vortexed and centrifuged at 13,000 rpm for 2 minutes, and the aqueous layer saved.

The complementary forward and reverse probes were annealed to form double-stranded probes. Labelled probes were used at 5 pmol/µl and unlabelled probes at 100 pmol/µl. Equal volumes of forward and reverse primer were mixed and $1/10^{th}$ total volume 10× annealing buffer added. Primers were annealed on the thermocycling block: thermocycling started at 95°C for 3 minutes, and decreased by 5°C every three minutes until reaching a temperature of 35°C, held for 5 minutes. Annealed primers were stored at -80°C.

2.4.1.3 Polyacrylamide gel preparation

Polyacrylamide gel solution was prepared using the volumes shown in Table 3. The gel solution was poured into a gel frame of 1.5 mm thickness, the well comb added and the gel left to polymerise. Electrophoresis was run with the gel submerged in 0.5 × TBE at 4°C.

Table 3: Volumes for 1×1.5mm polyacrylamide gel.

	Volume
37.5:1 acrylamide	12ml
dH_2O	47ml
10×TBE	3ml
TEMED	50µl
25% ammonium persulphate	500µl

2.4.1.4 EMSA binding reaction and electrophoresis

EMSA binding reactions consisted of 2µl 10×binding buffer (100 mM Tris, 500mM KCl; pH 7.5), 1µl poly dIdC, 0.5µl 50mM $MgCl_2$, 3µl nuclear extract, 1.5µl labelled probe, and dH_2O to 20µl. For competitor reactions, unlabelled competitor probes were added to the reaction and left for 30 minutes at 4°C to allow nuclear proteins to bind. For both standard and competitor reactions, labelled probes were then added and left to bind at 25°C for 50 minutes.

Samples were loaded onto the 6% polyacrylamide gel with 6×loading buffer (0.25% bromophenol blue, 0.25% xylene cyanol FF, 30% glycerol in water) and electrophoresis carried out at 4°C in 0.5×TBE for 4.5 hours at 120V. The electrophoresed gel was transferred to Hybond nylon membrane (Fermentas) overnight, and the membrane crosslinked with Stratagene's UV Stratalinker 2400. The location of biotin-labelled probes was visualised according to the instructions of the ThermoScientific Lightshift Chemiluminescent Nucleic Acid Detection module. Membrane was exposed to film and the film developed using the Konica SRX101A tabletop processor.

2.4.1.5 Multiplex competitor EMSA

For multiplex competitor EMSAs, competitor probes consist of known transcription factor binding sites rather than the sequence around the SNP[209]. 70 well-characterised transcription factor consensus sequences were used for multiplex competitor EMSA. Seven sets of ten consensus sequences were used as competitors as per the competitor reactions previously. When the band was found to be competed out by a particular set of competitor sequences, the probe was incubated with each of these sequences individually. The multiplex competitor sequences used and their

corresponding transcription factors can be seen in Table 4. Multiplex competitor EMSA was carried out on the rs4888378 probe to look for a protein family competing out the protein-binding band.

Table 4: Consensus competitor sequences for multiplex competitor EMSA. F and R refer to forward and reverse oligonucleotides respectively.

Competitor probe	Sequence (5' – 3')	Competitor probe	Sequence (5' – 3')
AP1_F	CGCTTGATGACTCAGCCGGAA	Pax5_R	CGGTGGTCACGCCTCAGTGCCCCATTC
AP1_R	TTCCGGCTGAGTCATCAAGCG	Pbx1_F	CTCCAATTAGTGCATCAATCAATTCG
AP2a_F	GATCGAACTGACCGCCCGCGGCCCGT	Pbx1_R	CGAATTGATTGATGCACTAATTGGAG
AP2a_R	ACGGGCCGCGGGCGGTCAGTTCGATC	Pit1_F	TGTCTTCCTGAATATGAATAAGAAATA
AR_F	GAAGTCTGGTACAGGGTGTTCTTTTTG	Pit1_R	TATTTCTTATTCATATTCAGGAAGACA
AR_R	CAAAAAGAACACCCTGTACCAGACTTC	PPAR_F	AGGTCAAAGGTCA
Brn3_F	CACAGCTCATTAACGCGC	PPAR_R	TGACCTTTGACCT
Brn3_R	GCGCGTTAATGAGCTGTG	PR_F	GATCCTGTACAGGATGTTCTAGCTACA
CBP_F	AGACCGTACGTGATTGGTTAATCTCTT	PR_R	TGTAGCTAGAACATCCTGTACAGGATC
CBP_R	AAGAGATTAACCAATCACGTACGGTCT	RAR_F	AGGGTAGGGTTCACCGAAAGTTCACTC
CDP_F	ACCCAATGATTATTAGCCAATTTCTGA	RAR_R	GAGTGAACTTTCGGTGAACCCTACCCT
CDP_R	TCAGAAATTGGCTAATAATCATTGGGT	RXR_F	AGCTTCAGGTCAGAGGTCAGAGAGCT
CEBP_F_R	TGCAGATTGCGCAATCTGCA	RXR_R	AGCTCTCTGACCTCTGACCTGAAGCT
cMyb_F	TACAGGCATAACGGTTCCGTAGTGA	SIE_F	GTGCATTTCCCGTAAATCTTGTCTACA
cMyb_R	TCACTACGGAACCGTTATGCCTGTA	SIE_R	TGTAGACAAGATTTACGGGAAATGCAC
CREB_F	AGAGATTGCCTGACGTCAGAGAGCTAG	Smad_F	GTCTAGACCA
CREB_R	CTAGCTCTCTGACGTCAGGCAATCTCT	Smad_R	TGGTCTAGAC
CTCF_F	GGCGGCGCCGCTAGGGGTCTCTCT	Smad34_F	TCGAGAGCCAGACAAAAAGCCAGACATTTAGCCAGACAC
CTCF_R	AGAGAGACCCCTAGCGGCGCCGCC	Smad34_R	GTGTCTGGCTAAATGTCTGGCTTTTTGTCTGGCTCTCGA
E2F1_F	ATTTAAGTTTCGCGCCCTTTCTCAA	Smuc_F	GGATCCCCCAACACCTGCTGCCTGA
E2F1_R	TTGAGAAAGGGCGCGAAACTTAAAT	Smuc_R	TCAGGCAGCAGGTGTTGGGGGATCC
Egr_F	GGATCCAGCGGGGGCGAGCGGGGGCGA	Sp1_F	ATTCGATCGGGGCGGGGCGAGC
Egr_R	TCGCCCCCGCTCGCCCCCGCTGGATCC	Sp1_R	GCTCGCCCCGCCCCGATCGAAT
ER_F	GGATCTAGGTCACTGTGACCCCGGATC	SRE_F	GGATGTCCATATTAGGACATCT
ER_R	GATCCGGGGTCACAGTGACCTAGATCC	SRE_R	AGATGTCCTAATATGGACATCC
Ets_F	GGGCTGCTTGAGGAAGTATAAGAAT	Stat1_F	CATGTTATGCATATTCCTGTAAGTG
Ets_R	ATTCTTATACTTCCTCAAGCAGCCC	Stat1_R	CACTTACAGGAATATGCATAACATG
Ets1_F	GATCTCGAGCAGGAAGTTCGA	Stat3_F	GATCCTTCTGGGAATTCCTAGATC
Ets1_R	TCGAACTTCCTGCTCGAGATC	Stat3_R	GATCTAGGAATTCCCAGAAGGATC
FAST1_F	TGTGTATTCA	Stat4_F	GAGCCTGATTTCCCCGAAATGATGAGC
FAST1_R	TGAATACACA	Stat4_R	GCTCATCATTTCGGGGAAATCAGGCTC
GAS_F	AAGTACTTTCAGTTTCATATTACTCTA	Stat5_F	AGATTTCTAGGAATTCAATCC
GAS_R	TAGAGTAATATGAAACTGAAAGTACTT	Stat5_R	GGATTGAATTCCTAGAAATCT
GATA_F	CACTTGATAACAGAAAGTGATAACTCT	Stat56_F	GTATTTCCCAGAAAAGGAAC
GATA_R	AGAGTTATCACTTTCTGTTATCAAGTG	Stat56_R	GTTCCTTTTCTGGGAAATAC
Gfi1_F	TAAATCACTGC	Tbet_F_R	AATTTCACACCTAGGTGTGAAATT
Gfi1_R	GCAGTGATTTA	TFE3_F	GATCTGGTCATGTGGCAAGGC
GR_F	AGAGGATCTGTACAGGATGTTCTAGAT	TFE3_R	GCCTTGCCACATGACCAGATC
GR_R	ATCTAGAACATCCTGTACAGATCCTCT	TFEB_F	CACGTG

HIF1a_F	TCTGTACGTGACCACACTCACCTC	TFEB_R	CACGTG
HIF1a_R	GAGGTGAGTGTGGTCACGTACAGA	TFIID_F	GCAGAGCATATAAAATGAGGTAGGA
ISRE_F	AAGTACTTTCAGTTTCATATTACTCTA	TFIID_R	TCCTACCTCATTTTATATGCTCTGC
ISRE_R	TAGAGTAATATGAAACTGAAAGTACTT	TGIF_F	ACTCTGCCTGTCAAGCGAGG
HNF4_F	CTCAGCTTGTACTTTGGTACAACTA	TGIF_R	CCTCGCTTGACAGGCAGAGT
HNF4_R	TAGTTGTACCAAAGTACAAGCTGAG	TR_F	AGCTTCAGGTCACAGGAGGTCAGAGAG
IRF1_F	GGAAGCGAAAATGAAATTGACT	TR_R	CTCTCTGACCTCCTGTGACCTGAAGCT
IRF1_R	AGTCAATTTCATTTTCGCTTCC	USF1_F	CACCCGGTCACGTGGCCTACACC
MEF1_F	GATCCCCCCAACACCTGCTGCCTGA	USF1_R	GGTGTAGGCCACGTGACCGGGTG
MEF1_R	TCAGGCAGCAGGTGTTGGGGGGATC	VDR_F	AGCTTCAGGTCAAGGAGGTCAGAGAGC
MEF2_F	GATCGCTCTAAAAATAACCCTGTCG	VDR_R	GCTCTCTGACCTCCTTGACCTGAAGCT
MEF2_R	CGACAGGGTTATTTTTAGAGCGATC	YY1_F	CGCTCCCCGGCCATCTTGGCGGCTGGT
MIBP1_F	TCTTTTCCCA	YY1_R	ACCAGCCGCCAAGATGGCCGGGGAGCG
MIBP1_R	TGGGAAAAGA	ZEB_F	GATCTGGCCAAAGGTGCAGGATC
MycMax_F_R	GGAAGCAGACCACGTGGTCTGCTTCC	ZEB_R	GATCCTGCACCTTTGGCCAGATC
NF1_F	TTTTGGATTGAAGCCAATATGATAA	HNF1_F	GTTAATGATTAAC
NF1_R	TTATCATATTGGCTTCAATCCAAAA	HNF1_R	GTTAATCATTAAC
NFE2_F	TGGGGAACCTGTGCTGAGTCACTGGAG	ARP1_F	AGGTGACCTTTGCCCA
NFE2_R	CTCCAGTGACTCAGCACAGGTTCCCCA	ARP1_R	TGGGCAAAGGTCACCT
NFATc_F	CGCCCAAAGAGGAAAATTTGTTTCATA	NFY_F	ATCAGCCAATCAGAGC
NFATc_R	TATGAAACAAATTTTCCTCTTTGGGCG	NFY_R	GCTCTGATTGGCTGAT
NFkB_F	AGTTGAGGGGACTTTCCCAGGC	HNF3_F	GCCCATTGTTTGTTTTAAGCC
NF-kB_R	GCCTGGGAAAGTCCCCTCAACT	HNF3_R	GGCTTAAAACAAACAATGGGC
NR5A2_F	GATCAACGACCGACCTTGAG	BARP_F	TCACTCAAGTTCAAGTTATT
NR5A2_R	CTCAAGGTCGGTCGTTGATC	BARP_R	AATAACTTGAACTTGAGTGA
OCT1_F	TGTCGAATGCAAATCACTAGAA	SREBP1_F	TTTGAAAATCACCCCATGCAAACTC
OCT1_R	TTCTAGTGATTTGCATTCGACA	SREBP1_R	GAGTTTGCATGGGGTGATTTTCAAA
p53_F	TACAGAACATGTCTAAGCATGCTGGGG	HSF1_F	GATCTCGGCTGGAATATTCCCGACCTGGCAGCCGA
p53_R	CCCCAGCATGCTTAGACATGTTCTGTA	HSF1_R	TCGGCTGCCAGGTCGGGAATATTCCAGCCGAGATC
Pax5_F	GAATGGGGCACTGAGGCGTGACCACCG		

2.4.1.6 Supershift EMSA

Supershift EMSA was used on the rs4888378 probe to verify the protein binding to the sequence. The assay was carried out as in a competitor EMSA above, but instead of adding unlabelled probes at the incubation step, 2µl anti-FOXA2 or 2µl anti-HFH4 antibody was added and incubated as before. A positive control was performed by using a labelled probe consisting of the consensus sequence for the HNF3 (FOXA) family, both incubated with the antibody and without.

2.4.1.7 Oestrogen-related SNP EMSAs

SNPs in strong LD ($r^2 \geq 0.8$) with the IMT lead SNP (rs4888378) were again analysed using the bioinformatic tool Haploreg v3. SNPs were then selected for analysis if they had ChIP-seq evidence of ER-α (estrogen receptor alpha) binding to the SNP, or if they changed an ER-α binding motif. Four SNPs were taken forward for analysis: rs2161648, rs4888400, rs2285222 and rs11149832.

Forward and reverse probes were ordered for the four SNPs, and EMSAs were carried out using Huh7 nuclear extract as in section 2.4.1. The probe for oestrogen receptor (ER), used as a competitor in multiplex competitor EMSA, was also biotin-labelled and used as an EMSA probe.

2.4.1.8 EMSA with oestrogen-stimulated cell extract

Oestrogen-stimulated Huh7 extract was produced for EMSAs with the oestrogen-related SNPs. Huh7 cells were grown in a T175 flask with 30 ml medium to 90-100% confluency. 8 µl β-estradiol (20 µg/ml) was added to the cell medium, the medium swirled to mix, and the cells incubated as normal for 30 minutes. The cells were then washed with PBS and nuclear extraction proceeded as in 2.4.1.1. EMSAs were carried out on the four oestrogen-related SNPs as in 2.4.1, using both standard and oestrogen-stimulated Huh7 extract. NF-κB and ER labelled probes were used as positive controls.

2.4.2 Luciferase reporter assay

Luciferase reporter assays were used to assess the impact of rs4888378 allele on gene expression.

2.4.2.1 Primer design

The In-Fusion Primer Design Tool[210] was used to design primers to amplify a region of 378 bp around the SNP (fragment 1), and to create *SalI* and *BamHI* restriction sites at each end of the sequence. After analysis of results, Genomatix's ElDorado tool[211] was used to predict binding sites in fragment 1 for proteins classified as repressor or repressor-related, and three further fragments were designed to include or exclude these binding sites. Primer sequences can be seen in Table 5.

Table 5: Cloning primers for luciferase sequence fragments. Primers were designed to incorporate restriction sites for *BamHI* and *SalI*; the pGL3-promoter vector and sequence fragments were digested with these enzymes to allow ligation.

Fragment	Forward/ reverse	Sequence (5' – 3')	Restriction site
1	Forward	AATCGATAAGGATCCATGTCAGTACAAGGCGAGCA	*BamHI*
1	Reverse	AAGGGCATCGGTCGACGGGTTCCCAGATACAGGTTG	*SalI*
2	Forward	AAATCGATAAGGATCCCTGAAGTCCTCCAAAACCAGA	*BamHI*
2	Reverse	AAGGGCATCGGTCGACTGCTTTTACGTCATCAGCAATCT	*SalI*
3 and 4	Forward	AAATCGATAAGGATCCCCTCCAGGCTTTGTGTATAAGG	*BamHI*
3	Reverse	AAGGGCATCGGTCGACATGCTTGGGACTGGAAGTGT	*SalI*
4	Reverse	AAGGGCATCGGTCGACATTTGCTTTTACGTCATCAGCA	*SalI*

2.4.2.2 Amplification of rs4888378 fragment

As no polymorphic variants are present in the region flanking rs4888378, the fragment was amplified from DNA of AA and GG genotype. rs4888378 creates an *EcoRI* restriction site; therefore, the strand can be cut when the A allele is present, but not the G allele. A diagnostic digest with *EcoRI* was therefore used to genotype DNA samples and AA and GG homozygotes were selected for PCR. A 96-well plate of test DNA was amplified using the volumes shown in Table 6.

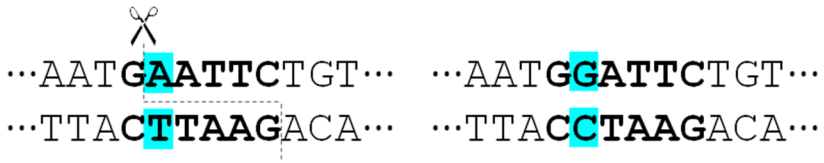

Figure 16: *EcoRI* cutting site disrupted by rs4888378.

Table 6: Volumes for amplification of test DNA.

Reagent	Volume per reaction (µl)
5×GC buffer	2
dNTPs	0.2
Primer luc1_for	0.05
Primer luc1_rev	0.05
DNA	10ng (dried)
Phusion polymerase	0.1
H_2O	7.6

Amplified DNA was digested with EcoRI as per Table 7, and digested for 2 hours at 37°C. Digested DNA was visualised on an agarose gel to identify AA and GG homozygotes.

Table 7: *EcoRI* digest.

Reagent	Volume per reaction (µl)
10× EcoRI buffer	1
EcoRI	0.15 (3 U)
H_2O	0.85
DNA	8

Samples of AA and GG genotype were selected. The samples were amplified in a 50 µl reaction using the appropriate primer pairs to produce fragment 1, 2, 3 and 4 (Table 8).

Table 8: Volumes for amplification of luciferase fragment.

Reagent	Volume per reaction
5×GC buffer	4
dNTPs	0.4
Forward primer	0.1
Reverse primer	0.1
Phusion	0.2
H_2O	10.2
DNA	5

2.4.2.3 Digest of pGL3-promoter

The pGL3-promoter vector was digested with *SalI* and *BamHI* to allow ligation of the insert fragment at the polylinker site. The digest was set up according to Table 9 and incubated at 37°C overnight. Digest of the insert fragments was not required, as PCR with the In-Fusion primers introduces the appropriate sticky ends at the ends of the rs4888378 fragments.

Table 9: *SalI* and *BamHI* digest.

Reagent	Volume (µl)
pGL3-promoter DNA (400ng/µl)	6
SalI	2.5 (50 U)
BamHI	2.5 (50 U)
NEB buffer 3	5
BSA	0.5
H_2O	33.5

2µl CIP (calf intestinal alkaline phosphatase) was added to the reaction after digestion to remove phosphate groups from the 5' end of the DNA strand, preventing recircularisation of the vector.

The digested vector and amplified AA and GG fragments were run on a 1% and 1.8% gel respectively, and gel-purified according to the instructions of the GE Healthcare illustra DNA and gel band purification kit.

2.4.2.4 In-Fusion cloning reaction

To insert the fragments into the linearised pGL3-promoter vector, the reaction in Table 10 was set up for each fragment. The cloning reaction was incubated for 15 minutes at 50°C.

Table 10: In-Fusion cloning reaction.

Reagent	Volume (µl)
5× In-Fusion HD Enzyme Premix	2
Linearised pGL3-promoter	50ng
rs4888378 fragment	10ng
dH_2O	To 10µl

2.4.2.5 Transformation into NEB 5-alpha competent cells

NEB 5-alpha competent *E. coli* cells (DH5α derivative) were thawed on ice and 50 µl was aliquotted into pre-chilled thin-walled tubes for the A-allele, G-allele and negative control. 4.5 µl of the A, G, or control (no insert) ligation product was added to each tube and the reactions incubated on ice for 20 minutes. The reactions were heat-pulsed in a 42°C water bath for 1 minute, then immediately placed on ice for 2 minutes. The contents were added to 500 µl LB broth warmed to 37°C in a 1.5 ml tube and the contents shaken at 37°C for 1 hour. Transformed cells in LB were then spread onto LB agar plates with 50 µg/ml ampicillin, then incubated at 37°C overnight. Plates were then checked for

colonies; if colonies were present on the A and G plates and not on the negative control, A and G colonies were picked for plasmid purification.

2.4.2.6 Plasmid purification (mini-prep)

Four colonies each were picked from the A and G plates for each fragment. Each colony was inoculated into 500µl LB containing 100 µg/ml ampicillin, and shaken overnight at 37°C. Bacterial cells were first harvested by centrifugation at 8000 rpm at room temperature for 3 minutes. DNA was then purified following the instructions of the Qiagen Spin Mini-prep kit, and eluted in 30 µl dH_2O.

2.4.2.7 Diagnostic digest

A diagnostic digest was performed on the mini-prep DNA with *BamHI* and *SalI* to verify that the fragment had been successfully inserted into the plasmid, as per Table 11. Digests were carried out in a reaction of 10µl for 2 hours at 37°C. Digest products were visualised on an agarose gel to check that two products of the expected sizes were present (5000 bp for the cut plasmid and 100-380 bp for the inserted fragments). Samples were then sent for sequencing to confirm the correct allele of rs4888378 was present and that the remainder of the sequence was correct. For each sample, 10 µl DNA was sent for sequencing at a concentration of 100 ng/µl, with primers diluted to 3.3 pmol/µl.

Glycerol stocks were made of bacterial cultures the pGL3-promoter DNA with the four inserted fragments of A and G genotype, for long-term storage of plasmid DNA. 500 µl bacterial culture was added to 500 µl glycerol and the stocks stored at -80°C.

Table 11: Diagnostic digest for luciferase fragment insertion.

Reagent	Volume per reaction (µl)
Plasmid DNA	2
SalI	0.5 (20 U)
BamHI	0.5 (20 U)
Buffer 3	1
BSA	0.1
H_2O	5.9

2.4.2.8 Plasmid purification (maxi-prep)

Maxi-prep was carried out to obtain large quantities of DNA for transfection. Transformed NEB5α culture (one of each for allele A and G) was streaked out on LB-agar plates and incubated at 37°C

overnight. A colony from each plate was inoculated into 150ml LB with 300µl ampicillin, and shaken at 37°C overnight. Bacterial culture was centrifuged at 5000 g for 10 minutes. Plasmid DNA was then purified according to the instructions of the GenElute HP Plasmid Maxiprep Kit, and eluted in 3ml elution solution. To concentrate DNA, the solution was ethanol precipitated (as section 2.1.7) and DNA resuspended in 300 µl H_2O. DNA concentration was measured using the Nanodrop.

2.4.2.9 Transfection

Huh7 cells were seeded into a 96-well plate 24 hours before transfection at 1.6×10^6 cells per well, a density previously ascertained to produce 80-90% confluency in this time. Cell medium was removed and the cells washed with PBS. 2ml trypsin was added for 3 minutes and 12ml EMEM then added to stop the reaction. Cells were counted and 1.6×10^6 cells added to each well.

Plasmid DNA and controls for transfection were used at a concentration of 400ng/µl. puc18 was used as a background reading (negative control), and pGL3-promoter as a positive control. pRL-TK was used as a co-transfection control, at 1:200 ratio, in order to normalise for variable transfection efficiency.

Two Opti-MEM based mixes were made up: master mix was made up with 12 µl pRL-TK (10 ng/µl) and 2928 µl Opti-MEM, and lipid mix was made up with 66 µl Lipofectamine 2000 and 3234 µl Opti-MEM. 343 µl master mix or pure Opti-MEM was added to 7 µl of each vector DNA stock as in Table 12. Lipid mix was prepared and incubated at room temperature for 5 minutes. 350 µl lipid mix was then added to each tube of vector DNA, creating a 700 µl transfection mix for each construct, and the resulting transfection mixes incubated at room temperature for 20 minutes. One row of a 96-well plate of Huh7 cells was used for each construct. 50 µl transfection mix was added to each well of the appropriate row. The cells were then cultured in standard conditions for 48 hours.

Table 12: Volumes of vector DNA and Opti-MEM mixes added for luciferase reporter assay transfection. Opti-MEM rather than master mix was used for the puc18 control, which was assayed without co-transfectant. Vector DNA was used at a concentration of 400 ng/µl.

Construct	Volume	Mixes
puc18	7µl	343µl Opti-MEM
pGL3control	7µl	343µl master mix
pGL3promoter	7µl	343µl master mix
Fragments 1-4, alleles A and G (8 total)	7µl	343µl master mix

2.4.2.10 Measurement of luciferase expression

LARII substrate was reconstituted using LARII buffer and Stop & Glo reagent according to the manufacturers' instructions. Cells were washed twice with 1×PBS, and lysed by addition of 20µl 1× Promega lysis buffer and 20 minutes' shaking. Luciferase was measured using Applied Biosystems' TR717 Microplate Luminometer. T-tests were used to compare expression between insert fragment and control DNA, and between alleles of the insert fragment.

2.5 Chapter 6: chromosome conformation capture

2.5.1 Optimisation of protocols

As current 4C protocols differ widely, many aspects of the method were tested and optimised before a final protocol was decided on. Areas of optimisation are described below. Where not otherwise stated, methods following the finalised protocol (section 2.5.2).

2.5.1.1 Crosslinking protocol

Cells were cultured and trypsinised as in section 2.5.2.1, then resuspended in 11.2 ml cell medium. To this solution, 312 µl or 624 µl 37% formaldehyde was added to the cells for a final concentration of 1% or 2% respectively. After 10 minutes the formaldehyde crosslinking reaction was quenched with the addition of 625 µl 2.5 M glycine for a final concentration of 129 mM. Treatment of crosslinked cells then proceeded as per the final protocol. After cell lysis, the molecular weight of crosslinked DNA was visualised on a gel and compared.

2.5.1.2 Primary restriction enzyme

SDS is required for the permeabilisation of nuclei in cross-linked cells, but its presence inhibits the action of restriction enzymes. While addition of detergents such as Triton X-100 quenches SDS to an extent, remaining SDS still impairs efficient digestion. Various primary restriction enzymes were therefore tested for their ability to digest crosslinked DNA in the presence of quenched SDS. Restriction enzymes were initially tested on genomic DNA in the presence of SDS and Triton X-100. *HindIII*, *BamHI* and *DpnI* were tested in 50 µl reactions consisting of 1 µg plasmid DNA (Promega's pGL3-promoter reporter vector), 1 µl enzyme and variable SDS and Triton X-100 (0-0.3% SDS; 0-1.8% Triton X-100).

Enzymes were also tested in various combinations on crosslinked cell DNA. *HindIII*, *BamHI* and *EcoRI* were tested on Hek293 and HUVEC crosslinked DNA, following the method in the final protocol, and digestion efficiency assessed by running undigested and digested controls on an agarose gel.

Subsequent restriction enzyme testing focused on the 4-cutter enzymes *DpnII*, *CviQI* and *Csp6I*. Hek293 and HUVEC DNA was again digested with these enzymes and the digestion products run on an agarose gel.

2.5.1.3 SDS and Triton X-100 concentration

Different concentrations of SDS and Triton X-100 were tested in the primary restriction digest in order to find concentrations that provided a sufficient yet quenchable volume of SDS. SDS was tested at a concentration of 0% (as a control), 0.1%, 0.2% and 0.3%, while Triton X-100 was used at a concentration of 0.6%, 1.2% and 1.8%. Digestion was tested on genomic and crosslinked cell DNA.

2.5.1.4 First ligation

A number of parameters were tested for the first ligation; in each case, ligation products were compared on an agarose gel. The reaction (as in the final protocol) was tested with and without the presence of 40 µl 10 mg/ml BSA. Ligation efficiency was compared on an agarose gel.

2.5.1.4.1 Primer design (initial)

After viewpoints and secondary fragments had been chosen and the efficacy of the restriction enzymes verified, primers for inverse PCR were designed. For each secondary fragment a forward and reverse primer were designed, referred to as the 'read' (primary restriction site) and 'non-read' (secondary restriction site) primer. In contrast to standard PCR, the primers were designed facing outwards from the secondary fragment so that when they have ligated into PCR circles, they amplify the captured fragment and the rest of the circle (Figure 17).

Primers were designed using Primer3Plus[212] in order to create primers with suitable length, GC content and melting temperatures (Tms), avoid primer dimer and hairpin formation, and check that primers do not match anywhere else in the genome. To create primers facing outwards, the sequence was rearranged: the 50 bp flanking the primary restriction site and the 100 bp flanking the secondary site were taken, swapped around and separated by a 100 bp stretch of N nucleotides to represent the captured sequence (Figure 17). Primer3plus was used to design a pair of primers to amplify the stretch of Ns. The read primer was designed as close as possible to the primary restriction site, while the nonread primer could be anywhere in the 100 bp region. Secondary fragments for which a suitable pair of primers could not be designed were discarded. For each sequence that allowed suitable primers, two pairs were ordered for testing on 4C library DNA.

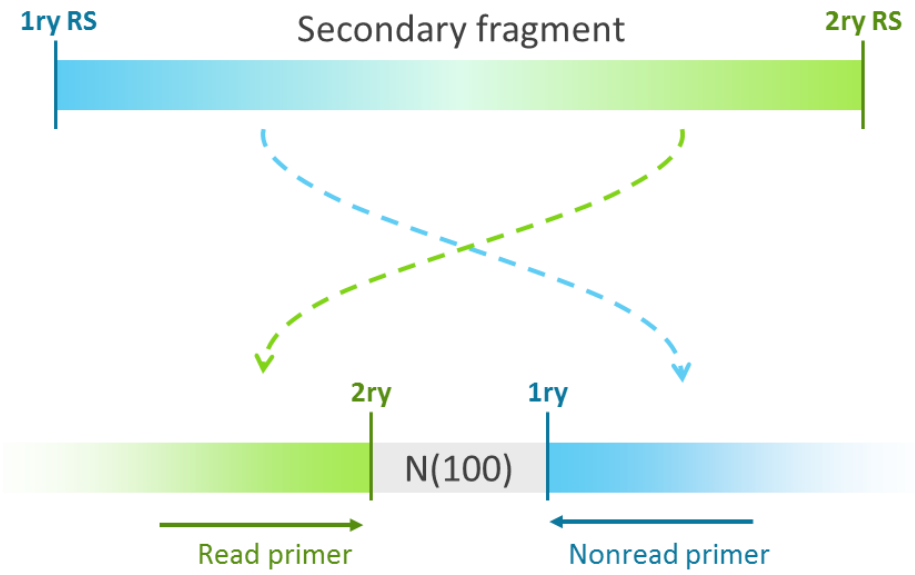

Figure 17: Sequence rearrangement before primer design. The sequence of the secondary fragment is taken and rearranged into a form suitable for conventional primer design before being used in Primer3Plus. The 50 bp adjoining the primary restriction site and the 130 bp adjoining the secondary restriction site are swapped over and connected with a 100 bp stretch of N-nucleotides, which serve as the placeholder for the unknown ligated fragment in the DNA circle. Primers are then designed to amplify this stretch of Ns. The available sequence for primer design by the secondary restriction site is longer because this primer will not be used to produce a sequencing read, so does not need to be as close to the restriction site.

2.5.1.5 4C PCR

2.5.1.5.1 Primer testing without adapters

Many variables were tested for amplification of 4C library DNA. Test reactions were carried out in reaction volumes of 50 µl, following the proportions in the final protocol. Test PCR reactions were carried out on the appropriate 4C library DNA using three different polymerases: NEB's Phusion High-Fidelity DNA Polymerase, Acqua Science's PCR Master Mix and Roche's Expand Long Template PCR System.

PCR reactions were tested with template 4C library DNA at volumes of 10 ng, 20 ng and 40 ng, in order to find the optimum amount of template DNA (the point beyond which the amount of PCR product does not markedly increase with an increase in template). Different thermocycling conditions were assessed. A gradient of annealing temperatures from 58-66°C were tested, and total number of cycles was varied from 30 to 35. An NTC was included in each reaction to check for contamination and primer dimer. Genomic DNA was also included as a control; as primers face outwards from the secondary fragment, the product should not amplify from genomic DNA. Controls with only the "read" primer were also included to check for non-specific binding of primers. Again,

no product should amplify using only this primer. The final set of primers before addition of sequencing adapters is shown in Table 13.

Table 13: Primer names and properties. Two pairs of primers were tested for each secondary fragment, where the sequence was amendable to primer design.

Viewpoint	Primer name	Site	Read/non-read	Secondary fragment side	Sequence (5' – 3')	Length	Tm
BCAR1-1 *DpnII*	11_*DpnII*_R_*NlaIII*_Read	*DpnII*	Read	Left	CATTGTGCCGACATCTGG	18	59.6
	11_*DpnII*_R_*NlaIII*_Non	*NlaIII*	Non	Right	CCCCTTTCCCTCTAGATTCA	20	58.2
	11_*DpnII*_R_*NlaIII*_Non2	*NlaIII*	Non	Right	GAGCTGTGGTGGTGATGATT	20	58.5
BCAR1-1' *Csp6I* (left side)	11p_*Csp6I*_L_*DpnII*_Read	*Csp6I*	Read	Right	AAGATGTCCGTGCCTGTG	18	58.1
	11p_*Csp6I*_L_*DpnII*_Read2	*Csp6I*	Read	Right	CAAGATGTCCGTGCCTGT	18	58.1
	11p_*Csp6I*_L_*DpnII*_Non	*DpnII*	Non	Left	ACCTAGGCCTTTTTCTGTCG	20	58.5
BCAR1-1' *Csp6I* (right side)	11p_*Csp6I*_R_*DpnII*_Read	*Csp6I*	Read	Left	GTGGCTCCTTATCTCCCTGT	20	58.2
	11p_*Csp6I*_R_*DpnII*_Read2	*Csp6I*	Read	Left	GGGAAGTGGCTCCTTATCTC	20	57.8
	11p_*Csp6I*_R_*DpnII*_Non	*DpnII*	Non	Right	TTCCCTATAGGACGGAGTGA	20	57.2
SNP *DpnII*	SNP_*DpnII*_R_*Csp6I*_Read	*DpnII*	Read	Left	CACAAAAACACCTGGTCTCC	20	58
	SNP_*DpnII*_R_*Csp6I*_Read2	*DpnII*	Read	Left	ACAAAAACACCTGGTCTCCA	10	58
	SNP_*DpnII*_R_Csp6_Non	*Csp6I*	Non	Right	AGAAACTGCCCTTCCAGTCT	20	58
SNP *Csp6I*	SNP_*Csp6I*_L_*NlaIII*_Read	*Csp6I*	Read	Right	GGGACCCAAGTTTAAACAAA	20	57.9
	SNP_*Csp6I*_L_*NlaIII*_Read2	*Csp6I*	Read	Right	GGACCCAAGTTTAAACAAACA	21	56.7
	SNP_*Csp6I*_L_*NlaIII*_Non	*NlaIII*	Non	Left	TAGTCACCCTGGCTGAAGTC	20	56.6

2.5.1.5.2 PCR optimisation with adapters

After confirmation of working primer pairs, these primers were ordered from Eurofins Genomics with the appropriate "read" or "nonread" Illumina sequencing adapter on the 5' end.

Read adapter: AATGATACGGCGACCACCGAGATCTACACACACTCTTTCCCTACACGACGCTCTTCCGATCT

Nonread adapter: CAAGCAGAAGACGGCATACGAGATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC

Primer concentration was tested using the new adapter primers. Primers were used in the PCR reactions at 200 pmol to 750 pmol. Reactions were again tested using genomic DNA template control and NTCs, and with 10-40 ng of template DNA. Thermocycling extension time was varied from 2 to 3 minutes per cycle.

2.5.1.6 Next-generation sequencing

Next-generation sequencing of amplified 4C libraries was tested with and without custom sequencing primers. Custom sequencing primers used were the applicable read primers for each amplified viewpoint.

2.5.2 Final protocol

2.5.2.1 Cell culture and crosslinking

Hek293 and Huh7 cells were cultured in DMEM with 10% FBS, while HUVECs were grown in Promocell Endothelial Growth Medium with serum-containing growth supplement. Between 12 and 18 T-175 flasks were cultured to 90-100% confluence. Fresh 2.5 M glycine was prepared and filtered through 0.22 µm filter membrane. Medium was removed from cells and PBS added for 3 minutes to wash the cell surface. Cells were trypsinised and collected in 15 ml tubes, then spun down at 400 g for 4 minutes. Cells were resuspended in 2 ml medium each, then consolidated into one tube of cells and spun down again as before. Cells were resuspended in 11.2 ml medium for a final volume of 11.25 ml. 50 µl was removed and set aside for counting cells.

312 µl 37% formaldehyde was added to the cells for a final concentration of 1%, and the solution mixed for 10 minutes by gentle rocking on the rocking mixer. While cells were crosslinking, some of the cells removed for crosslinking were diluted 1 in 10 in cell medium in order to increase the range of countable cells. The diluted and undiluted cells were counted. After 10 minutes, the formaldehyde crosslinking was quenched with the addition of 625 µl 2.5 M glycine for a final concentration of 129 mM. The solution was mixed for 5 minutes on the rocking mixer, then placed on ice for 15 minutes. The cells were centrifuged at 480 g for 10 minutes. Supernatant was removed and the cells resuspended in PBS. The cell solution was then divided into 1.5 ml tubes in order to contain 10-20 million cells per tube. The tubes were centrifuged at 480 g for 5 minutes at 4°C. Supernatant was removed and the pellets flash-frozen in dry ice, then stored at -80°C.

2.5.2.2 Cell lysis and first digestion

Cell lysis buffer was made up with 10 mM Tris-HCl (pH 8.0), 10 mM NaCl, 0.2% Igepal CA-630 and 1× Thermo Halt Protease Inhibitor Cocktails. 1 × digestion buffer was also made up and pre-chilled, using the relevant digestion buffers for the restriction enzymes being used. Primary restriction enzymes used were DpnII (with its unique DpnII buffer) and Csp6I (with Thermo Fisher Buffer B). Cell pellets were thawed and resuspended in 660 µl chilled lysis buffer, and the solution mixed by pipetting, then homogenised using a plastic microcentrifuge tube pestle. The solution was left on ice for 15 minutes.

A sample of 3 µl of the solution was taken and mixed with 3 µl methyl green-pyronin stain on a microscope slide and covered with a coverslip. Cells were then visualised under the microscope to check that individual nuclei (stained blue-green) had been isolated. If clumps of cells were still

visible, the cell solution was homogenised with the pestle and left for another 15 minutes. The cell solutions were then centrifuged at 2300 g at 4°C for 5 minutes. Supernatant was removed and the pellet washed with 500 µl chilled 1 × digestion buffer and spun again. The wash step was repeated once more. The final pellet was resuspended in 392 µl 1 × digestion buffer, to which 10 µl 10% SDS was added, to a final concentration of 0.2%. The mixture was incubated at 37°C for 1 hour, then in a 65°C water bath for 10 minutes, with agitation every 2 minutes. 44.4 µl 20% Triton X-100 to a final concentration of 2% was added to quench SDS, and the solution shaken at 37°C for 1 hour.

20 µl **Undigested control sample** was taken at this point to compare sizes of digested and undigested DNA, and stored at -20°C until use. Here 200 U of the primary restriction enzyme was added, and the solution incubated for 4 hours at 37°C with shaking. 200 U more enzyme was added at this point, and the solution incubated at 37°C with shaking overnight. Parafilm was used to seal the top of the tube to prevent evaporation.

In the morning, 20 µl digested control was taken. Controls were processed as in 2.5.2.3. Digestion was stopped using heat-inactivation, if the primary restriction enzyme could be heat-inactivated, or addition of SDS. For heat-inactivation, the solution was placed in a 65°C water bath for 20 minutes, manually mixing the tube every few minutes. For SDS inactivation, SDS was added to a final concentration of 1.5% and the solution again heated at 65°C for 20 minutes with occasional agitation.

2.5.2.3 Controls

To 20 µl controls the following were added: 76.5 µl 10 mM Tris-HCl, 2.5 µl Proteinase K (20 mg/ml) and 1 µl RNase A (100 mg/ml). Crosslinks were reversed at 65°C for 4 hours. Controls were then phenol/chloroform extracted following the instructions of the 5 Prime heavy phase lock gel. Controls were then run on a 1% agarose gel to confirm that the undigested control produced a single high-weight band, while the digested control produced a smear of lower weights.

2.5.2.4 First ligation

Digested DNA was transferred to a 15 ml tube and the reagents shown in Table 14 added. The ligation was then incubated at 16°C overnight.

Table 14: Volumes for first 4C ligation.

Reagent	Volume (µl)
20% Triton X-100	1200
10× NEB ligation buffer	800
BSA	80
dH_2O	4900
T4 DNA ligase	20 (120 U)
Total	7000

25 µl 20 mg/ml Proteinase K was added to the solution and the tube lid wrapped with parafilm to prevent evaporation. The tube was incubated for 6 hours or overnight at 65°C to reverse cross-links. 2.5 µl 100 mg/ml RNase A was added and the tube incubated at 37°C for 45 minutes. Ligated DNA was then phenol/chloroform extracted following the instructions of the 5 Prime 15 ml heavy phase lock kit. The result was ethanol precipitated and resuspended in 150 µl 10 mM Tris-HCl pH 7.5. This DNA was referred to as the "3C library". 2 µl 3C library was run on an agarose gel to check ligation efficiency.

2.5.2.5 Second digestion

After controls were processed to check for successful primary digestion and ligation, ligated DNA was digested with the secondary enzyme. At this point DNA was no longer cross-linked and no SDS was present, so restriction digestion could proceed as standard. 1 µl of 3C library DNA was run on a 1.5% agarose gel alongside reference samples of human genomic DNA (200, 500 and 1000 ng) and bands quantified using SynGene GeneTools, in order to estimate DNA concentrations of the 3C library.

25-50 µg of 3C library DNA was digested overnight with the secondary restriction enzyme: *DpnII*, *Csp6I* or *NlaIII*. Digests contained 50 µl 10× applicable digestion buffer, 25-50 µg 3C library DNA, 5 µl secondary restriction enzyme and dH_2O to a total volume of 500 µl. The digest reactions were incubated at 37°C overnight.

5 µl second digest control was taken. 35 µl 10 mM Tris-HCl was added to dilute the sample and half of this run on a 1.5% agarose gel to check digestion efficiency.

If DNA had digested sufficiently, the reaction was phenol/chloroform extracted using a 1.5 ml heavy phase lock tube, then ethanol precipitated, adding glycogen to a concentration of 0.4 µg/µl before

addition of sodium acetate and ethanol. DNA was resuspended in 100 µl dH_2O, and the concentration measured on the Nanodrop.

2.5.2.6 Second ligation

The digested DNA was transferred to a 50 ml tube, to which the reagents in Table 15 were added. The ligation reaction was incubated at 16°C for 4 hours/overnight.

Table 15: Volumes for second 4C ligation.

Reagent	Volume (µl)
Digested 3C library	100
10 × T4 ligase buffer	1400
T4 ligase	20 (120 U)
Milli-Q	12480
Total	14000

After ligation, the solution was phenol/chloroform extracted using 50 ml heavy phase lock tubes. To increase yield, the solution was back-extracted: the aqueous layer was transferred to a new tube, the phase lock gel pierced with a glass rod and an equal volume of dH_2O added, mixed and centrifuged as before. The aqueous layer was transferred to the new tube. The solution was ethanol precipitated, adding 0.4 µg/µl before addition of sodium acetate and ethanol. DNA was resuspended in 100 µl 10 mM Tris-HCl.

The final ligated DNA was purified using the illustra GFX PCR DNA and Gel Band Purification Kit, eluting in 50 µl 10 mM Tris-HCl. This is referred to at the 4C library. The DNA concentration was measured using the Nanodrop and the 4C library samples stored at -20°C.

2.5.2.7 Amplification of 4C library

After optimisation of 4C-PCR, amplification reactions were scaled up. Each 4C library was amplified in 16 reactions of 50 µl, using the Sigma-Aldrich Expand Long Template PCR System. The reaction was set up as shown in Table 16. Thermocycling conditions are shown in Table 17.

Table 16: Volumes for large-scale amplification of 4C library.

Reagent	Volume (µl)
Expand Long Template Buffer 1	80
dNTPs (10 mM)	16
Primer 1 (10 pmol/µl)	80
Primer 2 (10 pmol/µl)	80
Expand Long Template Polymerase Mix	12
4C library (10 ng/µl)	160
dH_2O	372
Total	800

Table 17: 4C PCR thermocycling conditions.

Cycle step	Temperature (°C)	Time (s)	Number of cycles
Initial denaturation	95	120	1
Denaturation	94	10	
Annealing	55	60	35
Extension	68	180	
Final extension	68	300	1

2.5.2.8 Quantification of amplified 4C library

Amplified 4C samples were quantified on the Agilent 2100 BioAnalyzer using the Agilent DNA 1000 kit. 4C DNA was diluted to 20 and 40 ng/µl, and the chip was loaded and prepared according to the kit instructions. Data was recorded and analysed using Agilent 2100 Expert Software. After confirming that suitable peaks were recorded for the DNA ladder and upper and lower markers, DNA size and amount was recorded.

2.5.2.9 Quantification of sequenceable DNA

To check for the presence of a DNA library that can be suitably sequenced with Illumina next-generation sequencing chemistry, the New England Biolabs NEBNext Library Quant Kit for Illumina was used. This qPCR kit checks for the presence of the Illumina sequencing adapters, and allows quantification of the sequenceable library using supplied DNA standards.

Following the instructions of the kit, amplified 4C library was diluted in 1× NEBNext Library Quant Dilution Buffer to 1:1,000, 1:10,000 and 1:100,000. The qPCR reactions were prepared in a 384-well plate as per the instructions. The qPCR assay was run on the ABI 7900HT Fast Real-Time PCR System using the thermocycling conditions shown in Table 18. The DNA standard curve was produced using

the amplified DNA standard data, and the concentration of undiluted sequenceable DNA in each sample calculated using the standard curve.

Table 18: Thermocycling protocol for qPCR of 4C libraries. Thermocycling followed the instructions of the New England Biolabs NEBNext Library Quant Kit for Illumina.

Cycle step	Temperature (°C)	Time (s)	Cycles
Initial Denaturation	95	60	1
Denaturation	95	15	35
Extension	63	45	

2.5.2.10 Next generation sequencing of amplified 4C libraries

The final amplified 4C libraries were sequenced to produce sequence reads that could be mapped to the genome to identify interacting genomic regions. Sequencing was carried out on the Illumina MiSeq Desktop Sequencer, with the MiSeq Reagent Kit v2.

Following the kit instructions, amplified 4C samples were diluted to 4 nM in 0.2 N NaOH, then to a final concentration of 12 pM in HT1 buffer. After the first sequencing attempt, 20% and later 50% 2nM PhiX control was added to the samples to increase base diversity at the start of the sequence reads. One sequencing run used custom primers: the PCR "read" primers with linked Illumina adapter sequence. These were diluted to 0.5 µM and loaded onto the cartridge as per the manufacturer's instructions. Subsequent sequencing runs used only the standard Illumina sequencing primers supplied in the cartridge. After loading the DNA samples and custom primers onto the reagent cartridge, the cartridge was inserted into the MiSeq and the sequencing run started.

2.6 Chapter 8: protein assays

2.6.1 General methods for protein assays

2.6.1.1 Cell lysis

Immediately after treatment with media or growth factors, cell plates were placed on ice and medium removed by aspiration. Cells were washed twice with PBS. Cells were then lysed with addition of RIPA buffer (30mM Tris/HCl, pH 7.4, 150mM NaCl, 1% Nonidet P40, 0.5% sodium deoxycholate and 2 mM EDTA) with 1 × protease inhibitor cocktail and 1 × phosphatase inhibitor cocktail. 80 µl RIPA was used for each well of a six-well plate and 50 µl for each well of a 12-well plate. Cell were scraped and pipetted into 1.5 ml microcentrifuge tubes on ice. Samples were

vortexed for ten seconds and centrifuged at 4°C for 10-15 minutes at 16,000 g to pellet cell debris. Supernatant was transferred to a fresh tube.

2.6.1.2 Protein measurement and standardisation

The BioRad DC Protein Assay was used to measurement. Solution A' was made up with 1ml Solution A and 20µl Solution S. To the wells of a 96-well plate, 25µl solution A', 200µl solution B, and 5µl sample or RIPA buffer (as a blank) were added and the plate left for 15 minutes at room temperature. Protein concentrations were measured using Tecan GENios Microplate Reader and Magellan 3 software, and the results used to standardise protein concentration.

2.6.1.3 Western blot

NuPAGE Sample Reducing Agent and NuPAGE LDS Sample Buffer were added to 1× concentration to each sample. The solution was vortexed briefly, incubated at 98°C for 3 minutes and placed on ice. If not being loaded immediately, samples were stored at -20°C.

The electrophoresis module was assembled with a NuPAGE 4-12% Bis-Tris 1.0mm gel. 1× NuPAGE® MOPS SDS Running Buffer was added to fill the inner chamber and half-fill the outer chamber. 10 µl sample was loaded onto the gel, with 10 µl Spectra Broad Range Protein Ladder as a marker. Electrophoresis was carried out at 200V for 50 minutes.

Invitrolon PVDF membrane (0.45 µm pore size) was soaked in methanol, rinsed in dH_2O and soaked in 1× NuPAGE Transfer Buffer with 10% methanol. Two filter paper sheets, sponges for the transfer module and the gel were also soaked in transfer buffer. The sponges, gel, membrane and filter paper were assembled in the transfer module as shown in Figure 18, the module placed in the tank and the chamber filled with transfer buffer. The transfer was run at 35V for one hour.

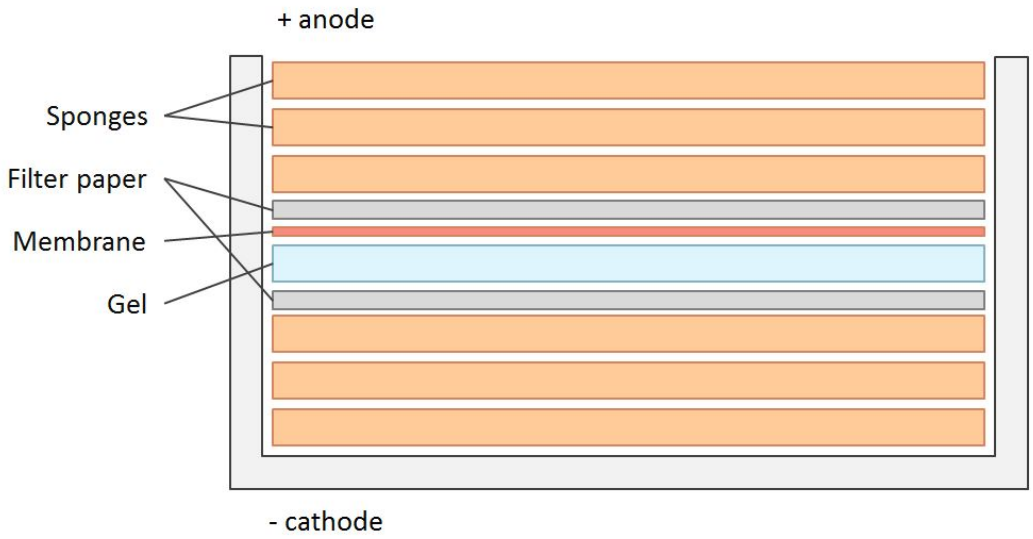

Figure 18: Assembly of components for membrane transfer. All components are soaked in transfer buffer before assembly. Charge running from the cathode to the anode stimulates transfer of proteins from the gel to the PVDF membrane.

After transfer, the membrane was blocked for one hour in 5% skimmed milk in PBST. The membrane was cut into sections to probe for the proteins of interest. Sections were incubated overnight with the primary antibodies to the proteins of interest diluted to 1/1000 (or as per manufacturer's instructions) in 5% milk in PBST. Membranes were subsequently washed with PBST for 10 minutes three times and incubated for one hour with the appropriate secondary antibody diluted to 1/10,000. Membranes were again given three ten-minute washes.

2.6.1.4 Western blot: protein detection

Proteins were detected using the Bio-Rad Clarity Western ECL Blotting Substrate. For each membrane, 0.5 ml peroxide reagent and 0.6 ml luminol/enhancer reagent were mixed and added to the membrane for five minutes. Protein location was captured by either exposing the membrane to film and developing the film with the Konica SRX101A tabletop processor, or by directly capturing fluorescence with the Syngene G:BOX Chemi XRQ and GeneSys v1.5.0.0 software.

2.6.2 Assays with GFP-BCAR1 fusion expression vector

A GFP-BCAR1 fusion construct was obtained from the Cardiovascular Biology and Medicine group. The fusion construct was produced by cloning the *BCAR1* gene into the pEGFP-C2 vector at *EcoRI* and *BamHI* restriction sites, as shown in Figure 19.

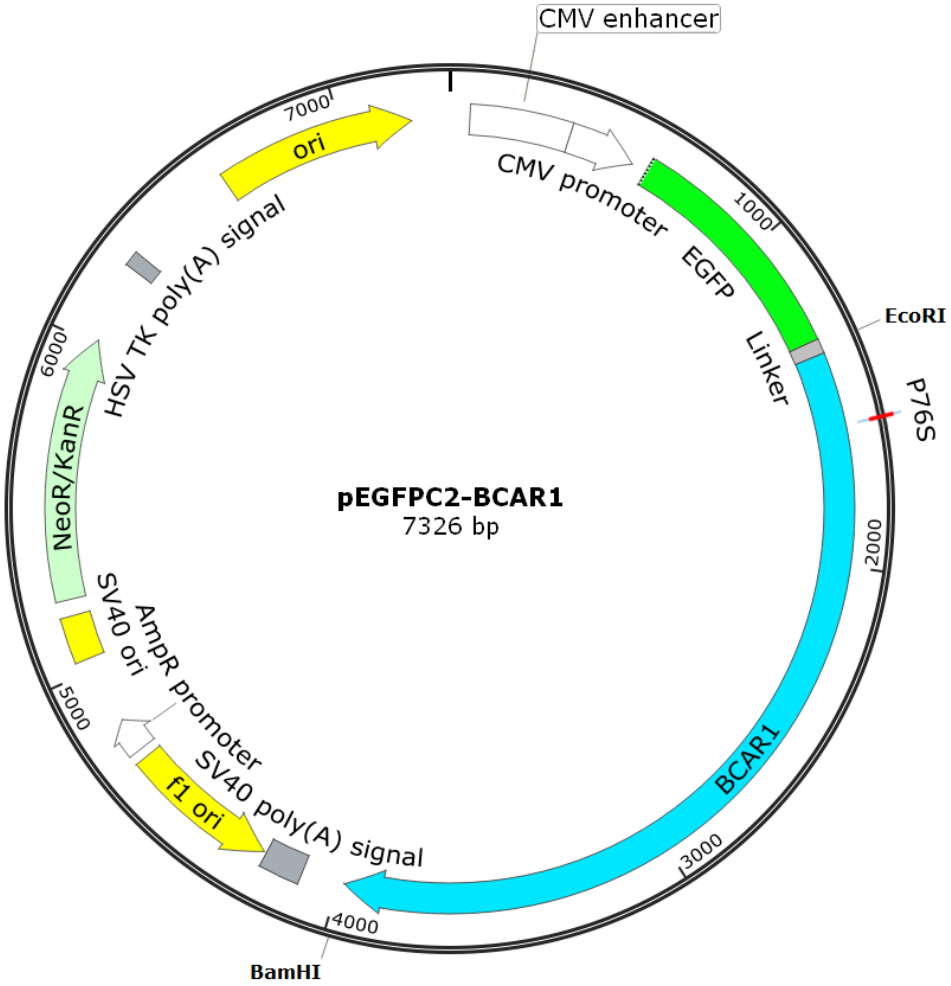

Figure 19: Map of pEGFPC2-BCAR1 plasmid. The plasmid was used as the wild-type BCAR1 expression vector, and for site-directed mutagenesis to create the mutant vector. The BCAR1 gene had previously been cloned into the plasmid. The position of the base change equivalent to rs1035539, changing a coded proline to serine, is marked by "P76S". Figure created in Snapgene[213].

2.6.2.1 Site-directed mutagenesis

Site-directed mutagenesis was performed on the pEGFP-C2/BCAR1 construct using Agilent Technologies' QuikChange Lightning mutagenesis kit, in order to change the base at the location of the rs1035539 SNP. Primers with the rs1035539 A allele were designed using Agilent Technologies' QuikChange Primer Design tool[214]. Primer sequences are shown in Table 19.

Table 19: Site-directed mutagenesis primer sequences.

Primer	Sequence (5' – 3')
SDM_for	CAGGGCCTGGCTCCGGCCCTCCC
SDM_rev	GGGAGGGCCGGAGCCAGGCCCTG

The mutagenesis reaction was prepared according to Table 20 and amplified according to the thermocycling conditions shown in Table 21.

Table 20: Volumes for QuikChange reaction.

Reagent	Volume (µl)
10× QuikChange Lightning Buffer	2.5
QuikSolution	1.5
dNTP mix	1
QuikChange Lightning Enzyme	1
H_2O	16
dsDNA template (100ng/µl)	1
Forward primer (100ng/µl)	1
Reverse primer (100ng/µl)	1

Table 21: Site-directed mutagenesis thermocycling conditions.

Cycles	Temperature (°C)	Time (min:sec)
1	95	2:00
30	95	0:20
	60	0:30
	65	3:30
1	65	5:00

After PCR, 1 µl *DpnI* endonuclease was added to the reaction and incubated for 10 minutes at 37°C. *DpnI* is specific for methylated and hemi-methylated DNA, and therefore digests the parental DNA template: DNA isolated from *E. coli* is dam methylated, while the mutated new templates are not.

2.6.2.2 Transformation and mini-prep plasmid purification

The original (referred to as wild-type) and mutated pEGFP-C2/BCAR1 construct was transformed into OneShot TOP10 chemically competent cells, following the manufacturer's protocol. 50µl aliquots of the cells were thawed on ice and 4µl vector DNA added to each aliquot. The cells were left on ice for 30 minutes, heat-pulsed at 42°C for 30 seconds and immediately placed on ice for 2 minutes. 250 µl room-temperature SOC medium was added and the vials shaken at 37°C for one hour.

The transformed cells in LB were then spread onto LB agar plates with 50 µg/ml kanamycin and incubated at 37°C overnight. Four colonies were picked from the wild-type and mutant plates and inoculated into 5 ml LB containing 50µg/ml kanamycin, and shaken overnight at 37°C.

Bacterial cells were harvested by centrifugation at 6000 g at room temperature for 3 minutes. DNA was then purified following the instructions of the Qiagen Spin Mini-prep kit, and eluted in 50 µl dH_2O. Construct DNA was sequenced by Source BioScience to confirm the correct mutation of the rs105339 allele. DNA was sent for sequencing to confirm the presence of the mutant.

After confirmation of correct sequence, 50 µl bacterial culture for the wild-type and mutant constructs was inoculated into 150 ml LB with 50 µg/ml kanamycin, and shaken at 37°C overnight. The cultures were centrifuged at 5000 g for 10 minutes. Plasmid DNA was then purified according to the instructions of the GenElute HP Plasmid Maxiprep Kit, and eluted in 3ml elution solution. DNA was isopropanol precipitated and resuspended in 300µl H_2O. Concentration was measured using the Nanodrop. DNA was then sent for sequencing across the whole gene, and the results compared to the reference sequence to check that no base changes had been introduced, with the exception of the mutated rs1035539 variant.

2.6.2.3 Cell culture and transfection

HEK293 (human embryonic kidney 293) and COS-7 cells were used for transfection of the BCAR1 plasmid. HEK293 and COS cells were cultured in Dulbecco's modified Eagle medium (DMEM) containing 10% FBS, and maintained at 37°C, 5% CO_2.

For transfection, cells were seeded into a 12-well plate containing 1 ml medium per well. For each well, 3 µl Lipofectamine 2000 (Invitrogen) was mixed with 250 µl Opti-MEM, and 1 µg plasmid DNA mixed with 250 µl Opti-MEM. Each mix was incubated at room temperature for 5 minutes. The DNA mix was then added to the Lipofectamine mix, and incubated at 20 minutes at room temperature. Solutions were added to the wells containing the cells. Cells were then left at 37°C 5% CO_2 for 24 hours and then serum-starved overnight (DMEM with 10% FBS was replaced with DMEM with 0.5% FBS) in preparation for treatment.

Four plasmids were used to transfect COS and HEK293 cells for BCAR1 assays. The four plasmids are referred to in this section as "GFP", "wild-type", "mutant" and "15F" respectively.

1. **pEGFP-C2:** basic GFP vector without *BCAR1* inserted. The vector was used as a negative control and for normalisation of transfection efficiency.
2. **pEGFP-C2/BCAR1(WT):** pEGFP-C2 vector with wild-type BCAR1/GFP fusion protein (proline at position 76).

3. **pEGFP-C2/BCAR1(M):** pEGFP-C2/BCAR1 with mutation coding serine at position 76, corresponding to the amino acid substitution caused by the SNP rs1035539.
4. **pEGFP-C2/BCAR1(15F):** pEGFP-C2/BCAR1 coding for an "unphosphorylatable" version of the BCAR1 protein. 15 key tyrosine residues in the substrate domain are mutated to phenylalanine and can no longer be phosphorylated.

2.6.2.4 Signalling assays

COS cells transfected with the four plasmids were used for signalling assays. Cells were cultured in standard conditions as described in section 2.1.6, then transfected and serum-starved overnight as per 2.6.2.3. Cells were treated with serum-free medium (SFM), or SFM with 2 ng/ml epidermal growth factor (EGF), 5 mM manganese chloride ($MnCl_2$) or 10 µM Nilotinib. Treatment times were 5, 5, 10 and 20 minutes respectively. Cells were lysed and proteins detected as per 2.6.1. Proteins detected were total BCAR1, phosphorylated BCAR1 (tyrosine 410) and GAPDH.

2.6.2.5 Protein localisation assay

HEK293 cells were transfected with the GFP, wild-type and mutant plasmids. After serum-starvation overnight, cells were plated into 8-well cell culture microscope slides and treated for five minutes with serum-free medium or medium with 5% FBS. After treatment, medium was removed and 4% formaldehyde added to fix cells. Fixed cells were stored at 4°C. Location of the GFP-BCAR1 fusion proteins was visualised using confocal microscopy.

2.6.2.6 Wound healing assay

COS cells were used for the wound closure assay. Cells were transfected with the four GFP plasmids as per 2.6.2.3. After transfection they were plated into an Essen ImageLock 96-well plate at a density of 3.5×10^4 per well. Cells were serum-starved overnight. The Essen WoundMaker was used to create 700-800 µM wide scratch wounds in the cell layer in each well of the plate. Distilled water and 70% ethanol were placed in the two wash boats, and the WoundMaker pin block was placed in the water and ethanol wash boats for five minutes each. The 96-well plate was then placed in the plate holder with the lid removed, and the pin block was placed on top of the plate. The lever was pressed to create the wounds, the pin block removed, and the pin block washed as before. The 96-well plate was then placed in the Essen BioScience IncuCyte ZOOM incubator, where cells were incubated at 37°C, 5% CO_2 as normal. Scans of the cell surface were scheduled every 2 hours, allowing quantification of the rate of wound closure.

2.6.2.7 siRNA knockdown of endogenous BCAR1

The Accell Human BCAR1 (9564) siRNA (Dharmacon) (target sequence CCAGGAAUCUGUAUAUAUU) was transfected into HUVEC cells to knock down endogenous BCAR1. The BCAR1 siRNA was initially resuspended in 100 µl RNase-free water for a concentration of 50 nmol/µl. A control siRNA with scrambled sequence was resuspended in the same conditions. HUVECs were seeded into two wells of a 6-well plate at 1.75×10^5 per well. The two siRNA mixes were made by mixing 4 µl (0.2 nmol) resuspended siRNA (BCAR1 or control) with 176 µl Opti-MEM. Oligofectamine mix was made by mixing 21 µl Opti-MEM with 21 µl oligofectamine, and incubated for 10 minutes at room temperature. 20 µl of oligo mix was then added to each siRNA mix, and the mixture incubated at room temperature for 25 minutes. Cells were then washed and 800 µl pre-warmed Opti-MEM added to each well. 200 µl of the oligo-siRNA mixes was added to the BCAR1 and control wells, and the cells incubated as normal for 4 hours. 500 µl 30% FBS in Opti-MEM was added for a final concentration of 10% FBS. After 24 hours, cell medium was changed to endothelial cell medium with growth supplement, and cells were assayed after 48 hours.

2.6.3 HUVEC transfection

Successful transfection of HUVECs was desired for two areas of the thesis: to perform luciferase reporter assays to assess the intronic SNP rs4888378's effect on gene expression, and to express wild-type and mutant BCAR1 protein for signalling and functional assays. Various methods were tested to try and achieve successful transfection of HUVECs, involving lipid reagents or electroporation.

2.6.3.1 Lipofectamine 2000

HUVECs were seeded in a six-well plate at 1.5×10^5 cells per well. Medium was changed 24 hours later. Four lots of 150 µl Opti-MEM were measured out, to which Lipofectamine 2000 (ThermoFisher) was added at 6, 9, 12 and 15 µl to create the lipid mixes. 10 µg of pEGFP DNA was diluted in 600 µl Opti-MEM, and 150µl of the DNA mix added to each lipid mix. Solutions were incubated for 5 minutes at room temperature, then 250 µl of each mix added to a well of cells. GFP expression was monitored using fluorescence microscopy.

2.6.3.2 Lipofectamine 3000

Transfection was carried out on HUVECs in 24 wells of a 96-well plate, following the Lipofectamine 3000 (ThermoFisher) protocol. Two volumes of Lipofectamine 3000 and two volumes of pEGFP DNA were tested. Lipofectamine 3000 was tested at 0.15 and 0.3 µl per well, and DNA at 0.1 and 0.5 µg

per well. Lipid mix was made up by diluting 0.9 µl or 1.8 µl Lipofectamine in 30 µl Opti-MEM. DNA mix was made by diluting 0.1 µg or 0.5µg DNA and 1.2 µl P3000 reagent in 30 µl Opti-MEM. The DNA mix was then added to the lipid mix, and the DNA-lipid complex incubated for 5 minutes at room temperature. 10 µl DNA-lipid complex was then added to each well, and cells checked for transfection efficiency with fluorescence microscopy.

2.6.3.3 Comparison of Lipofectamine 3000, jetPEI-HUVEC and electroporation

Transfection efficiency of Lipofectamine 3000, jetPEI-HUVEC (Polyplus transfection) and electroporation were compared using both the pEGFP and pEGFP-BCAR1(WT) plasmids. Transfection took place in a 12-well plate, in which 4 wells of 7×10^4 cells per well were plated out. 24 hours later, cell medium was replaced with fresh ECM, and cells were transfected with the two plasmids using Lipofectamine 3000 and jetPEI-HUVEC (one plasmid and lipid reagent combination per well).

Lipofectamine lipid mix was made up using 3 µl Lipofectamine with 100 µl Opti-MEM. Each Lipofectamine DNA mix contained 1 µg plasmid DNA, 2 µl P3000 reagent and 50 µl Opti-MEM. 50 µl lipid mix was added to each DNA mix, and the solutions were incubated at room temperature for 5 minutes. 100 µl was then added dropwise to each well.

JetPei lipid mix was made up with 8 µl JetPei and 100 µl NaCl. Each JetPei DNA mix contained 2 µg plasmid DNA and 50 µl NaCl. 50 µl lipid mix was added to each DNA mix, and the solutions vortexed and spun down briefly, then incubated at room temperature for 30 minutes.

Electroporation used 1.5×10^5 cells per well to account for cell death as a result of electroporation. For electroporation, cells were not plated out in advance. 3×10^5 cells were spun down and resuspended in 200 µl Mirus buffer. 100 µl of the cell solution was added to two 0.2 cm Mirus cuvettes containing 2 µg GFP and GFP-WT DNA, and electroporated using the Amaxa Nucleofector I on setting V-001. The electroporated solutions were each added to 1.5 ml medium and all added to wells. Cells were monitored for GFP signal on the IncuCyte.

2.6.4 BCAR1 virus production and infection

Two adenoviruses were produced using the Gateway system (ThermoFisher) in order to examine the effects of wild-type and mutant BCAR1. Adenoviruses are non-enveloped viruses with a double-stranded DNA genome; on infection, viral DNA is introduced into the host cell but is not incorporated into the host genome. Adenoviruses were used as the delivery vehicle for BCAR1 due to their ability to infect a wide variety of cell types, and in particular their ability to infect HUVECs[215].

The Gateway cloning system allows a DNA fragment to be inserted into an entry vector (here pENTR3C), then transferred into a larger virus-specific destination vector which can be used for viral infection.

2.6.4.1 Virus entry vector construct

The BCAR1 gene had been cloned into the pENTR-3C Gateway dual selection vector by the Cardiovascular Biology and Medicine group (Figure 20). DNA sequences of choice can be easily cloned into this vector, then transferred into the destination vector.

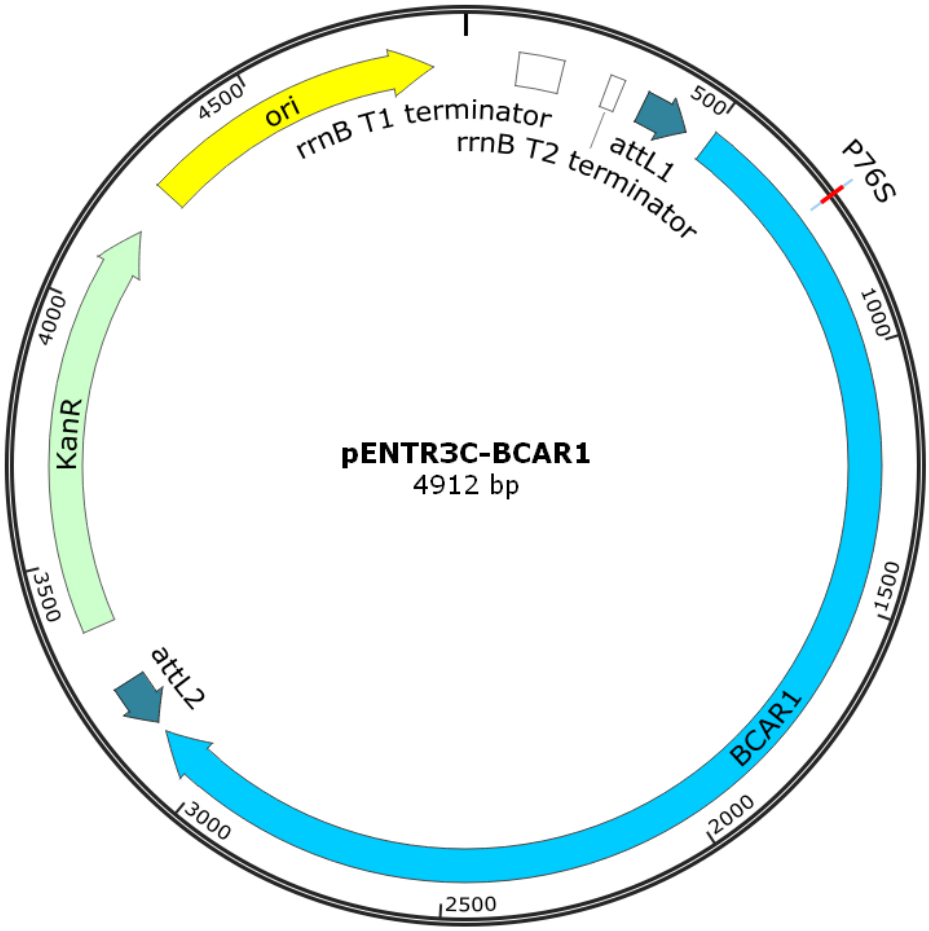

Figure 20: Map of pENTR3C-BCAR1 plasmid. The plasmid was used as the wild-type entry vector for production of adenovirus, and basis for site-directed mutagenesis to create the mutant vector. The BCAR1 gene had previously been cloned into the plasmid. The position of the base change equivalent to rs1035539, changing a coded proline to serine, is marked by "P76S". Figure created in Snapgene[213].

2.6.4.2 Site-directed mutagenesis

The gene sequence in the plasmid is identical to that used for the pEGFP-BCAR1 expression vector. Therefore, the same site-directed mutagenesis primers and protocol was used as in 2.6.2.1.However, the thermocycling protocol was altered due to the shorter length of the vector (4912bp to the

pEGFP-BCAR1 vector's 7326bp), reducing the extension time to 2m30s. 1 µl *DpnI* was added to the mutagenesis product and incubated for 10 minutes at 37°C to digest the parental DNA template.

2.6.4.3 Transformation and plasmid purification

The mutagenesis product was transformed into XL10-Gold ultracompetent cells. Cells were thawed on ice and 45 µl added to pre-chilled thin-walled tubes. 2 µl β-mercaptoethanol mix was added to each aliquot of cells and the cells kept on ice for 2 minutes. 2 µl vector DNA was added to each aliquot and the tubes gently swirled and incubated on ice for 30 minutes. Cells were heat-pulsed at 42°C for 30 seconds, then placed on ice for 2 minutes. Cells were then added to 500 µl preheated (37°C) LB, and shaken at 37°C for 1 hour.

The pENTR3C vector contains a kanamycin resistance gene which allows selection of positive transformants. The cells in LB were spread onto LB agar plates with 50 µg/ml kanamycin and incubated at 37°C overnight. Four colonies were picked from each plate and inoculated into 5 ml LB containing 50 µg/ml kanamycin, and shaken overnight at 37°C. Bacterial cells were harvested by centrifugation at 6000 g at room temperature for 3 minutes. DNA was then purified following the instructions of the Qiagen Spin Mini-prep kit, and eluted in 30 µl dH_2O. A diagnostic digest was performed to confirm that the gene had been cloned correctly into the pENTR3C vector. *SalI* and *EcoRV* were used to digest the vector. The vector should therefore be cut at both the polylinker sites and halfway through the BCAR1 gene, where a *SalI* site is present. The reaction was set up as in Table 22 and digested for 90 minutes at 37°C.

Table 22: Volumes for diagnostic digest of pENTR3C clone.

Reagent	Volume (µl)
DNA	6
SalI	0.5 (10 U)
EcoRV	0.5 (10 U)
NEBuffer 3	1
BSA	0.1
dH_2O	1.9
Total	10

When the digest had confirmed the presence of the *BCAR1* gene, mutated DNA was sent for sequencing. After confirmation of the correct sequence, 50 µl of the mutated bacterial culture was inoculated into 150 ml LB with 50 µg/ml kanamycin, and shaken at 37°C overnight. The cultures were

centrifuged at 5000 g for 10 minutes, and plasmid DNA purified according to the instructions of the GenElute HP Plasmid Maxiprep Kit, eluting in 3ml elution solution. DNA was isopropanol precipitated and resuspended in 300µl H_2O. The clone sequenced across the whole gene, and the results compared to the reference sequence to check that no additional base changes had been introduced.

2.6.4.4 Cloning into pAd/CMV/V5-DEST destination vector (LR recombination reaction)

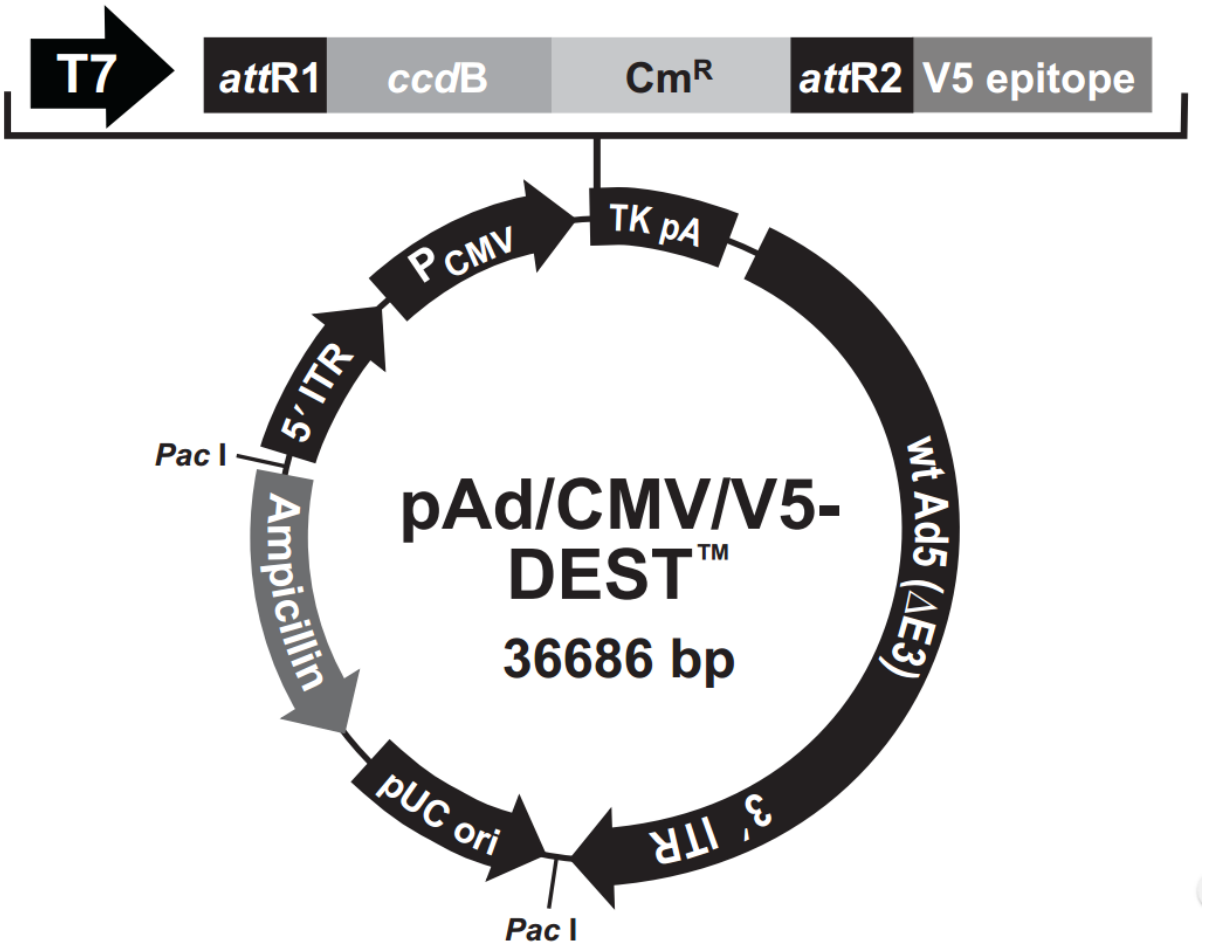

Figure 21: Map of pAd/CMV/V5-DEST destination vector. Image from ThermoFisher: pAd/CMV/V5-DEST™ and pAd/PL-DEST™ Gateway® Vectors manual.

After the production of the entry clone, an LR recombination reaction is carried out to transfer the sequence of interest into the destination vector. The destination vector used here was pAd/CMV/V5-DEST (Figure 21), which contains genes essential for synthesis of adenoviral proteins. Both the entry and destination vectors contain *attL/attR* sequences, which react specifically with each other to transfer the sequence.

The LR Clonase II reaction was set up for the wild-type and mutated vectors, and a control vector (pENTR-GUS). Each reaction was set up with 1 µl pENTR vector (50 ng/µl stock), 2.75 µl

pAd/CMV/V5-DEST vector (50 ng/µl stock), 2 µl LR Clonase II mix, and 4.25 µl TE (pH 8). Reactions were incubated at 37°C for 1 hour. 1 µl supplied Proteinase K was added and the reaction incubated for 10 minutes at 37°C.

The *attR* sequences in the destination vector flank the *ccdB* gene, which is toxic to the cell. This provides negative selection, as a successful recombination reaction produces an expression vector with the gene of interest rather than the *ccdB* gene. The product of the LR recombination reaction must be transformed into competent cells that do not contain the F' episome, as these cells contain the gene *ccdA*, which compensates for the negative selection provided by the *ccdB* gene. As the destination vector contains the ampicillin resistance gene rather than the kanamycin resistance gene, selection plates use ampicillin to prevent growth of cells with the entry clone.

The LR Clonase II reaction product was therefore transformed into One Shot OmniMAX 2 T1 Phase-resistant Cells. Three vials of cells were thawed on ice for the wild-type, mutant and control reactions. 1 µl DNA was added to each vial and mixed gently, then incubated on ice for 30 minutes. Cells were heat-shocked at 42°C for 30 seconds, then placed on ice for 2 minutes. 250 µl pre-warmed (37°C) SOC medium was added to each vial, and shaken at 37°C for 1 hour. The solution was then plated onto LB agar plates with 50 µg/ml ampicillin, and incubated at 37°C overnight.

The next day, after confirming that colonies were present on the wild-type and mutant plates but not negative control, DNA was purified according to the instructions of the GenElute HP Plasmid Maxiprep Kit, and ethanol precipitated as before. DNA concentrated was measured using the Nanodrop and sent for sequencing to confirm the correct base, correct *BCAR1* sequence, and alignment with the pDEST vector after the attR1 site.

2.6.4.5 *PacI* digest

pDEST vectors were digested with *PacI*. Digestion with *PacI* exposes the left and right viral ITRs, allowing proper viral replication and packaging, and removes the bacterial sequences. 10 µg DNA was digested in a 100 µl reaction, with 10 µl NEB CutSmart buffer and 2 µl *PacI*. The DNA was then purified with phenol/chloroform extraction, following the instructions of the 5 Prime Heavy Phase Lock Gel. DNA was then ethanol precipitated and resuspended in 100 µl.

2.6.4.6 Preparation of adenoviral stock

PacI-digested destination vectors were transfected into HEK293 cells for large-scale production of adenovirus. HEK293 cells were plated into two 6-well plates. The following day, the pAd/CMV/V5-

DEST-BCAR1 wild-type and mutant vectors were transfected into the cells using Lipofectamine 2000. 4 µg WT and mutant DNA was mixed with 1 ml Opti-MEM, while for each the WT and mutant, 12 µl Lipofectamine 2000 was diluted in 1 ml Opti-MEM. These were incubated at room temperature for 5 minutes, then the DNA mix added to the lipid mix and these incubated for 20 minutes. 500 µl of the mix was added dropwise to 4 wells each for the WT and mutant.

After 1-3 days, when cells were confluent, each 6-well plate was split into six T175 flasks. Culture medium was replaced every 2-3 days until visible regions of cytopathic effect were seen. The viral infection was left to proceed until about 80% cytopathic effect was seen.

Cells were removed from the plate by pipetting with a 10 ml pipette and transferred to a 50 ml tube. Freeze/thaw cycles were then performed to prepare a crude viral lysate, allowing the cells to lyse and release viral particles. Tubes of harvested cells were placed at -80°C for 30 minutes, removed and thawed in a 37°C water bath for 15 minutes. These two steps were repeated twice for a total of three freeze-thaw cycles.

2.6.4.7 Virus purification

Purified virus was produced from the crude viral lysate using the PureSyn Adenopure Adenovirus Purification Kit. After the final thaw of the three freeze-thaw cycles, cell debris was pelleted by centrifuging the samples at 2000 g for 5 minutes at room temperature. 50 µl of 25 U/µl Benzonase was added to each lysate and the solution mixed and incubated at 37°C for 30 minutes. Lysates were then diluted, equilibrated, loaded onto the virus module and eluted following the instructions of the Adenopure kit. The final elution volume was 4.5-5 ml.

2.6.4.8 Virus dialysis

Eluted wild-type and mutant BCAR1 viruses were dialysed in order to replace the elution buffer with a glycerol-containing TE buffer more suitable for long-term storage. Dialysis was performed using 3-12 ml Thermo Scientific Slide-A-Lyzer Dialysis Cassettes (10,000 molecular weight cut-off). Two cassettes were immersed in the dialysis buffer, sterile TE for 1-2 minutes to hydrate the membrane. A 18-gauge syringe was used to load the wild-type and mutant virus solution into the cassettes. Each cassette was floated in 1 L TE and the solution mixed using a magnetic stirrer for 4 hours at room temperature. The TE buffer was then replaced with 1 L TE with 10% glycerol, and the solutions with cassettes mixed overnight at 4°C. A syringe was then used to remove the viral solution from the

cassettes, and both solutions were filter-sterilised with a 0.2 µm filter. Viral solutions were subsequently stored at -20°C.

2.6.4.9 Measurement of viral titre

Viral titre was measured using the QuickTiter Adenovirus Quantitation Kit. DNA standards were prepared by serial dilution of the supplied standard, and final solutions for measurement were prepared in duplicate following the manufacturer's instructions. Fluorescence was measured using the TECAN GENios Fluorescence, Absorbance and Luminescence Microplate Reader with 480/520 nm filter, and recorded using TECAN Magellan data analysis software. Adenovirus titre was calculated using the viral solution RFUs and the standard curve.

2.6.5 BCAR1 protein assays in HUVECs

Production of adenovirus expressing wild-type and mutant BCAR1 allowed assays to be carried out on these proteins in HUVECs.

2.6.5.1 HUVEC culture and infection

Pooled HUVECs were obtained from Promocell and cultured in Promocell Endothelial Cell Medium with added serum-containing growth supplement. Cells were maintained at 37°C, 5% CO_2. Before viral infection, cells were seeded into a 6-well plate. 24 hours later, 1 µl viral solution was added to each well. After 48 hours, cells were serum-starved overnight by replacing cell medium with Endothelial Cell Medium containing 0.5% serum.

2.6.5.2 Signalling assays

For signalling assays, HUVECs were infected with wild-type or mutant BCAR1 virus, control virus (GFP or lacZ-expressing virus) or no virus.

After 48 hours, cells were treated with 25 ng/µl VEGF. 10 µl VEGF (50 µg/ml) was added to wells at timepoints of 0 (i.e. no VEGF), 5, 10, 20, 40 and 60 minutes. During treatment cells were incubated at 37°C, 5% CO_2. After treatment, cells were immediately placed on ice and lysed as in 2.6.1.1. Samples were run on a gel and proteins detected as in 2.6.1.3. Proteins detected were: total BCAR1, phosphorylated BCAR1 (Y410) and phosphorylated BCAR1 (Y249), total paxillin, phosphorylated paxillin (Y119), Akt and GAPDH. Total protein amounts were quantified using GeneSys v1.5.0.0 software and compared using two-way ANOVA.

2.6.5.3 Well migration assays

HUVECs were infected with wild-type or mutant BCAR1 virus, control virus (GFP or lacZ-expressing virus) or no virus. 750 µl ml serum-free endothelial cell medium was added to each well of a 24-well plate. Control wells contained only serum-free medium, while treatment wells contained 25 ng/µl VEGF. Transwell inserts made of low-pore-density polyethylene terephthalate (8 µm pore size) (Falcon, BD Biosciences) were then added into each well. Cells were trypsinised and resuspended at a concentration of 3×10^5 cells/ml. 500 µl of cell suspension was added into each well insert, giving 1.5×10^5 cells per well. Cells were incubated in standard conditions for 4 hours. The inserts were then removed, and cells on the upper side of the inserts that had not migrated through the membrane were removed with cotton buds, leaving only those that had successfully migrated through to the other side.

Membranes were then fixed and stained following the instructions of the Reastain Quik-Diff kit. Cells that had migrated to the lower side of the membrane were therefore stained, and were counted at 400× magnification using a microscope. After four independent experiments containing 2-3 repeats of each treatment/plasmid combination, number of migrated cells were compared between treatment/no treatment and plasmid expressed using two-way ANOVA.

3 Bioinformatics analysis of the *CFDP1-BCAR1-TMEM170A* locus

3.1 Introduction

Genome-wide association studies have identified many novel loci for complex diseases such as CHD, but implicating a genetic locus is only the first step to discovering the basis for these associations. As discussed in the introduction, the presence of linkage disequilibrium provides the means by which tag SNPs can represent a haplotype containing many additional SNPs. However, this means that in areas of strong LD, GWAS signals cannot implicate a single variant with the trait, as all variants in LD will be correlated. Of these correlated variants, it is often unclear which is likely to be directly causing a change in the phenotype.

Strong phenotypic changes created by a single SNP are often the result of a non-synonymous coding variant that alters the amino acid and resultant protein structure. For example, sickle cell anaemia is characterised by homozygosity for the minor T allele of rs334, coding for a valine instead of glutamic acid at position six of the haemoglobin protein[216]. Identifying these non-synonymous risk variants can be relatively straightforward as they have a predictable effect on protein structure. However, the majority of disease-associated variants found by GWAS are in non-coding areas of the genome[217,218]. In the CARDIoGRAMplusC4D meta-analysis looking for CAD risk loci, for example, 77% of associated variants occurred in intergenic or intronic regions[6]. Being unable to alter protein structure, such variants are therefore likely to exert the phenotypic effect through disruption of gene expression.

While the non-coding regions that make up the majority of the genome was traditionally largely regarded as unused "junk" DNA, genomic annotation projects like ENCODE are assigning increasing functionality to these intronic and intergenic regions[219]. Some variants in these non-coding areas have been shown to have a functional effect on cardiovascular disease risk[220–222], but genetic regulation is still an area consisting of many unknowns, and numerous methods can be required to investigate the possible effects on regulation at an associated locus.

Gertow and colleagues identified a locus on chromosome 16 that was associated with cIMT and CAD[119], but the causal variant at the *CFDP1-BCAR1-TMEM170A* locus, and the basis behind the genetic association, are not known. The aim of this chapter was to investigate the chromosome 16 cIMT locus to determine which variants might be good candidates for functionality, and to explore which gene or genes at the locus they might be acting on to cause the effect.

In Gertow et al's study, genotyping did not follow a genome-wide approach, where a genotyping array examines SNPs spaced across the genome in order to tag as many SNPs as possible. Instead the Metabochip was used; as described in section 1.2.5.1, this is a custom genotyping array that tests 200,000 SNPs concentrated at loci already known to be of metabolic and cardiovascular interest[114]. While it allows fine-mapping of known cardiovascular loci, it leaves relatively sparse coverage at non-focused regions, as was the case for the *CFDP1-BCAR1-TMEM170A* locus, which had not previously been implicated in cardiovascular traits from GWAS. The discovery of the lead SNP at the locus therefore results in the possibility for a potentially large number of functional variants. The functional variant tagged by the lead SNP, rs4888378, could be any of those in strong LD with it. The first stage of analysis was therefore to find SNPs in LD with the lead SNP.

The online tool SNAP (SNP Annotation and Proxy Search) was used to find such SNPs, and these taken forward for further analysis. SNPs in strong LD with the lead SNP were identified using the 1000 Genomes Pilot 1 data set[196], which at the time of analysis was the most up-to-date population available for LD calculation. After selection of candidate SNPs and functional analysis, more comprehensive 1000 Genomes Phase I data became available, and variants in LD were re-calculated and analysed. Bioinformatics tools were used to prioritise the likely functionality of these SNPs. The likelihood of functionality of a SNP was evaluated by examining its location (for example, whether it is present in protein-coding regions) and by looking at regulatory characteristics annotated by ENCODE[219] and the NIH Roadmap Epigenomics Mapping Consortium[132].

Data from chromatin immunoprecipitation (ChIP-seq) assays show regions bound by DNA-binding proteins; variants in these areas might be expected to affect the binding and action of transcription factors[223]. The histone modifications present on the chromatin also provide information about the potential function of a region. For example, the methylation and acetylation marks H3K4me3 and H3K9ac mark areas consistent with promoters, while H3K4me1 and H3K27ac are generally found in active regulatory elements such as enhancers. Other marks indicate functions such as transcription start (H3K79me2) or repressive elements (H3K9me3)[219,224]. Variants in these areas are therefore more likely to affect regulation of gene expression.

Hypersensitivity to the DNase I enzyme is a mark of chromatin accessibility, and correlates with the presence of regulatory elements such as promoters and enhancers[225]. Mapping of DNase I hypersensitivity by DNase-seq in ENCODE allowed over 2 million unique DNase I hypersensitive sites

to be mapped to the genome. Variants in these areas are again more likely to have an effect on regulation.

In addition to looking for the variant that is responsible for the association, the phenotypic effect of the locus was explored on a wider scale by examining genes whose expression was associated with the lead SNP. Variants associated with the expression of genes are known as expression quantitative trait loci (eQTLs)[226]. Data from eQTLs does not help to distinguish functional variants from a haplotypic block; if eQTL data is from the same population, the same patterns of LD will be present and all SNPs in LD will be correlated with expression. Such data therefore cannot prove that a SNP is directly influencing expression. However, by implicating the gene or genes that are associated with the measured phenotype, we can understand more about how the functional variant might be causing the effect. This can also give clues about the location of the functional variant; for example, if the lead SNP is associated with expression of a certain gene at the locus, it may be useful to look at any variants in strong LD that are present in the promoter of this gene.

Gertow and colleagues' discovery study conducted eQTL analyses on the lead IMT SNP in the Advanced Study of Aortic Pathology (ASAP) and Biobank of Karolinska Endarterectomies (BiKE) data sets. ASAP comprises gene expression data from biopsies of liver, mammary artery and ascending aorta tissue from patients undergoing aortic value surgery, while BiKE has expression from atherosclerotic tissue from patients undergoing endarterectomies. All surgeries were carried out at Karolinska University Hospital in Stockholm, Sweden, and SNP genotype was obtained from patients' DNA samples using the Illumina Human 610W-Quad Beadarray. Using tissues relevant to the cIMT phenotype (thoracic aorta intima-media, aortic adventitia, mammary artery, heart, liver, human plaque and peripheral blood mononuclear cells), Gertow and colleagues looked at the relationship between the lead SNP and expression of the nine genes within 200 kb of the locus. Allele-specific expression was observed for *TMEM170A*, for which the association was strongest, *BCAR1* and *LDHD*, suggesting that functional variation at the locus may be acting through the expression of one of these genes (Figure 11).

To extend the analysis, looking at whether allele-specific expression was replicated in a different data set and different tissues, expression by rs4888378 allele was here examined using the publically available Genotype-Tissue Expression (GTEx) portal[155], and other publicly available eQTL resources were examined for variants of interest at the locus, to look for insight into active regulatory elements.

3.2 Results

3.2.1 Selection of candidate SNPs

While the lead SNP, rs4888378, may itself be exerting a functional effect causing increased cIMT, it may also be tagging a functional variant or variants with which it is in LD. LD analysis using 1000 Genomes project data indicated 214 variants in strong LD ($r^2 \geq 0.8$) with rs4888378 (Figure 22), implicating any of these as a potential candidate for the functional variant. These variants spanned the genes *CFDP1*, *TMEM170A* and *CHST6*. None of these SNPs were non-synonymous coding SNPs, located within splice junctions or in predicted miRNA binding sites, thus any function of these SNPs affecting cIMT was likely to be a result of effects on gene regulatory elements.

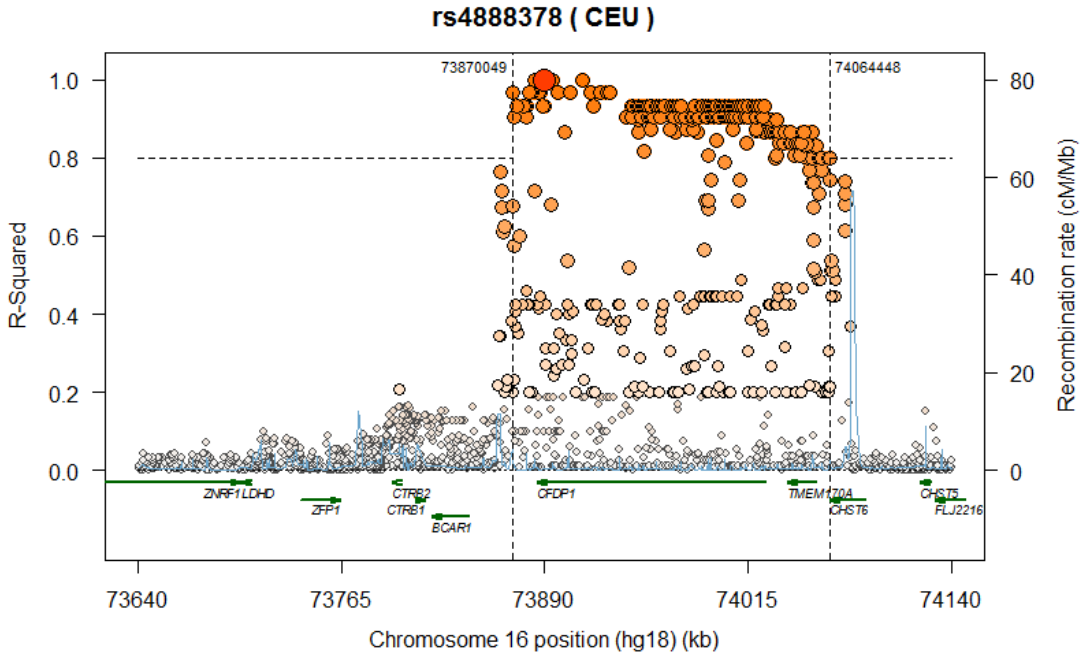

Figure 22: Linkage disequilibrium plot for lead SNP rs4888378. Data from 1000 Genomes (Pilot 1); graph plotted with data and R script from SNAP[197]. Lead SNP rs4888378 is shown in red; other variants are shown relative to their position at the locus (x-axis) and LD (r^2) with rs4888378 (y-axis). The majority of variants in strong LD lie within *CFDP1* and *TMEM170A*, and they may affect these genes, or lie within enhancer regions which affect expression of other genes. Chromosome coordinates are based on NCBI Build 36.1 (hg18).

Potential regulatory effect of the variants in strong LD was assessed using ENCODE, Roadmap and the ElDorado tool, as described in section 2.2. SNPs with certain regulatory features were assumed to be more likely to have a functional effect. Variants were taken forward for further analysis if they fulfilled all of the following conditions: located within protein binding sites (based on ChIP-seq data), located in or within 200 bp of a DNaseI hypersensitive site, located in promoter or enhancer histone signatures, and disrupt a predicted transcription factor binding motif. These regulatory marks can be visualised using online tools such as the UCSC Genome Browser[127] and HaploReg[133].

Five variants passed the criteria, which were selected for further analysis along with the lead SNP, rs4888378: rs4888379, rs4888392, rs2865530, rs3743609, and rs11643207. Table 23 shows each selected SNP with the strengths of associated regulatory marks. Figure 23 shows the position of each chosen SNP on the UCSC Genome Browser in relation to annotated histone marks, DNaseI hypersensitivity and bound proteins. Figure 24 shows an example of a SNP in strong LD with rs4888378 with insufficient regulatory marks, and that was therefore not chosen for further analysis.

Table 23: Chosen SNPs with associated regulatory marks.
* Indicates DNaseI hypersensitive sites within 200 bp of the SNP

					Histone marks		DNaseI hypersensitivity			Altered binding motifs	
SNP	Position (hg19)	Genetic location	Alleles	MAF (EUR)	Promoter (H3K4me3, H3K9ac)	Enhancer (H3K4me, H3K27ac)	Cell types	Score	Bound proteins (UCSC)	HR	ElDorado
rs4888378	75332041	CFDP1 (intronic)	A/G	0.57	0	8	8*	239	1	3	9
rs4888379	75340231	CFDP1 (intronic)	A/T	0.56	12	19	51	776	23	3	4
rs4888392	75412262	CFDP1 (intronic)	C/T	0.57	4	18	83*	1000	10	6	5
rs2865530	75414376	CFDP1 (intronic)	G/T	0.57	1	12	23	763	29	2	3
rs3743609	75467021	CFDP1 (intronic)	C/G	0.56	24	0	77	1000	41	9	2
rs11643207	75498793	TMEM170A (5' flanking region)	C/T	0.63	20	3	125*	1000	74	5	2

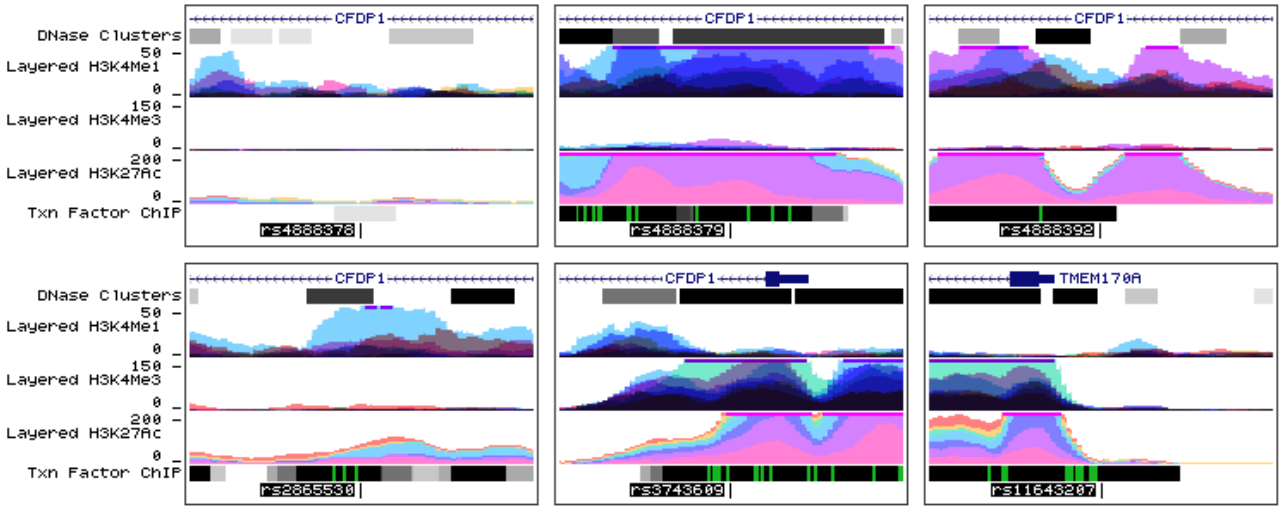

Figure 23: Selected SNPs with annotated bound proteins, promoter- and enhancer-associated histone marks, and DNaseI hypensensitivity clusters. H3K4me3 histone marks signify areas consistent with promoters, while H4K4me1 and H3K27ac marks signify areas consistent with enhancers. SNPs in strong LD with the lead SNP were viewed on UCSC Genome Browser[127] with these relevant regulatory tracks to visualise their location in regards to genes at the locus and other SNPs in strong LD.

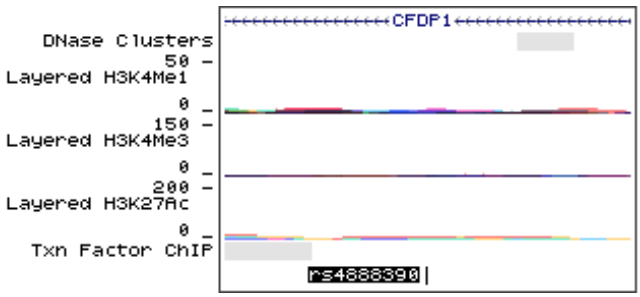

Figure 24: Example of an unselected SNP. SNPs such as rs4888390 were not selected for further analysis due to their lack of bound proteins, promoter- and enhancer-associated histone marks, and proximity to DNaseI hypersensitivity clusters. Figure adapted from UCSC Genome Browser[127].

The final six selected SNPs with their locations in relation to the genes at the locus, and reasons for selection, are shown in Figure 25. These SNPs were taken forward for functional analysis to determine their likelihood of affecting gene expression. Analysis of 1000 Genomes Phase I data which had become available after the initial analysis showed a further 56 variants in strong LD with rs4888378. These were analysed using regulatory data as above, but none passed the threshold for further analysis.

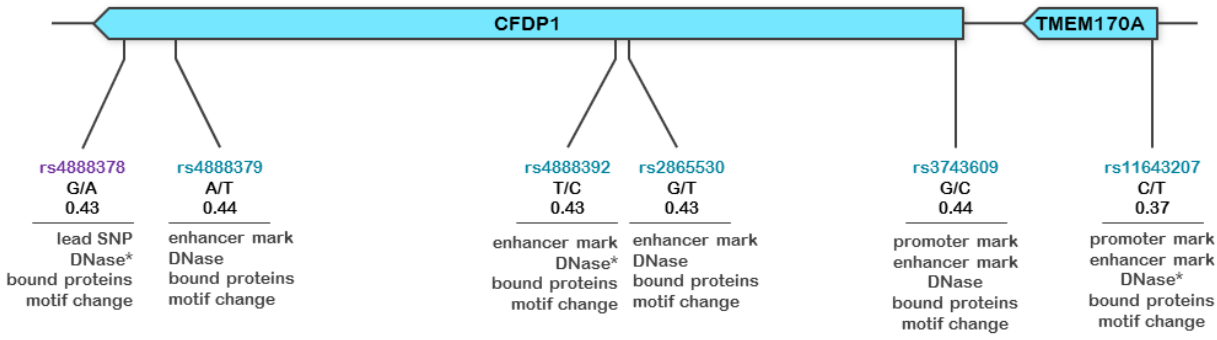

Figure 25: Location of chosen SNPs at the *CFDP1-BCAR1-TMEM170A* locus. SNP locations are shown in relation to the genes CFDP1 and TMEM170A, with the reasons for choice highlighted. *DNaseI hypersensitive site within 200 bp of the SNP.

3.2.2 eQTL analysis

It has previously been shown in Gertow and colleages' study that the lead SNP rs4888378 was associated with at least nominal significance with expression of *TMEM170A*, *BCAR1* and *LDHD* in aortic media, aortic adventitia and carotid plaque[119]. The gene expression dataset GTEx was used to extend the analysis and find out which relationships could be replicated. Again, the association between rs4888378 and transcripts from the nine genes within 200 kb of the lead SNP were tested. The blood vessel tissues aortic artery, coronary artery and tibial artery were selected for study.

Of the gene transcripts examined, three showed nominally significant allele-specific expression with rs4888378: *BCAR1*, *LDHD* and *CFDP1*, while the association with *TMEM170A* was not replicated (Figure 26 and Figure 27, Table 24). Only the *BCAR1* association was present in more than one tissue (aortic artery and tibial artery) and remained significant after correction for multiple testing. It also showed the highest effect sizes (β = 0.22 to 0.26 per G allele on normalised gene expression). Association in coronary artery tissue was borderline significant before correction. As seen in Figure 26, expression of *BCAR1* increases with each G allele – the allele associated with higher IMT and CAD risk in the discovery study.

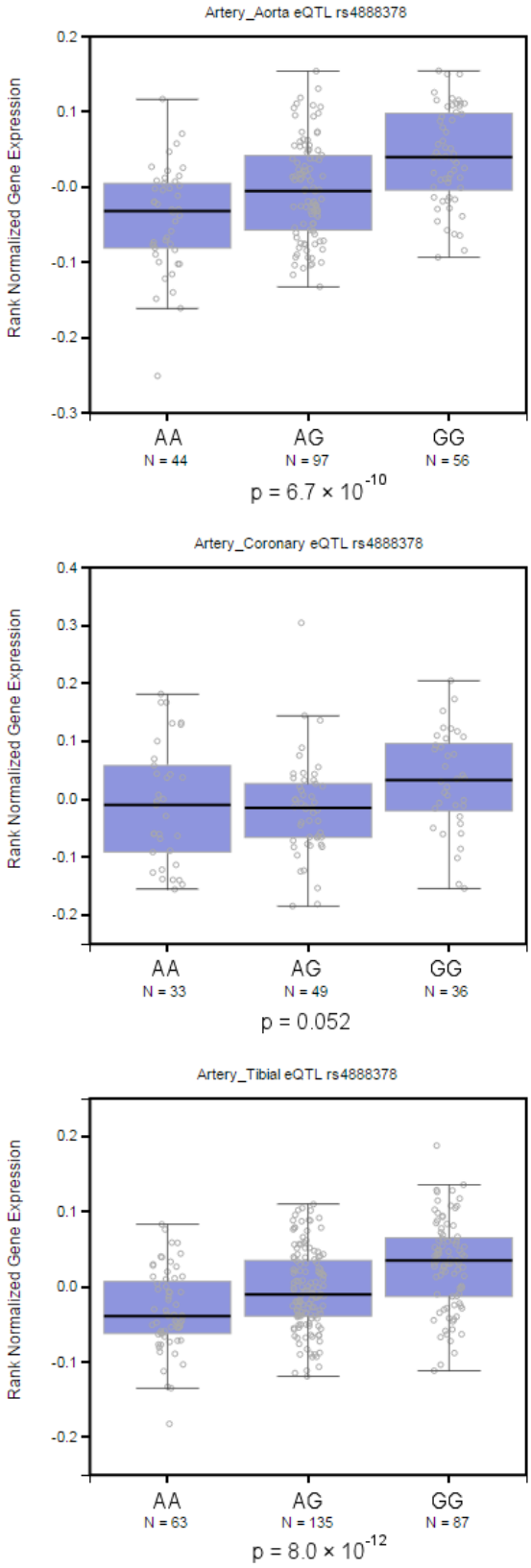

Figure 26: *BCAR1* expression by rs4888378 genotype in aortic, coronary and tibial artery tissues. Data and graphs from Gene-Tissue Expression (GTEx) Portal[155], accessed February 2016. Of the nine genes analysed at the locus, only associations with *BCAR1* remained significant after correction for multiple testing. Expression of *BCAR1* was shown to increase with each high-risk allele of the SNP.

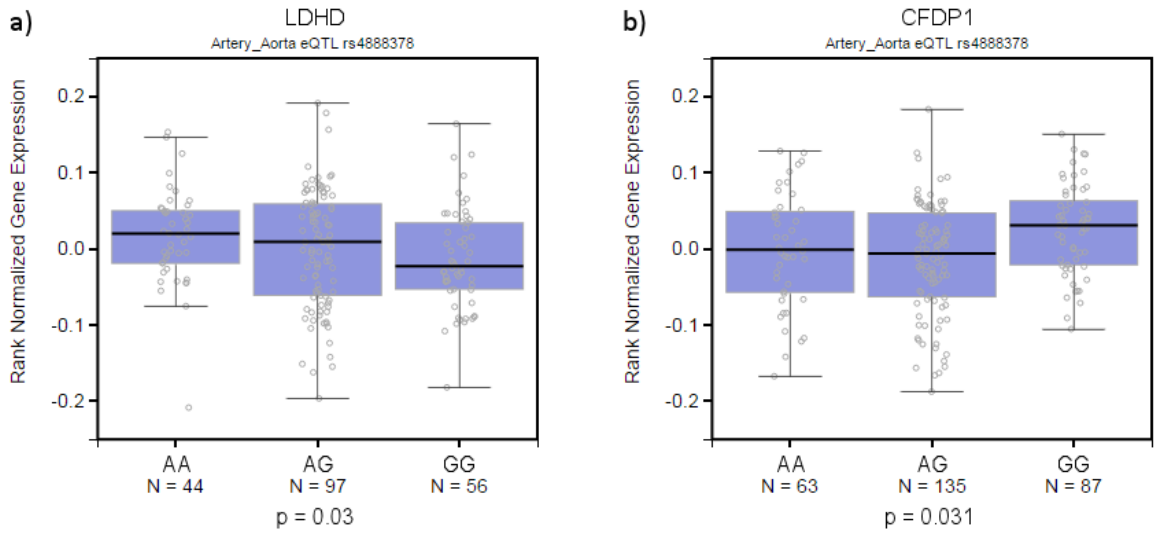

Figure 27: (a) *LDHD* and (b) *CFDP1* expression by rs4888378 genotype in aortic artery. Data and graphs from Gene-Tissue Expression (GTEx) Portal[155], accessed February 2016. Expression of these genes was associated with rs4888378 genotype in aortic artery only, but the association did not remain after correction for multiple testing.

Table 24: Association between rs4888378 and expression for genes at the *CFDP1-BCAR1-TMEM170A* locus. Data from GTEx Browser[155]. Effect size is expression per G allele.
*Represents significant association ($p < 0.05$) between genotype and expression (*BCAR1, CFDP1, LDHD*).
**Represents association significant after correction for multiple testing (*BCAR1*).

Gene	Tissue					
	Aortic artery		Coronary artery		Tibial artery	
	p-value	Effect size	p-value	Effect size	p-value	Effect size
BCAR1	6.70×10^{-10}**	0.26	0.052	0.13	8.90×10^{-12}**	0.22
CFDP1	0.031*	0.086	0.067	-0.13	0.94	-0.0029
CHST6	0.62	-0.035	0.19	0.1	0.23	0.055
CTRB1	0.099	0.16	0.1	0.21	0.31	0.085
CTRB2	0.32	0.1	0.26	0.15	0.84	0.016
LDHD	0.03*	-0.11	0.29	-0.075	0.94	0.0026
TMEM170A	0.6	0.023	0.78	-0.024	0.16	-0.054
ZFP1	0.94	-0.0044	0.32	0.065	0.085	0.083
ZNRF1	0.28	0.072	0.97	-0.0035	0.11	-0.079

Analysis of the Gilad/Pritchard eQTL browser, which shows eQTL identified by various studies[227] indicated 7 eQTL variants at the locus. These variants were present in non-blood-vessel tissues, and as GTEx now provides a more up to date eQTL resource than this browser, were not taken forward for further analysis. However, two DNase sensitivity QTLs (dsQTLs) were present, which give distinct information from eQTLs, showing variants at which chromatin accessibility is associated with allele. rs73605136 near *BCAR1*, and rs247454 in *CHST6* were identified as dsQTLs in Yoruban lymphoblastoid cell lines (Figure 28). These loci represent variants likely to directly affect chromatin

accessibility and transcription factor binding or nucleosome occupancy, and were therefore chosen for further analysis to explore their function. The two variants were in only weak LD with the lead SNP rs4888378; however, as with eQTL analysis, they were chosen for investigation as studying the effects of different regulatory elements at the locus may indicate through which genes the functional variation is working.

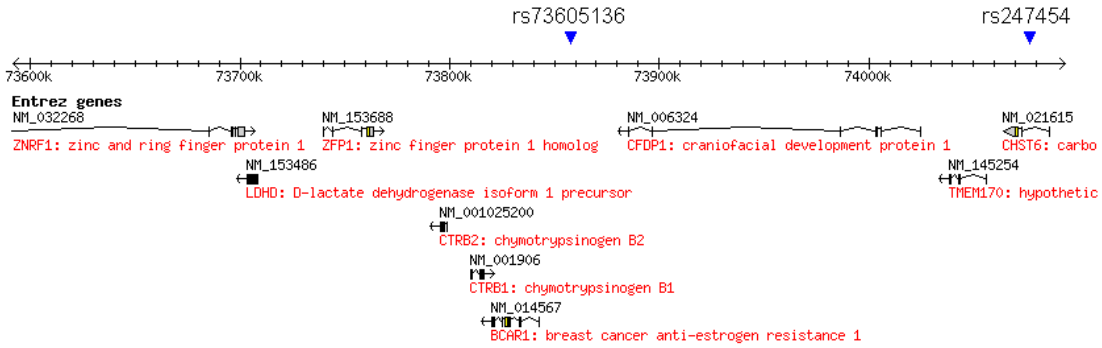

Figure 28: dsQTLs at the *CFDP1-BCAR1-TMEM170A* locus. Data shown is from the Gilad/Pritchard eQTL browser[227]. rs73605136 is shown to affect local DNaseI sensitivity

3.3 Discussion

This chapter used bioinformatics tools to investigate the chr16 cIMT locus, with the aim of prioritising variants most likely to have a regulatory effect and investigating allele-specific gene expression at the locus.

3.3.1 Selection of candidate SNPs

The first obstacle in investigating GWAS results is often presented by the LD blocks that mask the functional variant. The *CFDP1-BCAR1-TMEM170A* locus had a particularly strong extent of LD, with 214 variants in strong LD with the lead SNP. These were all selected for investigation of regulatory elements.

Regulation data from ENCODE, RoadMap Epigenomics and ElDorado was used to evaluate the 214 variants in strong LD for their regulatory potential. One criterion was the presence of the variant in a proven transcription factor binding site, as assayed by ChIP-seq data, with the idea that such a variant is more likely to disrupt binding, with possible downstream effects. ChIP experiments have previously been used to verify functional SNPs in disease; for example, a SNP in the gene *FGFR2* (fibroblast growth factor receptor 2) was found to affect binding of the transcription factor Runx2 *in vitro* (using EMSA) and *in vivo* (using ChIP), subsequently regulating *FGFR2* regulation and risk of

breast cancer[228]. Disruption of a predicted TF binding motif was therefore also used as a condition of candidate SNPs, using the tools Haploreg and Genomatix's ElDorado. Of the SNPs with the features above, these were used to select variants with changes in strong predicted transcription factor binding motifs, as assessed by core similarity (nucleotide similarity to the four consecutive highest conserved bases in the motif) and matrix similarity (nucleotide similarity across the whole motif) in ElDorado.

Location in or close to a DNase-I hypersensitive site was also deemed to be important. Risk variants for disease have been shown to be enriched in these sites[229,230], and as they act as markers for regulatory elements, variants here are more likely to disrupt regulatory activity.

Regulatory elements are also indicated by different histone signatures, characterised by methylation and acetylation present on specific residues of histone tails. It has been shown that using these histone marks can be used to assign certain functions to genomic regions[231], so they could be used to select only variants present in areas consistent with promoters (H3K4me3 and H3K9ac) and enhancers (H3K4me1 and H3K27ac). Fulfilment of the above regulatory conditions determined the shortlist of candidate SNPs to be taken forward for functional analysis (chapter 5).

Previous studies have used similar methods of examining regulatory annotations to select candidate functional SNPs, such as a study that used ENCODE regulatory data to select SNPs in regulatory elements in the 9p21 cardiovascular risk locus, finding SNPs in an enhancer impair binding of STAT1 and drive *CKDN2BAS* expression[232]. Another study identified a variant altering *ANGPTL3* expression and CAD risk through characterisation of regulatory marks and assays on DNA-protein binding and chromatin accessibility[233].

Bioinformatics data were used in this chapter to aid in the selection of potential regulatory SNPs, but it should be considered that decisions on candidate SNPs cannot be made with complete confidence. It is first necessary to choose the r^2 LD threshold over which variants are said to be in strong LD. An r^2 value of 0.8 was chosen here to follow the general consensus among researchers and online tools; Haploreg, PLINK and SNAP use 0.8 as the default threshold of strong LD and tagging SNPs[133,197,200]. Any threshold r^2 value will necessarily be arbitrary to an extent, and it cannot be ruled out that the functional variant is in lower LD with the lead SNP than the cut-off used, particularly considering the low coverage of the region by the CardioMetabochip. However, with such a large pool of potential SNPs, such thresholds are necessary in order to provide a manageable pool of the most likely

candidates. It is also possible that the functional variant is not present on the 1000 Genomes panels used for LD analysis and therefore not included in this study. However, re-analysis with the 1000 Genomes Phase I data did not show that a potentially functional variant had been overlooked.

To investigate whether a functional SNP may have been missed by imposing an LD cut-off of 0.8, the coverage of the locus by the CardioMetabochip was later examined.

Variant data for all SNPs at the locus (defined as the 305 kb region containing rs4888378 bordered by recombination hotspots) was used to identify Metabochip SNPs, and pairwise LD calculated between these and all other SNPs. Figure 29 shows the maximum LD of all common SNPs (MAF >0.05) with SNPs on the Metabochip. 60.7% of common SNPs (MAF > 0.05) were in strong LD ($r^2 \geq 0.8$) with a Metabochip SNP and therefore tagged by it. SNPs above the $r^2 = 0.8$ threshold and tagged by the chip are on average more common, so these are more likely to be captured by the chip.

These calculations were carried out after functional analysis of chosen SNPs, so did not influence the method of selection. It should also be noted that this was a basic analysis that did not consider IMT-association values of the other Metabochip SNPs as this was not available from the original association study. With this additional data, SNPs could have been prioritised or rejected based on strength of LD with other SNPs with different phenotypic association strengths. The calculations nevertheless provide an indication that although the majority of common SNPs at the locus were covered, it is possible that a functional variant was present in those not examined.

Figure 29: Common SNPs (MAF > 0.05) at the *CFDP1-BCAR1-TMEM170A* locus and their highest LD with any SNP on the Metabochip. SNPs are shown with their location relative to the genes at the locus, their highest LD with any SNP on the Metabochip (this SNP indicated by colour), and their minor allele-frequency (shown by size of point). SNPs in strong LD with Metabochip SNPs tend to be more common, but some common SNPs still lie below the threshold.

Pairwise LD calculations between SNPs are dependent on the cohort from which the samples are taken. A large number of genotyped individuals gives better confidence to LD calculations. The 1000 Genomes Pilot I data used a subset of the final cohort, so pairwise LD calculations were not finalised, and were a less accurate estimate of the real value. The more complete 1000 Genomes Phase I data was later used when this became available and examined with HaploReg, slightly altering the candidate list: 39 SNPs were removed from the "strong LD" band and 56 new ones were introduced. The changing nature of such results is likely to be an issue in bioinformatics until large enough datasets are produced and become established as standards. Larger sequencing efforts, such as Genomics England's 100,000 Genome Project, are likely to provide a robust catalogue of variation in UK subjects[234].

The CEU (Northern and Western European) 1000 Genomes data set was used to calculate LD with the lead SNP, to match IMPROVE, the discovery cohort. Linkage disequilibrium is strongly influenced by population substructure, due to the different demographic histories of ethnic groups. European and Asian populations generally have more extensive LD blocks than African populations[235,236]. SNPs in LD with the study's lead SNP should therefore be calculated from a similar population to the one genotyped to obtain the correct pool of SNPs tagged by the genotyped SNP.

The different patterns of LD observed between ethnic groups may provide useful advantages for research. Carrying out GWAS in ethnic groups other than Europeans would produce different association signals surrounding functional variants, allowing the pool of potential functional variants to be narrowed down. Previous studies have used this method to refine trait-associated loci; for example, Willer and colleagues carried out refinement on Teslovich's 2010 blood lipids GWAS in Europeans[60] to fine-map loci of interest[116]. Additional genotyping was carried out using the Metabochip in European, African, East Asian and South Asian populations; despite the relatively small sample sizes of the non-European populations, the differences in LD allowed fine-mapping of signals that in Europeans had been broader. Another study carried out trans-ethnic fine-mapping of blood lipid loci in African, East Asian and European populations, finding additional signals at some loci that increased explained phenotypic variance by 1.3-1.8×, and narrowing the lists of candidate variants at multiple genes[237].

Fine-mapping using multiple ethnic groups is of particular value in African populations, who have with shorter haplotype blocks and, on average, fewer SNPs in LD with a lead variant[236]. However, to obtain such benefits, future cohorts will need to diverge from the current model of largely

European-based cohorts: currently only 3% of published GWAS consist of African Americans, and less than 1% for other ethnic groups[238]. Additionally, for sparsely-covered loci such as the chr16 IMT locus, using a genotyping chip with denser coverage at the locus, or whole genome sequencing, would be advisable in order to separate out the association signal.

Using regulatory data to prioritise SNPs for analysis, as was performed here using ENCODE and Roadmap regulatory data, necessarily involves some degree of speculation, especially when the pool of potential candidates is large. A semi-quantitative method was used to choose SNPs, requiring presence of multiple regulatory features and changes in predicted transcription binding motifs above a particular threshold. The online tool Haploreg v3 was used to consolidate regulatory data into an easily viewable format, but a degree of subjectivity remains, meaning there is a chance that some suitable candidates are missed. A fully quantitative method of scoring SNPs would be ideal. At the time of analysis, RegulomeDB[134] only contained early data from the ENCODE project, but it has now been brought up to date with current ENCODE and Roadmap releases and represents a viable choice for ranking SNPs. The NIH's LDlink[199], which was also not available at the time of analysis, is also a useful tool that provides information about SNPs in LD and scoring their regulatory potential. As more data is produced about regulatory elements, particularly using new assays for understanding genetic regulation, variant annotation tools such as these are likely to be able to score variants with increased confidence.

3.3.2 eQTL analysis

Expression analyses are valuable in the investigation of a trait-associated locus in order to understand how functional variation is causing a phenotypic change. Expression assays in relevant tissues from the ASAP and BiKE studies revealed associations of rs4888378 with *TMEM170A* (in aortic intima-media and adventitia), and nominal significance with *BCAR1* (in carotid plaque) and *LDHD* (in aortic intima-media and adventitia). The protective minor allele was associated with lower *TMEM170A* and *BCAR1* expression, and higher *LDHD* expression. While there is a possibility of false positives due to multiple testing, the association with multiple genes supports the idea that rs4888378 may be present in an enhancer acting on multiple promoters.

To further investigate allele-specific expression with rs4888378, tissue expression data was examined from the GTEx portal, a resource combining gene expression from multiple tissues with information about genetic variation, in order to identify eQTLs. Expression data was analysed from aortic artery, coronary artery and tibial artery. Of the genes analysed, *LDHD* and *CFDP1* showed

nominally significant allelic differences in expression in one tissue, but only *BCAR1* remained significant after multiple testing, and showed differences in more than one tissue (aortic artery and tibial artery). The relationship was in the same direction as that seen previously with BiKE data, with the protective A allele being associated with lower expression.

These expression data back up the idea that the functional variant is having a phenotypic effect through the regulation of a gene or genes at the locus. The strongest candidate from these sets of data, *BCAR1*, is a particularly interesting gene for IMT phenotypes. In endothelial cells, growth factors such as VEGF stimulate phosphorylation of the BCAR1 protein, which is required for cell migration[166,185], whereas in vascular smooth muscle cells, BCAR1 is required for cell contraction through actin polymerisation[186], with phosphorylated BCAR1 being linked to the actin cytoskeleton through tensin 1[162]. Growth factor-stimulated migratory responses of VSMCs, important in the development of atherosclerosis, are mediated through BCAR1, and its expression and phosphorylation promote formation of neotima[187]. Due to these roles in blood vessel tissues, Gertow and colleagues speculate that the causal variant may be regulating transcription of *BCAR1* in these cells.

This hypothesis is supported by the eQTL data, which shows the protective A allele to be associated with lower expression of *BCAR1*. Further work would investigate whether candidates for a functional variant directly cause a change in gene expression.

eQTL analysis with the Gilad/Pritchard eQTL browser does not provide means for specified variant-gene expression associations to be calculated as GTEx does, but only shows eQTLs with genome-wide significance. However, examination of the browser revealed 7 general eQTL variants at the locus in lymphoblastoid cell lines, liver and monocytes. As GTEx now provides more comprehensive eQTL analysis with user-specified variant-transcript calculations, and as the cells lines studied from the Gilad/Pritchard eQTL browser were of limited interest for blood vessel phenotypes, these variants were not taken forward for further analysis. However, two dsQTL loci were chosen for functional analysis because genetic variants altering chromatin accessibility may directly indicate functional variants that lead to differences in gene expression[239]. DNaseI sensitivity, as a quantitative marker of open chromatin, has been shown to correlate with active regulatory regions. As the majority of dsQTLs act on nearby areas of chromatin, more confidence can be placed in a dsQTL carrying out the functional effect, compared to an eQTL variant where LD can span longer distances.

The source of samples must be taken into account regarding eQTL data. Ethical and practical issues preclude the option of taking multiple tissue biopsies from healthy subjects; ASAP and BiKE biopsies are from patients undergoing aortic valve surgery and endarterectomies. GTEx biopsies are taken from autopsies within 24 hours of death, from donors aged 21-70 who have died from cerebrovascular, liver, renal, respiratory or neurological causes or traumatic injury. It should therefore be considered that gene expression in these data sets may not be representative of that in healthy individuals.

3.3.3 Conclusions and further work

In this chapter, over 200 SNPs were identified in LD with the *CFDP1-BCAR1-TMEM17A* locus lead SNP. Bioinformatic tools with ENCODE, Roadmap and other data were used to assess regulatory features around these SNPs, from which a shortlist of SNPs for further analysis was drawn up.

Previous eQTL data at the locus, and new eQTL data from the GTEx portal were combined to identify a gene of particular interest at the locus, and to give clues about how we would expect the functional SNP to affect gene expression. Two dsQTLs were identified for further analysis, to find out how modification of chromatin accessibility might affect genetic regulation.

Further work will follow four main areas:

1. Having identified candidate functional SNPs, functional analysis should be carried out on these to determine their effect on protein binding, and on whether they directly affect gene expression.
2. To investigate further through what mechanism the functional variant is affecting IMT, association of the locus with other traits should be investigated in other cohorts. Identification of other phenotypes may give clues as to the pathways involved in the association.
3. The locus under investigation is gene-dense, with functional variation that may be affecting a number of genes, not necessarily the one it is closest to. An alternative method or methods of studying genetic regulation, particularly across distance, should be investigation.
4. Other methods of investigation of genetic regulation, particularly across distance, should be investigated.
5. This chapter has focused on the regulatory potential of SNPs to affect gene expression, due to the absence of protein-coding SNPs in the high-LD candidate list. However, looking at how

any SNPs affecting protein structure may disrupt phenotype may also give clues about through which gene we expect the functional variant to act.

4 Genotyping and association analyses of regulatory variation

4.1 Introduction

This chapter investigates further the association between the *CFDP1-BCAR1-TMEM170A* locus and IMT phenotypes, by looking at IMT and the additional phenotype IMT progression. It was hypothesised that studying additional locus-phenotype relationships may provide more information about the mechanism of the locus's association with IMT. Thus association analysis was carried out using a cohort with the related phenotype IMT progression, Progressione della Lesione Intimale Carotidea (PLIC). This was a cohort of general population subjects, in contrast to IMPROVE with its high-risk participants. Analyses were also carried out in the original discovery cohort, IMPROVE, and for replication, three cohorts used for replication in Gertow and colleagues' original study were used (WhiteHall II, Edinburgh Artery Study, Malmö Diet and Cancer Study)[119].

The speed at which carotid intima-media thickness is increasing is a distinct variable from a static value, and may differ from IMT in its relationships with other risk factors and genetic loci. Unlike single measures of IMT, progression variables are the combination of atherosclerotic burden at multiple time points. Cardiovascular biomarkers have been shown to vary over time, and measuring these over multiple time points should reduce the chance of recording atypical values and improve assessment of CHD risk[240]. IMT progression is associated with vascular risk factors and atherosclerosis [241,242], but there are conflicting reports on whether there is a significant relationship between progression and vascular events; while multiple studies report an association[243–245], a recent large meta-analysis failed to detect one[246].

Results from the IMPROVE study suggest that while standard IMT progression variables in different segments of the carotid tree are not associated with vascular events, a composite of the fastest IMT-max progression over all segments did have predictive value[247]. Such findings highlight one of the complications of using IMT phenotypes, as discussed in section 1.2.2.2: many different variables can be created by measuring different sections of the carotid tree, and by taking mean or maximum values.

The PLIC cohort comprises 2015 general population subjects attending the Atherosclerosis Centre in Bassini Hospital, University of Milan, and was designed to study the presence and progression of atherosclerotic lesions in the common carotid artery[203]. IMT measures of the common carotid artery were taken at baseline and at a 6-year follow-up for calculation of progression. As subjects were

from the general population, their cardiovascular risk was lower than those in IMPROVE, who had been selected for higher risk.

For this chapter, the lead SNP rs4888378 was genotyped in PLIC using a high-throughput TaqMan allelic discrimination assay, and the results analysed to study the relationship between the SNP and IMT-progression. Analysis was also split by sex to test for any differences in genetic effect by sex. It was noted that additional analyses increase the risk of type I error, but it was considered that these analyses were valuable, as sex is a risk factor of particular relevance to CHD. Men have a higher overall risk of CHD[23], and though differences between sexes in traditional risk factors contribute to this disparity[248], sex is one of the most significant risk factors over and above these differences[249].

The distinct physiology of men and women has led some studies to look for sex-specific effects of genetics on cardiovascular phenotypes. For example, *APOE* genotype has a greater impact on triglycerides and HDL in women than men[250], variants in *SLC2A9* affect uric acid levels (a risk factor for CVD) to a much greater extent in women[251], and variations in the platelet-derived growth factor D (*PDGFD*) gene increase the risk of CHD in women[252]. Despite this, many genetic association studies dealing with CVD do not consider that there may be an interaction between genotype and sex[253].

Therefore, in this chapter it was considered appropriate to carry out sex-stratified analyses. A different effect of the locus on IMT phenotypes between men and women could suggest differences between atherosclerotic pathways between men and women, and may indicate which gene or genes at the locus are more likely to be involved in the phenotype.

4.2 Results

4.2.1 PLIC cohort characteristics

PLIC comprises 2144 general population participants attending the Centre for the Study of Atherosclerosis, Bassini Hospital (Cinisello Balsamo, MI, Italy). The study was designed to study atherosclerotic lesions and IMT in the common carotid artery, and relationships with cardiovascular risk factors. Characteristics are shown in Table 25. It can be seen that there are differences in average levels of many of the cardiovascular risk factors between men and women. Where there are differences, risk factors are higher for men than women, with the exception of total cholesterol.

Table 25: Characteristics of PLIC.
Values expressed as mean (SEM). Characteristics are for subjects used in the analysis; i.e. those with phenotypes and rs4888378 genotypes available.

	Men (n= 795)	Women (n= 1076)	p-value (difference between sexes)
Age (years)	54.0 (0.4)	55.0 (0.3)	0.064
BMI (kg/m^2)	27.12 (0.12)	26.07 (0.14)	< 0.001
Systolic BP (mmHg)	134.3 (0.6)	129.3 (0.5)	< 0.001
Diastolic BP (mmHg)	83.4 (0.3)	80.4 (0.3)	< 0.001
Total cholesterol (mmol/l)	5.66 (0.04)	5.79 (0.03)	0.006
HDL-C (mmol/l)	1.28 (0.01)	1.53 (0.01)	< 0.001
Triglycerides (mmol/l)	1.39 (0.03)	1.10 (0.01)	< 0.001
LDL-C (mmol/l)	3.74 (0.03)	3.75 (0.02)	0.965
apoA-I (mg/dL)	141.5 (0.89)	154.6 (0.79)	< 0.001
apoB (mg/dL)	115.0 (0.98)	112.6 (0.77)	0.115
Remnant-C (mmol/L)	0.64 (0.02)	0.51 (0.01)	< 0.001
Glucose (mmol/l)	5.42 (0.03)	4.97 (0.02)	< 0.001
CCA-IMT (mm)	0.663 (0.005)	0.643 (0.004)	0.002

4.2.2 IMPROVE characteristics

IMPROVE (IMT and IMT-Progression as Predictors of Vascular Events) is a prospective multicentre longitudinal study set up to investigate carotid intima-media thickness in individuals at high risk of CVD[201]. 3711 participants (54-79 years) with at least three vascular risk factors were recruited in seven centres in Finland, France, Italy, the Netherlands and Sweden. Vascular risk factors were defined as: male sex, or female at least 5 years post-menopausal, hypercholesterolaemia, hypertriglyceridaemia, hypoalphalipoproteinaemia, hypertension, diabetes or impaired fasting glucose, smoking habits and family history of cardiovascular diseases.

As the subjects were selected for higher risk of cardiovascular disease, rather than chosen from the general population, it can be seen that risk factors such as age, blood pressure and triglycerides are higher on average than in PLIC. Characteristics of the cohort are shown in Table 26. Again, many risk factors are higher in men than in women, although more women are on anti-hypertensive therapy than men.

Table 26: Characteristics of IMPROVE.
Values expressed as mean (SEM) or percentage (frequency). Difference between sexes calculated using t-test or chi-squared test for mean and frequency data respectively.
*Significant difference between men and women.
†Log-transformed due to non-normality; SEM is approximate

	Men (n=1772)	Women (n=1931)	p-value (difference between sexes)
Age (years)	64.0 (0.13)	64.4 (0.13)	0.03*
BMI (kg/m^2)	27.4 (0.09)	27.1 (0.11)	0.03*
Systolic BP (mmHg)	142.4 (0.44)	141.6 (0.42)	0.21
Diastolic BP (mmHg)	83.1 (0.24)	80.9 (0.22)	2.16×10^{-12}*
Hypertension	67.9% (1203)	69.9% (1349)	0.21
Total cholesterol (mmol/l)	5.25 (0.02)	5.71 (0.03)	3.7×10^{-36}*
LDL-C (mmol/l)	3.39 (0.02)	3.68 (0.02)	1.7×10^{-18}*
HDL-C (mmol/l)	1.14 (0.01)	1.38 (0.01)	2.8×10^{-91}*
Triglycerides (mmol/l) †	1.42 (0.02)	1.29 (0.01)	2.8×10^{-8}*
Combined vascular event (cardiac, cerebrovascular or peripheral)	7.6% (134)	4.2% (81)	1.657×10^{-5}*
Diabetes prevalence	29.4% (521)	20.3% (392)	1.7×10^{-10}*
Lipid-lowering therapy	47.6% (827)	50.7% (965)	0.07
Anti-hypertensive therapy	54.6% (967)	59.1% (1142)	0.01*

4.2.3 Other cohort characteristics

The Whitehall II (WHII) study consists of 10,308 men and women between 1985 and 1989 from the civil service in London. Clinical measurements were taken at intervals of 5 years, and carotid measurements made in 2003-2004. The 2,138 subjects used by Gertow and colleagues were included in the analysis[205].

The Edinburgh Artery Study recruited 1,592 men and women (55-74 years old) in 1988 from general practices in Edinburgh[204]. IMT was measured 5 years after recruitment. 630 individuals had valid IMT measurements and appropriate CardioMetabochip genotyping data and were used for analysis.

The Malmö Diet and Cancer Study is a population-based prospective study that recruited 28,449 men and women of 45-73 years between 1991 and 1996[206]. IMT ultrasound measurements were taken in a random sample of 6,103 subjects, the "cardiovascular arm" of the study. 2,143 non-diabetic subjects were genotyped with the CardioMetabochip.

4.2.4 rs4888378 genotyping in PLIC

PLIC was genotyped for the lead SNP at the *CFDP1-BCAR1-TMEM170A* locus, rs4888378, using a TaqMan allelic discrimination assay. Genotyping was successful with a call rate of 90%. The minor

allele frequency of rs4888378 in PLIC was 0.44, not significantly different to the frequency of 0.43 observed in IMPROVE (χ^2 = 0.072, p = 0.788). The genotype distribution is shown in Table 27. Heterozygotes were present less often than expected, but there was no significant deviation from Hardy-Weinberg equilibrium (p = 0.41).

A number of samples could not be called with confidence due to inadequate amplification, despite optimisation of the assay with varying concentrations of $MgCl_2$ and DMSO (Figure 30).

Table 27: Observed and expected rs4888378 genotypes in PLIC. Fewer heterozygotes than expected were observed, but there was no significant deviation from Hardy-Weinberg equilibrium (p = 0.41).

Genotype	GG	GA	AA
Observed	367	908	607
Expected	358	926	598

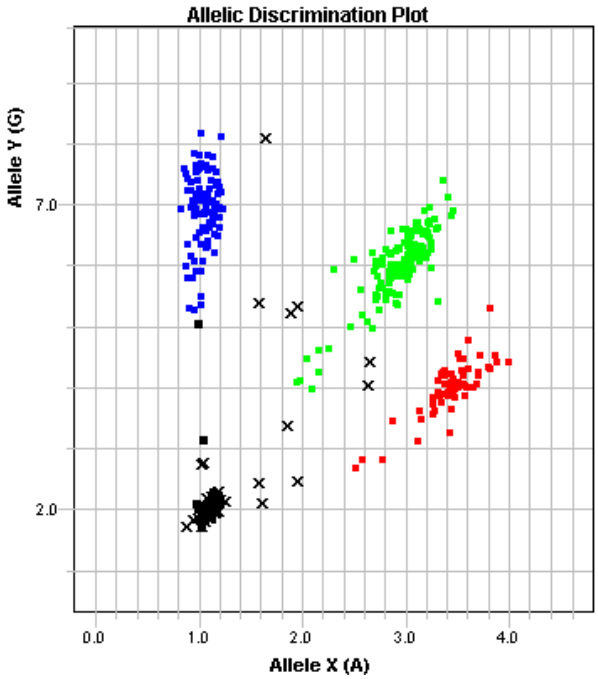

Figure 30: Example of TaqMan allelic discrimination amplification plot. The plot shows an example plate with clusters of called genotypes: GG (blue), GA (green) and AA (red). Black crosses represent samples that did not cluster clearly so could not be called with confidence.

4.2.5 Association analyses

4.2.5.1 Association of rs4888378 with basal IMT in PLIC

The relationship between rs4888378 and IMT was investigated in the PLIC cohort to allow for the investigation of the related phenotype of IMT progression. There was no significant association between rs4888378 genotype and basal CC-IMT (p=0.52; ANCOVA, adjusted for age, sex and

smoking; Figure 31), in contrast to the significant association seen in Gertow and colleagues' original study.

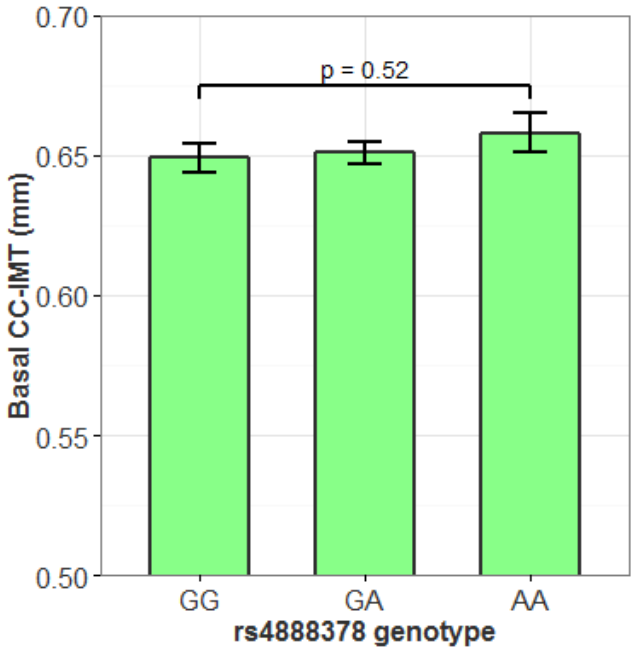

Figure 31: Basal CC-IMT by rs4888378 genotype in PLIC. No significant association between genotype and basal CC-IMT was observed (p=0.52; ANCOVA, adjusted for age, sex and smoking).

4.2.5.2 Association of rs4888378 with IMT progression in PLIC

rs4888378 was also not significantly associated with 6-year progression of CC-IMT in PLIC (p = 0.45, ANCOVA, adjusted for age, sex and smoking). A non-significant trend for lower CC-IMT progression with each A allele was observed.

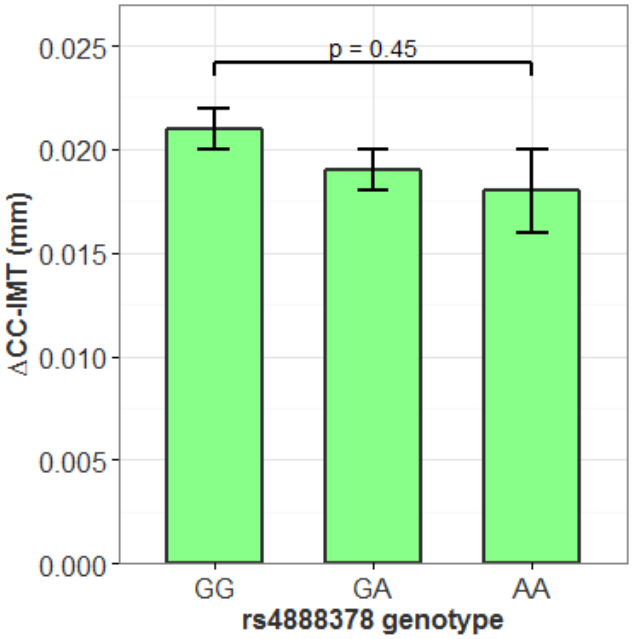

Figure 32: 6-year progression of CC-IMT with rs4888378 genotype in PLIC. No significant association between genotype and CC-IMT progression was observed p = 0.45, ANCOVA, adjusted for age, sex and smoking).

4.2.5.3 Association with cardiovascular risk factors

rs4888378 was associated with the cardiovascular risk factors of waist circumference (p=0.030) and waist/hip ratio (p=0.007) (Table 28, model adjusted as above). The minor A allele was associated with higher values of both risk factors. Taking into account that waist and waist/hip ratio positively correlate with carotid IMT[254,255], the relationship with the A allele (protective in regards to IMT), this relationship is in the opposite direction to that expected.

Table 28: Cardiovascular risk factors and CCA-IMT by rs4888378 genotype in PLIC.
P(trend) values were adjusted for age, sex, and smoking habits. Values are expressed as mean (SEM).
*Log-transformed due to non-normality; SEM is approximate.

Parameter	GG (n=603)	GA (n=901)	AA (n=367)	P (trend)
Basal CCA-IMT (mm)*	0.649 (0.005)	0.651 (0.004)	0.658 (0.007)	0.518
Annual ΔCCA-IMT (mm)*	0.021 (0.001)	0.019 (0.001)	0.018 (0.002)	0.451
BMI (kg/cm^2)	26.2 (0.13)	26.7 (0.16)	26.7 (0.22)	0.157
Waist (cm)	89.3 (0.52)	91.1 (0.43)	91.6 (0.66)	**0.030**
Hip (cm)	103.4 (0.34)	104.5 (0.30)	103.5 (0.57)	0.410
Waist/hip ratio	0.86 (0.03)	0.87 (0.03)	0.88 (0.04)	**0.007**
Triglycerides (mmol/l)	1.04 (0.02)	1.06 (0.02)	1.06 (0.03)	0.078
Glucose (mmol/l)	5.10 (0.03)	5.19 (0.03)	5.21 (0.04)	0.825
Remnant-C (mmol/l)	0.54 (0.01)	0.58 (0.01)	0.57 (0.02)	0.078

4.2.5.4 Sex-specific analysis in PLIC

Genetic analyses in PLIC were separated by men and women to test for genotype-phenotype associations that may differ between sexes. Association between rs4888378 and basal CC-IMT remained not significant in both men and women (p = 0.93 and 0.27 respectively; Figure 33).

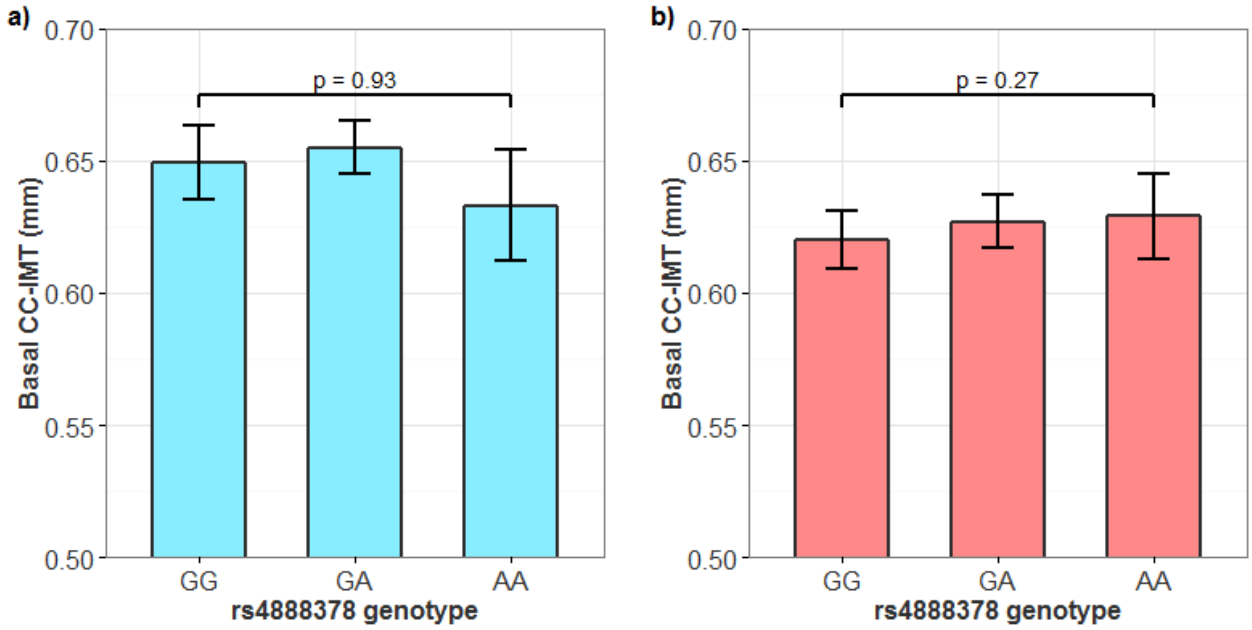

Figure 33: Basal CC-IMT by rs4888378 genotype in PLIC in (a) men and (b) women. No significant association between genotype and basal CC-IMT was observed (p = 0.93 and 0.27 for men and women respectively; ANCOVA, adjusted for age and smoking).

Progression of IMT did show a difference in association with rs4888378 between sexes. In women, CC-IMT progression was 10% lower for each protective A allele (AA vs GG, p=0.04, ANCOVA and Bonferroni post-Hoc analysis), while no association was seen in men (p = 0.94) (Figure 34).

Direct validation of this IMT progression finding could not be carried out, as the phenotype had not been measured in the available cohorts. However, data on basal IMT was split by sex to test for any male/female differences.

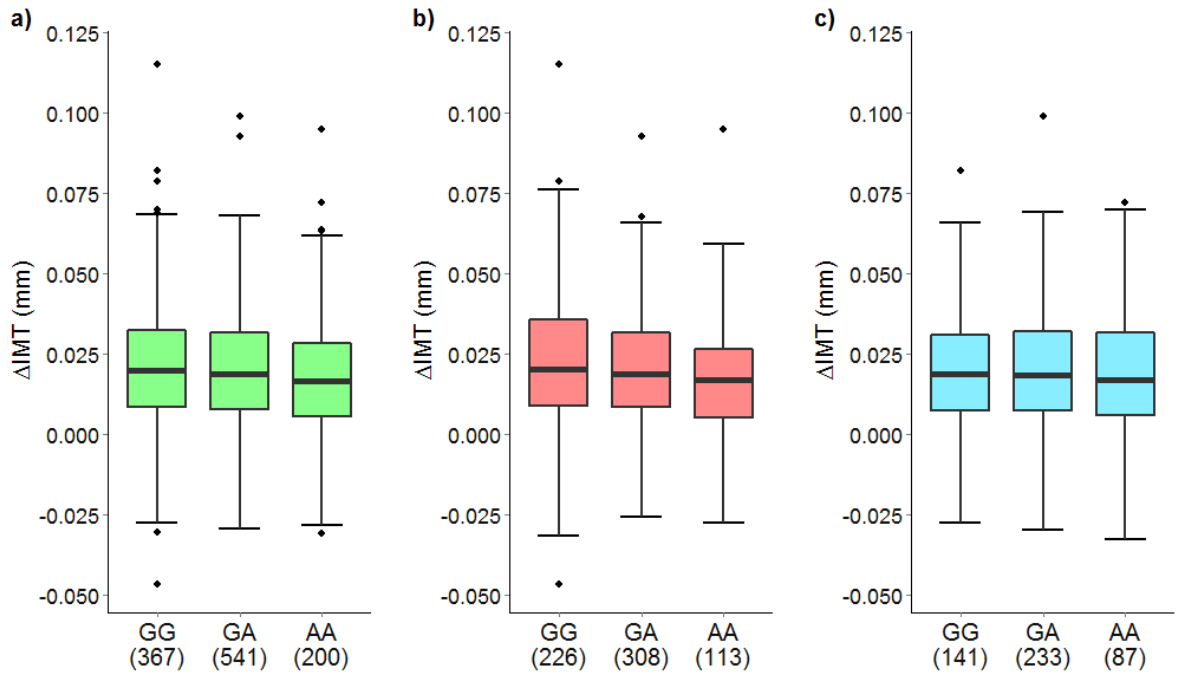

Figure 34: Annual change in IMT in PLIC. a) Whole cohort: despite a trend for lower ΔIMT with each A allele, no significant association was present with genotype (p = 0.45). **b) Women:** ΔIMT in AA genotype was lower than in GG genotype (p = 0.04), although overall trend was not significant (p = 0.09). **c) Men:** no significant association between genotype and ΔIMT (p = 0.94)

4.2.5.5 Sex-specific: meta-analysis for association with common-carotid IMT

Additional analyses into sex-specific associations were carried out. A meta-analysis was carried out on the association of rs4888378 with CC-IMT (chosen as it was the variable measured in all cohorts) in men and women. The cohorts used were PLIC, IMPROVE and three replication cohorts used in Gertow and colleagues' original study: WHII, EAS and MDC. Overall 5119 men and 4369 women were included in the analysis.

Linear regression beta-values and SEs, using an additive genetic model, were obtained for each study in men and women separately. Models were adjusted for age and, in IMPROVE, MDS coordinates, to account for population substructure. Meta-analysis was carried out using a random-effects model.

In women, rs4888378 was associated with CC-IMT in the expected direction (a decrease of 0.0047 mm for each A allele; $p=1.63\times10^{-4}$). However, in men, no significant association was found (β=-0.0029, p=0.07) (Figure 35). It can also be seen in Figure 35 that the association found originally in IMPROVE attained significance only in women. It therefore appears that the functional variation is having an effect more strongly, or only, in women.

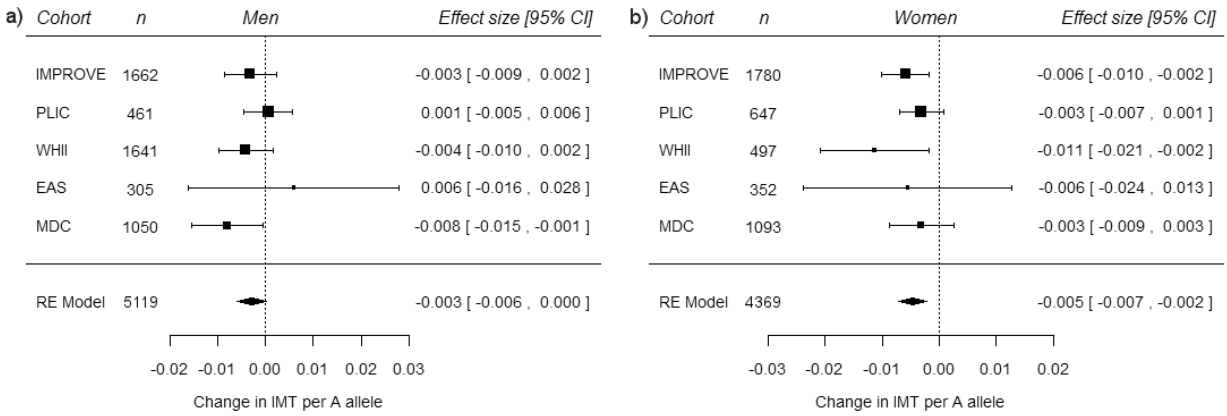

Figure 35: Forest plot showing meta-analysis of CC-IMT by rs4888378 allele in (a) men and (b) women. (a) No overall association between the SNP and IMT is observed in men (β=-0.0030, p=0.0591). **(b)** In women, the A allele is associated with a decrease in IMT for each A allele (β=-0.0046, p=1.59×10-4).

4.2.5.6 Sex-specific: other IMT phenotypes in IMPROVE

IMT phenotypes other than the common-carotid variable were not focused on as they were not available in all cohorts for meta-analysis. The IMT values for different segments of the carotid tree in IMPROVE are shown in

Table 29. It can be seen here that some IMT segments are associated with rs4888378 genotype in men (in addition to all segments overall and in women); particularly, the bifurcation and internal segments and those composite measures including them. With the exception of mean internal carotid arteries, the association effect size is greater in women for all variables, and the p-values for association are between 1 and 4 orders of magnitude smaller in women than men.

Table 29: IMT phenotypes by rs4888378 genotype in IMPROVE. Phenotype values shown are mean (standard error). P values calculated by linear or logistic regression as appropriate; adjusted for age, MDS coordinates and sex (where applicable). All IMT variables were log-transformed before analysis.

	Overall					Men					Women				
Variable	**GG**	**GA**	**AA**	**Effect size**	**p**	**GG**	**GA**	**AA**	**Effect size**	**p**	**GG**	**GA**	**AA**	**Effect size**	**p**
Mean common carotid	0.749 (0.004)	0.745 (0.003)	0.729 (0.005)	-0.005	5.71×10^{-3}	0.772 (0.007)	0.775 (0.006)	0.758 (0.008)	-0.004	0.185	0.726 (0.005)	0.717 (0.004)	0.705 (0.006)	-0.006	5.06×10^{-3}
Mean common carotid (cm closest to bifurcation)	0.808 (0.005)	0.796 (0.004)	0.789 (0.006)	-0.005	1.29×10^{-2}	0.828 (0.007)	0.823 (0.006)	0.824 (0.010)	-0.002	0.380	0.787 (0.007)	0.772 (0.005)	0.760 (0.007)	-0.006	6.48×10^{-3}
Mean internal carotid arteries	0.903 (0.011)	0.870 (0.008)	0.840 (0.012)	-0.012	5.06×10^{-4}	0.973 (0.017)	0.948 (0.013)	0.907 (0.020)	-0.013	0.013	0.832 (0.014)	0.799 (0.010)	0.784 (0.015)	-0.011	0.014
Mean bifurcation	1.167 (0.012)	1.150 (0.010)	1.096 (0.015)	-0.013	5.58×10^{-5}	1.240 (0.017)	1.226 (0.015)	1.182 (0.023)	-0.012	0.012	1.093 (0.015)	1.080 (0.012)	1.025 (0.019)	-0.014	1.32×10^{-3}
Mean whole carotid tree	0.907 (0.006)	0.890 (0.005)	0.862 (0.007)	-0.010	1.97×10^{-6}	0.954 (0.009)	0.943 (0.007)	0.917 (0.011)	-0.009	4.01×10^{-3}	0.859 (0.008)	0.842 (0.006)	0.817 (0.009)	-0.010	9.47×10^{-5}
Maximum whole carotid tree	2.113 (0.025)	2.029 (0.019)	1.919 (0.030)	-0.020	1.34×10^{-7}	2.264 (0.037)	2.184 (0.029)	2.093 (0.047)	-0.017	1.29×10^{-3}	1.957 (0.031)	1.888 (0.025)	1.776 (0.038)	-0.022	2.42×10^{-5}
Mean of maximal values in each segment	1.279 (0.009)	1.253 (0.007)	1.210 (0.011)	-0.011	1.57×10^{-7}	1.345 (0.013)	1.333 (0.011)	1.297 (0.017)	-0.008	0.014	1.211 (0.011)	1.180 (0.009)	1.138 (0.013)	-0.013	7.96×10^{-6}
Presence of plaque	0.734 (0.013)	0.685 (0.011)	0.620 (0.019)	-0.283	1.88×10^{-7}	0.789 (0.017)	0.772 (0.014)	0.715 (0.026)	-0.221	0.010	0.677 (0.020)	0.607 (0.016)	0.542 (0.026)	-0.304	2.29×10^{-5}

4.2.5.7 Sex-specific: association with cardiovascular event rate

Following the discovery of the difference between rs4888378-IMT association in men and women, the genotypic association with events was studied to see if it showed the same pattern. As event data were not available in the replication cohorts, combined vascular event rate (cardiac, cerebrovascular or peripheral) was analysed by sex in IMPROVE.

As with the genotypic association, the association between rs4888378 and vascular events was significant only in women. Overall each A allele was associated with a reduction in hazard ratio of 23% (Cox proportional hazard model; p=0.01); in women a reduction of HR of 36% (p=0.01) and in men no significant association (reduction in HR 14%, p=0.24) (Figure 36).

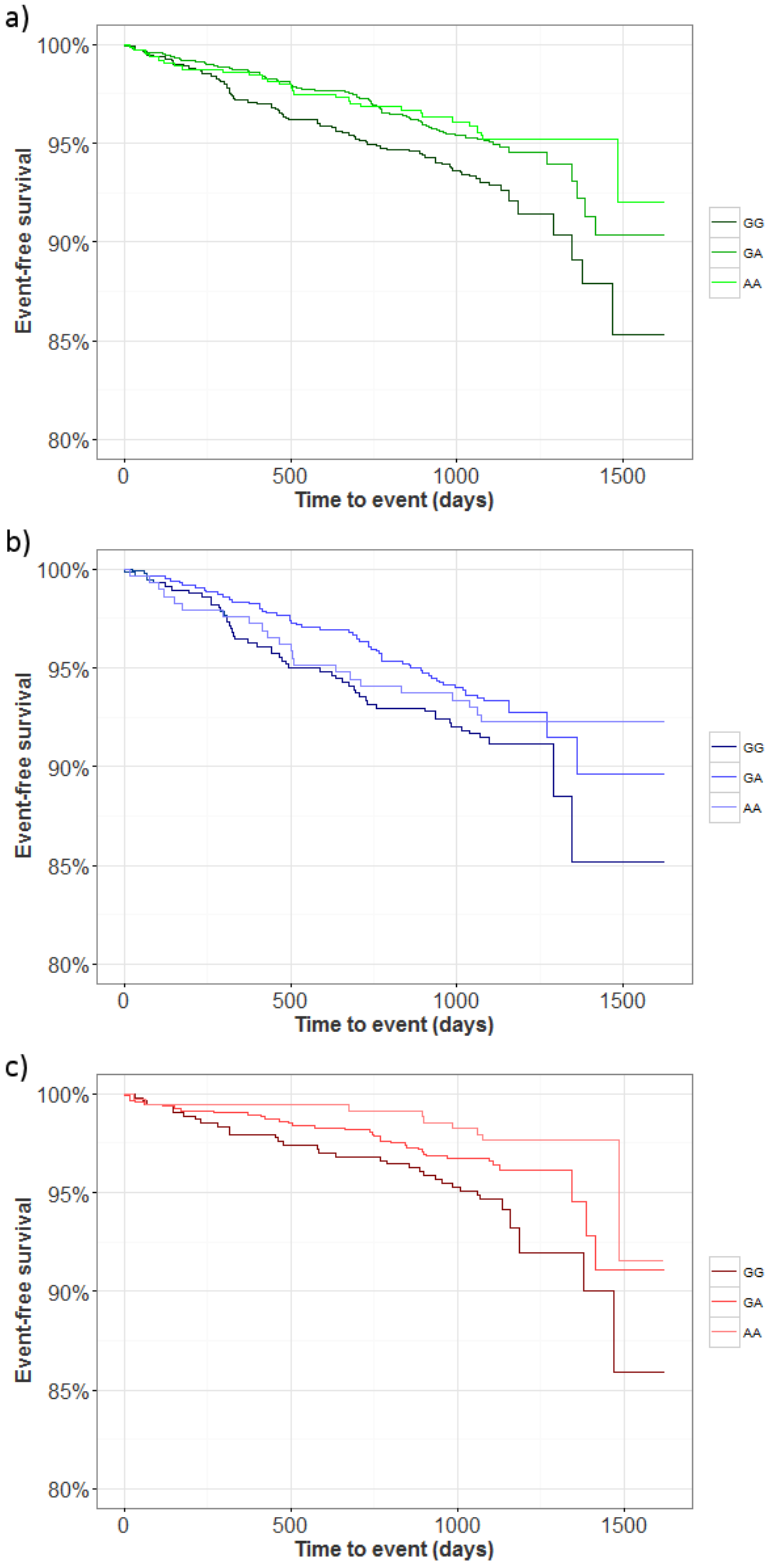

Figure 36: Vascular-event-free survival by rs4888378 genotype in (a) all subjects, (b) men and (c) women. Vascular events comprise cardiac, cerebrovascular and peripheral events; data from IMPROVE. **a)** Overall, each A allele is associated with a significant reduction in HR of 23% (Cox proportional hazard model; p=0.01). **b)** In men, no significant difference is seen (reduction in HR 14%, p=0.24) **c)** In women, each A allele is associated with a significant reduction in HR of 23% (reduction in HR 36%, p=0.01).

4.2.5.8 Interaction between genotype and sex

Heterogeneity between men and women was tested during the meta-analysis of rs4888378 on vascular events; no significant heterogeneity between the two was detected (Q(df = 1) = 1.31, p = 0.25). It was therefore considered to be unlikely that there would be sufficient power to detect an interaction between genotype and sex. Due to this lack of power, and as interaction statistics were not readily available for each cohort, a formal test for interaction was not carried out.

4.2.5.9 Sex-specific association with cardiovascular risk factors (PLIC)

As seen in 4.2.5.3, rs4888378 was overall associated with the risk factors of waist circumference and waist-hip ratio, with the A allele being associated with high levels of the risk factor. In women, the SNP was also associated in the same direction with BMI, hip circumference and glucose (Table 30). In men, only hip circumference showed an association (Table 31), with the lower-risk A allele being associated with smaller circumference.

Table 30: Cardiovascular risk factors and CCA-IMT by rs4888378 genotype in PLIC – women only.

Parameter	GG	GA	AA	p (trend)	p (post-hoc)		
					GG vs GA	GG vs AA	GA vs AA
BMI (Kg/cm^2)	25.4 (0.20)	26.2 (0.20)	26.8 (0.30)	0.001*	0.020*	0.001*	0.052
Waist (cm)	84.30 (0.61)	86.4 (0.55)	88.2 (0.97)	0.003*	0.035*	0.002*	0.082
Hip (cm)	102.9 (0.48)	104.3 (0.43)	104.7 (0.72)	0.012*	0.015*	0.009*	0.614
Waist/hip ratio	0.82 (0.01)	0.83 (0.01)	0.84 (0.01)	0.023*	0.340	0.006*	0.036*
Triglycerides (mmol/L)	1.05 (0.02)	1.11 (0.02)	1.16 (0.04)	0.075	0.110	0.039*	0.330
Glucose (mmol/L)	4.92 (0.03)	4.95 (0.03)	5.07 (0.05)	0.025*	0.361	0.007*	0.037*
Remnants-C (mmol/L)	0.48 (0.01)	0.51 (0.01)	0.52 (0.02)	0.075	0.111	0.039*	0.321
Basal CCA-IMT (mm) *	0.620 (0.011)	0.627 (0.010)	0.629 (0.016)	0.274	0.799	0.962	0.868
Annual ΔCCA-IMT (mm) *	0.026 (0.001)	0.020 (0.001)	0.017 (0.002)	0.246	0.405	0.040*	0.201

***P(trend)* values were adjusted for age and smoking.**

*Significant with $p < 0.05$.

†Log-transformed due to non-normality; SEM is approximate.

Table 31: Cardiovascular risk factors and CCA-IMT by rs4888378 genotype in PLIC – men only.

Parameter	GG	GA	AA	p (trend)	p (post-hoc)		
					GG vs GA	GG vs AA	GA vs AA
BMI (Kg/cm^2)	27.2 (0.23)	27.3 (0.17)	26.5 (0.23)	0.139	0.757	0.118	0.050*
Waist (cm)	96.8 (0.64)	97.6 (0.72)	96.2 (0.50)	0.331	0.382	0.541	0.154
Hip (cm)	104.7 (0.45)	104.6 (0.37)	101.7 (0.86)	0.001*	0.833	0.001*	0.001*
Waist/hip ratio	0.92 (0.01)	0.93 (0.01)	0.93 (0.01)	0.150	0.218	0.055	0.316
Triglycerides (mmol/L)	1.35 (0.05)	1.43 (0.04)	1.34 (0.06)	0.572	0.289	0.719	0.602
Glucose (mmol/L)	5.34 (0.05)	5.48 (0.05)	5.35 (0.08)	0.234	0.105	0.727	0.298
Remnants-C (mmol/L)	0.62 (0.02)	0.66 (0.02)	0.59 (0.03)	0.571	0.298	0.719	0.602
Basal CCA-IMT (mm) †	0.649 (0.014)	0.655 (0.010)	0.633 (0.021)	0.934	0.500	0.473	0.852
Annual ΔCCA-IMT (mm) †	0.019 (0.001)	0.020 (0.001)	0.019 (0.002)	0.519	0.829	0.943	0.733

P(trend) values were adjusted for age and smoking.
*Significant with $p < 0.05$.
†Log-transformed due to non-normality; SEM is approximate.

4.3 Discussion

4.3.1 Overview

In this chapter, the locus on chromosome 16 previously associated with carotid intima-media thickness and CAD was investigated using additional cohorts and further analyses on IMT progression. The association between the lead SNP rs4888378 and baseline IMT was not replicated, and no overall association was seen with IMT progression. However, an analysis split by sex found an association between the SNP and progression in women. To further investigate this apparent difference between sexes, a meta-analysis was carried out on CC-IMT by rs4888378 allele using IMPROVE, PLIC, and three of the replication cohorts used in Gertow's study (WHII, MDC and MDC). While the genotype-IMT association was seen as expected in women (a decrease of 0.0047 mm for each A allele; $p=1.63\times10^{-4}$), no significant association was seen in men (Figure 35).

This difference between sexes was robust, also being seen for vascular events: in IMPROVE, each A allele was associated with a 36% lower hazard ratio (p=0.01), while no significant relationship was observed in men.

IMT values in other segments, studied only in IMPROVE, did show some association with rs4888378 in men, for variables containing the bifurcation or internal carotid artery, although these were weaker than those seen in women. These variables could not be studied in the meta-analysis and would require replication, but if confirmed, the results would suggest the pathways implicated at the locus that differ most between the sexes might be more important at certain segments of the carotid tree.

Previously eQTL data from the ASAP study and GTEx browser implicated *BCAR1* as one of the genes affected by the variation at the *CFDP1-BCAR1-TMEM170A* locus. This finding becomes particularly interesting when we consider the sex-specific associations with CC-IMT, IMT-progression and vascular events seen in this chapter.

It is unclear what differences in CVD pathology exist between the sexes, such that some genetic variants have a stronger effect on the phenotype of women. Of the cardiovascular risk factors, only female hormonal status has been shown to be sex-specific[248]. Oestrogen is thought to have a protective effect, and with its decrease in the body after menopause, the levels of LDL, total cholesterol and triglycerides increase, promoting atherogenesis[256]. In addition to its effects on the blood lipid profile, oestrogen may affect endothelial cell function directly, as shown by the fact that oestrogen therapy is associated with enhanced fibrinolytic potential[257], a possible protective effect – reduced plasma fibrinolytic activity has been shown to be a marker of CVD risk[258].

It could be hypothesised that there is an interaction between the effect of oestrogen on endothelial cells during plaque formation, and the functional variant's effect on plaque in the arteries. If *BCAR1* is indeed the gene causing the effect, this interaction could occur through *BCAR1*'s known role in endothelial and vascular smooth muscle cells[141]. The BCAR1 protein's connection with oestrogen is of particular relevance here, it having been characterised under the name *breast cancer antiestrogen resistance 1* after its overexpression was found to confer antiestrogen resistance in breast cancer cells[190]. In addition to its conceivable role in atherosclerosis, it is a strong candidate for a gene having an interaction with sex, and will be studied further in later chapters.

Additional differences in association seen here may also help to elucidate the role of the functional gene – *BCAR1* or otherwise – in affecting IMT. The weaker, but statistically significant, rs4888378-IMT association seen in some carotid IMT segment measures seen in IMPROVE suggest that the implicated protein is involved in processes that differ from the common-carotid artery to the bifurcation.

One key difference between the common-carotid artery and the bifurcation is the stresses experienced by the vessel walls. The intima at the bifurcation experiences much greater shear stress due to the increased friction of blood at the separation of the vessel walls (Figure 37). Shear stress has been shown to affect atherosclerosis: plaques form more readily in regions of low stress[259] and intima-media thickness is greater[260]. It has been proposed that this is due to slower blood flow allowing greater deposition of atherogenic particles[86], and/or effects on endothelial function[261,262]. Shear stress modulates expression of proteins involved in vascular remodelling such as VEGF and PDGF[182], and of particular relevance here, to tyrosine phosphorylation of BCAR1[183]. Mechanical stretch produced by shear stress has been suggested to expose the protein's substrate domain, promoting the phosphorylation of tyrosine residues[184]. BCAR1 at focal adhesions downstream of blood flow is exposed to greater shear stress and thus experiences more phosphorylation to protein upstream of blood flow[263]. This raises the question of whether shear stress in the carotid artery could influence the effect of BCAR1 in remodelling of endothelial cells to produce atherosclerotic lesions.

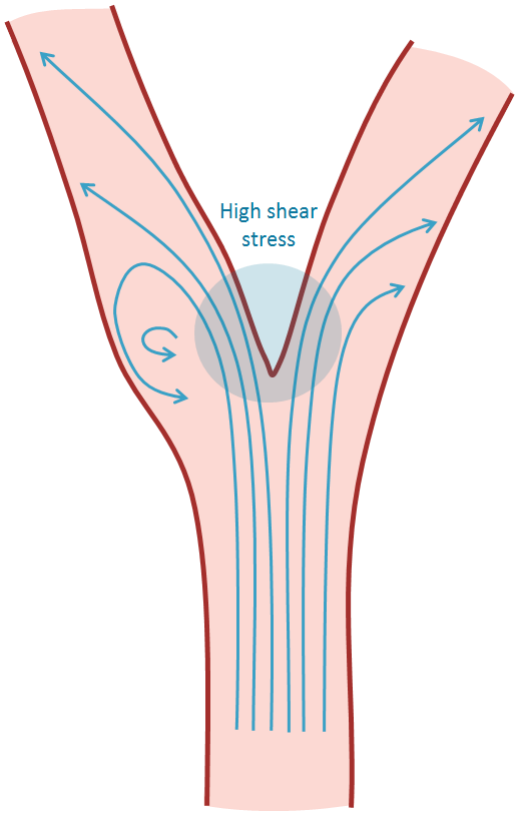

Figure 37: Blood flow at the carotid bifurcation, showing area of high shear stress. Stress is highest at the bifurcation itself, and lowest friction occurs on the outside walls of the carotid bifurcation and bulb[259].

4.3.1.1 rs4888378 and basal IMT in PLIC

An unexpected finding of the analysis was that the relationship of rs4888378 with basal CC-IMT was not replicated in PLIC. This lack of significance may be a consequence of study power: the number of women in PLIC was smaller than that in IMPROVE (1076 compared to 1931), an effect compounded by a genotyping assay with a sub-optimal call rate of 90%.

It is also possible that the low call rate might have introduced bias in the results; certain genotypes, especially heterozygotes, are less easy to call when they have not amplified well. These genotypes may therefore be absent from analysis more often than by chance. (Heterozygotes were indeed present less often than expected – 908 compared to 925.8 – but the deviation from Hardy-Weinberg equilibrium was not significant ($p = 0.41$)). Unfortunately more robust genotyping assays could not be used; genotyping was also trialled using a KBioscience KASP allelic discrimination assay but this produced a poorer call rate.

The lack of association is most likely to be related to patient characteristics. Patients with higher levels of risk factors, as in IMPROVE, are more likely to accumulate higher IMT and suffer

cardiovascular events, increasing the chance of seeing SNP-associated differences. It is worth noting that the other general-population cohorts (EAS and MDC) did not individually show a significant association in women, but contributed to the overall significance in the meta-analysis (Figure 35). Patients with more severe atherosclerosis may also experience different effects of the SNP on IMT to those with little atherosclerosis, depending on the phenotypic mechanism.

4.3.1.2 rs4888378 and IMT progression in PLIC

As with basal IMT, IMT-progression was not seen to associate with rs4888378 overall in PLIC, although a trend was seen in the expected direction: a slower progression of IMT for each A allele. This may be an effect of chance or an undetectable relationship due to reduced power. The lack of significant relationship here is not particularly informative, as none was found for basal IMT; if there is insufficient power to detect a small effect for basal IMT, we would expect similar results here.

Stratification by sex, however, produced an interesting finding, with women with two A alleles having slower progression of IMT than those with two G alleles. However, it was noted that an overall trend as measured by logistic regression was not present; significance was only reached when comparing AA and GG genotypes. The finding was therefore treated with caution and investigated further with the meta-analysis and events analysis before drawing conclusions.

The significance of finding an association with IMT progression depends on a number of factors. Progression values will be different depending on whether they are adjusted for baseline or not; an increase in IMT might be more relevant in terms of risk of a CVD event if IMT is already high, or it might be considered that a faster increase in IMT would be expected when atherosclerosis has already progressed a significant amount. Values in PLIC were not adjusted for baseline. It can also be expected that progression rates will differ depending on when the baseline value is taken. Baseline measurements taken in younger subjects with fewer cardiovascular risk factors are likely to be smaller with more potential for progression. For these reasons IMT progression at this stage is not a well-defined variable for association analysis.

4.3.1.3 IMT meta-analysis

To further investigate the effect of stratifying by sex, the data originally used to identify the locus and used with the PLIC cohort to assess the effect of sex stratification on association with the locus. WHII, EAS and MDC provided three additional cohorts for meta-analysis. Two cohorts used in the original study, Rotterdam-I and Rotterdam-II, were not used as sex-specific data was not readily

available. A random-effects model was chosen after considering the heterogeneity in IMT measurement methods between cohorts.

As discussed in section 1.2.2.2, IMT variables may vary between different segments as they reflect different processes and stresses. For the meta-analysis, common-carotid IMT was used as it was the variable available in all cohorts; it is also the variable most commonly measured and studied[83]. Progression data were not available in other cohorts to do so, but the difference between sexes in association with IMT progression should also be replicated in future.

4.3.1.4 Sex differences in event rate

It is of particular interest that the sex difference in rs4888378-phenotype association was present for event rate as well as IMT. IMT has been shown to predict cardiovascular events[72,77,264], and the SNP was associated with CAD risk in the original study[119], but it is desirable to be sure that the observed sex differences are present through to the relevant endpoint of CVD, the phenotype that is ultimately of clinical importance.

Cardiovascular event data were unfortunately not available in the replication cohorts, meaning this analysis had to be restricted to IMPROVE. To replicate such data, event definition would also have to be considered. Here combined vascular events (cardiac, cerebrovascular or peripheral) were used, this being the variable generally used as the outcome in the main IMPROVE papers[202].

4.3.1.5 Association with cardiovascular risk factors in PLIC

As no association was found between rs4888378 and cardiometabolic risk factors in IMPROVE[119], the association with weight, BMI, glucose and hip circumference in PLIC was unexpected. So too was the direction of effect, with the protective IMT allele being associated with higher-risk parameters.

For example, we expect BMI to be associated with IMT, being strongly associated with CHD over and above the effect of its correlation with other cardiovascular risk factors[46,265], but here the protective IMT genotype was associated with higher BMI. In addition, no such association was seen in IMPROVE. Hip circumference may confer higher risk if acting as a measure of adiposity[266], but greater hip circumference after BMI adjustment has been associated with protection against heart disease[267,268].

The unexpected directions of effect may be due to IMPROVE and PLIC's different characteristics (high risk versus general population), but the unexpected direction and lack of relationship in

IMPROVE suggest the observed association may have occurred by chance, and that the variant's mechanism of action is not related to adiposity or related measures such as BMI. If this is the case, it would suggest the IMT phenotype is affected through non-lipid or weight-related factors. Increased IMT could be occurring though inflammatory factors affecting formation of plaque in vessel walls, or through factors affecting cells' ability to remodel and form the plaque. If the observations here are true, it seems unlikely that decreased BMI would here be causal in increased IMT, as the opposite relationship between the two is known[269]. It would be expected that decreased IMT and BMI would be phenotypes caused by unrelated pathways.

It is possible that the SNP is present in an enhancer that acts upon several genes, in which case it could be affecting unrelated pathways that cause differences in BMI. To test this theory, published GWAS data on BMI[270] was visualised through LocusZoom[271], showing no variant-BMI associations with rs4888378 or any other variants at the locus.

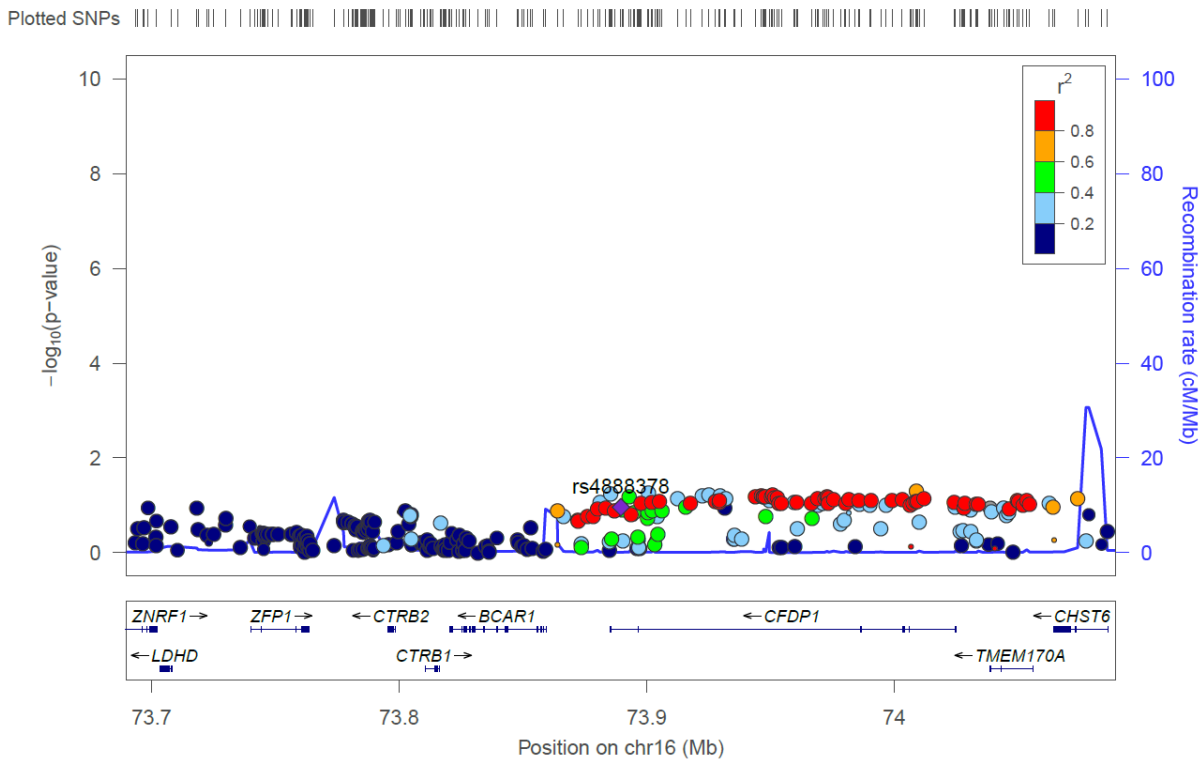

Figure 38: Regional association results for BMI at the *CFDP1-BCAR1-TMEM170A* locus. Figure produced using LocusZoom[271]; GWAS data from the GIANT consortium[270]. Published GWAS results show no association with BMI for any variant at the locus.

4.3.2 Conclusions and further work

In this chapter, samples from a cohort studying the progression of IMT were genotyped for the lead SNP at the *CFDP1-BCAR1-TMEM170A* locus to investigate relationship with IMT phenotypes. Results from this cohort and from a meta-analysis of five cohorts suggest the effect of the variation on common-carotid IMT and vascular events is present only in women. At the bifurcation and internal carotid artery, the difference may be less pronounced, with men showing a weaker rs4888378-IMT association.

This difference was considered in light of the findings of chapter 3, which implicated *BCAR1* as the gene that might be causing the phenotype. The protective effect of oestrogen on atherosclerosis, and the role of both oestrogen and *BCAR1* in endothelial cell function, suggest a possible interaction between the two that may be involved in the mechanism of the observed effect on atherosclerosis and risk of CVD.

Further work will examine the *BCAR1* gene and protein, to investigate its possible role in intima-medial thickening and atherosclerosis. Genetic variation will be studied for characterised or previously unknown variants that might affect relevant processes and disrupt the IMT phenotype. Coding variants, yet unexplored at the locus, may provide more information about how genes at the locus have an effect.

5 Functional analysis of regulatory variation at the *CFDP1-BCAR1-TMEM170A* locus

5.1 Introduction

The *CFDP1-BCAR1-TMEM170A* locus on chromosome 16 had previously been associated with IMT and CAD risk[119]. In order to understand the reason for such an association, and ultimately gain clinical value, the next step is to discover the causal variant and characterise the mechanism by which the variant exerts its effect. While the lead SNP is intronic in *CFDP1*, it is not clear whether this is the gene through which the functional variation at the locus influences IMT: there are numerous genes in proximity to the lead SNP, and many SNPs in LD with it, spanning multiple genes (Figure 22).

In chapter 3 the locus was examined to determine which variants are most likely to be the functional variant. Various bioinformatics data were used to evaluate the regulatory potential of variants in strong LD with the lead SNP, rs4888378, from which a shortlist of candidate functional SNPs was drawn up. These candidate SNPs are associated with features that might be found at areas of genetic regulation. For example, they have histone marks associated with active promoters or enhancers, have evidence of proteins binding to their sequence, or alter a predicted transcription factor binding motif.

After selecting candidates on this basis, this chapter examined whether proteins actually bind to the sequences and whether this binding differs by allele, and attempted to identify these proteins. If a SNP is shown to affect the binding of a protein by allele, the action of the protein may be affected, causing a difference in gene regulation.

Chapter 3 also examined expression data from relevant tissues, finding that *BCAR1* was most robustly associated with the lead SNP genotype and implicating it as the gene most likely to be involved. This association also needs to be verified for the candidate SNPs: we know them to be associated *in vivo*, but do they actually cause a difference in expression?

5.1.1 DNA-protein interactions

In this chapter, binding of proteins to the candidate SNPs was investigated using electrophoretic mobility shift assays (EMSA). EMSA is a functional technique used to detect protein binding to a DNA sequence. Forward and reverse DNA oligos are generated for each allele of the SNP of interest, consisting of a short section of the sequence surrounding the SNP (here 25 bp). The oligos are labelled with biotin and annealed to create double-stranded probes for each allele. Nuclear extract is

produced from cells cultured *in vitro* in order to obtain a solution of nuclear proteins. The labelled oligos are incubated with nuclear extract, allowing nuclear proteins to bind to the DNA in a simple model of how they may act *in vivo*. This reaction is run on a polyacrylamide gel, and the biotin-labelled probes are visualised using a suitable chemiluminescent detection module. Probes with bound nuclear proteins migrate more slowly on the gel than free probes and are identifiable on the gel as a "shifted" band, and allele-specific protein binding can be identified.

Multiplex competitor EMSA is a technique that uses a cocktail of unlabelled probes, consisting of known transcription factor binding motifs, to compete for binding with nuclear proteins[209]. An excess of these unlabelled probes are added to nuclear extract and incubated before addition of labelled SNP sequence. The protein that binds to a labelled SNP sequence under standard assay conditions here is competed out with the excess of unlabelled binding motif instead, causing the labelled DNA-protein band to disappear, and implicating this motif as the one belonging to the transcription factor of interest. The technique was here used to identify a protein that showed differential binding to a candidate SNP. EMSAs were also used to evaluate protein binding of the two DNAse-I sensitivity QTL variants at the *CFDP1-BCAR1-TMEM170A* locus identified in chapter 3.

5.1.2 Luciferase reporter assays

The effect of SNP genotype on gene expression was also assessed using luciferase reporter assays. This *in vitro* assay is used to quantify and compare gene expression. A reporter vector containing the luciferase gene is transfected into the cell line of choice. A DNA fragment for study – here the region of DNA around the SNP – is inserted into the reporter vector at an appropriate location (e.g. in the promoter region for variants in promoters, or the 3' region for variants in enhancers). Two days after transfection, cells are lysed and light emission measured: the amount of light emitted is proportional to the expression of the reporter gene, and can thus be used to compare levels of expression between alleles. Here a candidate SNP identified by EMSAs was analysed using the luciferase reporter assay to compare expression between the two alleles.

5.1.3 Effect of oestrogen

In chapter 4, a difference between the sexes was found of the effect of genotype on IMT and vascular events at the *CFDP1-BCAR1-TMEM170A* locus, with the genetic effect largely being seen only in women. In light of *BCAR1* being implicated by eQTL data and given the gene's full name *breast cancer antiestrogen resistance 1*, this difference between sexes appeared of particular interest.

In light of this, functional analysis was designed also to examine SNPs at the locus that might be influenced by oestrogen. Oestrogen, the primary female sex hormone, consists of three major subtypes: estrone, estradiol and estriol. Estradiol is the primary form of oestrogen in women during reproductive years, though after menopause this role shifts to estrone[272]. Oestrogens enter the cell and bind to the oestrogen receptor located in the cytosol (Figure 39). This complex translocates to the nucleus and binds to response elements in the DNA, thus regulating gene expression[273]. The complex may also bind to other DNA-bound transcription factors to modulate their action[274].

To investigate whether oestrogen interactions with SNPs might be involved in the phenotype, SNPs in LD with the lead SNP were again examined, to see if any were located in sequences binding oestrogen receptors. SNPs with potential oestrogen interactions were subjected to functional analysis as with the candidate SNPs.

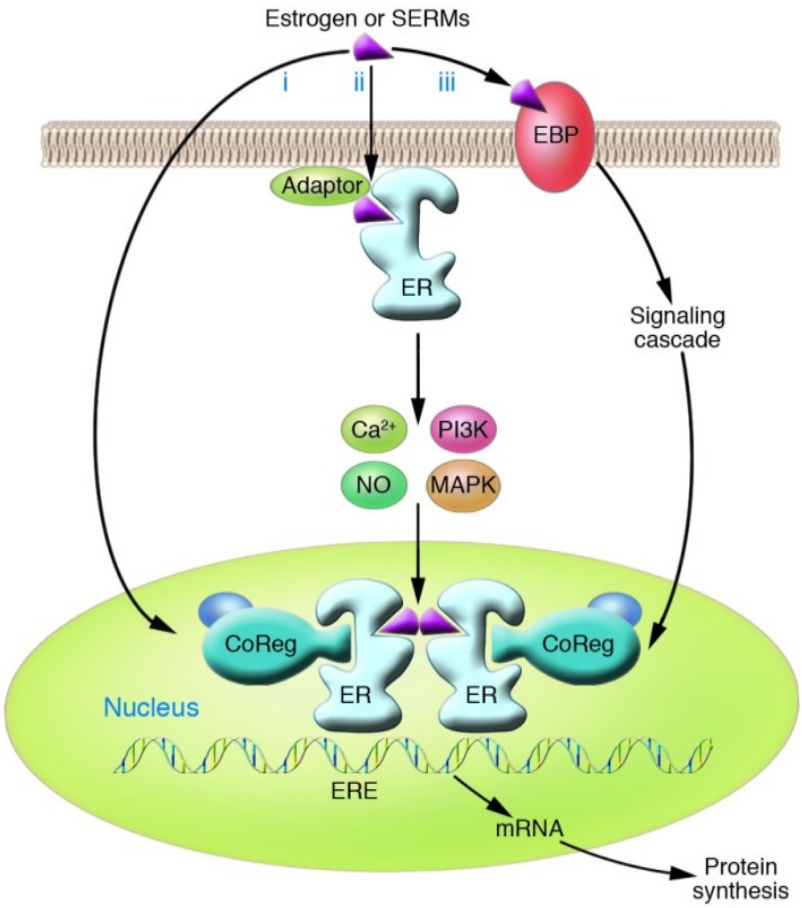

Figure 39: Oestrogen receptor pathways. Figure from Deroo and Korach[274].Oestrogen or selective oestrogen receptor modulators (SERMs) bind to the oestrogen receptor (ER), which binds to oestrogen response elements (EREs) in target genes and recruits coregulatory proteins. Oestrogen may also bind to ER adjacent to the plasma membrane, stimulating signalling cascades which also affect transcription. Non-ER proteins (oestrogen-binding proteins or EBPs) may also trigger signalling cascades.

5.2 Results

5.2.1 6 candidate SNPs

5.2.1.1 Effect of candidate SNPs on DNA-protein interactions

EMSAs were performed on the six candidate SNPs within potential regulatory elements, selected in chapter 3. Nuclear extract was produced from the hepatoma cell line Huh7 and incubated with double-stranded biotin-labelled oligos consisting of the 25 bp surrounding the SNP. Labelled probes for transcription factors NF-κB or Sp1 were used as a positive control, as protein binding had been shown to be successful for these probes. For competitor assays, an excess of unlabelled competitor oligos (the same sequence as the labelled ones) were first incubated with the nuclear extract before addition of the labelled oligos. A binding protein should completely bind to the unlabelled competitor in this case, and no band for the DNA-complex would be seen.

One SNP, rs4888378, demonstrated differential protein-binding by allele (Figure 40). The G allele (the major and risk allele, in terms of IMT and CAD risk) bound a number of proteins strongly, while the A allele had much weaker protein binding. This suggests that the G>A change weakens a binding site for a protein or complex of proteins. The other candidate SNPs did not show protein binding (Figure 41).

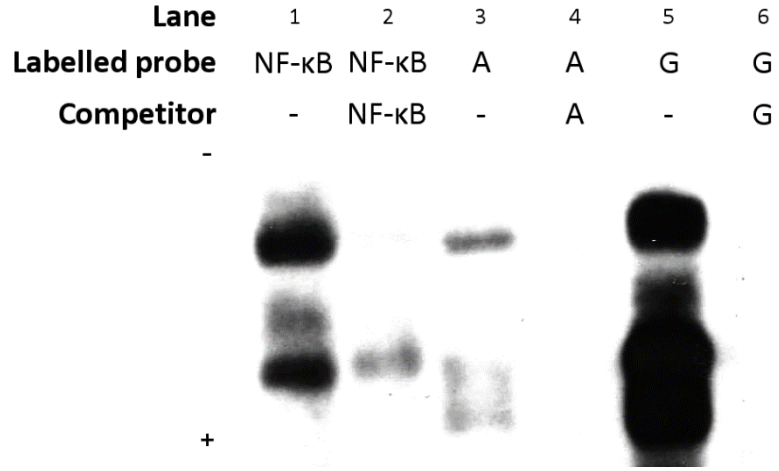

Figure 40: EMSA shows differential protein binding for rs4888378. Lane 1 shows binding of the control protein, NF-κB, to the control band (NF-κB consensus); lane 2 shows band being competed out by unlabelled NF-κB probe. Lanes 3 and 5 show weak and strong protein binding to the A and G probes respectively. Lanes 4 and 6 show the binding to be competed out by respective unlabelled probes. Bands show the position of biotin-labelled probe on the gel (direction from top to bottom). Probes unbound by DNA run fastest and are not visible on the picture; bands here are protein-bound probe which migrates more slowly. "A" and "G" refer to the respective alleles of rs4888378. Positive and negative signs show electrode charge. Image is representative of 4 replicates.

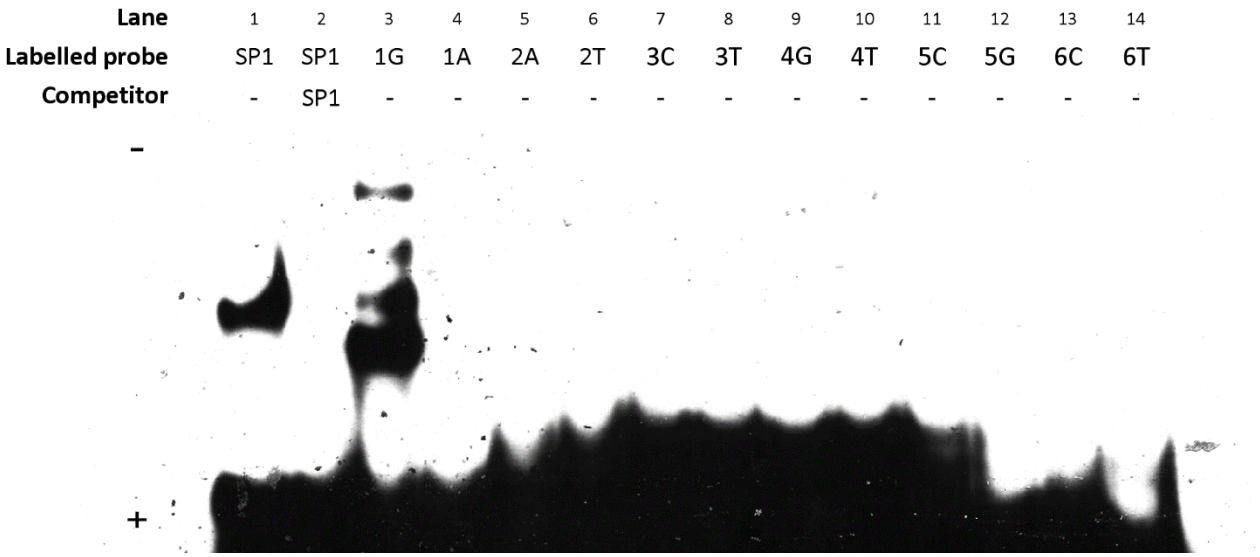

Figure 41: EMSA shows differential protein binding for one of the six tested SNPs: rs4888378. Labelled probes 1-6 refer to SNPs rs4888378, rs4888379, rs4888392, rs2865530, rs3743609 and rs11643207 respectively; the letter denotes the allele. Bands show the position of biotin-labelled probe on the gel (direction from top to bottom). Probes unbound by DNA migrate fastest and are present at the bottom of the picture; bands above are protein-bound probe which migrates more slowly. Cell extract is from Huh7 cells. Image is representative of 3 replicates.

5.2.1.2 Identification of binding protein

To characterise the DNA-protein interaction, multiplexed competitor EMSA (MC-EMSA) was performed[209]. This involved the incubation of the nuclear extract with 70 unlabelled competitor DNA consensus sequences for well-characterised transcription factors, proteins predicted to have a binding motif altered by the variants, and proteins bound to the regions determined by ChIP-seq through ENCODE. These were incubated prior to addition of the rs4888378 labelled probe. To allow for the study of many potential binding proteins at once, these competitors are incubated initially as "cocktails" of 10 competitors. When one cocktail is found to compete out the band, the assay is repeated using each component separately.

MC-EMSA showed the protein binding to be competed out on addition of the FOXA consensus sequence (Figure 42). While some other competitors caused a reduction in band intensity, they did not consistently compete out the band upon repeating the assay.

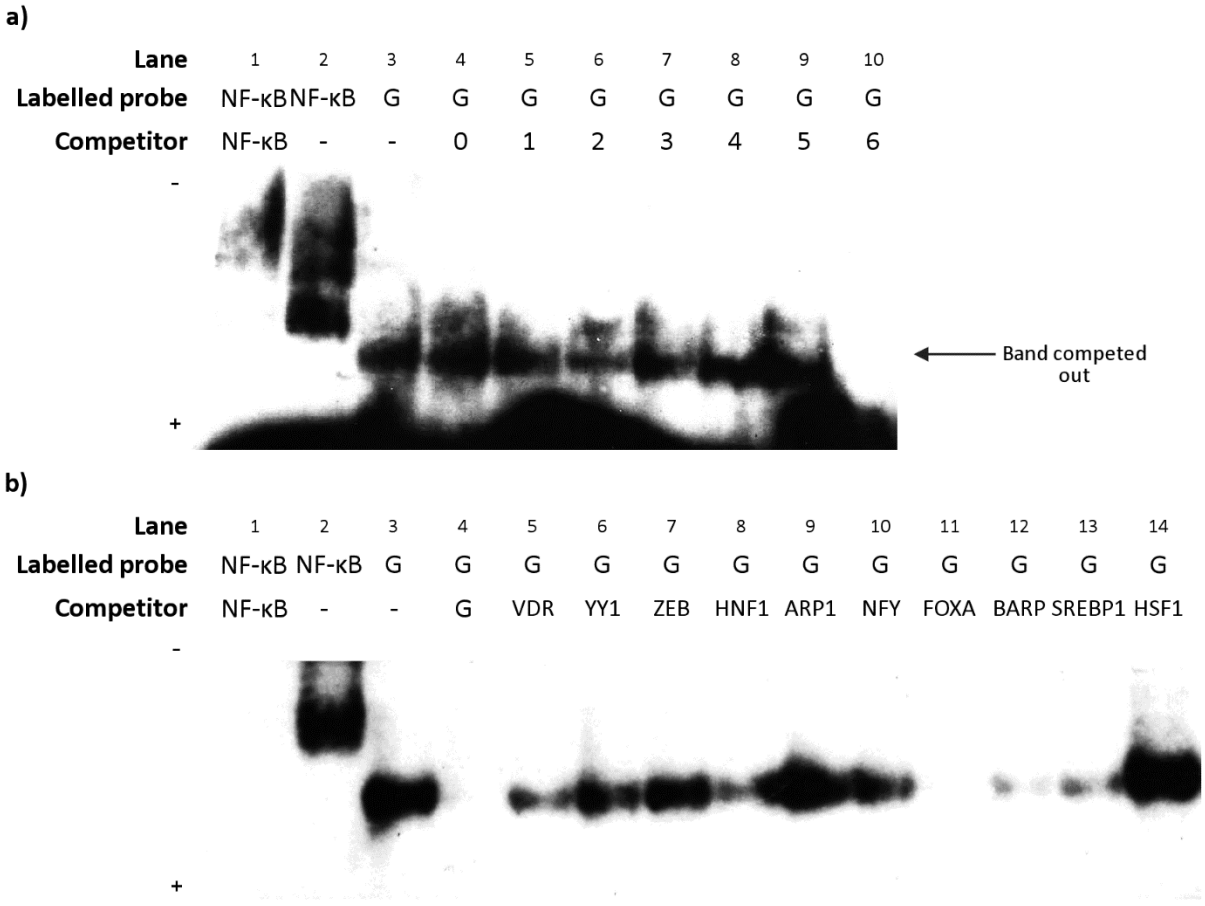

Figure 42: Multiplex competitor EMSA showing competition of the protein-binding allele of rs4888378. "G" refers to the rs4888378 allele. Cell extract is from Huh7 cells. Images are representative of 3 replicates.
a) MC-EMSA with six competitor cocktails. Lanes 3 to 10 show protein binding to the G probe is competed out by one or more components of cocktail six. Signal at the bottom shows free (unbound) probe.
b) MC-EMSA with the ten individual components from cocktail 6. Lanes 3 to 14 show protein binding to the G probe is competed out only by the unlabelled G probe, as expected, and by the FOXA competitor.

To further test this finding, the FOXA binding motif was compared to the genomic sequence around rs4888378. This revealed a close match, and that the A allele weakens the binding motif (Figure 43).

Figure 43: FOXA consensus sequence used in multiplex competitor EMSA compared with genomic sequence around rs4888378. The G allele of the SNP forms a closer match with the consensus sequence, reinforcing the implication that FOXA is the protein binding to the SNP.

5.2.1.3 DNA-protein binding in HUVECs

EMSAs were initially carried out using nuclear extract from the Huh7 cell line, but the SNP's association with expression in vascular cells was later discovered. Assays were therefore repeated using extract from human umbilical vein endothelial cells (HUVECs) in order to look for DNA-protein binding in a relevant cell model. Differential binding was again seen with rs4888378 (Figure 44), and competed out with the FOXA sequence (Figure 45).

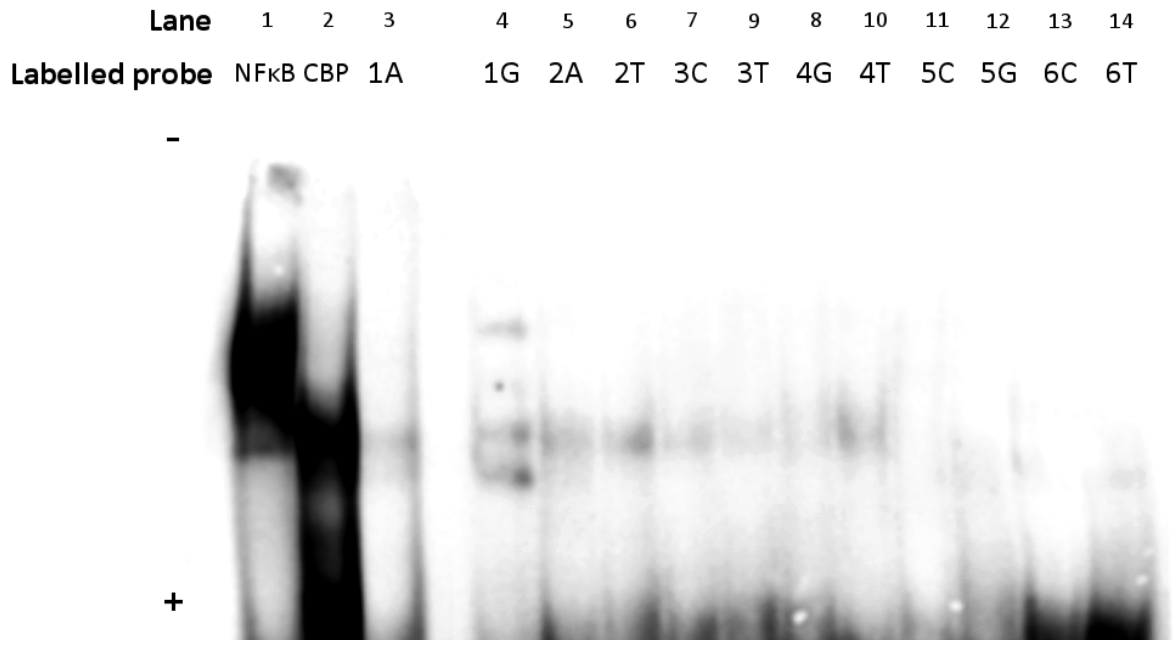

Figure 44: EMSA with HUVEC extract shows differential protein binding for rs4888378. Lanes 1 and 2 show binding of the positive control proteins NFκB and CBP to the control bands (NFκB and CBP consensus). Some protein binding is observed with SNPs 2, 3 and 4. As with Huh7 extract, the only SNP that shows differential protein binding is SNP 1, rs4888378. Labelled probes 1-6 refer to SNPs rs4888378, rs4888379, rs4888392, rs2865530, rs3743609 and rs11643207 respectively; the letter denotes the allele. Bands show the position of biotin-labelled probe on the gel (direction from top to bottom). Probes unbound by DNA migrate fastest and are present at the bottom of the picture; bands above are protein-bound probe which migrates more slowly. Image is representative of 3 replicates.

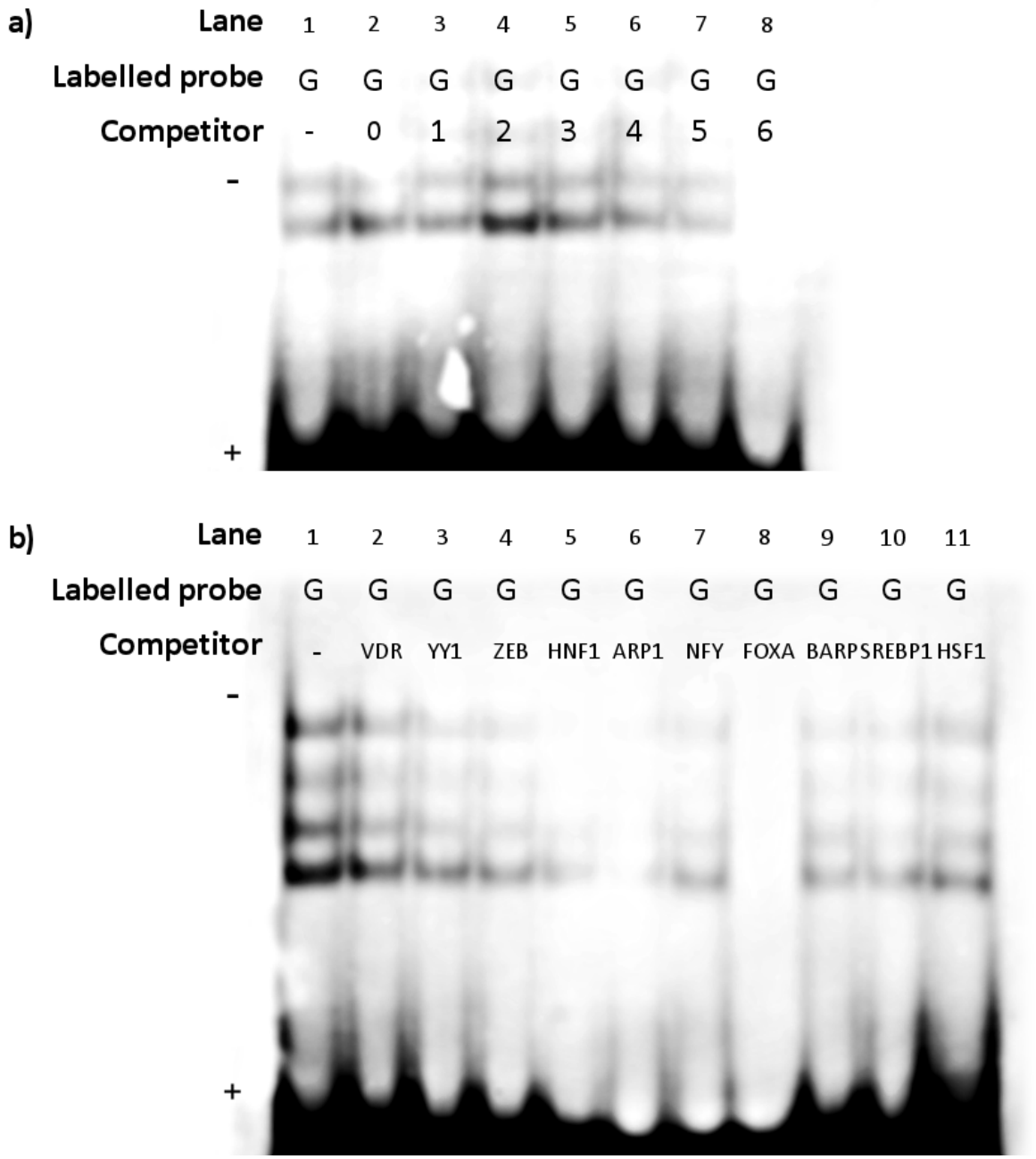

Figure 45: Multiplex competitor EMSA with HUVEC extract, showing competition of the protein-binding allele of rs4888378. "G" refers to the rs4888378 allele. Cell extract is from Huh7 cells. Images are representative of 3 replicates.
a) MC-EMSA with six competitor cocktails. Lanes 2 to 8 show protein binding to the G probe is competed out by one or more components of cocktail six. Signal at the bottom shows free (unbound) probe.
b) MC-EMSA with the ten individual components from cocktail 6. Lanes 2 to 11 show protein binding to the G probe is competed out by the FOXA competitor.

5.2.1.4 Supershift EMSA for confirmation of protein binding

MC-EMSA implicated the protein FOXA protein family as binding to the rs4888378 SNP, but to truly verify the protein, a supershift EMSA is required. Here an antibody to the protein of interest is incubated with the nuclear extract before addition of labelled probes. The antibody binds to the

protein which subsequently binds to the labelled probe, creating a larger antibody-protein-DNA complex that migrates more slowly through the gel. A successful shift confirms the identity of the protein.

Supershift EMSA was performed on the rs4888378 G probe to identify the protein binding to it in the assay. The proteins in the FOXA family share very similar consensus sequences, so MC-EMSA could not identify which one was binding. Two antibodies were examined: anti-FOXA2 and anti-FOXJ1 antibody. These were incubated with the nuclear extract before addition of rs4888378 labelled probe and a positive control probe (consensus sequence for the FOXA family).

However, while the rs4888378 G-allele probe produced a band as expected, no shift was visible on addition of FOXA2 or FOXJ1 antibody (Figure 46). A positive control tested the binding capacity of the antibodies, using the FOXA motif as the labelled probe. However, again no band shift was seen upon addition of the antibodies (Figure 46). As the antibody is expected to bind to the FOXA-family protein, and the protein to the known FOXA motif, the lack of shift suggests the antibodies used do not bind to the protein under EMSA conditions.

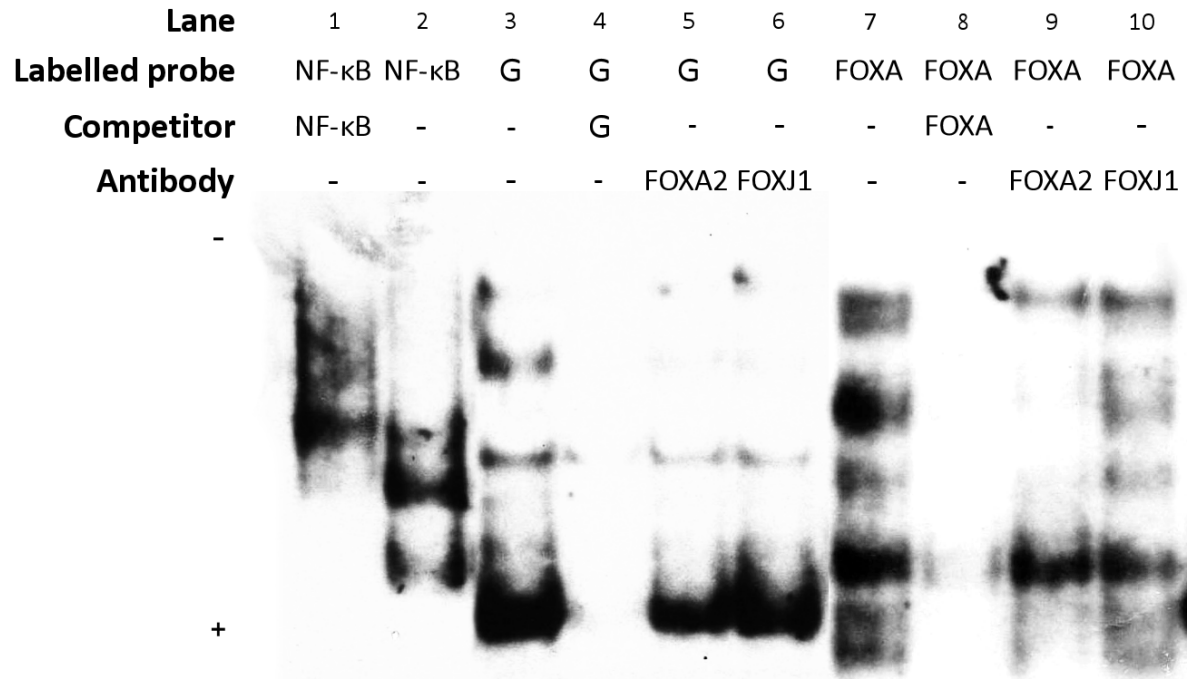

Figure 46: Supershift EMSA showing the effect of addition of FOXA2 and FOXJ1 antibody on protein binding. Lanes 3-6 show the rs4888378 G allele probe ("G"). A band is seen and competed out with its own sequence, as expected, but the addition of antibodies produces no band shift. Lanes 6-10 (positive control) show binding to the FOXA motif. The antibodies would be expected to bind to the protein that binds to this sequence, but no supershift is seen. This gel is representative of three replicates, all of which failed to demonstrate antibody binding to the FOXA consensus sequence.

5.2.1.5 DNA-protein binding for oestrogen-related SNPs

The suggestion of *BCAR1* as a gene of interest at the locus, and the sex difference in association, suggested a possible role of oestrogen in the phenotype. Additional bioinformatics work was carried out to look for variants that could potentially interact with oestrogen. Again, all the variants in LD with the lead SNP were analysed. Haploreg[133] was used to find any SNPs shown to bind an oestrogen receptor, or to change the binding consensus sequence for an oestrogen receptor. Four SNPs were identified: rs2161648, rs4888400, rs2285222 and rs11149832. All had an alteration of the oestrogen receptor alpha (ERα) motif, and one showed ERα binding by ChIP-seq (Figure 47).

chr	pos (hg38)	variant	Ref	Alt	Proteins bound	Motifs changed	GENCODE genes	dbSNP func annot
16	75280731	rs2161648	T	G	HAE2F1,GATA2,ERALPHA_A,GATA3,P300	ERalpha-a,HEN1,LRH1,NR4A,RAR,RXRA,SF1,TAL1	6kb 3' of U6	
16	75391698	rs4888400	G	C		AP-4,CTCF,E2A,ERalpha-a,LBP-1,Lmo2-complex,Nanog,RP58,TCF12,ZEB1	CFDP1	intronic
16	75395449	rs2285222	A	T		ERalpha-a,FXR,HDAC2,NR4A,RAR,RORalpha1,RXRA,SF1	CFDP1	intronic
16	75428815	rs11149832	C	T		ERalpha-a,Evi-1	CFDP1	intronic

Figure 47: SNPs in LD with lead SNP rs4888378 that bind an oestrogen receptor or change its binding motif. All oestrogen-related interactions are with ERα. Image adapted from HaploReg v3[133].

As additional variants of interest, these four SNPs were also investigated using EMSA to look for differential protein binding. In addition to the protocol with standard nuclear extract, assays were performed using nuclear extract from cells stimulated with oestrogen. Before carrying out the nuclear extraction protocol, 16 ng β-estradiol was added to one T175 flask of Huh7 cells in 30 ml medium. This was incubated at 37°C for 30 minutes, and the nuclear extraction protocol was carried out as before. EMSA was carried out on the four SNPs with stimulated and unstimulated Huh7 nuclear extract. None of the variants demonstrated binding in this instance.

5.2.1.6 Luciferase reporter assay

EMSA identified rs4888378 as the candidate SNP with allele-specific protein binding; therefore, it was then necessary to see whether this difference has an effect on gene expression. The SNP is located in the 3' intron of the gene *CFDP1*, so its effect is likely to be through enhancer rather than promoter activity. The genomic sequence surrounding rs4888378 was therefore inserted in the 3' downstream region of the luciferase gene to mimic the effects of an enhancer rather than a promoter (Figure 48).

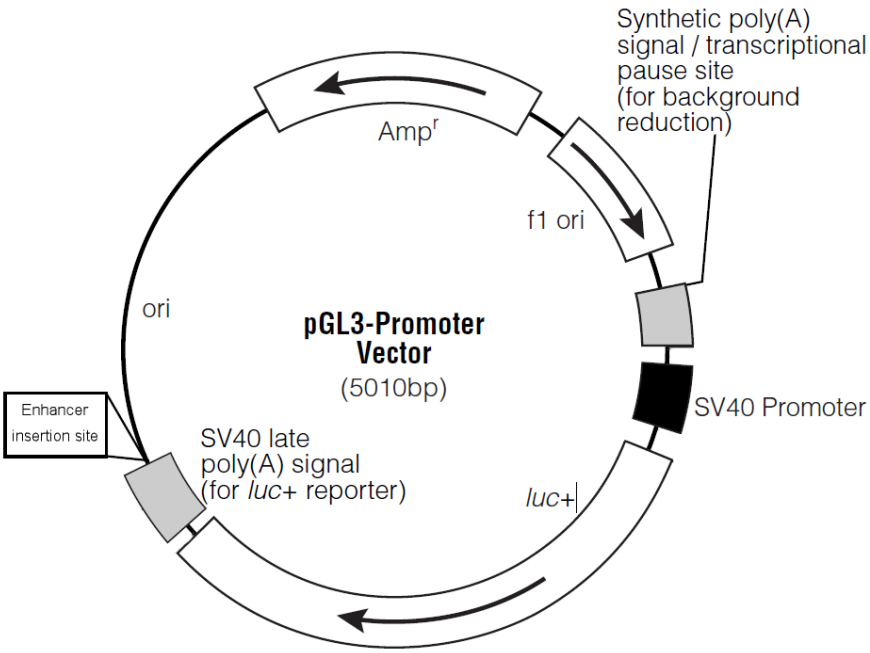

Figure 48: pGL3-promoter reporter vector. Enhancer insertion site shows where DNA sequences were inserted for study. Image adapted from Promega[275].

Selection of suitable restriction sites for cloning (*BamHI* and *SalI*) in the sequence resulted in a fragment of 379 bp around rs4888378. This sequence was amplified from genomic DNA of AA or GG genotype and digested with *BamHI* and *SalI* enzymes to leave the fragment of choice. After sequencing confirmed that no other bases differed between the A and G fragments, the fragments were cloned into the pGL3-promoter vector using the In-Fusion cloning system, and the vector transfected into Huh7 cells.

The luciferase assay revealed a decrease in expression compared to the control vector (pGL3-promoter) for both alleles of rs4888378 (Figure 49). It was considered that this decrease may be caused by the binding of a repressor protein to part of the sequence of the inserted fragment.

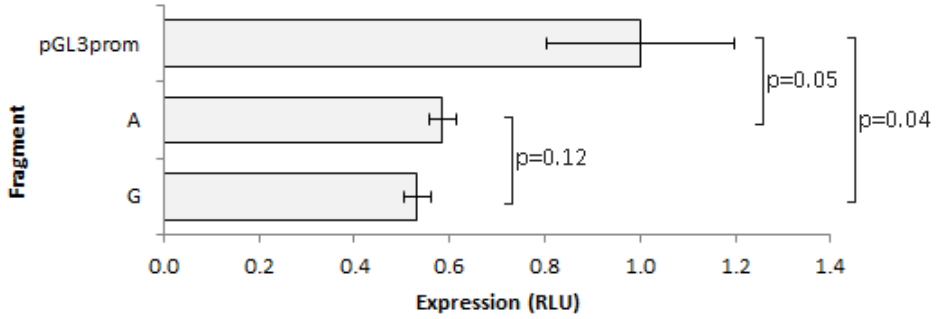

Figure 49: Enhancer fragments with both the A and G alleles of rs4888378 show decreased expression compared to pGL3-promoter control. Luciferase reporter assay data. P-values derive from unpaired t-test. Error bars show SEM. (12 replicates each.)

The decrease in expression prompted evaluation of the inserted fragment to look for possible binding of repressor proteins. Bioinformatics sequence analysis using the MatInspector tool[276] predicted the binding of three repressors and repressor-related proteins to the sequence fragment: E4BP4, CHR and PRDM1 (positions shown in Figure 50a). Three further cloning fragments were designed to investigate the effect of these potential DNA-protein interactions on gene expression. These fragments eliminated the segment with potential E4BP4 binding, the segment with CHR and PRDM1 binding, or all three of these binding sites (Figure 50b-d).

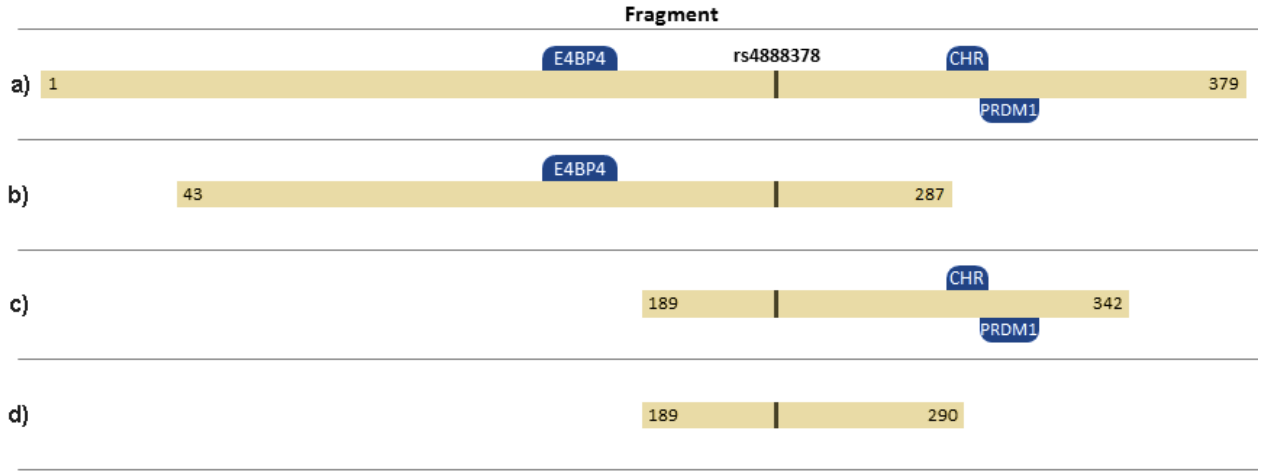

Figure 50: Predicted binding positions of three repressor-related proteins to the rs4888378 fragment, and additional fragments created to avoid these binding sites. Position data from MatInspector[276]. Repressor proteins binding to this sequence may affect expression of the luciferase gene during the assay. Fragments shown in b, c, and d were therefore designed to test the effect of avoiding these binding sites.

The additional fragments were cloned into the reporter vector and assayed as before. Expression results differed between insert fragments. Allele-specific reporter expression was observed from the vectors with insertions 3 and 4 (Figure 51c,d; $p=5.7\times10^{-2}$; $p=4.0\times10^{-22}$), with a non-significant trend in the third (Figure 51b). The effect size and direction was dependent upon fragment size: while

expression with allele G remained similar in all cases, allele A showed a 35% and 92% reduction in expression compared to G in the shorter fragments.

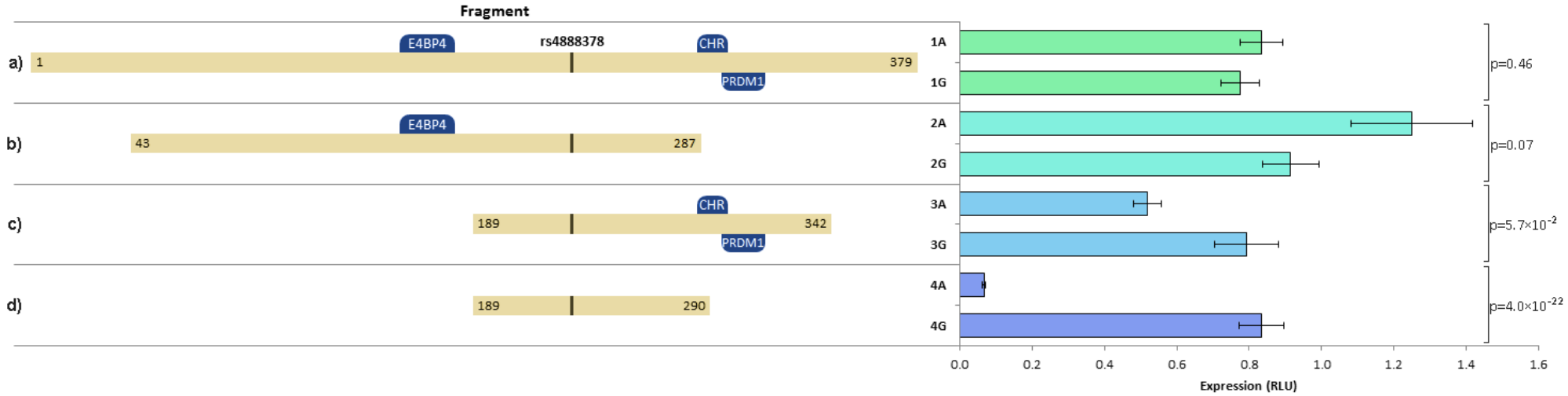

Figure 51: rs4888378 allele is associated with differential expression depending on sequence elements. Luciferase reporter assay data. Fragments 1 and 2 show no significant differential expression (a and b). Fragments 3 and 4 show a decrease in expression for the A allele, with the decrease in fragment 4 being much more pronounced (c and d). Expression values are normalised to pGL3-promoter control. P-values derive from unpaired t-test. Error bars show SEM. (4 plates of 12 replicates each.)

5.2.1.7 Transfection of HUVECs for luciferase reporter assay

Luciferase reporter assays were carried out in Huh7 cells, as these had previously been used reliably for luciferase reporter assays, and at the time of initial assays the association with expression in vascular cells was unknown. However, human umbilical vein endothelial cells (HUVECs) were a preferred cell type, as the IMT phenotypes under study relevant to blood vessel endothelial cells. Transfection assays were therefore carried out on HUVECs using the reporter vector.

Transfection was tested using the lipid reagent Lipofectamine 3000, and using the Amaxa Nucleofector I. Vectors for transfection were the pGFP plasmid (green fluorescent protein allows easy detection of expression under a microscope), pGFP with a pRL-TK co-transfectant (to mimic the co-transfectant conditions in the luciferase assay), and pGL3-promoter with pRL-TK co-transfectant (as in a standard assay).

Transfection using Lipofectamine 3000 was not successful, but as seen in Figure 52a, transfection of the GFP plasmid using electroporation achieved some success. Co-transfection with pRL-TK resulted in lower transfection efficiency (Figure 52b). Expression peaked 14-20 hours after transfection (8-14 hours after first timepoint on graph).

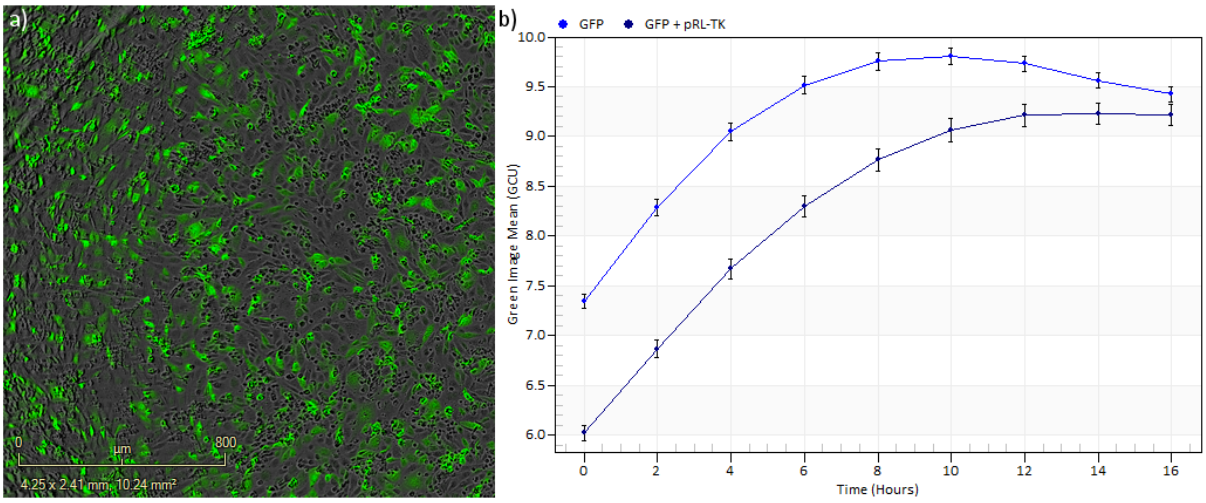

Figure 52: GFP expression in HUVECs after electroporation of GFP plasmid. (a) Detection of GFP expression shows cells with successful transfection of plasmid. **(b)** GFP emission in cells transfected with GFP plasmid or GFP + pRL-TK plasmids. Graph starts 6 hours after electroporation. It can be seen that expression of GFP is lower with the co-transfectant.

Transfection of the reporter vector into these cells could not be measured via GFP emission. HUVECs were electroporated with the vectors puc, pGL3-control and pGL3-promoter, with pRL-TK co-transfectant, and luciferase expression was measured by light emission. This indicated that transfection of the pGL3-control and pGL3-promoter vectors was unsuccessful (Figure 53): readings

of luciferase and renilla were extremely low compared to those seen in Huh7 cells (rightmost data points; note logarithmic scale), and would not be adequate for reporter assays. Repeated experiments did not achieve successful transfection.

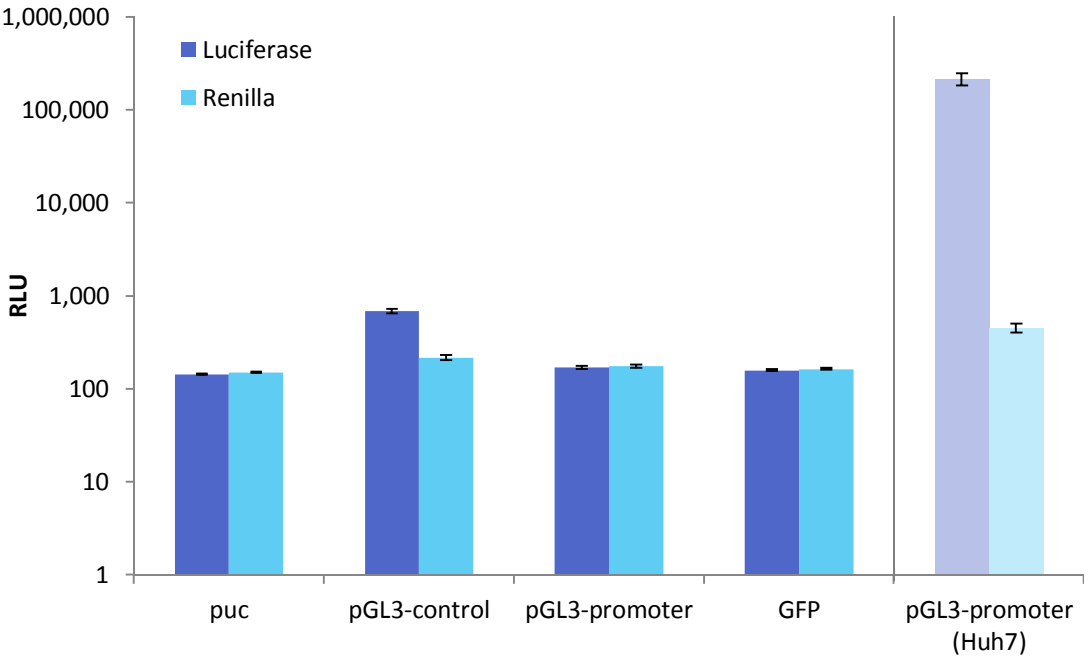

Figure 53: Luciferase (main transfectant) and renilla (co-transfectant) readings for vectors transfected into HUVECs. Relative light units are a measure of luciferase expression. Luciferase readings would be expected to be many times higher than renilla readings. Although transfection vector readings for pGL3-control were higher than those for the co-transfectant, these values were too low for reliable comparison of signal between fragments. "pGL3-promoter (Huh7)" indicates average readings for luciferase readings of this vector in Huh7 cells; note logarithmic scale.

5.2.2 DNA-protein binding for dsQTL variants

In chapter 3, analysis of the Gilad/Pritchard eQTL browser found two DNase sensitivity QTLs (dsQTLs), where chromatin accessibility is associated with SNP allele (Figure 28): rs73605136 and rs247454. EMSAs were carried out on these two SNPs to analyse how the variants may be affecting chromatin accessibility. 25 bp probes were created as before and incubated with Huh7 extract. However, no protein-binding bands were observed (Figure 54).

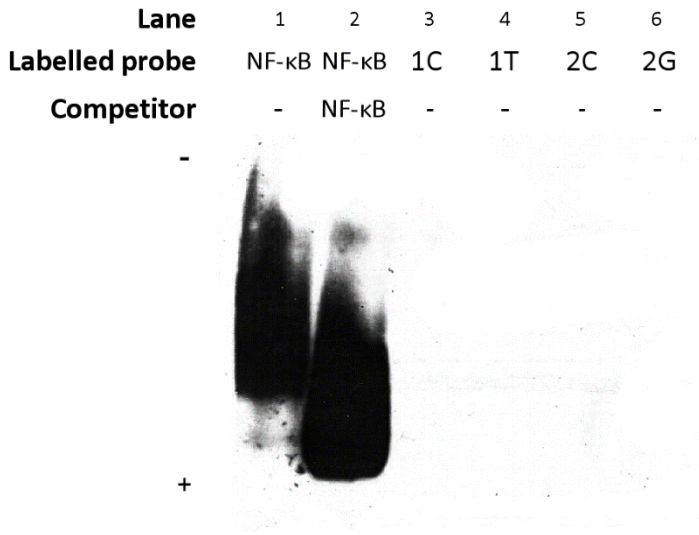

Figure 54: EMSA shows no observed protein-binding for dsQTL SNPs. Labelled probes 1 and 2 here refer to the dsQTLs rs73605136 and rs247454 respectively, and the letter to the allele. Control bands are here overexposed as film was overexposed to check no signal was present for the dsQTL SNPs. Image is representative of 3 replicates which showed no protein binding.

5.3 Discussion

5.3.1 Overall

In chapter 3, investigation of the *CFDP1-BCAR1-TMEM170A* locus with bioinformatics data identified six candidate functional SNPs that may potentially be affecting IMT and CAD. In this chapter, DNA-protein interactions surrounding the SNPs were tested, identifying that the lead SNP rs4888378 altered allele-specific protein binding. Competitor assays indicated a member of the FOXA protein family may be responsible for this allele-specific binding. This SNP was taken forward for expression analysis using luciferase reporter assays. Expression varied widely according to the length of DNA sequence cloned into the reporter vector, but where allele-specific expression was present, the G allele of the SNP produced significantly higher expression than the A allele.

Other aspects of interest at the locus were examined: two dsQTLs near *BCAR1* and *CHST6* were assayed for allele-specific protein binding, but no proteins were found to bind. Four SNPs in LD with rs4888378 that bound ERα or were predicted to change its binding site were assayed.

5.3.2 Effect of candidate SNPs on DNA-protein interactions

Of the candidate SNPs assayed using EMSA, only the lead SNP rs4888378 showed allele-specific protein binding, with a protein or proteins binding strongly to the risk G allele. As multiple bands were visible for the rs4888378 G probe when film was subjected to a longer exposure, it appears that a complex of proteins may be binding here. If this is indeed the functional SNP, it could therefore be hypothesised that a transcriptional activator binds more strongly to this sequence when the G allele is present, increasing expression of an atherogenic gene at the locus, resulting in increased IMT. Alternatively, a repressor may bind more strongly, reducing expression of a gene that is atheroprotective. Expression results from the luciferase reporter assay suggest the former possibility, as the G allele here causes increased expression.

Using multiplex competitor EMSA it was found that the FOXA binding motif competed out the G band of the rs4888378 probe, implicating this as the protein family that may be binding. FOXA is a subfamily of hepatocyte nuclear factors consisting of FOXA1, FOXA2 and FOXA3, which share a common consensus sequence. The proteins act as transcriptional activators for liver-specific genes, and are involved in regulation of metabolism and glucose homeostasis[277,278]. They have been shown to bind sites in compacted chromatin and mediate its change into an open, more active state[279].

The FOXA family contains three proteins and numerous isoforms that share similar binding motifs, and so it was not practical to perform a supershift assay for all proteins. The FOXA2 antibody was chosen as this protein has been implicated in cardiovascular-related phenotypes: it has been suggested that it is critical in diabetes and metabolic syndrome[278,280]. FOXJ1 (forkhead box J1) is a transcription factor related to the FOXA family (previous names for FOXA and FOXJ1 are *hepatocyte nuclear factor 3* and *hepatocyte nuclear factor 3 forkhead homolog 4* respectively). FOXJ1 was specifically identified in chapter 3.2.1 as one of the motifs altered by rs4888378, and its binding motif is indeed similar to that of FOXA2[281], so it was also chosen as a priority for supershift.

Unfortunately, the identity of the protein binding could not be verified with a supershift assay. No supershift was seen with the addition of FOXA antibodies to the probe containing rs4888378-G; however, a lack of band shift for the positive consensus sequence indicated that the antibodies were not suitable for this assay. Unfortunately, no EMSA-verified FOXA antibodies were available; therefore, conclusions could not be drawn about the identity of the protein.

EMSAs have utility in detecting allele-specific effects without the hindrance of LD, but it must also be considered that they examine variants outside their natural chromatin environment and cannot take into account *in vivo* properties such as chromatin state. It should also be considered that the nuclear protein profile is that of the cells used to produce the nuclear extract. Here Huh7 cells were initially used, and subsequently HUVECs, which were a more appropriate model for vascular phenotypes.

5.3.3 DNA-protein binding for oestrogen-related SNPs

Of the 214 SNPs in strong LD, four were found to bind ERα or change its binding site, so these were chosen for analysis with EMSA. No binding was observed for the variants in this assay. As ERα can induce an effect without binding to the DNA sequence itself, other SNPs could also be involved in an effect of oestrogen[274].

These oestrogen-stimulated EMSAs were preliminary experiments that would require more thorough analysis if being followed up. The 30-minute treatment time with β-estradiol was chosen based on research by Dominguez and Micevych, who found stimulation with estradiol induced accumulation of ERα with a peak at 30 minutes[282]. Future work would vary concentration and treatment time to investigate what conditions produced the greatest binding of ERα.

Different cells would also be examined for creating the nuclear extract in future work. ERα is expressed in liver cells such as Huh7, although at a lower level than in target tissues such as uterine

and mammary tissue (gene ESR1, GTEx[155]). Future assays should use HUVEC and smooth muscle cell extract to examine the possible effect in vascular tissues.

5.3.4 DNA-protein binding for dsQTL SNPs

EMSA analysis demonstrated no protein-DNA interactions for the two dsQTL SNPs. These SNPs were reported to be associated with different accessibility of chromatin depending on allele. This mechanism may be through the action of proteins binding directly to the genomic sequence, or by another method, such as histone modifications A different method to investigate the relevance of such dsQTLs might be through assaying whether the loci interact with other loci of interest, such as promoters or enhancers of genes (explored in chapter 6).

5.3.5 Allelic effect on gene expression

Previous examination of genotype-expression data found multiple associations between the lead SNP and gene expression, so it is likely that such effects on expression may be involved in producing the carotid IMT phenotype. The disadvantage of these datasets is that they can only report associations rather than causality, so SNPs in LD show similar findings to each other. Conversely, reporter assays test causality, as only the allele of the SNP under investigation differs between constructs. This allowed the direct evaluation of whether a chosen SNP affects gene expression.

Expression data from ASAP and GTEx had shown the G allele to be associated with higher expression; in the luciferase reporter assay, the same direction of effect was seen. However, the difference in expression varied greatly depending on the size of the DNA fragment inserted. For the larger fragments, no significant difference was seen, while for the others, the magnitude of the difference was greater for the smallest fragment. This highlights the strong effect that proteins binding *in vitro* to proximal sequence elements can have on enhancer or repressor activity. The different fragments were chosen to include or exclude the predicted binding sites for three repressor-related proteins (Figure 50). The two fragments that excluded the predicted E4BP4 (NFIL3) binding site showed a decrease in expression with allele A. E4BP4 is a transcriptional repressor that binds to activating transcription factor (ATF) sites in many promoters[283].

The effect on expression from sequence fragment length and genotype implies a strong effect and potential for interaction with the FOXA element from proteins binding to these additional elements close to the SNP. While these results do not prove an effect of E4BP4 at the locus, they do suggest that this or another binding protein may suppress the effect of the SNP under certain circumstances.

Selection of a representative fragment for assaying expression is complicated by the artificial nature of the assay. As with the EMSA, the luciferase reporter assay cannot measure long-distance interactions between loci, or chromatin state. This emphasises the value of assaying multiple sequence fragments around the SNP, as was done here. Another drawback was the cell line used; Huh7 cells were initially used due to their reliable performance in reporter assays, the fact that the genes of interest at the locus (*BCAR1*, *CFDP1*, *TMEM170A* and *LDHD*) are expressed in liver cells[155], and the similar pattern of ENCODE regulatory signals such as open chromatin observed between endothelial and hepatoma cells. However, HUVECs may be a preferred model due to their vascular nature. HUVECs are difficult to transfect[284,285], and tests with different transfection methods could not achieve sufficient transfection efficiency of the reporter vector, so this was not possible. Further work would carry out the reporter assays using these cells: one possibility would be the production of a viral expression vector.

It would be ideal to test all of the candidate SNPs with the reporter assay, but the time required to clone each SNP made this impractical for this study. Therefore rs4888378 was prioritised due to its effects on protein binding.

Another limitation of this reporter assay was that the SV40 promoter was used, rather than those at the gene locus. Tissue expression data implicated the gene *BCAR1* as the gene possibly causing the association. However, at the time of performing the cloning, this was not known. As many genes are present at the locus, a single promoter could not be chosen, and the SV40 promoter present in the pGL3-promoter vector was used, a known strong promoter that should provoke high levels of expression and allow for comparison. Future work would carry out the reporter assay with the *BCAR1* promoter (see chapter 6 for discussion of putative *BCAR1* promoters).

To investigate the regulatory effects of SNPs in an *in vivo* context and look at regions much larger that the fragments used in luciferase, further work will use chromosome conformation capture, allowing examination of interactions between regulatory regions over long distances.

6 Circular chromosome conformation capture for investigation of the *CFDP1-BCAR1-TMEM170A* locus

6.1 Introduction

As discussed in Chapters 3, 4 and 5, regulation of gene expression is often not as simple as a regulatory element affecting the gene it is closest to. The regulatory SNP rs4888378, which is intronic in the gene *CFDP1*, instead appears to regulate the nearby gene *BCAR1*. Although gene expression is often thought of as occurring proximal to genes, it is becoming increasingly evident that long-range interactions play an important role in the regulatory landscape[286]. Many promoters and enhancers are now known to engage in long-range interactions[287], physically looping to their target genes[288]. Such interactions can also occur between different chromosomes (*trans*- rather than *cis*- interactions)[289].

These long-range interactions are of particular importance in the context of GWAS studies. Most lead variants for cardiovascular GWAS (and other complex diseases) occur within intronic or intergenic regions, and thus are likely to be exerting their effect through distal regulatory elements[218]. Certain GWAS hits are even found in gene deserts: large genomic regions (sometimes defined as larger than 500 kb[290]) with no annotated genes. The most well-known cardiovascular example is that of the 9p21 locus, which has the strongest known association with CAD and MI[107,108,291], yet is over 40 kb away from the nearest protein-coding genes (*CDKN2B* and *CDKN2A*, involved in cell cycle control). Long-range interactions are likely to be the primary mode by which GWAS hits in such regions affect cardiovascular traits.

Understanding more about the basis of this regulation at cardiovascular and other loci is a crucial objective in understanding the genetic basis to CHD and other diseases. For this purpose, a number of methods have been developed in recent years to test long-range interactions between DNA. The basis of the se is Chromosome Conformation Capture (3C), a method that captures the arrangement of chromosomes in living cells, allowing interactions between loci to be detected[289]. In these methods, formaldehyde is used to crosslink-DNA to proteins, “fixing” the current chromosomal interactions. Crosslinking occurs as a nucleophilic group on an amino acid or DNA base forms a covalent bond with formaldehyde, forming a methylol adduct, which is converted to a Schiff base and finally a methylene bridge[292]. In 3C, the fixed DNA is cut into fragments along the whole genome using restriction enzymes. These fragments, still linked at interaction sites with formaldehyde, are ligated together, with ligation occurring preferentially between fragments brought into proximity by

these physical interactions (Figure 55). In this way an excess of ligation junctions are formed between loci that interact *in vivo*. PCR across the ligation junctions and sequencing of the amplification product can identify ligated sequences and quantify the number of junctions.

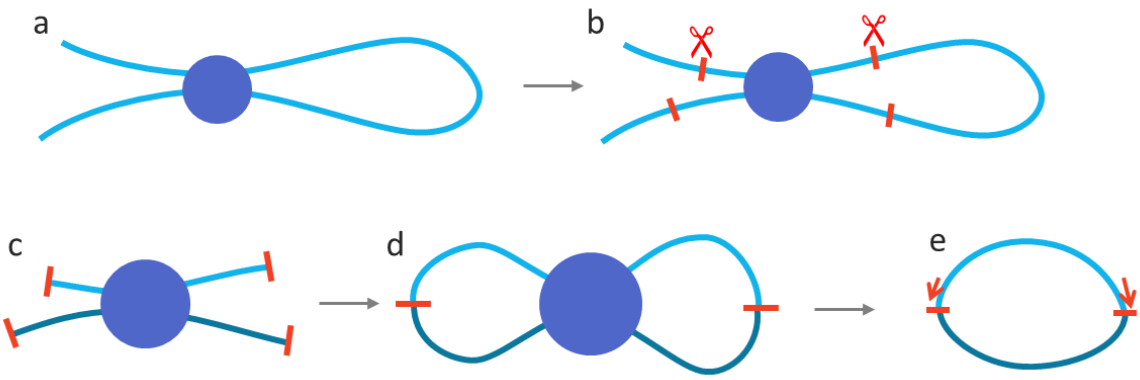

Figure 55: Principle of 3C. (a) Formaldehyde is used to crosslink DNA to proteins, linking interacting regions. **(b/c)** DNA is cut into fragments with a restriction enzyme. **(d)** Matching sticky ends ligate in a proximity-mediated manner, joining interacting fragments. **(e)** PCR and sequencing identifies the unknown linked sequence.

Many random ligations may also occur between restriction enzyme cut sites, so some interactions are likely to be detected between any two loci, with more expected the closer the two sites are in the genome. However, when ligations are detected more than would be expected for random interactions, it can be concluded that the two sites interact *in vivo*.

6.1.1 3C technologies

Chromosome conformation capture (3C) is a "one-versus-one" approach: genetic loci are selected *a priori* and the assay tests whether pairs of loci interact more than would be expected by chance[289]. Related methods include Circular Chromosome Conformation Capture (4C)[293], carbon-copy chromosome conformation capture (5C)[294], chromatin interaction analysis by paired-end tag sequencing (ChIA-PET)[295], and Hi-C, a genome-wide method that uses biotin labelling to pull down ligated fragments[296].

In contrast to 3C, 4C is a "one-versus-all" method: the sequence of the site of interest ("bait") is needed, but interactions with any other site of interest ("prey") can be captured[293]. For this to work, ligated DNA has crosslinks removed and is digested again with a second restriction enzyme. An additional ligation step causes self-circularisation of these shorter fragments, producing small circles of DNA containing the junction between the bait and prey. Inverse PCR to amplify the circular DNA uses PCR primers directed outwards from the bait sequence, amplifying a section of the unknown

sequence. Next-generation sequencing allows the unknown sequence to be identified and mapped to the genome. This allows unbiased searching for loci that interact with a sequence of interest, without *a priori* expectations.

5C is a multiplex "all-versus-all" protocol that combines 3C with ligation mediated amplification (LMA) to detect interactions on a large scale. A 3C library is created, and LMA carried out with primers that bind to the 3' end of predicted restriction fragments. Only primers annealed next to each other can be ligated. The primers also contain universal tails, allowing large-scale amplification of ligated primers to create a 5C library, which can be analysed using micro-arrays or next-generation sequencing.[294,297].

Hi-C is more high-throughput than the previous methods. After the restriction digest stage of 3C, biotin-labelled nucleotides are annealed to the overhangs at ligation sites, and the resulting blunt ends ligated. Biotin pull-down is used to isolate ligated fragments, and the library mapped to the genome[296]. Chromatin interaction analysis by paired-end tag sequencing (ChIA-PET) combines 3C and chromatin immunoprecipitation to screen for interactions occurring via a proposed protein of interest. Crosslinked chromatin is digested and then pulled down using an antibody against the protein. Linker sequences are ligated to the ends of DNA, allowing them to ligate, and the interacting fragments are sequenced and mapped to the genome[295,298].

6.1.2 4C to analyse the *CFDP1-BCAR1-TMEM170A* locus

In investigation of the *CFDP1-BCAR1-TMEM170A* locus, previous chapters identified rs4888378 as a potential functional variant, and *BCAR1* as the gene likely to be the gene causing the phenotypic change. It is clear that at this locus, regulatory elements may not act only on the genes they are closest to, making it a good target for chromosome conformation capture to define the long-range regulatory network at this locus.

The specific chromosome capture method to be used depends on the research question. One regulatory scenario here, fitting with the hypothesis that rs4888378 affects *BCAR1* regulation, is that the regulatory SNP is located within an enhancer, which interacts with the promoter to modulate its expression. 3C could be used to test this hypothesis, but for unbiased detection of interactions, 4C is considered the preferred strategy to assess individual loci[299]. Using 4C allows searching for any regions that interact with these loci without making assumptions before testing. As few specific loci

are being studied, multiplex methods such as 5C would not be suitable. Therefore, in this chapter a 4C protocol was developed for analysis of the *BCAR1* promoter and the regulatory SNP rs4888378.

Although the design and optimisation steps produced a strategy to analyse the regulatory SNP and *BCAR1* promoter and produced four 4C libraries, challenges with the next-generation sequencing precluded any sequences from being obtained. Experiments are ongoing to obtain sequencing data from the 4C libraries, and the experimental design of this 4C protocol is discussed here, with the results presented from individual 4C steps.

6.2 Protocol and results

6.2.1 Choice of loci

The two primary areas of interest for 4C were rs4888378 and the *BCAR1* promoter. Before selecting the restriction enzymes for the protocol, the *BCAR1* promoter location had to be characterised.

The gene is shown to have multiple transcripts on databases such as UCSC Genome Browser and Ensembl[127,300], raising the question of what the true and commonly-expressed transcripts are. Kumbrink and Kirsch demonstrate the presence of two *BCAR1* mRNAs expressed in normal tissues with alternative first exons[301]. These alternative exons encode only three different starting amino acids, Asn-His-Leu and Ser-Val-Pro. The two transcripts are referred to as *BCAR1*-1 and *BCAR1*-1' (Figure 56a), and they coincide with the two areas of strong H3K4me3 signal (histone signals associated with promoters), and with CpG islands (Figure 56b). These two promoters were therefore chosen as the true promoters for investigation with 4C.

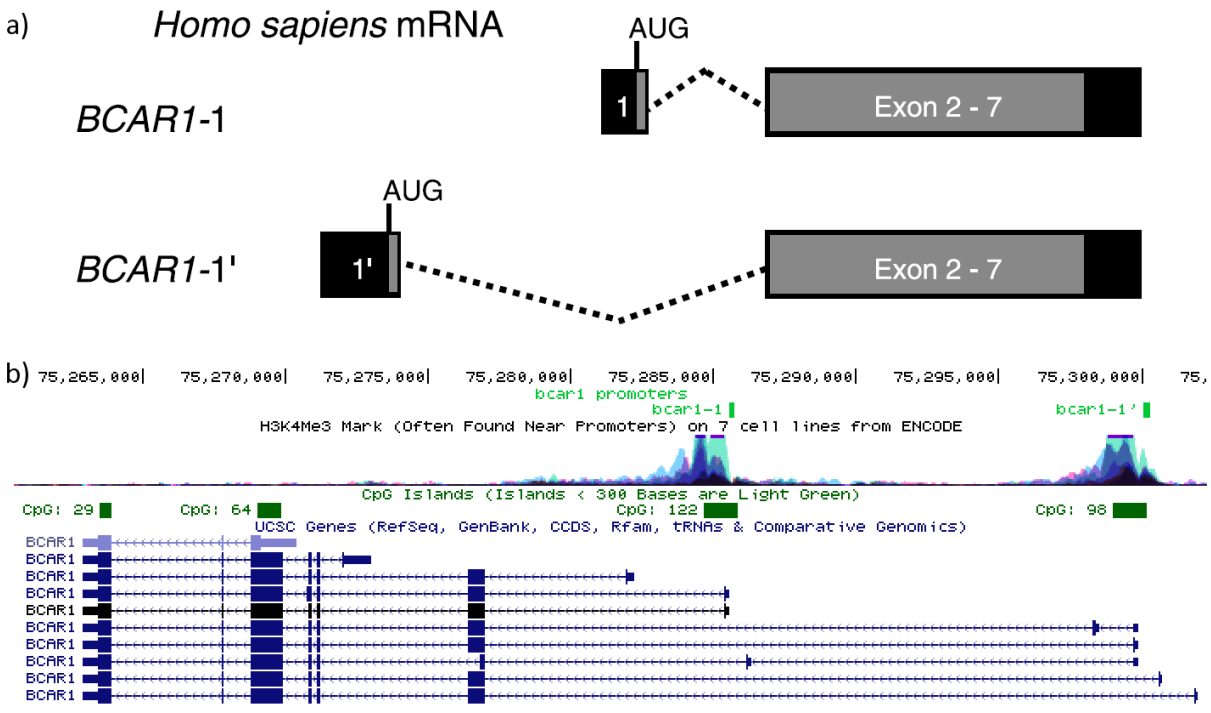

Figure 56: Presence of two promoters for the *BCAR1* gene. (a) There are alternative first exons for *BCAR1* with distinct promoters (figure from Kumbrink et al[301]). **(b)** The location of these two promoters maps to two regions of promoter-associated histone marks and CpG islands as shown on UCSC Genome Browser[127]. While multiple transcripts are shown on the genome browser, these can be considered to be the important promoters.

6.2.2 Choice of cells

As discussed in previous chapters, endothelial or smooth muscle cells are likely to be more relevant to the plaque phenotype, so HUVECs were considered as suitable cells for analysis of chromosome

conformation. However, the 4C protocol requires a large number of cells (10 million cells are suggested by van de Werken[302] and Stadhouders[303]). As HUVECs can only grow to a limited passage number before differentiation or senescence[304], culturing large numbers of HUVECs was inefficient for protocol optimisation. Therefore, HEK293 cells were used for initial development of the protocol, as they can easily be cultured quickly to large numbers, providing sufficient chromatin to optimise subsequent stages of the 4C protocol.

6.2.3 Choice of restriction enzymes

The principle of 3C techniques is based on the digestion of crosslinked chromatin with a restriction enzyme, ligation to join interacting sequences, and PCR to detect ligation junctions (Figure 55). Original 3C studies used this method of a single digest and ligation step[289]. For 4C, circular DNA is required for amplification using inverse PCR. Due to the complexity of interactions *in vivo*, the digest step often forms aggregates of many fragments joined together. Ligation and removal of crosslinks causes these to link together and form large circles of DNA, which are often too large to amplify using PCR (Figure 57). For this reason, a secondary digest is necessary for 4C. The large circles are trimmed with a different restriction enzyme (often a more frequent cutter) to create short fragments. A second ligation circularises these fragments, due to proximity, and the small DNA circles are then suitable for amplification.

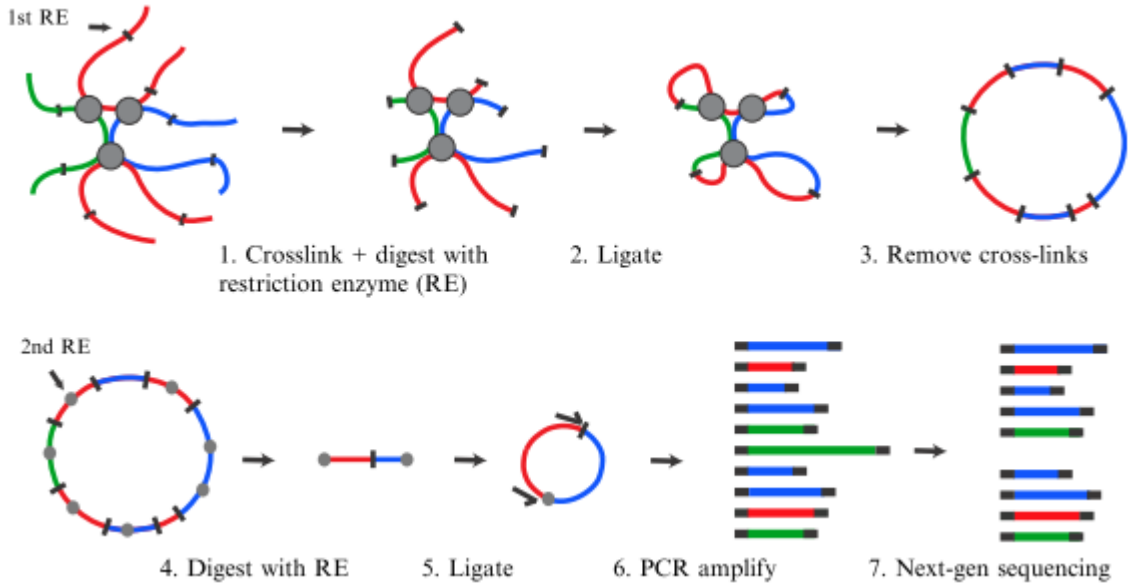

Figure 57: Overview of 4C protocol. Image adapted from Van de Werken et al[302]. Chromatin crosslinking joins many strands of DNA together, leaving large DNA aggregates on digestion, which become large DNA circles on ligation and removal of crosslinks. Trimming the circles with another restriction enzyme allows the creation of smaller DNA circles. These can be amplified more easily by PCR and the ligation junctions analysed using next-generation sequencing.

To design a 4C experiment, suitable primary and secondary restriction enzymes must be chosen. The two main considerations for enzyme choice are:

1. The ability of the enzymes to digest the DNA in the required conditions. This is primarily a concern for the first digestion, where the restriction enzyme must digest crosslinked chromatin in the presence of reaction-inhibiting SDS. (Discussed in 6.2.6.)
2. The location of the enzymes' restriction sites in the locus. As discussed below, the restriction sites for the primary enzyme must cut to create a fragment around the point of interest of suitable size to capture interactions. The restriction sites for the secondary restriction enzyme must cut so that a DNA circle is formed that is large enough to be ligated and sequenced easily, but not too large as to not amplify with PCR. (Discussed in 6.2.3.1.)

Many published 4C techniques use a primary restriction enzyme that cuts at a 6 bp site ("6-cutter"), and a secondary restriction enzyme that cuts at a 4 bp site ("4-cutter")[303]. As there are four DNA bases, the chance of a specific 6 bp sequence coming up at a random point in the genome is roughly $(¼)^6$ = 1/4096; i.e. every 4096 base pairs on average. The secondary restriction enzyme is usually a 4-cutter, which cuts on average every 256 $((¼)^4)$ bases. The resulting smaller fragments are more likely to ligate to form suitable circular DNA.

The base composition of the genome is of course not truly random; for example, C and G bases are present at higher frequency in GC-rich areas such as CpG islands. For this reason, it is ideal to choose restriction enzymes that have cutting sites with a balanced GC/AT content to avoid bias.

In some cases a 4-cutter can be used as the primary restriction enzyme. This produces a 16 times smaller initial fragment size, on average, and makes it less likely that the fragment around the locus of interest will be large enough. Nevertheless, for 4C assays where few loci are being studied, loci can be examined individually to see if a primary restriction enzyme will produce a suitably sized fragment. In general, the smaller fragments produced with 4-cutters increase resolution of the assay; that is, an interaction between two genetic sequences can be narrowed down to a smaller area, because the fragments containing them are smaller. A disadvantage is the cost: 4-cutter enzymes tend to be much more expensive per unit than 6-cutters, and hundreds of units are required for the initial digestion step.

6.2.3.1 Locus digestion strategy

Restriction enzyme choice is dependent upon the location of the restriction sites at the loci of interest. The primary restriction enzyme must cut to create a DNA fragment large enough to be involved in crosslinked interactions. Van de Werken et al and Splinter et al recommend that viewpoint fragments are at least 500 bp, "a size arbitrarily considered large enough to frequently be subject to cross-linking"[302,305]. These viewpoints should also not be too large (here chosen as 10 kb), in order to maintain an acceptable resolution for the assay. As viewpoint fragments increase in size, the number of possible interactions with regions distinct from the loci of interest increases, decreasing the confidence in any interactions that are found. For longer sequences like the *BCAR1* promoters, it is also necessary to ensure that it does not get cut by the primary restriction enzyme.

The size of the secondary fragment is also important. As mentioned above, the fragment must be small enough so that, on self-circularisation with its interacting fragment, it forms a DNA circle small enough to be amplified by PCR. 1 kb was chosen as the maximum ideal secondary fragment size. The larger the circular DNA formed with the interacting fragment is, the less well it is likely to amplify, introducing possible size bias into the PCR reaction.

The minimum size should be large enough so that it can physically circularise easily. The sequence with which it is ligated adds to the size of the circle. As shown in Figure 58, the primers for inverse PCR face outwards from the secondary fragment. These have long overhangs containing the Illumina adapter sequences, used later for next-generation sequencing. The secondary fragment should therefore be larger than about 200 bp to prevent the primers hybridising their ends to each other. To maximise the ability of the fragments to ligate, both the primary and secondary restriction enzymes should create sticky ends rather than blunt ends.

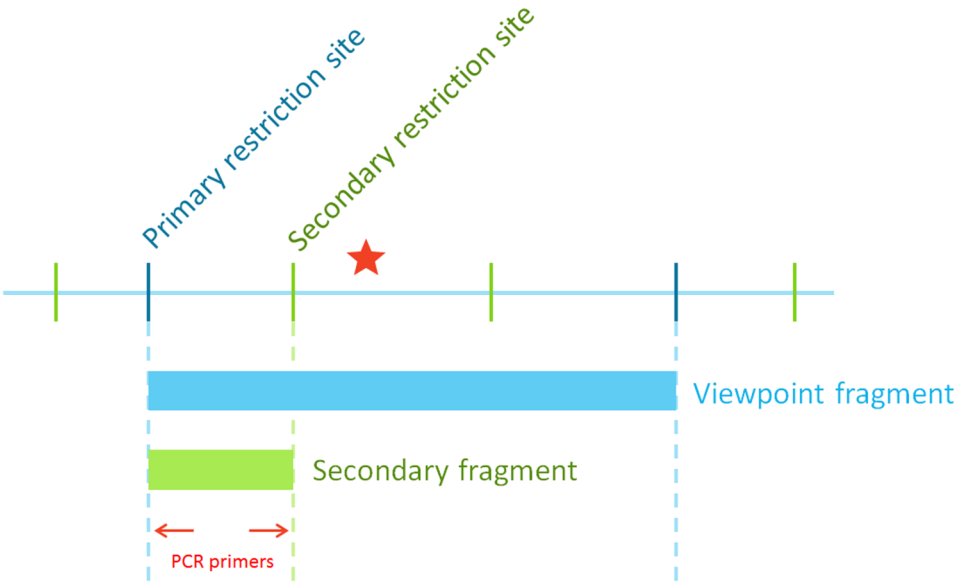

Figure 58: Primary and secondary fragments designed around a locus of interest. Locus of interest marked by ★. The larger primary fragment must contain the locus of interest. At the first ligation step, proximity-mediated ligation will cause it to ligate with other primary fragments with which it is crosslinked. The secondary digest will cut the fragment at the green-marked points. The secondary fragment will be sequenced following the arrow, allowing the identification of the sequence on the other side of the primary ligation junction. This interacting sequence can then be located in the genome. Note that the locus of interest does not need to be covered by the secondary fragment. At this point the fragment serves as an identifier of the primary fragment, for which only a short segment of its sequence is needed.

With the above requirements in mind, rs4888378 and the two *BCAR1* promoters were examined to design a suitable digestion strategy. Potential 6-cutter primary restriction enzymes were chosen based on those recommended by previous studies as being able to work in the presence of SDS (6.2.6) and those containing a relative balance of AT/GC, giving a shortlist of *HindIII*, *BamHI*, *BglII* and *EcoRI*.

It is also desirable that enzyme action is not blocked by CpG methylation, as this would decrease the possibility of fragment interactions being detected in certain regions such as CpG islands, introducing bias into the results. It was noted that despite its recommendation in some studies, *EcoRI* is blocked by some CpG methylation, and thus would not be an ideal primary restriction enzyme.

All the corresponding restriction sites were mapped to the *CFDP1-BCAR1-TMEM170A* locus using UCSC Genome Browser[127] (Figure 59). Using these, for each location of interest – rs4888378, and the promoters BCAR1-1 and BCAR1-1' – the ability of each enzyme to create a primary restriction enzyme with the above parameters was assessed.

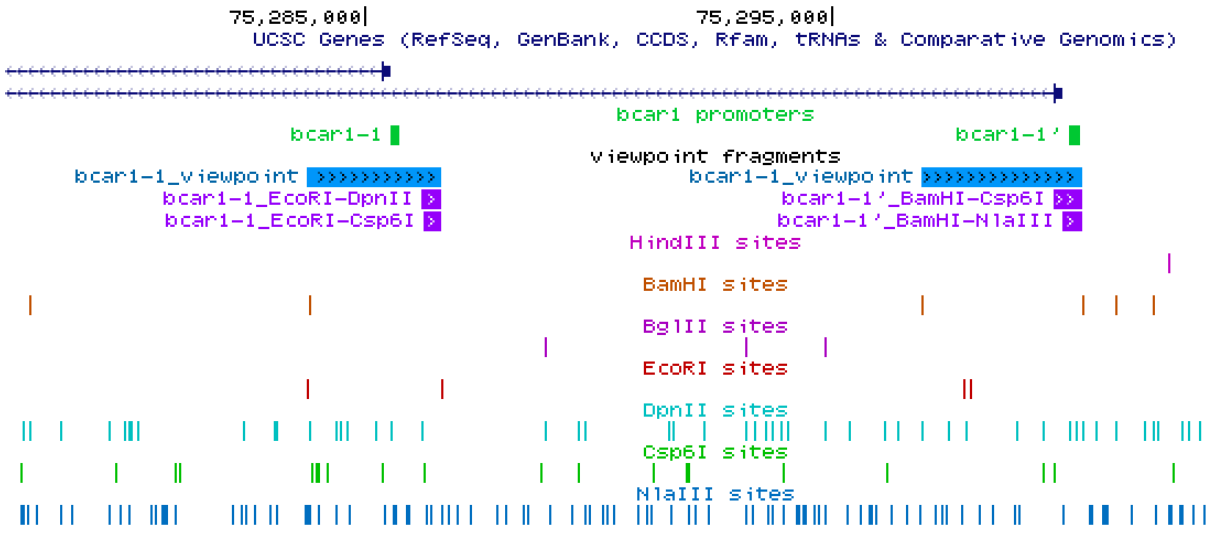

Figure 59: Viewpoint fragments created for the two BCAR1 promoters with 6-cutter restriction enzymes. Created using UCSC Genome Browser[127]. *EcoRI* and *BamHI* can be seen to create viewpoint fragments of suitable sizes for the two promoters, and various 4-cutter enzymes to create smaller secondary fragments.

These enzymes were then tested on crosslinked DNA. As detailed in 6.2.6, neither *BamHI* nor *EcoRI* were able to sufficiently digest chromatin in the presence of SDS, which is required for the first digestion. As no 6-cutter enzymes were suitable here, a second approach at designing a suitable digestion strategy was attempted, using 4-cutters as the primary restriction enzyme. *DpnII*, *Csp6I* (*CviQI*) and *NlaIII* restriction sites were mapped to the locus and the suitability of each to create viewpoint fragments was again assessed.

DpnII and *Csp6I* were selected as the two primary restriction enzymes, as these were able to produce viewpoint fragments of suitable sizes (Figure 60). For each viewpoint, restriction sites for the remaining 4-cutters were examined to see if a suitable secondary fragment could be obtained. If it was possible, this digest combination was taken forward.

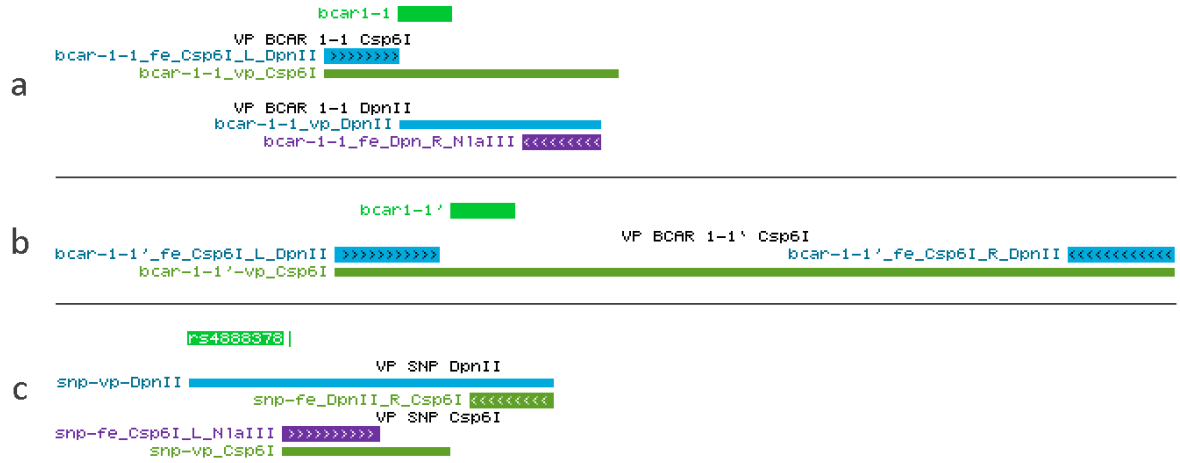

Figure 60: Viewpoint fragments created with 4-cutter restriction enzymes. Figure adapted from UCSC Genome Browser[127]. *DpnII* and *Csp6I* create viewpoint fragments around **(a/b)** the two promoters and **(c)** the SNP, and *DpnII*, *Csp6I* and *NlaIII* create secondary fragments.

Six viewpoint/secondary fragment combinations could be designed (Figure 60). Two were designed for each location of interest, to allow for failure of a viewpoint due to poor digestion or failure to design suitable PCR primers. To analyse these viewpoints, four digestion strategies were required, using two primary restriction enzymes and two subsequent secondary enzymes each (Figure 61). Digestion testing therefore proceeded with these enzymes.

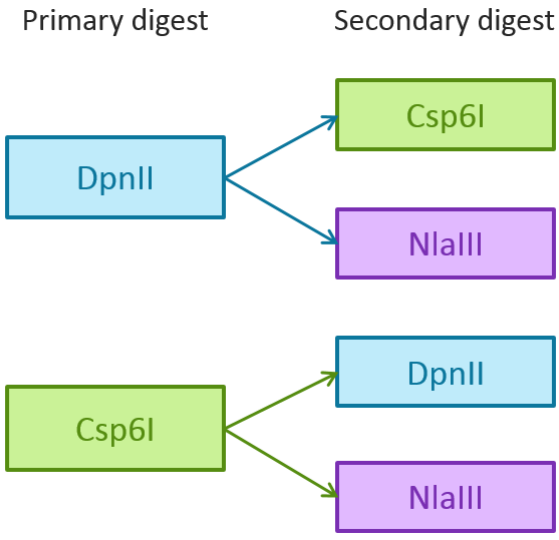

Figure 61: Digest strategy for designed viewpoints. Crosslinked cells will be digested with either *DpnII* or *Csp6I*. These cells will be split into two and digested with each of the appropriate secondary enzymes.

6.2.4 Formaldehyde crosslinking

Previous 4C studies use different methods of crosslinking living cells with formaldehyde. When using adherent cells, as all the cells under study here are, formaldehyde can either be added directly to adherent cells[306], or the cells trypsinised and formaldehyde added to the cell suspension[302,303]. Crosslinking cells in suspension may allow formaldehyde to access the chromatin more easily[303]. The disadvantage is the possibility of the trypsinisation process affecting chromatin structure in the cell, and producing an inaccurate representation of interactions in the cell nucleus. However, as numerous published 3C studies have successfully used this a approach[302,303,305], this was used here.

The percentage of formaldehyde also varied between published studies[303,305,307], with final formaldehyde percentage in crosslinking solutions generally ranging from 1-2%. Crosslinking was therefore tested at final concentrations of 1 and 2%. Both concentrations resulted in a high molecular weight band of DNA when visualised on a gel, so crosslinking was deemed to be suitably successful for both (data not shown). As stronger crosslinking has been shown to "over-fix" chromatin, impeding its digestion with the primary restriction enzyme[307], 1% was chosen as the formaldehyde concentration.

It is also important to consider the stage at which cells are crosslinked. 4C analyses the state of the chromatin at the point in time when crosslinking occurs; cells that are actively dividing are likely to be actively expressing different genes to those that have stopped growing. Therefore it was decided that cells would always be crosslinked at a confluence of 90%, where cells had not yet reached confluence and stopped growing, but a large number of cells could still be harvested from each flask.

6.2.5 Cell lysis

Two different types of lysis buffers are required for the 4C protocol. The first must break the cytoplasmic membrane to release intact nuclei. Published studies use varying preparations of cytoplasmic lysis buffer, but all require a salt, protease inhibitors and a non-ionic detergent. The detergent found to be suitable here was Igepal.

The efficiency of lysis can be assessed with methyl-green pyronin, which stains cytoplasm pink and nuclei blue. A small sample of cells is stained and viewed under a microscope to check for the presence of isolated nuclei (Figure 62).

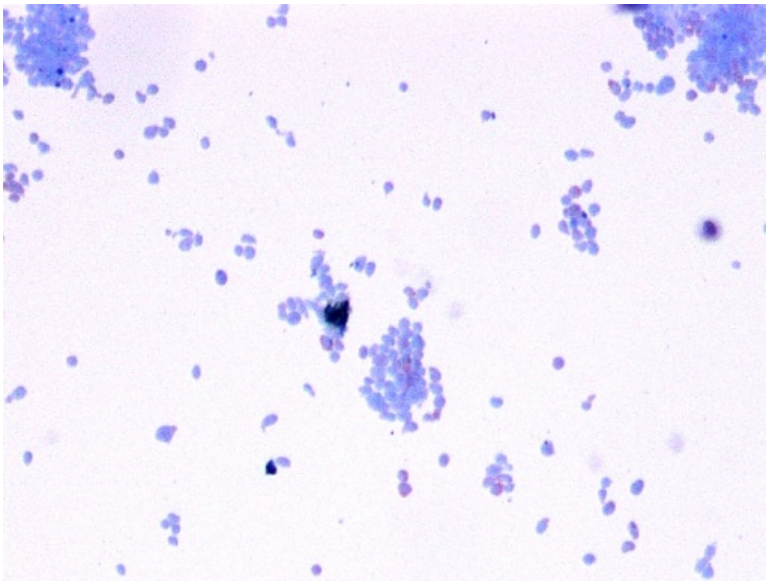

Figure 62: Example of HEK293 nuclei stained with methyl-green pyronin after lysis of cytoplasmic membrane. After lysis of the cytoplasmic membrane, only nuclei (strained purple) are visible.

Initial lysis was tested for various timepoints, and with and without douncing of the solution with a plastic microcentrifuge pestle. Douncing was deemed to be necessary to avoid the of cell clumps during lysis, and 15 minutes of lysis was chosen as the optimum period before checking lysis efficiency.

In order to release the fixed chromatin, the nuclei must then be lysed using an ionic detergent. Published studies use sodium dodecyl sulphate (SDS) at this point[302,303,305]. However, SDS strongly denatures proteins, and therefore its presence in solution inhibits the action of restriction enzymes, and was a major obstacle in the 4C protocol (6.2.6). As it is crucial that the proteins remain crosslinked to the chromatin during digestion and ligation, the DNA cannot be purified to remove the SDS. Instead, a non-ionic detergent such as Triton X-100 is usually added to quench the SDS[302,303]. Previous 4C studies use SDS at a concentration of 0.3%[302,303,305]. The concentration of SDS is a balance between having enough to sufficiently permeabilise the nuclei, and little enough to allow the restriction digest reaction to process. Lower concentrations of SDS (0.1 – 0.3%) were also trialled here, but none were found to aid restriction enzyme action.

6.2.6 Restriction enzyme testing

The action of a number of primary restriction enzymes was tested on crosslinked DNA. Previously published 4C studies suggested that the 6-cutter enzymes *EcoRI*, *HindIII* and *BamHI*[302,303] are effective on crosslinked DNA and in the presence of SDS. 6-cutter restriction enzymes cut on average every 4096 bp. The peak DNA size for digested crosslinked DNA is likely to be higher, as even the

most effective enzymes have up to only 70% digestion on average[307]. As the actual distance between restriction sites varies widely, the digestion product is expected to run as a large smear rather than a band (Figure 63). A substantially higher-weight smear would indicate considerably incomplete digestion.

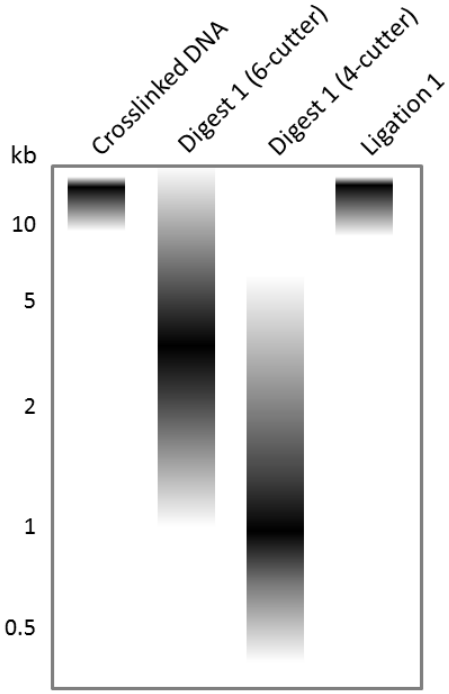

Figure 63: Simulated expected band sizes for crosslinked DNA, digested DNA with 6- and 4-cutter restriction enzymes, and ligated DNA. Smear weights are expected to very with enzyme digestion efficiency.

As shown in Figure 64, the 6-cutter restriction enzymes were tested but only *HindIII* was able to achieve some digestion of crosslinked chromatin. This result fits in with the fact that many 3C-related studies use *HindIII*, finding it to digest best in the presence of SDS[302,308,309]. Unfortunately, while this proved the principle that digestion would work in the experimental conditions (implying no fundamental problem with conditions), *HindIII* was not a suitable primary restriction enzyme for any of the designed viewpoints.

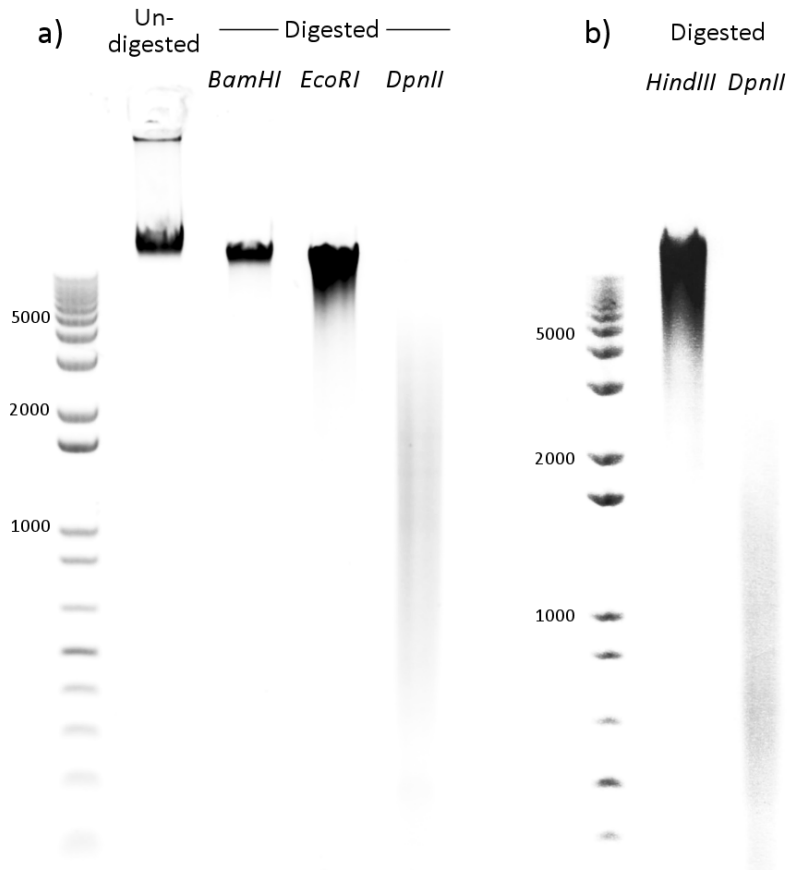

Figure 64: Crosslinked DNA with and without digestion with restriction enzymes. Undigested DNA run on the gel as a high-weight band, as expected. Little digestion is seen with the 6-cutter enzymes, with the possible exception of *HindIII*; however, the peak is still well above 5 kb. The 4-cutter enzyme *DpnII* was run as a comparison, which produces a smear of DNA around roughly 1 kb.

Attempts to improve the digestion reaction in the presence of SDS included increasing the initial volume of restriction enzyme, adding 100-200 units of additional restriction enzyme every 4 hours of digestion, varying the amount of Triton X-100 used to quench the reaction between 0.6-1.8%, varying quenching time between 10 minutes and 1 hour, and increasing the digestion shaking speed to 900 RPM (as recommended by Splinter et al[305] and van de Werken et al[302]). When it was determined that the necessary 6-cutter enzymes could not be used to digest the crosslinked chromatin, viewpoint design was reconsidered, using 4-cutters as the primary enzyme, which had been used successfully in other 3C experiments[310]. *DpnII* and *Csp6I* were found to be the best choices for creating suitable viewpoint fragments (6.2.3.1). Digestion of crosslinked chromatin in the presence of SDS was found to be successful with *DpnII* and *Csp6I* (Figure 65).

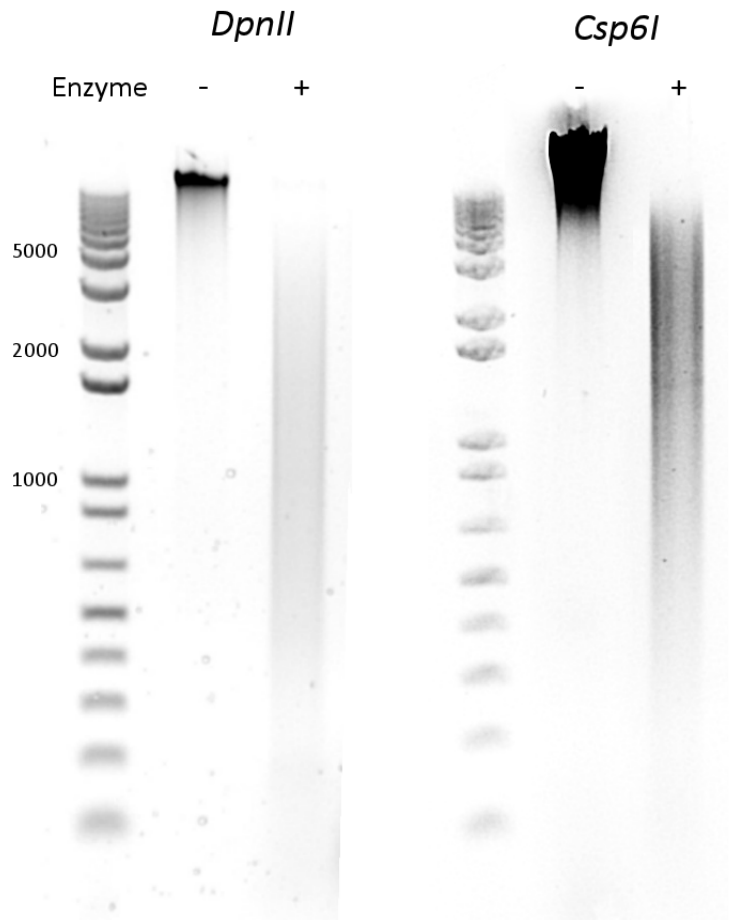

Figure 65: Crosslinked DNA with and without digestion with the restriction enzymes *DpnII* and *Csp6I*. DNA without added enzyme is undigested and present as a high-weight band. (More DNA was present in the *Csp6I* sample, explaining the thicker band.) *DpnII* appears to have higher digestion efficiency than *Csp6I*.

6.2.6.1 First ligation reaction

After crosslinked chromatin is cut into fragments with the primary restriction enzyme, the sticky ends are ligated in order to join interacting fragments. This involves inactivation of the primary restriction enzyme. Both *DpnII* and *Csp6I* are sensitive to heat-inactivation, so heating to 65°C was used to stop digestion before ligation. In contrast to digestion, ligation must be carried out at low DNA concentrations. This increases the chance that ligation will occur between fragments that are physically crosslinked together, rather than those that are in close proximity due to chance. Ligation was therefore carried out in a 7 ml total volume.

The ligation reaction generally produces large aggregates of ligated fragments (Figure 55) which are seen on the gel as a high-weight band. Using 100 units of T4 ligase, the tested ligation reactions successfully produced this band as expected (Figure 66). As some studies use BSA in the ligation

step, ligation was tested with and without the addition of 40 µl 10 mg/ml BSA. Little difference was seen between samples with and without BSA (Figure 66).

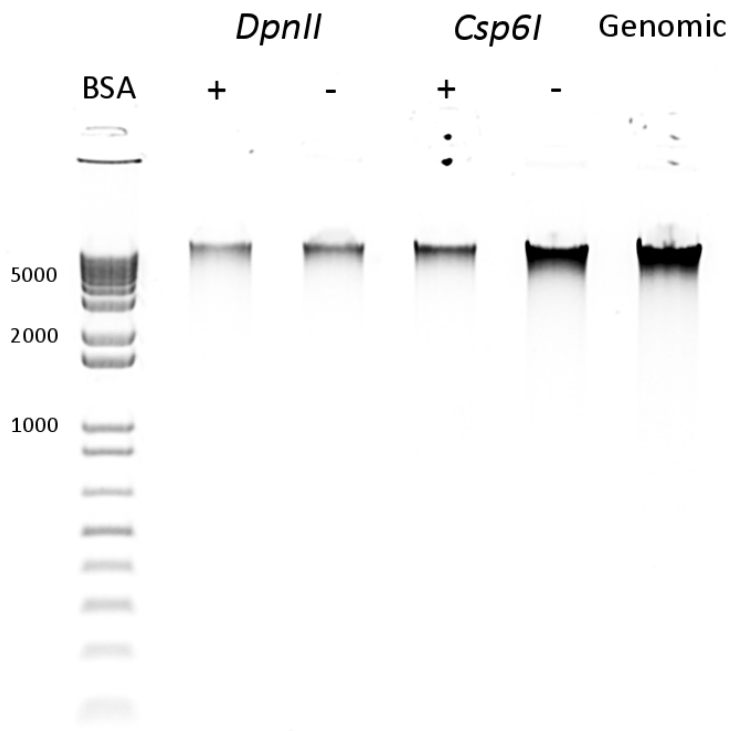

Figure 66: Ligated DNA after digestion with *DpnII* and *Csp6I*. After crosslinked DNA was digested with the primary enzymes and ligated, it was purified by ethanol precipitation and 1 µl run on a gel. Each sample was ligated with and without BSA (+ and -). All ligated samples ran as high-weight bands, expected for large aggregates of crosslinked DNA. Samples were run alongside 250 µg reference genomic DNA which also runs as a high-weight band.

After ligation, crosslinks are reversed before the secondary restriction digest. This is done by heating the reactions to 65°C for 6 hours in the presence of proteinase K. At this point, there are no crosslinks and DNA can be purified, which is done by phenol/chloroform extraction and ethanol precipitation.

6.2.6.2 Secondary restriction enzymes

As the secondary restriction digest was carried out on purified DNA in standard digestion conditions, no special conditions had to be tested. Secondary restriction enzymes were thus chosen on their ability to produce a secondary fragment of suitable length, dependent on the location of their restriction sites at the locus, and ability to create sticky ends. The secondary enzyme also cannot cut at the primary restriction site (as is the case for enzymes such as *BamHI* and *DpnII*). As discussed in 6.2.3.1, the three enzymes used were *DpnII*, *NlaIII* and *Csp6I*.

The chosen enzymes were found to successfully digest ligated DNA. *NlaIII* digested DNA more effectively than *DpnII* and *Csp6I* (Figure 67), but all smear sizes here were considered acceptable.

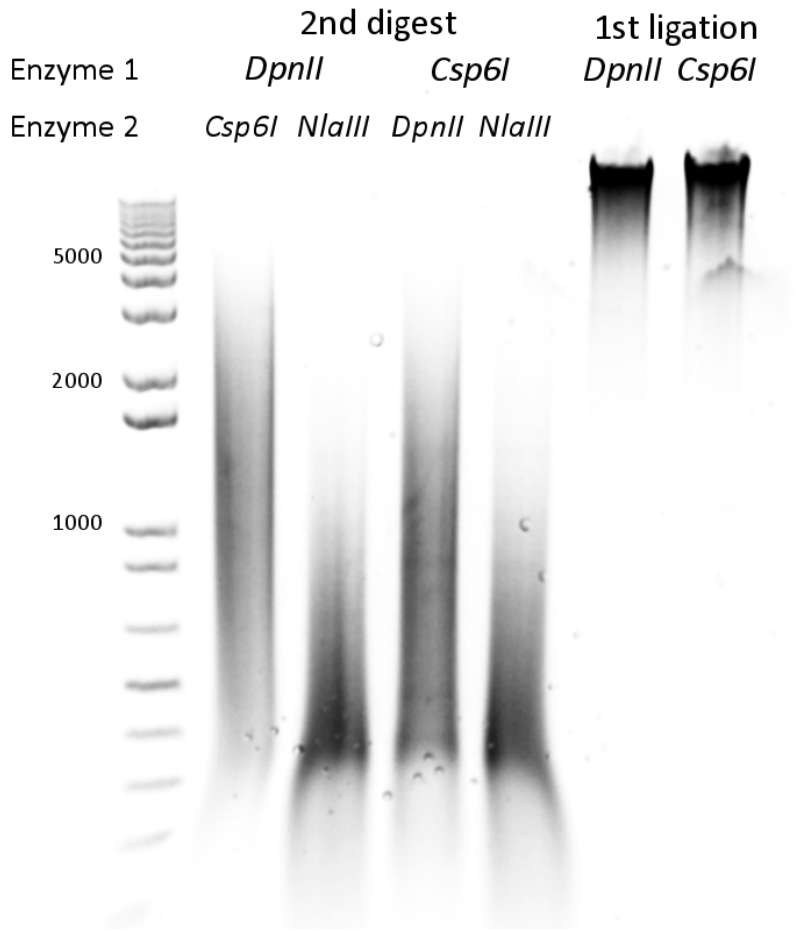

Figure 67: Ligated DNA digested with the secondary restriction enzymes. All secondary restriction enzymes digested ligated DNA to a lower-weight smear, with *NlaIII* showing the greatest digestion efficiency. Samples run next to ligated DNA for comparison of DNA fragment sizes.

6.2.6.3 Secondary ligation and DNA purification

As with the first ligation reaction, secondary ligation is carried out in diluted conditions to favour intra-molecular ligations between ligated fragments. Ligated fragments are here expected to largely circularise, producing varying fragments that do not differ largely from those seen in the second digestion (Figure 68). After phenol-chloroform extraction and ethanol precipitation, DNA was purified using the QIAquick gel purification kit, used for direct cleanup from enzymatic reactions.

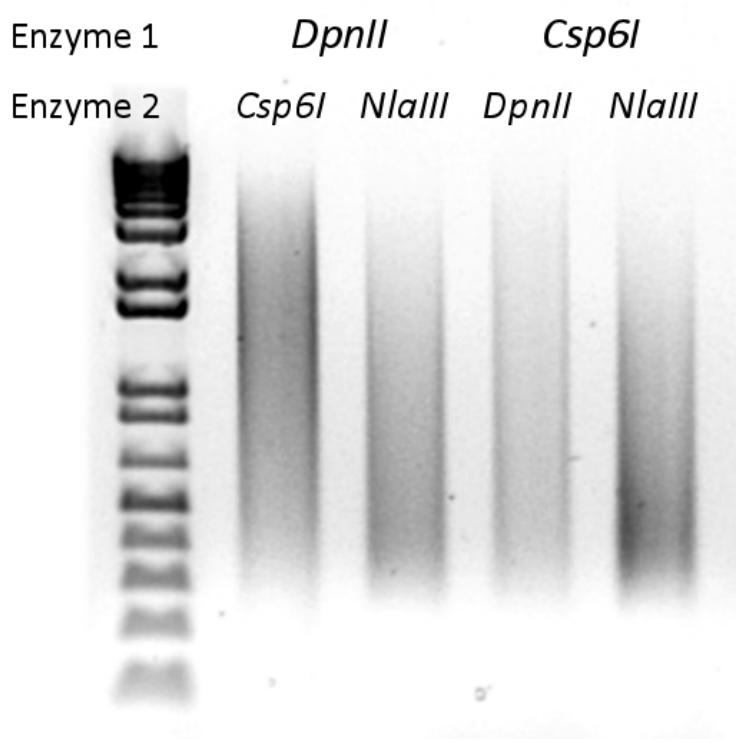

Figure 68: DNA produce of the second ligation step. The DNA smears are not expected to be greatly different in size to those in the second digestion, as most fragments self-circularise.

6.2.7 4C-PCR

6.2.7.1 PCR primer design

The aim of the PCR reaction is first to amplify up circles of DNA from the genomic locations of interest, and secondly to identify the sequences contained in these circles using next-generation sequencing. The PCR reaction will therefore ultimately be carried out with primers linked to Illumina sequencing adapters: these become incorporated into the PCR product, and allow sequencing using Illumina technology.

Only one primer is used for sequencing; this is the 'read' primer, while the other primer is 'non-read'. The primers are positioned close to the edge of the secondary fragment. The sequencing reaction reads from the read primer into the unknown sequence of the ligated fragment, allowing it to be identified and its location in the genome determined (Figure 69).

One of the principal questions in primer design is that of the location of the read primer. Recent 4C studies suggest it should be designed directly on the primary restriction site, so that the sequencing result is composed entirely of unknown interacting sequence, rather than uninformative sequence from the secondary fragment[302,303]. However, since the publications of these studies, sequencing technology has improved and sequence reads are substantially longer (150 bp reads were used

here), allowing for capture of sufficient informative sequence even if the primer binds further back from the restriction site.

Primers were designed to be no further than 30 bp from the restriction site; an informative sequence read of over 100 bp is sufficient for alignment of the sequence to the genome. This strategy of primer positioning allowed improved primer design, as restricting primers to an exact site considerably increases the chances of poor primers, due to potential mispriming, inappropriate melting temperatures (Tm) or secondary structure. The non-read primer was designed to be up to 100 bp away from the secondary restriction site. As this is not used for sequencing, the only concern was an unnecessary increase in size of the PCR product.

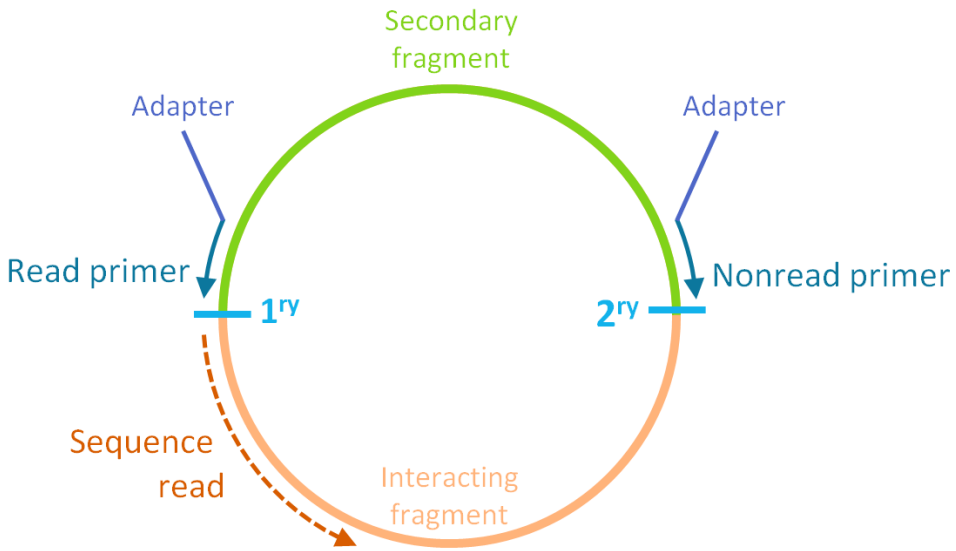

Figure 69: Primer design and position. The inverse PCR primers are designed to face outwards from the secondary fragment, close to the primary and secondary restriction sites. Both primers are designed with Illumina sequencing adapter overhangs. The read primer binds close to the primary restriction site and incorporates the reading sequence adapter here. When the amplified DNA circles are sequenced, the sequence read progresses through the primary restriction site into the unknown interacting fragment.

As many primers often have to be trialled, testing was initially carried out with primers without the attached adapter sequence. Successful primers were then designed with the adapter sequences.

Primers were designed using the primer design tool Primer3Plus[212], allowing the production of primers with the best conditions; i.e. no mispriming to elsewhere in the genome, balanced GC content, similar Tms and low self-complementarity. Each of the six secondary fragments from 6.2.3.1 were used to design primers for inverse PCR. As the primers here face outwards rather than inwards, each sequence was rearranged before inserting into Primer3Plus so that it resembles more typical PCR design (Figure 70).

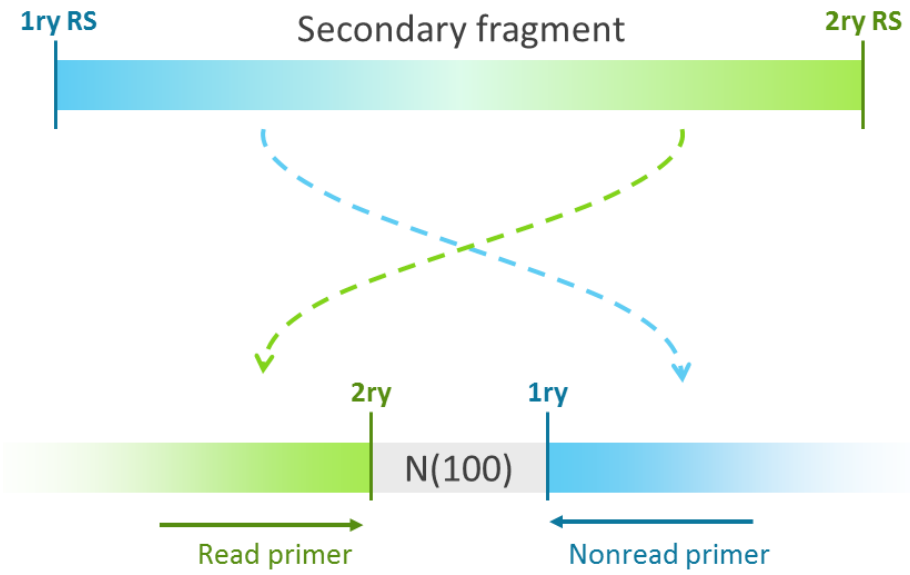

Figure 70: Sequence rearrangement before primer design. The sequence of the secondary fragment is taken and rearranged into a form suitable for conventional primer design before being used in Primer3Plus. The 50 bp adjoining the primary restriction site and the 130 bp adjoining the secondary restriction site are swapped over and connected with a 100 bp stretch of N-nucleotides, which serve as the placeholder for the unknown ligated fragment in the DNA circle. Primers are then designed to amplify this stretch of Ns. The available sequence for primer design by the secondary restriction site is longer because this primer will not be used to produce a sequencing read, so does not need to be as close to the restriction site.

No suitable primers could be designed for one of the *BCAR1*-1 promoter viewpoints (BCAR1-1 *Csp6I*, shown in Figure 60). This left primers for five viewpoints overall. For each, two read primers or two non-read primers were designed, to allow testing of two primer pairs each.

Multiple sets of primers were subsequently designed, as initial sets produced substantial non-specific binding in control genomic DNA templates (6.2.7.2). More stringent mispriming thresholds were used, and the final set of primers can be seen in Table 32.

Table 32: Primer names and properties. Two pairs of primers were tested for each secondary fragment, where the sequence was amendable to primer design.

Viewpoint	Primer name	Site	Read/non-read	Secondary fragment side	Sequence	Length	Tm
BCAR1-1 *DpnII*	11_*DpnII*_R_*NlaIII*_Read	*DpnII*	Read	Left	CATTGTGCCGACATCTGG	18	59.6
	11_*DpnII*_R_*NlaIII*_Non	*NlaIII*	Non	Right	CCCCTTTCCCTCTAGATTCA	20	58.2
	11_*DpnII*_R_*NlaIII*_Non2	*NlaIII*	Non	Right	GAGCTGTGGTGGTGATGATT	20	58.5
BCAR1-1' *Csp6I* (left side)	11p_*Csp6I*_L_*DpnII*_Read	*Csp6I*	Read	Right	AAGATGTCCGTGCCTGTG	18	58.1
	11p_*Csp6I*_L_*DpnII*_Read2	*Csp6I*	Read	Right	CAAGATGTCCGTGCCTGT	18	58.1
	11p_*Csp6I*_L_*DpnII*_Non	*DpnII*	Non	Left	ACCTAGGCCTTTTTCTGTCG	20	58.5
BCAR1-1' *Csp6I* (right side)	11p_*Csp6I*_R_*DpnII*_Read	*Csp6I*	Read	Left	GTGGCTCCTTATCTCCCTGT	20	58.2
	11p_*Csp6I*_R_*DpnII*_Read2	*Csp6I*	Read	Left	GGGAAGTGGCTCCTTATCTC	20	57.8
	11p_*Csp6I*_R_*DpnII*_Non	*DpnII*	Non	Right	TTCCCTATAGGACGGAGTGA	20	57.2

SNP *DpnII*	SNP_*DpnII*_R_*Csp6I*_Read	*DpnII*	Read	Left	CACAAAAACACCTGGTCTCC	20	58
	SNP_*DpnII*_R_*Csp6I*_Read2	*DpnII*	Read	Left	ACAAAAACACCTGGTCTCCA	10	58
	SNP_*DpnII*_R_Csp6_Non	*Csp6I*	Non	Right	AGAAACTGCCCTTCCAGTCT	20	58
SNP *Csp6I*	SNP_*Csp6I*_L_*NlaIII*_Read	*Csp6I*	Read	Right	GGGACCCAAGTTTAAACAAA	20	57.9
	SNP_*Csp6I*_L_*NlaIII*_Read2	*Csp6I*	Read	Right	GGACCCAAGTTTAAACAAACA	21	56.7
	SNP_*Csp6I*_L_*NlaIII*_Non	*NlaIII*	Non	Left	TAGTCACCCTGGCTGAAGTC	20	56.6

As all sequencing reactions are loaded onto the same flow cell, different 4C reactions can only be distinguished by the sequence itself. The sequence of the read primer can be used to distinguish between viewpoints. However, if reactions in different conditions – e.g. using crosslinked DNA from different cells – are to be sequenced on the same flow cell, different barcodes are needed on the sequencing primers. To allow for potential sequencing of reactions from HEK293 and HUVEC cells, adapter-linked read primers were ordered with a 2 bp barcode (GA or TC) between the primer and adapter.

6.2.7.2 PCR reaction

Numerous parameters were tested in the optimisation of 4C-PCR, including thermocycling extension temperature, amount of template DNA, volume of primer, and type of polymerase.

NEB's Phusion High-Fidelity DNA polymerase, AcquaScience's Acqua polymerase, and Roche's Expand Long Template PCR System were trialled. Expand Long Template was used as it was found to produce the best amplification (data not shown). It also amplifies large circular DNA well, decreasing possible size bias in amplification.

Multiple negative controls were used for the PCR reaction. A no-template control was used to check for the presence of primer dimer, which is a particular concern in the case of the long adapter-linked primers. Genomic DNA should also be used as a negative control. As the primers point outwards from the secondary viewpoint, no amplification would be expected with a genomic DNA template. Assuming satisfactory primer design, any that occurs would be due to non-specific binding.

For initial primers, many bands were produced with control genomic DNA, indicating mispriming to the genome (although online database tools such as UCSC In-silico PCR[127] did not indicate any templates in the genome to be present). Subsequent sets were designed with higher mispriming thresholds.

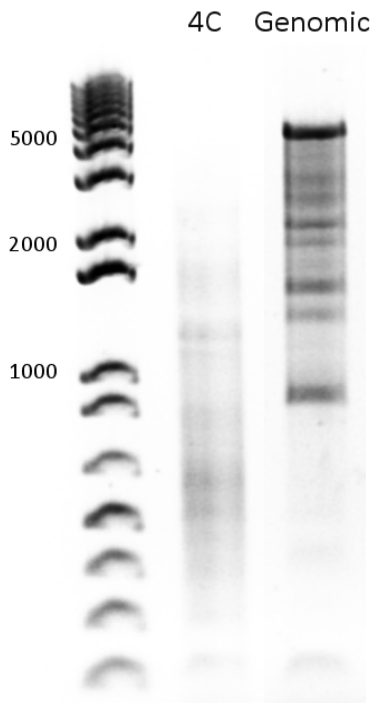

Figure 71: Example of PCR results on 4C DNA and genomic DNA (control). Amplification has occurred with genomic DNA, indicating mispriming.

After suitable primers had been verified, the initial primers were designed with linked Illumina sequence adapters. As these adapters are overhangs that do not bind the template, PCR conditions were calculated on the basis of the Tm from the initial primers.

Primers for four of the secondary fragments were found to amplify the 4C DNA successfully (Figure 72), and larger-scale PCR reactions were then used to amplify large amounts of DNA to create the final 4C libraries. As shown in Figure 72, the amplification product is generally a smear of DNA below 1000 bp, with a stronger band where the secondary fragment has re-ligated to itself.

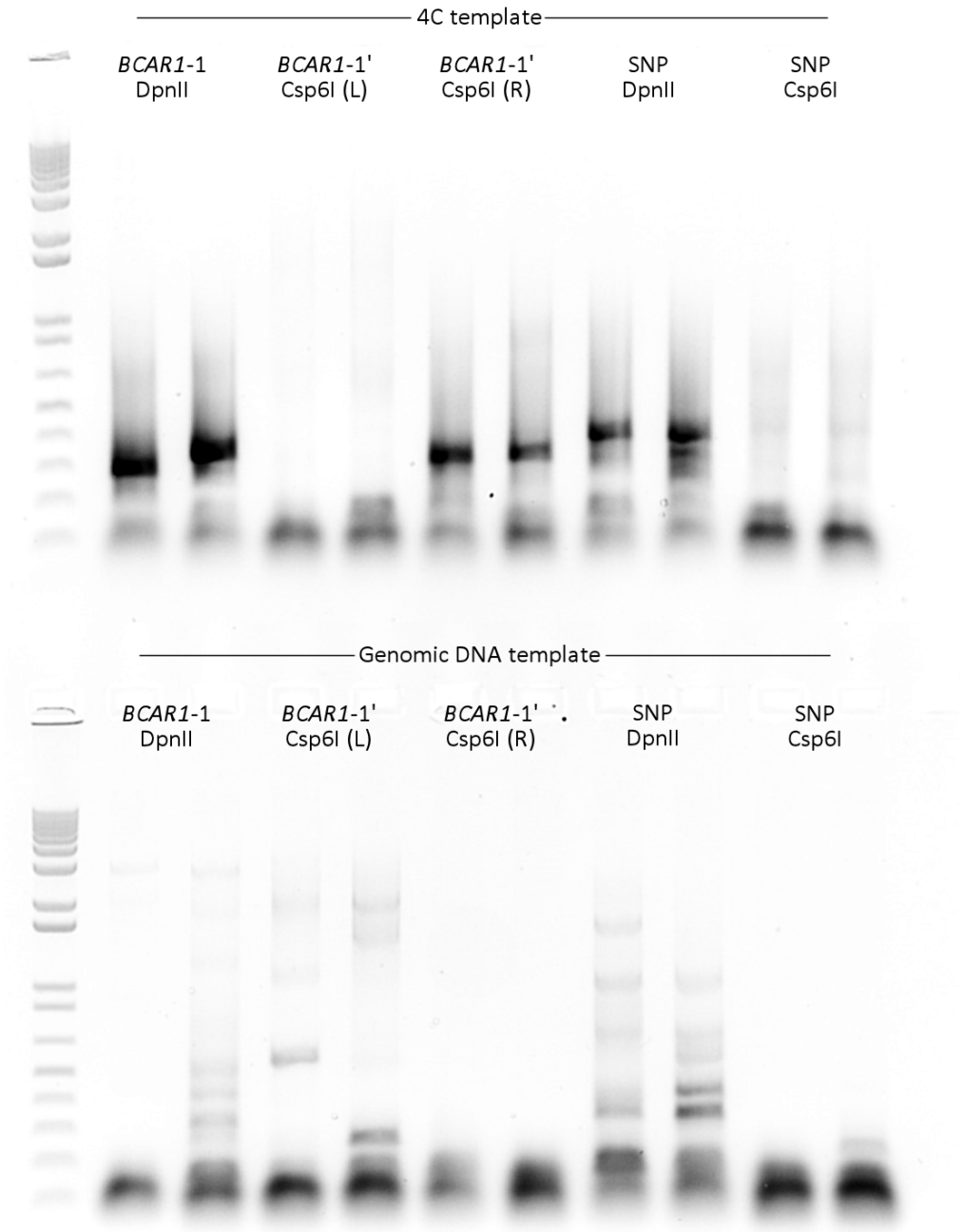

Figure 72: PCR amplification using final primers on 4C DNA and genomic DNA control. Some amplification is seen with genomic DNA, but amounts are lower than that seen with 4C DNA. Both pairs of primers were tested for each secondary fragment, and the more efficient primer pair taken forward to create the final 4C libraries. The *BCAR1*-1' *Csp6I* (L) primers could not achieve successful amplification and were not taken forward.

6.2.8 Verification of 4C PCR product

The size and quantity of PCR-amplified DNA was measured using the Agilent 2100 BioAnalyser in order to calculate DNA concentration. This allowed an estimation of the relative amounts of different circular DNA sizes (Figure 73). This data showed DNA to be present at a range of sizes, an effect largely masked by the large peak that represents self-ligation and recircularisation of the DNA

fragment. This method does not prove that this DNA will form sequenceable clusters an Illumina sequencing platform.

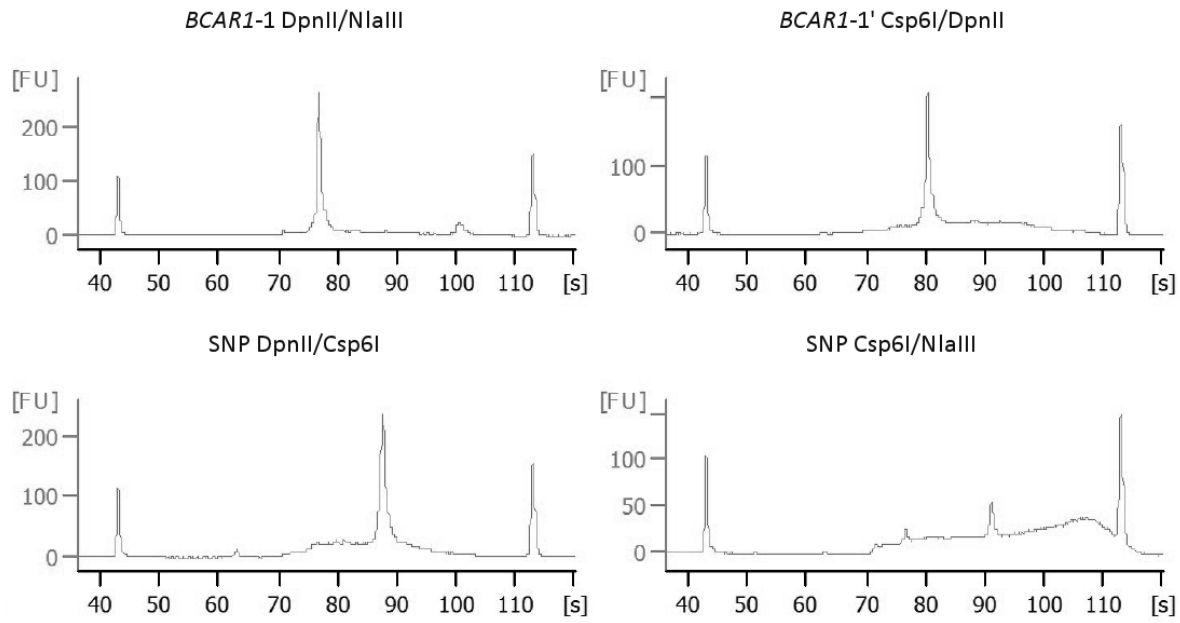

Figure 73: Amplified 4C DNA size and abundance for the four successfully amplified viewpoints at the locus. The peaks at 43 and 103 represent marker DNA, with the test DNA in the middle. For each sample, a peak of DNA of a certain size is seen, which represents the self-circularised fragment, expected to occur much more often than other ligations. The remaining DNA at lower abundance represents ligations with other fragments. [s] represents time for analysis and is proportional to fragment length. [FU] represents relative abundance.

The NEBNext Library Quant Kit for Illumina was used to assay the presence specifically of sequenceable DNA molecules. This is a qPCR kit that quantifies DNA containing the correct Illumina adapter sequences. The 4C libraries were diluted to 1:1000, 1:10,000 and 1:100,000. The qPCR reaction was carried out on these samples, DNA standards from 0.01 to 10 pM and NTCs. All samples were measured in triplicate (Figure 74). The DNA standards were used to create a standard curve by plotting concentration vs Ct (threshold cycle; the cycle number at which fluorescence crosses the fluorescence threshold), from which the concentration of samples could be determined (Figure 75). Undiluted library concentrations were found to be:

BCAR1-1 *DpnII/NlaIII* 408 nM

BCAR1-1' *Csp6I/DpnII* 405 nM

SNP *DpnII/Csp6I* 393 nM

SNP *Csp6I/NlaIII* 343 nM

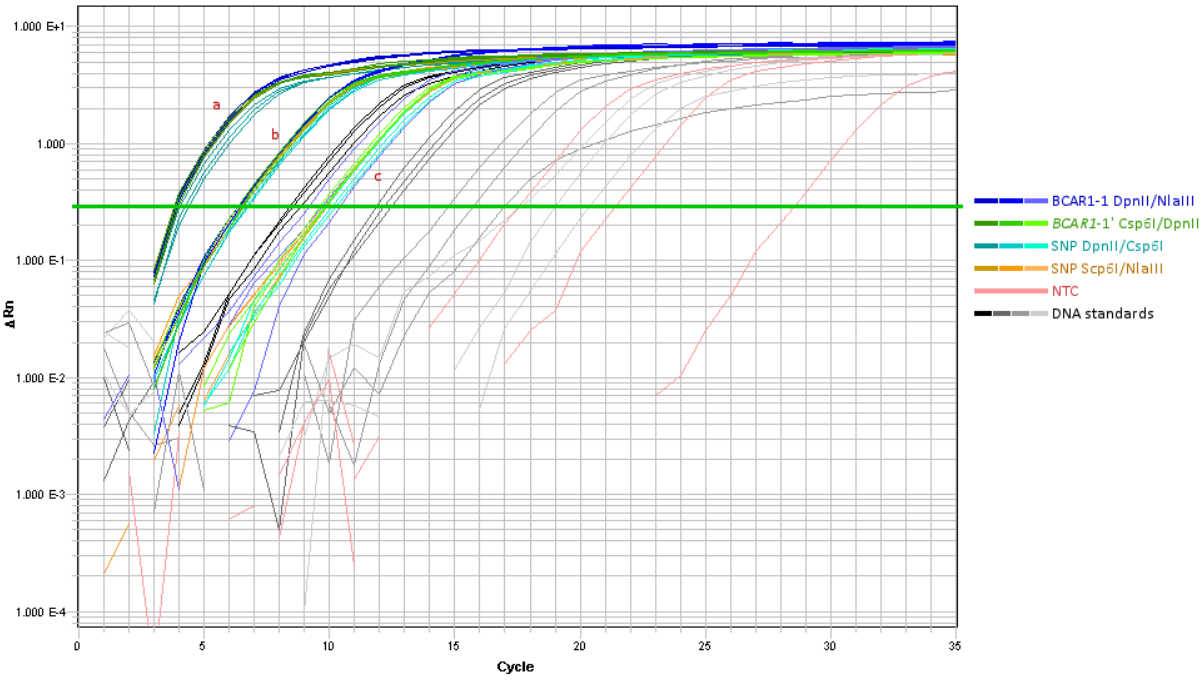

Figure 74: qPCR amplification plot for 4C libraries. Each sample was run at three dilutions: 'a', 'b' and 'c' mark the 1:1000, 1:10,000 and 1:100,000 dilutions respectively. It can be seen that there is substantial amplification for each of the samples, indicating the presence of sequenceable DNA. The green line represents the fluorescence threshold from which Ct values are calculated.

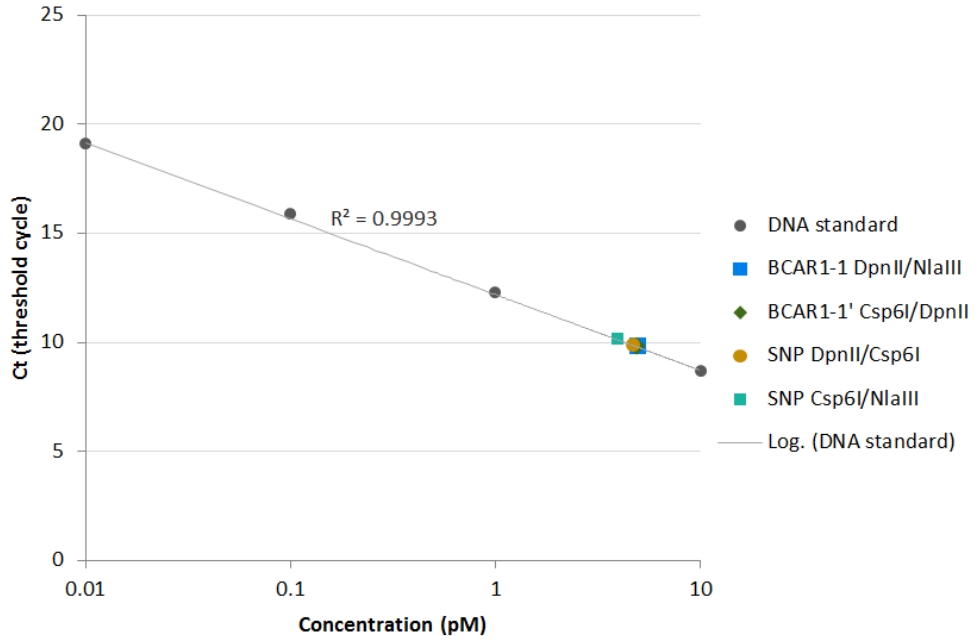

Figure 75: qPCR standard curve. The four DNA standard were plotted on a graph of concentration vs Ct, from which the Ct values of each sample were used to calculate their concentrations. The points seen on the graph here all refer to the smallest 1:100,000 dilutions; others were of too high a concentration to fit on the standard curve.

6.2.9 Sequencing of PCR libraries

Once the presence of DNA with the correct adapters has been verified, the 4C libraries can be sequenced using a high-throughput sequencing platform such as the Illumina HiSeq 2000. Libraries were sequenced on the MiSeq desktop sequencer to obtain initial results, with the intention of moving to the HiSeq if the number of reads produced was insufficient.

Preparation of the 4C libraries for sequencing involved consideration of possible sequencing problems. Sequence reads can be disrupted if an identical nucleotide is present in all reads at the same position. With the four read primers that were used, this situation is possible; therefore, a PhiX bacteriophage DNA control was added at 20% to increase nucleotide diversity.

However, the MiSeq reaction generated no clusters for DNA sequencing in multiple runs. PhiX control DNA was increased to 50% but this still produced no clusters. This result was unexpected, as control DNA should still cluster even if there were problems with the 4C library. Due to time constraints and no logical explanation for the lack of clusters, this experiment was stopped for the purpose of this thesis.

6.3 Discussion

As more is understood about the mechanism of long-range interactions in genetic regulation, chromosome conformation capture methods are likely to be used increasingly in understanding the regulation landscape. In this chapter the *CFDP1-BCAR1-TMEM170A* locus was used to create a 4C protocol with suitable viewpoints, digestion strategies and 4C-PCR methods. Final 4C libraries were produced and can be sequenced if further optimisation can be performed.

A number of obstacles hindered the 4C process. As the technique is relatively new, with as yet no single established protocol, there are many ways to carry out the assay and obtaining end results is not necessarily straightforward. Here the next-generation sequencing assay was repeatedly unable to generate clusters with the four 4C libraries, despite verification of sequenceable DNA using qPCR.

The nature of the 4C assay places restrictions on protocol design, as suitable restriction enzymes have to be chosen based on the location of restriction sites at the locus, limiting the number of suitable enzymes or even making a locus unsuitable for analysis. This difficulty is compounded by need for a secondary digestion step with secondary restriction enzymes, the restriction sites for which must also be located to give appropriately-sized digestion products and suitable sites for primer design. The primary digestion step itself also presents problems in the development of the assay, because SDS – crucial for the assay as it permeabilises cell nuclei –impedes the action of restriction enzymes, making the majority of enzymes unusable for the assay and requiring optimisation of the digestion protocol for the remaining candidates.

These requirements are a disadvantage of locus-focused methods like 4C. Under multiplex methods such as Hi-C, where a specific sequence is not being targeted, one digestion strategy can be used with whichever enzyme performs best under assay conditions. (For example, Mifsud 2015[308].)

6.3.1 What results do we expect?

In this chapter, sequencing was not successful for the prepared 4C libraries, meaning that the sequence reads could not be mapped to the genome. This is the step that should provide information about interactions with the loci under study.

After obtaining successful sequencing results, sequencing reads would have to be processed in order to generate interaction profiles. As all samples are sequenced on the same flow cell, sequence reads first have to be binned according to the reading primers. All primers are specific to their secondary viewpoint, so this allows reads to be binned according to the relevant 4C library being assayed. If

libraries from experiments in multiple cell types are being run on the same flow cell, the 2 bp barcode sequence in the read primer is used to differentiate reads from these different experiments.

Reads are then trimmed to remove the primer sequence and sequence before the restriction site, as these will prevent the read from aligning to the reference sequence. Reads are then mapped to the genome and this data converted into a format suitable for visualisation on UCSC GenomeBroswer, such as the bedGraph format. Quality controls methods should be carried out: the ratio of mapped reads in cis (same chromosome) vs in trans (a different chromosome) is indicative of the quality of the experiment; a low ratio indicates many random ligations, perhaps due to low crosslinking efficiency. A large peak is expected at the viewpoint itself where the fragment has self-circularised rather than ligating to interacting fragments. After removing these reads, the sequence read density should indicate the strength of interactions between any location and the viewpoint fragment.

Figure 76 and Figure 77 show results we might expect from sequencing of the 4C libraries under the hypothesis that rs4888378 is present in an enhancer that interacts directly with one or both of the *BCAR1* promoters. A large peak of sequencing reads is expected in close proximity to the viewpoint itself, as interaction frequencies generally decrease rapidly as genomic distance increases[307]. Peaks of denser sequence reads above this baseline indicate specific looping interactions. Figure 76a shows expected results if rs4888378 is present in an enhancer that interacts with both of the identified *BCAR1* promoters. Alternatively, it may be the case that the enhancer interacts with only one of the promoters; carrying out the assay in multiple cell types might provide information about whether the promoters are differentially active in different tissue types.

If the hypothesis that rs4888378 interacts directly with the *BCAR1* promoters is incorrect, sequencing results would be expected to show no clear peaks at these loci for the relevant viewpoints. Instead, they may suggest alternative models for the regulatory landscape at the locus; for example, they may show other enhancers with which the *BCAR1* promoter interacts, or indicate that rs4888378 exerts its effect through interactions with other enhancers.

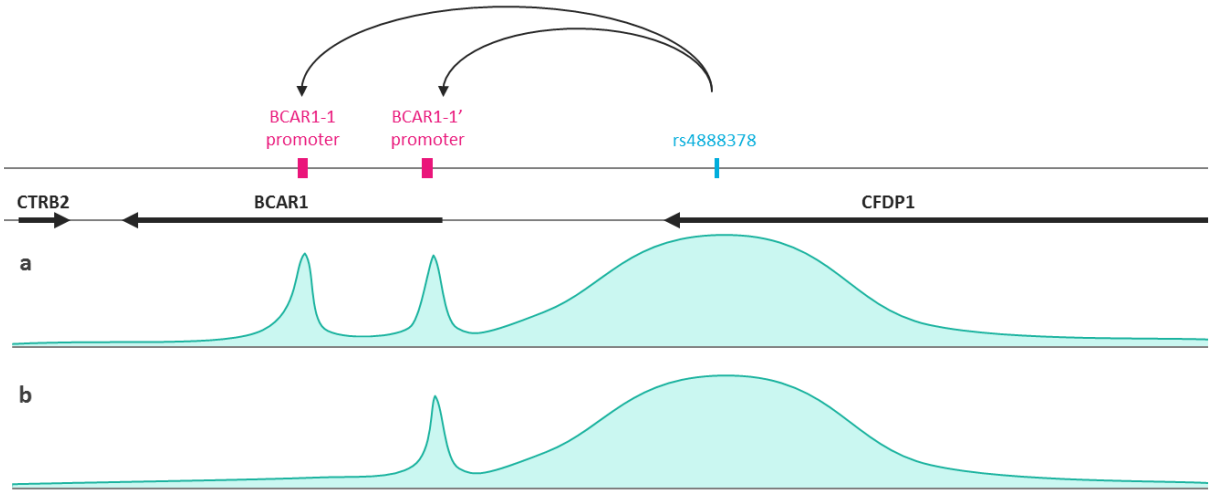

Figure 76: Theoretical results for mapping of sequence reads from the rs4888378 4C libraries. The height of the curve represents sequencing read depth at locations across the locus. **(a)** indicates a scenario in which rs4888378 is present in an enhancer which physically interacts with both of the *BCAR1* promoters; **(b)** indicates results if the enhancer interacts with only one promoter.

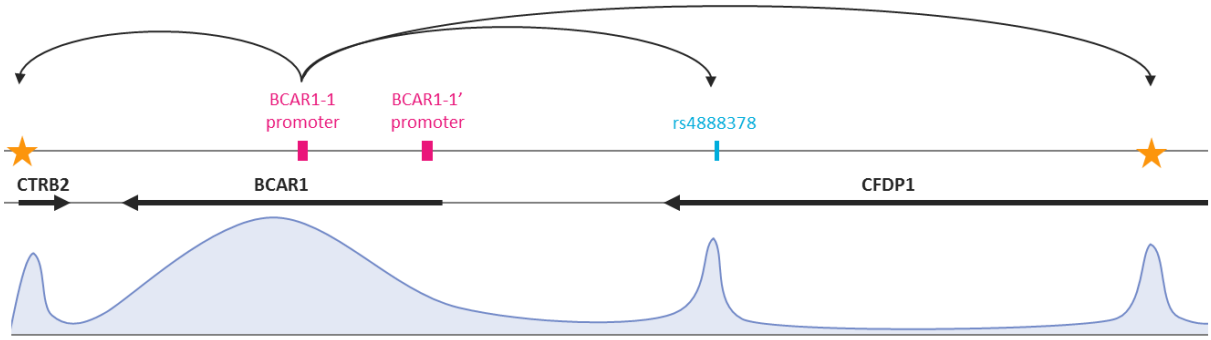

Figure 77: Theoretical results for mapping of sequence reads from a *BCAR1* promoter library. The height of the curve represents sequencing read depth at locations across the locus. The figure indicates theoretical interactions with the *BCAR1*-1 promoter: in additional to interacting with an enhancer containing rs4888378, the promoters may interacts with other enhancers either within the locus or located more distally.

If 4C results could be used to successfully identify regions interacting with both the rs4888378 and the *BCAR1* promoters, further work could investigate how genetic variants affect interaction and subsequently regulation.

For example, a result showing that rs4888378 interacts with the *BCAR1* promoters could be followed up with allele-specific chromosome conformation capture to test whether the variant actually disrupts interaction, as carried out in Visser et al[311]. 4C assays would be carried out using cell lines or tissues of known genotype for rs4888378 and the results compared to see whether interaction peaks different between genotypes. This assay would be improved with the use of genome editing via clustered regularly-interspaced short palindromic repeats (CRISPR)[312] to change only the base at

the location of the SNP (see chapter 9.4). This would ensure that differences in interaction were not due to other variants at the locus.

4C sequencing results might also be expected to reveal the presence of insulators at the *CFDP1-BCAR1-TMEM170A* locus. Insulators are genetic boundary elements that inhibit enhancer-promoter interactions. They may do so by blocking the action of the enhancer on the promoter, if situated between the two[313], or by preventing the spread of condensed chromatin which silences expression[314]. Insulator sequences in vertebrates are bound by the transcriptional repressor CCCTC-binding factor (CTCF)[315], and the insulator activity is thought to work primarily through regulation of the 3D chromatin structure through loop formation, mediated by CTCF[316]. Insulator sequences at the locus studied with 4C, defined as CTCF-enriched elements, are shown in Figure 78. It can be seen that in four of six of the cell types assayed, no insulators are present between the promoters and rs4888378, indicating that these are not prevented from interacting *in vivo*. It is also of interest that the presence of insulators varies between cell types, highlighting the value of carrying out 4C in different cell types to assess how interaction profiles may vary.

In this chapter HEK293 cells were used for the optimisation of the 4C protocol, with a view to carrying out the final protocol on HUVEC cells. Further work would use additional cell lines and tissues; it may be particularly valuable to assay vascular smooth muscle cells in addition to endothelial cells such as HUVECs, as these are the two main components of the blood vessel wall. Enhancers, such as those potentially interacting with the *BCAR1* promoters, may be tissue-specific[317], in which case assays in different cell lines might be expected to show different interaction profiles, providing information about in what tissue and which stage of atherosclerosis genetic variants under study are having an effect.

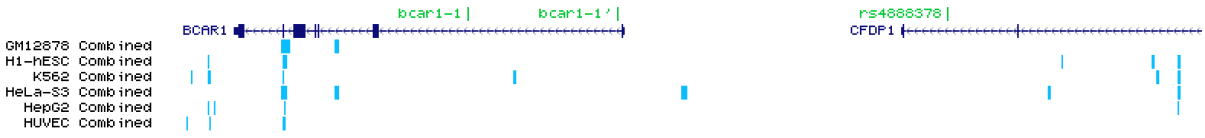

Figure 78: Insulators at the locus under study. Visualised on UCSC Genome Browser[127]. Insulators are defined at CTCF-enriched elements.

7 Coding variation at the *CFDP1-BCAR1-TMEM170A* locus

7.1 Introduction

The majority of GWAS hits for clinically relevant traits are located in non-coding regions of the genome[229], including the lead SNP at the *CFDP1-BCAR1-TMEM170A* locus and the variants in strong LD. Variation in these intronic or intergenic regions with a functional effect is likely to be regulatory, having an effect on gene expression. Variation in protein-coding regions is also of importance to cardiovascular disease, as many cardiovascular conditions are caused by a base change in an exon that alters protein structure. Familial hypercholesterolaemia (FH), for example, is largely caused by point mutations or minor insertions/deletions in the genes *LDLR*, *APOB* and *PCSK9*[88]. These mutations code protein alterations that inhibit uptake and degradation of LDL through the LDL-receptor pathway, raising circulating LDL levels and causing early cardiovascular disease. Hypertrophic cardiomyopathy is also due to base changes in coding regions, largely mutations in genes coding for proteins of the cardiac sarcomere, causing thickening of the myocardium and increasing risk of sudden cardiac death[318].

These diseases are caused by a single mutation with a large effect, and are inherited in a Mendelian pattern. Characterisation of such Mendelian diseases does not explain a large proportion of inherited variation for most traits, as the Mendelian mutations are rare, but finding the variants can provide new biological insights about the genes and phenotypes they are involved in. For example, the discovery of FH variants increased understanding of the LDL-receptor pathway and about the effects of LDL cholesterol on MI[5].

Common variants in the *CFDP1-BCAR1-TMEM170A* locus would not be expected to have large effects, as selection against strongly damaging mutations keeps them rare[319]. However, variants affecting protein sequence can still be used to understand the role of the gene in certain phenotypes.

Studying protein-coding variants is made easier by the fact that they can have a predictable and measurable effect on phenotype. At the protein sequence level, it can be easily determined whether a variant will either cause no change, alter an amino acid, alter many amino acids by changing the reading frame, or introduce a premature stop codon which truncates the protein. It is also clear which gene a variant affects, in contrast to those in non-coding regions.

The proportion of coding variants causing a phenotypic change is larger than for those in non-coding regions. The majority of rare missense mutations are predicted to be deleterious[320], and of all common variation, missense variants are the type most likely to affect phenotype[321].

In this chapter, the *CFDP1-BCAR1-TMEM170A* locus was examined for coding variants to determine whether such variants exist, and if they may result in observable phenotypic alterations. The cohorts used for phenotypic association analysis were IMPROVE and PLIC. The phenotypes studied were carotid IMT and carotid plaque. The variables are similar, with plaque being a binary variable derived from the common-carotid IMT variable. As discussed in 7.3.5, these variables are calculated in different ways. Studying the IMT variable involves looking at the relationship between the SNP and continuous IMT, and is vulnerable to the effects of extreme values, while plaque looks at either presence or absence of discrete plaques defined at a certain threshold. Due to the differences in genotype-IMT associations seen in chapter 4, analyses were also carried out in a sex-stratified manner.

It should be noted that in chapter 3, none of the variants in strong LD with the lead SNP were found to be in coding regions. Nevertheless, it is possible that there is a variant causing an effect that was not in LD with a variant on the Metabochip, especially considering the sparse coverage of this region. It is also possible that such a variant is in fact in strong LD, but was not present in the 1000 Genomes panel with which LD values were calculated.

Therefore the region was scanned for coding variants, which could then be tested for LD with the known IMT signal. It was also considered that there was the possibility of other signals being present, which could require larger GWAS studies for detection. The analysis was therefore carried out to investigate the possible presence of any common coding variants associated with the phenotypes of interest.

Exome Variant Server was here used, a browser using data from the NHLBI GO Exome Sequencing Project (ESP) to show rare and common variants in all coding regions in the genome. It was used to identify any protein-coding variants in the nine genes at the locus, and further association analyses carried out to investigate the results.

7.2 Results

7.2.1 Identification of a SNP in *BCAR1*

The nine genes within 200kb of the lead SNP were examined here, as in chapter 3 and in Gertow and colleagues' original study[119]. Data for all variants was filtered to leave only those causing a nonsynonymous change or located in a splice site. These were then filtered to remove all variants with a MAF under 5%: as the lead SNP is very common (MAF 48% in Europeans), a causal SNP in strong LD would also be expected to be common. If a SNP is not in strong LD with the lead SNP but were to be tested for its effect on phenotypes, a common SNP would be required to obtain sufficient numbers of individuals to perform the analysis.

After filtering, one SNP remained: rs1035539 in *BCAR1*. The SNP has a MAF of 32% in Europeans, and causes a serine/proline amino acid change at position 76 of the BCAR1 protein. This amino acid change is present in the proline-rich domain of the protein (Figure 79). As described in chapter 1.4, BCAR1 is an adapter protein with roles in cell migration and movement. Phosphorylation of its substrate domain is critical for many of its roles in signalling cascades, but less is known about the proline-rich region containing the amino acid change coded by rs1035539.

In Europeans, the major allele of the SNP is A, encoding the serine residue at this position. However, sequence alignment with various other species shows the residue at this location to be proline, encoded by the G allele (Figure 80). Moreover, in African populations, G is the major allele (Figure 81). This suggests that G is the ancestral allele, with A arising in humans and spreading further in groups that had already left Africa. The G allele and corresponding proline residue will thus be referred to here as the wild-type.

The structure of the proline and serine amino acids are shown in Figure 82 and Figure 83. Proline is a non-polar aliphatic amino acid whose alpha-amino group is attached directly to the side chain, forming a nitrogen-containing ring. This cyclic structure means that it does not fit in many of the main chain conformations adopted by other amino acids. It often forms tight turns in proteins, and introduces kinks when found in alpha helices. It has roles in molecular recognition: for example, the small protein domains WW (a domain containing two conserved tryptophan residues) and SH3 (SRC homology 3), found in many different proteins, recognise and bind to proline-containing peptides as part of signalling cascades[322]. However, it is otherwise non-reactive and rarely found in binding sites[323]. Serine is a small polar uncharged amino acid. Like proline, it is also found within tight turns

in the protein, but is more reactive, being able to form hydrogen bonds with many polar substrates. Serine is often phosphorylated by serine/threonine kinases[323].

Figure 79: Protein domain structure of BCAR1. The proline to serine amino acid change coded by rs1035539 is shown in pink in the proline-rich domain of the protein.

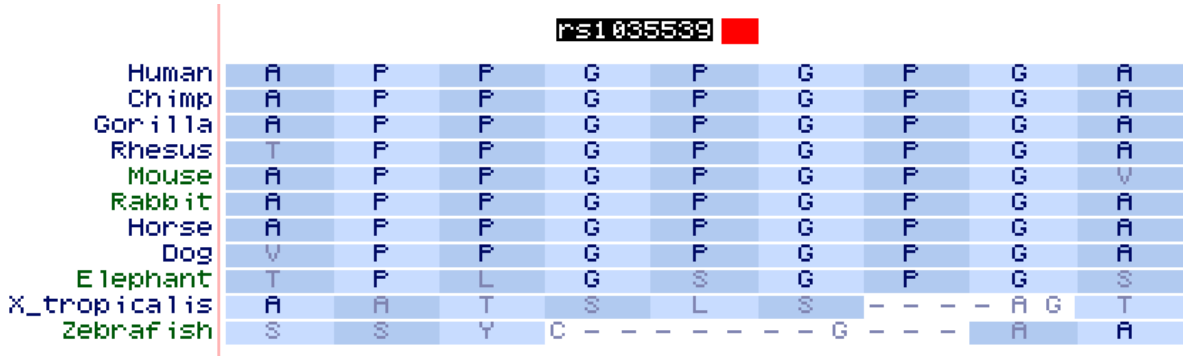

Figure 80: Amino acid sequence alignment at rs1035539 location. In most species (especially those related closely to humans), proline is present at this location, coded for by the A allele of rs1035539 (minor allele in Europeans). Figure adapted from UCSC Genome Browser[127].

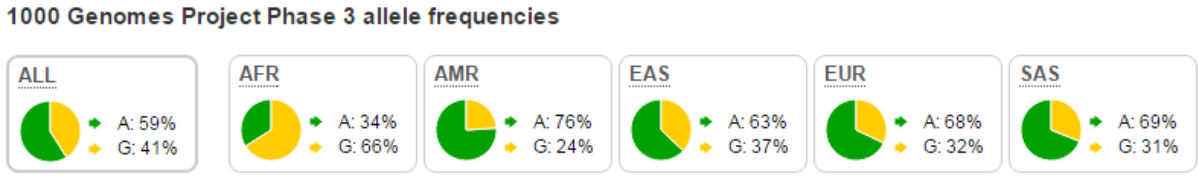

Figure 81: Allele frequencies of rs1035539 in populations of 1000 Genomes Phase 3. Populations are African (AFR), American (AMR), East Asian (EAS), European (EUR) and South Asian (SAS). The major allele is G in the African population, but A in each other population. Figure from Ensembl (release 83)[300].

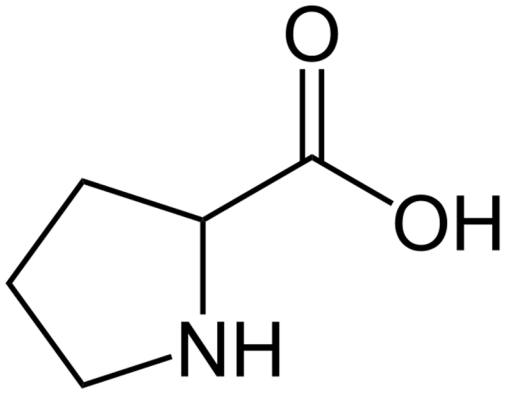

Figure 82: Structure of proline (P). Proline is a non-polar aliphatic amino acid. It is the only amino acid with a secondary amine, and the only amino acid with a cyclic structure, in which the alpha-amino group is attached directly to the side chain.

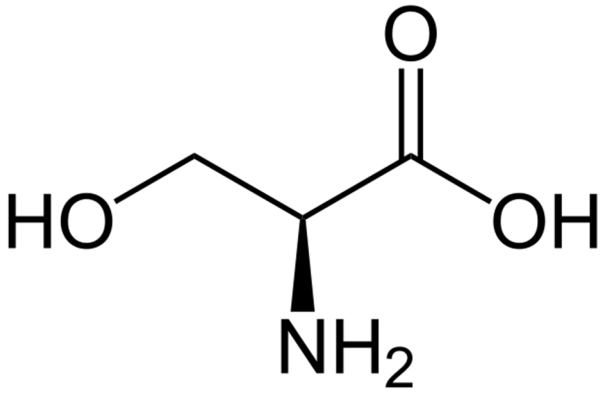

Figure 83: Structure of serine (S). Serine is a small polar uncharged amino acid. Its hydroxyl group can form hydrogen bonds with polar substrates.

7.2.2 Cohort characteristics

The characteristics of the IMPROVE and PLIC cohorts used for genotyping in this chapter are described in chapter 4.2.

7.2.3 Genotyping of rs1035539 in IMPROVE

The SNP rs1035539 was not present in the 1000 Genomes Pilot 1 panel, and therefore could not be used for LD calculations with Metabochip SNPs using online tools at the time of analysis. The SNP was therefore genotyped in IMPROVE to calculate LD with the lead SNP and assess phenotypic associations.

3232 samples in the IMPROVE cohort were genotyped using an LGC KASP genotyping assay. The call rate was 89.2%. Despite several attempts to optimise the assay, AA and AG samples could not be made to cluster with complete separation on the fluorescence plot, meaning certain samples could not be called with confidence (Figure 84). The minor allele frequency in IMPROVE was 30.5%, and the genotype frequencies were not out of Hardy-Weinberg equilibrium ($\chi^2 = 0.042$, $p = 0.84$).

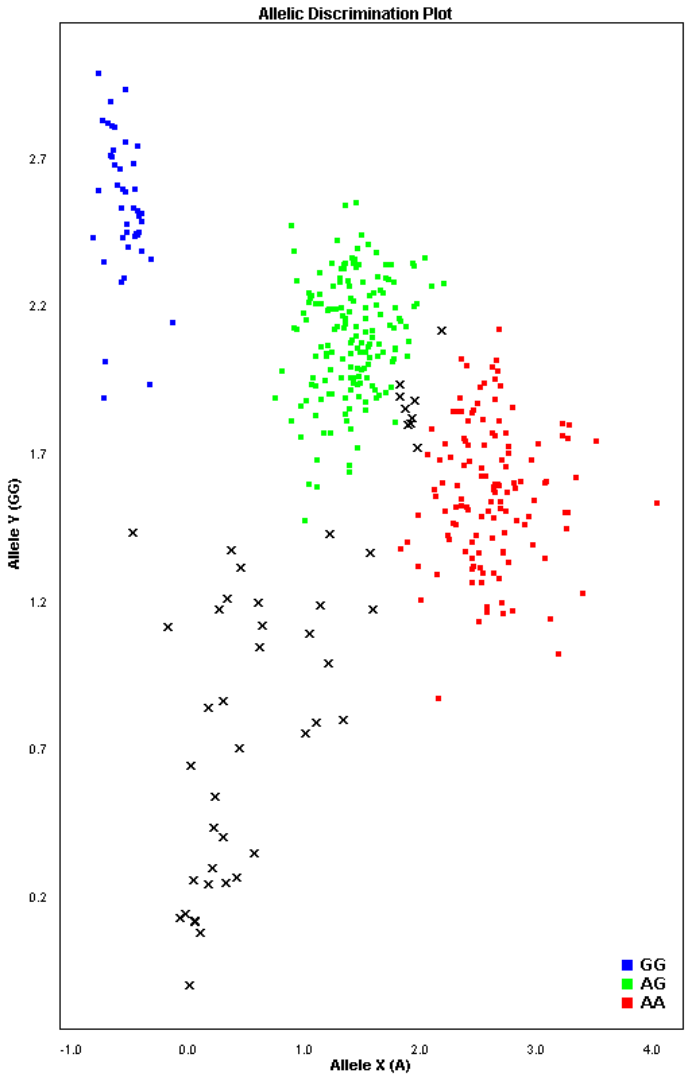

Figure 84: Example of KASP allelic discrimination amplification plot. Amplification of fluorescent labelled oligos allows separation of samples in to genotype clusters. Imperfect amplification in this assay caused incomplete clustering of genotypes, so some samples could not be called (represented by black crosses).

7.2.3.1 LD analysis of rs1035539

Pairwise LD between rs1035539 and the lead SNP rs4888378 was generated using PLINK. Pairwise r^2 was 0.005, and D' 0.088, indicating no significant LD between the two. It therefore appears that any effect of rs1035539 would not explain the association of the lead SNP with IMT.

With updated 1000 Genomes panels, LD data for rs1035539 later became available. This revealed that two SNPs were in $r^2 > 0.8$ with rs1035539, one intronic in *BCAR1* and one upstream. Any phenotypic associations with rs1035539 could be caused by any one of these SNPs, but the SNP altering protein structure was considered most likely to have an effect.

7.2.3.2 rs1035539 is not associated with IMT

LD analysis revealed that the original association between the lead SNP rs4888378 could not be explained by any functional effect of rs1035539. Association analysis therefore tested the possibility of an independent association signal. In IMPROVE, rs1035539 was not found to be associated with IMT variables, either overall in the carotid tree or in individual segments (Table 33). This was also true under dominant and recessive models. In each case, a non-significant trend was observed for lower IMT with each G allele (Table 33, example in Figure 85).

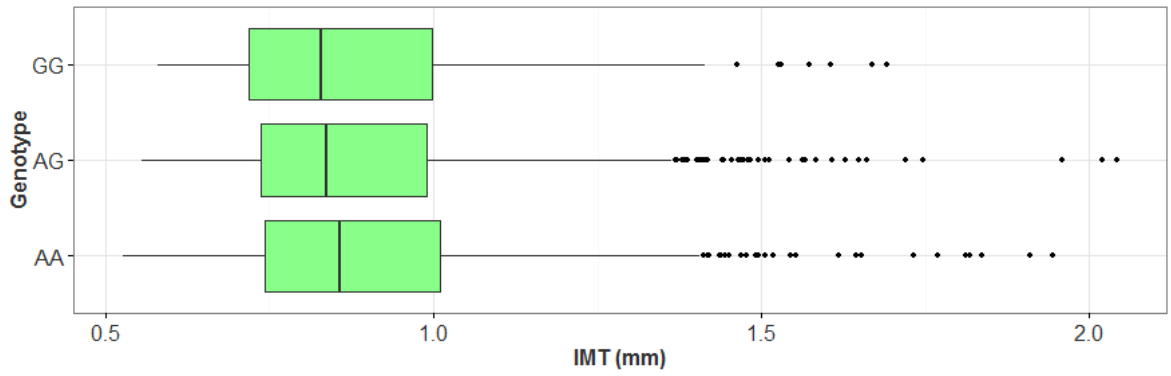

Figure 85: Mean IMT across carotid tree by rs1035539 genotype. As with IMT variables in individual segments, there was no significant association with genotype (β = -0.002, p = 0.39), but a trend was present for lower IMT with each G allele.

Table 33: IMT phenotypes by rs1035539 genotype in IMPROVE. All phenotypes are mean across the segment. Phenotype values by genotype are mean (standard error). P values calculated by linear regression; adjusted for age, sex and MDS coordinates, and log-transformed before analysis.

	Genotype			Association with genotype	
	AA	AG	GG	β value	p
Whole carotid tree	0.895 (0.005)	0.885 (0.006)	0.881 (0.012)	-0.0020	0.39
Common-carotid	0.747 (0.003)	0.744 (0.004)	0.733 (0.008)	-0.0020	0.29
Bifurcation	1.155 (0.010)	1.133 (0.011)	1.129 (0.023)	-0.0031	0.39
Common-carotid (cm closest to bifurcation)	0.803 (0.004)	0.795 (0.004)	0.791 (0.009)	-0.0016	0.44
Internal carotid artery	0.878 (0.009)	0.870 (0.010)	0.876 (0.022)	-0.0002	0.95

7.2.3.3 rs1035539 is associated with presence of plaque

rs1035539 showed a trend towards an association with plaque in IMPROVE, with the minor G allele being associated with lower presence of plaque (logistic regression: OR = 0.89; p=0.054, adjusted for age, sex and MDS coordinates). Graphical presentation of this effect suggested a possible dominant effect of the G allele (Figure 86), so this model was tested, indicating a significant association with

the G allele (OR = 0.81, p = 0.036). However, it was noted that this sub-hypothesis carried an increased chance of type I error and the initial hypothesis of an additive relationship was not confirmed. No significant effect was seen under a recessive model (Table 34).

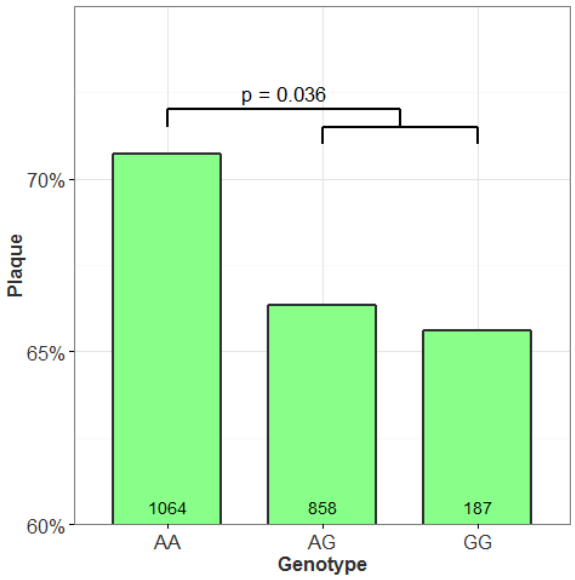

Figure 86: Prevalence of carotid plaque by rs1035539 genotype in IMPROVE. Presence of plaque is lower when a G allele is present (p = 0.036). Numbers with plaque shown on columns.

Table 34: rs1035539 is associated with presence of carotid plaque in IMPROVE. Results from logistic regression model.

	Plaque (adjusted; age, sex and MDS)	
Genetic model	OR (95% CI)	p
Additive	0.86 (0.79-1.00)	0.054
Dominant	0.81 (0.72-0.99)	0.036*
Recessive	0.87 (0.69-1.19)	0.463

7.2.3.4 rs1035539 is not associated with disease traits and CVD risk factors

rs1035539 was not found to be significantly associated with combined vascular events, nor with cardiac, cerebro-vascular or peripheral events in IMPROVE. It was also not associated with total cholesterol, LDL-cholesterol, HDL-cholesterol or triglycerides (Table 35).

It is likely that there was insufficient power to detect an effect of genotype on events. Power calculations show that under the study conditions, 80% power could be achieved to detect a genotype relative risk of 1.35 or greater on combined vascular events. The association with plaque gives a relative risk of 1.08 per A allele on plaque; a risk ratio of 1.22 of plaque on events gives an

expected relative risk of 1.32 of genotype on events, below the calculated threshold. Fewer cases for the individual event variables means there is less power to detect associations with these variables.

Table 35: rs1035539 is not associated with vascular events or lipid levels in IMPROVE. Results from logistic regression model, adjusted for age, sex and MDS coordinates.

Vascular event	OR (95% CI)	p
All combined	1.13 (0.90, 1.41)	0.31
Cardiac	1.26 (0.95, 1.67)	0.11
Cerebro-vascular	0.90 (0.61, 1.32)	0.58
Peripheral	1.02 (0.53, 1.94)	0.97
Lipid risk factors	Beta (95% CI)	p
Total cholesterol	0.008 (-0.049, 0.07)	0.78
LDL	-0.014 (-0.07, 0.04)	0.60
HDL	0.016 (-0.002, 0.03)	0.09
Triglycerides	-0.008 (-0.07, 0.06)	0.82

7.2.3.5 Sex-stratified: rs1035539's association with plaque is only present in men

In light of the sex differences seen in chapter 4, association analyses were also split by sex. Stratifying association into men and women revealed that the significant association of the G allele with plaque, under a dominant model, was present only in men. Again, plaque was lower when the G allele was present (logistic regression: OR = 0.77; p=0.039, adjusted for age and MDS coordinates). A trend in the same direction appeared present in women but there was no statistically significant effect (Figure 87,Table 36).

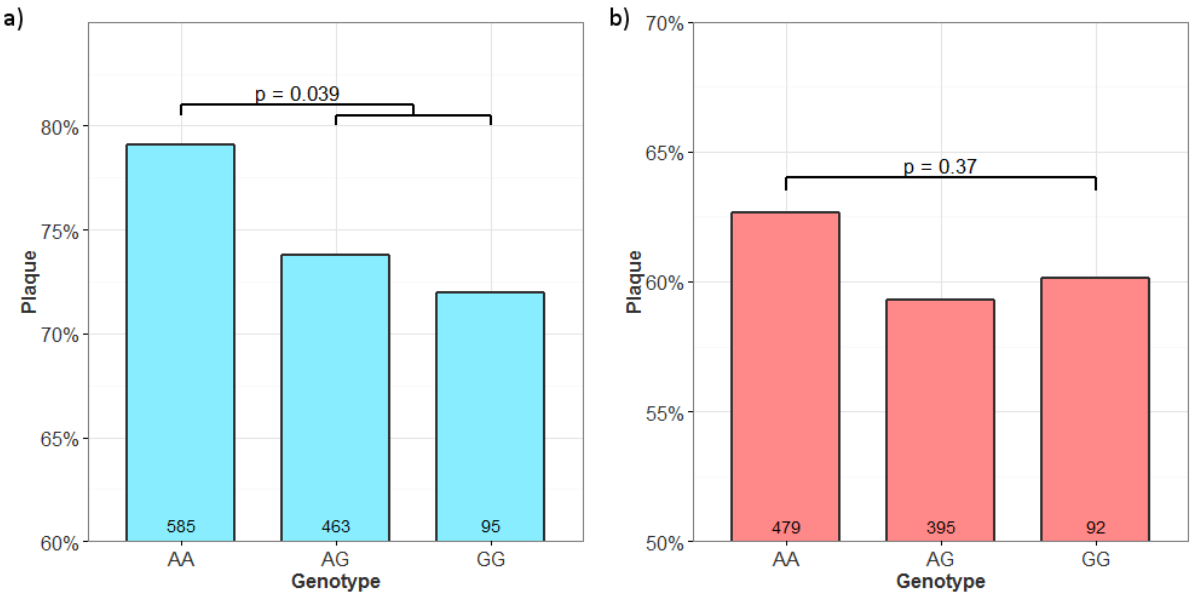

Figure 87: Prevalence of carotid plaque by rs1035539 genotype in men and women in IMPROVE. (a) In men, prevalence of plaque is lower with the G allele (p = 0.039). **(b)** No significant relationship is seen in women (p = 0.37). Numbers with plaque shown on columns.

Table 36: rs1035539 is associated with presence of carotid plaque in men in IMPROVE. Results from logistic regression model.

	Men (adjusted for age and MDS)		**Women (adjusted for age and MDS)**	
G-allele model	OR (95% CI)	p	OR (95% CI)	p
Additive	0.83 (0.69-1)	0.054	0.93 (0.79-1.09)	0.37
Dominant	0.77 (0.6-0.99)	0.039*	0.9 (0.73-1.11)	0.32
Recessive	0.85 (0.57-1.3)	0.44	0.95 (0.67-1.35)	0.77

7.2.3.6 Sex-stratified: rs1035539 associated with IMT?

As with non-stratified IMT variables, significant associations were largely not present between rs1035539 genotype and IMT variables in men and women. The same trend for lower IMT with the G allele appeared present, and to be stronger in men (e.g. Figure 88). One variable, IMT at the bifurcation, showed significance for genotype-IMT association in the expected direction in men (β = -0.0118, p = 0.028), although this did not remain significant after correction for multiple testing.

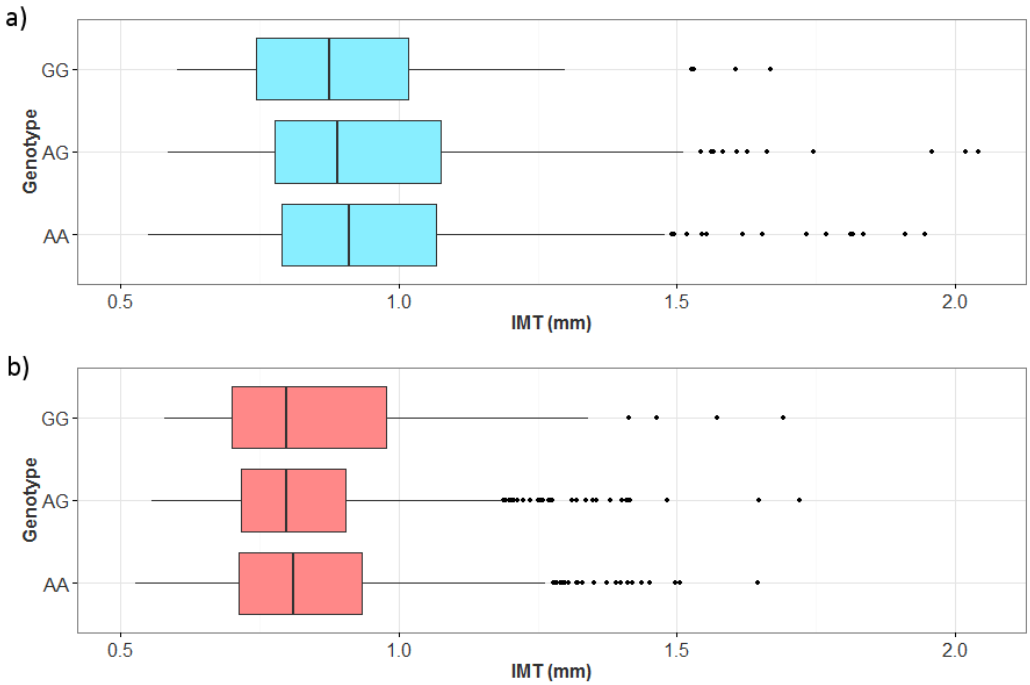

Figure 88: Mean IMT across carotid tree by rs1035539 genotype in (a) men and (b) women. There was no significant association with genotype, but IMT values again appeared lower with each G allele, an effect more pronounced in men.

Table 37: IMT phenotypes by rs1035539 genotype in IMPROVE, stratified by sex. All phenotypes are mean across the segment. Phenotype values by genotype are mean (standard error). P values calculated by linear regression; adjusted for age and MDS coordinates, and log-transformed before analysis.
*Denotes statistical significance.

	Men					Women				
	Genotype			Association with genotype		Genotype			Association with genotype	
	AA	AG	GG	β value	p	AA	AG	GG	β value	p
Whole carotid tree	0.950 (0.008)	0.940 (0.009)	0.903 (0.018)	-0.0054	0.12	0.844 (0.006)	0.832 (0.006)	0.864 (0.017)	0.0011	0.70
Common-carotid	0.770 (0.005)	0.778 (0.007)	0.753 (0.015)	0.0002	0.94	0.725 (0.004)	0.712 (0.004)	0.716 (0.009)	-0.0036	0.12
Bifurcation	1.247 (0.016)	1.205 (0.017)	1.146 (0.033)	-0.0118	0.028*	1.069 (0.013)	1.063 (0.014)	1.113 (0.033)	0.0046	0.35
Common-carotid (cm closest to bifurcation)	0.826 (0.007)	0.827 (0.007)	0.806 (0.014)	0.0001	0.99	0.780 (0.005)	0.764 (0.005)	0.779 (0.012)	-0.0028	0.29
Internal carotid artery	0.957 (0.014)	0.951 (0.016)	0.905 (0.031)	-0.0057	0.32	0.803 (0.011)	0.793 (0.011)	0.851 (0.031)	0.0044	0.39

7.2.4 Genotyping of rs1035539 in PLIC

The relationship between rs1035539 and carotid plaque suggested this SNP, or a SNP in LD, may also be having an effect on atherosclerosis in the carotid artery. This prompted additional study of the SNP in a general population, rather than high-risk, cohort. It was genotyped in the PLIC cohort in order to assess whether the same relationship was seen.

7.2.4.1 PLIC genotyping results

After multiple attempts to genotype PLIC with different assays (see 7.3.2), rs1035539 was genotyped by KBioscience using their KASP allelic discrimination assay. Genotyping was completed with a call rate of 82.0%. The minor allele frequency was 33.4% and the genotype frequencies were not out of Hardy-Weinberg equilibrium ($\chi^2 = 2.03$, $p = 0.15$).

7.2.4.2 rs1035539 is nominally associated with CC-IMT

The relationship between rs1035539 and IMT was also examined in PLIC. The variable available in the cohort was CC-IMT. IMT values here appeared higher in heterozygotes than in those of AA genotype, although was lowest in the GG group (Figure 89), as seen with plaque. Overall, the G allele was associated with lower IMT ($\beta = -0.003$; $p = 0.038$, adjusted for age and sex).

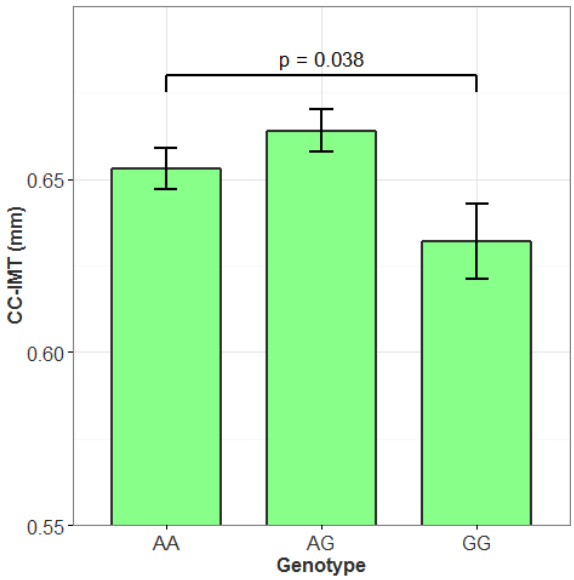

Figure 89: Baseline CC-IMT by rs1035539 genotype in PLIC. IMT in those of GG genotype is lower than those with AA or AG genotype (p = 0.038).

7.2.4.3 rs1035539 is not associated with carotid plaque

Unlike in IMPROVE, rs1035539 was not associated with presence of carotid plaque in PLIC ($\chi^2 = 2.31$; $p = 0.32$). However, the same trend was seen as in IMPROVE: prevalence of plaque was lower with each G allele (although here with an apparent recessive effect of the G allele, which was not significant: p =

0.13) (Figure 90). The percentage of subjects with plaque was much lower than that seen in IMPROVE, which was expected as the PLIC subjects are not high risk.

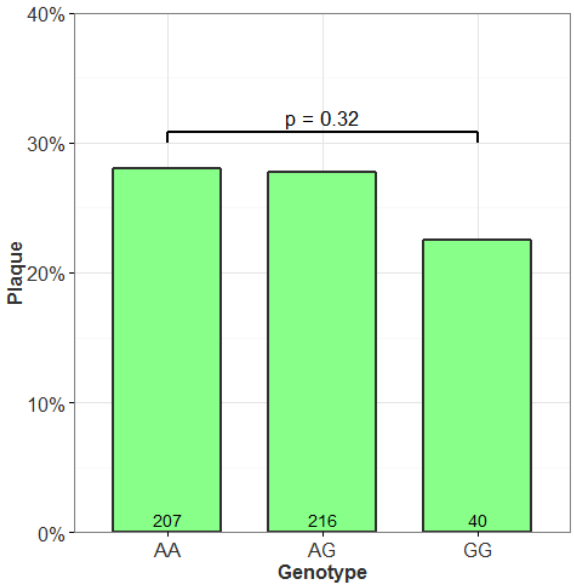

Figure 90: Prevalence of carotid plaque by rs1035539 genotype in PLIC. No significant association between genotype and plaque was observed, but the same trend was seen as in IMPROVE, with a lower percentage plaque with the GG genotype.

7.2.4.4 rs1035539 not associated with cardiometabolic parameters

The relationship between rs105339 and numerous cardiometabolic parameters was tested in PLIC. As in IMPROVE, no significant association was found between these and the SNP (Table 38).

Table 38: rs1035539 is not associated with cardiometabolic factors in PLIC. Results from linear or logistic regression, adjusted for age and gender. Values by genotype are mean (standard deviation).

	AA	AG	GG	p
Systolic blood pressure (mmHg)	131.02 (0.59)	132.26 (0.58)	131.22 (1.28)	0.31
BMI (kg/cm2)	26.41 (0.15)	26.67 (0.15)	26.53 (0.31)	0.44
Waist/hip	0.87 (0.002)	0.869 (0.002)	0.87 (0.005)	0.92
Total cholesterol (mmol/l)	5.80 (0.04)	5.70 (0.04)	5.73 (0.07)	0.18
HDL-C (mmol/l)	1.43 (0.01)	1.43 (0.01)	1.46 (0.03)	0.65
Triglycerides levels (mmol/l)	1.25 (0.03)	1.23 (0.02)	1.15 (0.05)	0.22
LDL-C (mmol/l)	3.79 (0.03)	3.71 (0.03)	3.74 (0.07)	0.17
Glucose levels (mmol/l)	5.22 (0.03)	5.18 (0.03)	5.1 (0.07)	0.23
apoB (mg/dL)	115.1 (1.0)	112.9 (0.9)	111.7 (2.0)	0.13
apoA-I (mg/dL)	149.8 (0.9)	148.4 (0.9)	151.2 (1.9)	0.34
Remnant-C (mmol/l)	0.57 (0.01)	0.57 (0.01)	0.53 (0.02)	0.22
Diabetes (n, %)	21 (3.8%)	23 (3.8%)	5 (3.6%)	0.74
Dyslipidaemia (n, yes)	318 (70%)	332 (67%)	80 (70%)	0.12
Hypertension (n, yes)	133 (29%)	167 (34%)	32 (28%)	0.26

7.2.4.5 Sex-stratified: rs1035539 not associated with IMT

Stratifying PLIC by sex did not change the appearance of the relationship between rs1035539 genotype and IMT (Figure 91). IMT appeared highest in heterozygotes, although lowest for GG genotypes. However, no significant associations were detected between genotype and IMT (β = 0.001; p = 0.36 and β = -0.002; p = 0.09 for men and women respectively).

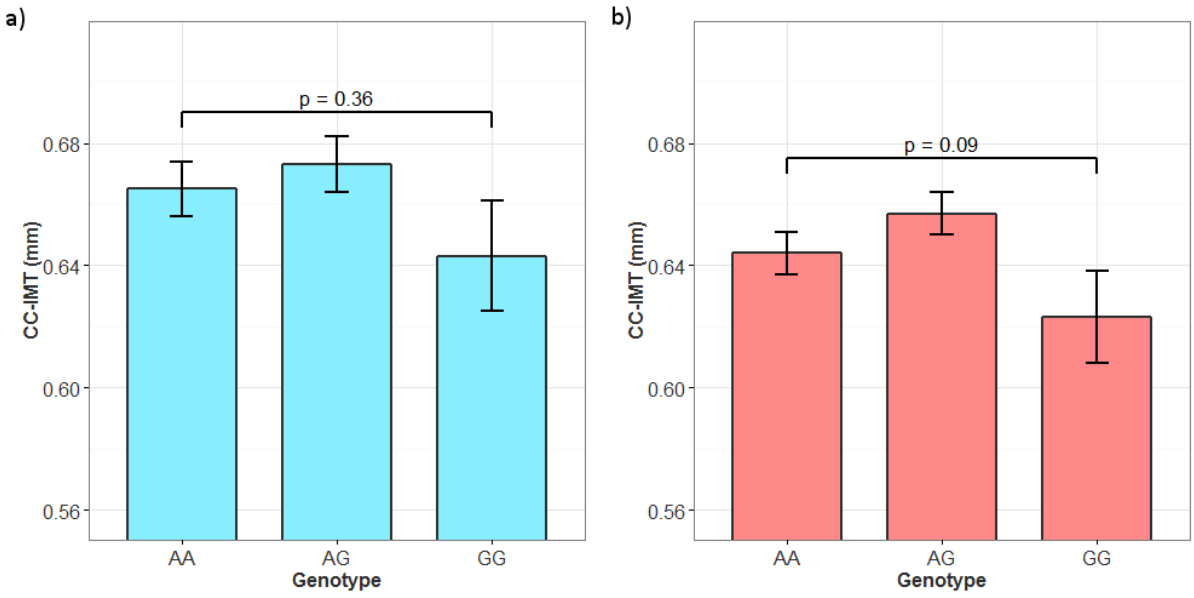

Figure 91: CC-IMT by rs1035539 genotype in (a) men and (b) women. Observed IMT by genotype appears similar between groups, although no significant association is present.

7.2.4.6 Sex-stratified: rs1035539 associated with plaque in men under a recessive model

Stratifying the PLIC cohort by sex revealed differences in genotypic association between the sexes, with an association again seen in men, but only under certain genetic models (Figure 92). No significant association was seen in women (χ^2 = 0.02, p = 0.99). In men, the G allele was associated with lower plaque under a recessive model (χ^2 = 4.51; p = 0.034), although the overall trend was not significant (p = 0.096).

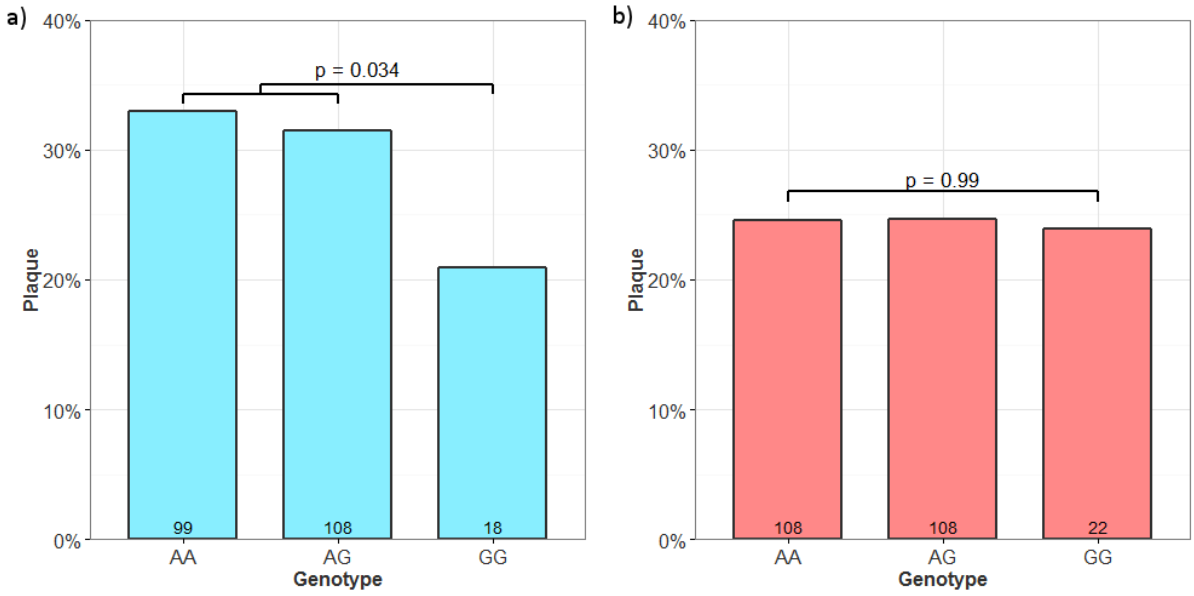

Figure 92: Prevalence of carotid plaque by rs1035539 genotype in (a) men and (b) women in PLIC. In men, prevalence of plaque is lower for GG genotypes than other genotypes (p = 0.03). No such association is seen in women (p = 0.99).

7.3 Discussion

While earlier chapters explored the possibility for regulatory variants at the *CFDP1-BCAR1-TMEM170A* locus to affect IMT, this chapter focused on the possibility of functional causal exonic variants at the locus. Investigation of exome sequencing data identified one common variant at the locus that altered a protein sequence, present in *BCAR1*, the gene that had previously been implicated in the IMT phenotype in chapters 3 and 4.

Genotyping of the SNP in IMPROVE did not show an association with IMT, but it did show an association with carotid plaque under a dominant model of the G allele. This finding should be treated with caution, bearing in mind that carrying out more tests under different models increases the chance of type I error. Nevertheless, the association is interesting considering that the phenotype is similar to that tested in Gertow et al's original study. This result may give more confidence to the theory that *BCAR1* is the causal gene involved in IMT. This effect appeared to be present only in men, in contrast to

the results seen with rs4888378 in chapter 4. Genotyping in the PLIC cohort supported the finding of the SNP's effect on plaque and IMT, although sex-specific effects could not be verified.

7.3.1 Identification of SNP rs1035539

Non-synonymous SNPs are present less frequently per base than synonymous or non-coding SNPs, because their increased likelihood of affecting the phenotype results in greater selection pressure[319,324]. Therefore, few missense variants at the *CFDP1-BCAR1-TMEM170A* locus were expected to be found using Exome Variant Server (EVS), and those that were found would be more likely to have a phenotypic effect than other types, making them good candidates for study.

SNPs conferring a large advantageous or disadvantageous effect are not present as frequently as those with smaller effects, as selection is here likely to drive the SNP towards homozygosity in the population[319,321]. Many rare missense and splice base changes were found in the EVS data, some of which may have phenotypic effects, but for rare variants there are few individuals with rare genotype, and thus they do not lend themselves to association testing. Future work may benefit from examining the effect of these rare variants on phenotype, but increased power to detect rare variants would be required. Using larger cohorts would be one way to increase power in this way; a different approach would be the use of a burden test, in which information about multiple genetic variants in a functional unit (such as a gene or locus) is collapsed into a single gene score. This increases the risk of detecting a true effect, but may take the assumption that all rare variants in a set are causal[325,326]. Alternatively, burden tests can be carried out using only those that are predicted to be damaging using bioinformatic tools, or by weighting variants according to their predicted functional effect[327].

Therefore, results were filtered only to include common SNPs disrupting protein function, leaving only rs1035539 in *BCAR1*. This variant seemed particularly interesting for further study, as *BCAR1* had been implicated as the gene potentially causing the IMT effect at the locus. It encodes a proline to serine change in a proline-rich domain of the protein (discussed in 7.3.6), and thus was of increased interest as a candidate for further analysis.

Later updates of EVS data indicated two new SNPs with MAF > 5% in Europeans: rs4737 in *CTRB2* (coding for alanine to threonine) and rs117435647 in *CHST6* (arginine to glycine). Neither were in LD with the lead SNP rs4888378. Using the online Protein Variation Effect Analyser (PROVEAN)[328] to predict the functional consequences of these SNPs showed the changes to both be predicted as neutral. Further work would examine these SNPs to assess their potential effects.

7.3.2 Genotyping of rs1035539 in PLIC and IMPROVE

A limitation of the analysis of this SNP was the unsatisfactory KASP genotyping assay for rs1035539, which genotyped with call rates of only 89% in IMPROVE and 82% in PLIC. This reduced the sample size and thus power for association analysis, and introduced potential bias where samples could have been incorrectly called (AA and AG genotypes in particular clustered closely together). However, tests for Hardy-Weinberg equilibrium did not indicate an excess or scarcity of heterozygotes. The difference between the cohorts may have been a result of differing DNA quality of samples, supported by the fact that call rate differed markedly in PLIC between individual DNA plates.

Other genotyping methods were tried in order to improve the genotyping rate. A TaqMan allelic discrimination assay was tested, but failed in development, possibly due to the presence of a rare variant 2 bp away (rs143543666).

After 1000 Genomes Phase I data became available, two SNPs were found to be in strong LD with rs1035539. One SNP, rs11645191 (intronic in *BCAR1*), had a suitably high r^2 value of 0.95 to use as a proxy. TaqMan and subsequently KASP assays were therefore also ordered for this SNP, but neither produced genotype clusters. The assay was therefore ultimately carried out by LGC with their KASP rs1035539 assay, producing the 82% call rate.

7.3.3 *BCAR1* retrogene

The failure of multiple assays around the rs1035539 SNP prompted investigation of this genomic region to understand the obstacles to genotyping variants here. Using the BLAST-Like Alignment Tool (BLAT)[329] to align the locus to matching sequences in the genome revealed two areas of high similarity on chromosome 15. These are part of two retroposed genes (retrogenes): repetitive DNA fragments inserted into chromosomes after being reverse transcribed from RNA molecules. The fragments are non-autonomous, unlike retrotransposons, as they do not encode reverse transcriptase, and do not have protein-coding ability. They are therefore a subset of pseudogenes (DNA sequences similar to genes without any function).

Aligning the sequence of and surrounding the two retrogenes to the *BCAR1* locus showed the area of homology to cover the exon containing rs1035539, and two intronic regions, including the proxy rs11645191 (Figure 93). At the relative position of the two SNPs, the corresponding bases in the retrogenes are monoallelic.

The presence of these homologous regions means that the probes used in KASP and TaqMan can misprime to these regions, amplifying the wrong sequence and resulting in a failure to produce

accurate genotyping clusters, explaining the poor genotyping results seen in this chapter. The presence of the retrogenes also has implications for sequencing of the region and *BCAR1* gene: PCR or sequencing primers would need to avoid the repeated regions to bind to the region with specificity.

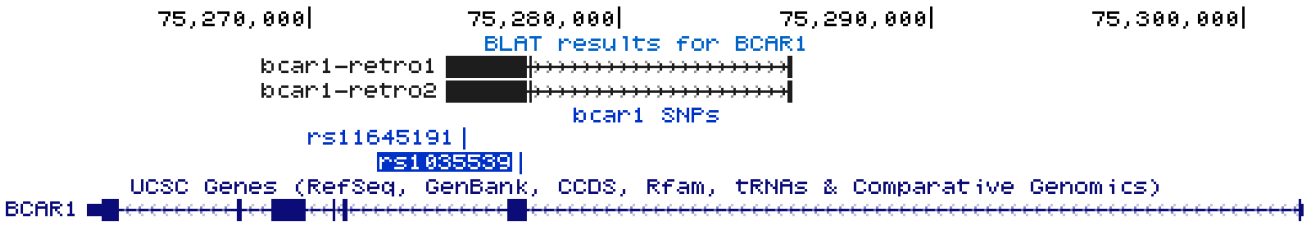

Figure 93: Alignment of the two *BCAR1* retrogenes over the *BCAR1* gene. Regions of homology encompass the exon containing rs1035539 and some intronic sequence. Image from UCSC Genome Browser[127].

Retrogenes can contain information about the evolutionary history of a sequence[330]. In the two retro-*BCAR1* genes, the base equivalent to rs1035339 is G, the 'minor' allele. This substantiates the theory that the minor allele is the ancestral allele (7.2.1).

7.3.4 LD analysis of rs1035539

Genotyping of the *BCAR1* coding SNP in IMPROVE revealed it not to be in LD with the lead SNP rs4888378. This was the expected outcome, as there was no previous evidence that the SNPs were in LD. However, in light of this lack of LD, the subsequent discovery of the coding SNP's association with carotid plaque was more noteworthy. As this association is genetically independent of that seen with the lead SNP, at least two distinct areas of functional variation at the locus are associated with carotid IMT phenotypes.

7.3.5 rs1035539 shows associations with plaque and IMT

After the lead and coding SNPs were found not to be in LD, there was no *a priori* expectation that rs1035539 would be associated with IMT-related phenotypes. However, an association was found between rs1035539 and carotid plaque under a dominant model in IMPROVE. If this association can be validated, it indicates the presence of a separate association signal at the locus.

This raises the issue of the functional variation causing the change in plaque. Unlike the regulatory lead SNP, rs1035539 is in strong LD with only two SNPs, one intronic in *BCAR1* and one in an intergenic region. Variants in weaker LD were also intronic or intergenic. As a missense change that alters protein structure is more likely to change the phenotype[319], it seemed probable that rs1035339 was the functional variant. This was later investigated in chapter 8 by assaying the effect of the amino acid substitution on protein function.

If this amino acid substitution in the *BCAR1* protein is indeed affecting carotid plaque, the similarity in phenotypes suggests the earlier regulatory variation captured by rs4888378 may be acting on the same gene, through regulation of expression rather than changing the protein. This reinforces the previous evidence for *BCAR1* as the functional gene seen in chapters 3 and 4.

While rs1035539 showed an associated with presence of plaque in IMPROVE under a dominant model, no association with IMT was found, although a trend in the same direction for that of plaque was seen in each segment of the carotid tree. Conversely, replication in PLIC showed the SNP to be associated nominally with IMT but not plaque. These findings should be taken with caution as they are the results of multiple tests on differing phenotypes.

These discrepancies in results may be a result partly of power. PLIC contains fewer subjects overall, a lower percentage of subjects with plaque (as it is a lower-risk cohort), and a lower percentage of successfully genotyped samples. This reduces power to detect an association with plaque. For example, there appears to be a lower prevalence of plaque in those of GG genotype, but with fewer individuals in this group, this was not statistically significant (Figure 90). The poorer genotyping rate of rs1035539 in PLIC makes results in this cohort less robust.

Power calculations were subsequently carried out, taking into account the number of subjects that had been successfully genotyped and phenotyped. These revealed that to detect the effect size seen in IMPROVE with plaque under the dominant model of the G allele (under which the significant association was seen), IMPROVE is powered to 94.7%, above the typical desired threshold of 80%. However, to detect the same effect size, PLIC is only powered to 20.9%. An improved genotyping assay and larger number of subjects would be required to replicate the associations with plaque.

Apart from sample size and power, other differences between the cohorts may have contributed to the different results. In IMPROVE, presence of plaque was defined as maximum IMT of greater than 1.5 mm across the whole carotid tree[202], whereas in PLIC, it was defined as IMT greater than 1.3 mm and/or presence of focal plaque as defined by longitudinal projection caliper thickness greater than 1.3 mm[203]. The cut-off for the definition of plaque may affect its value as a variable.

As plaque was a variable calculated from max IMT, it was unexpected that the SNP was associated with plaque in IMPROVE but not the variable upon which it is calculated, max IMT across the carotid tree. This may be due to the IMT variable distributions; all are right-skewed, with the majority of values above the mean. Log-transformed variables were used for analysis, making the regression results

reliable, but the SNP may be better associated with IMT at the lower ranges (around the plaque cut-off) than the higher ranges that also affect linear regression analyses.

7.3.6 Why was an association seen between rs1035539 and plaque?

The association results, if validated, raise the question of how the SNP could be causing the effect on plaque. In contrast to the regulatory SNP analysis in chapters 4 and 5, the nature of rs1035539 (a missense SNP altering protein structure) provides a clearer pathway to investigating the change.

As discussed in chapter 1, BCAR1 is an adaptor protein with roles in cell migration, adhesion and proliferation. As a scaffolding molecule, it binds numerous other proteins, many at its substrate-binding and SH3 domains[141]. The proline-rich domain, in which rs1035539 codes a proline to serine change, may also be of importance. Such proline-rich protein regions are common and often play a role in binding, often in complexes of multiple proteins[331] (such as those observed with the BCAR1 protein). These generally produce multiple weak binding sites, allowing rapid modulation of binding[332]. The presence of an amino acid change in the proline-rich domain might therefore affect the binding action of the protein, particularly if the original amino acid is proline.

The serine residue is associated with greater prevalence of plaque. It is possible that this residue disrupts the proline-rich region such that protein assembly is disrupted, affecting pathways downstream. As BCAR1 is involved in PDGF stimulation of VSMC migration[187], the amino acid change may alter this migration in such a way that formation of plaque is increased.

The SNP was only significantly associated with plaque in IMPROVE under a dominant model of the G allele, with the G allele conferring lower risk of plaque (Figure 86). This suggests one copy of the wild-type protein is sufficient to largely restore the “healthy” or low-risk function of the gene.

7.3.7 rs1035539 not associated with cardiometabolic parameters or disease

Despite its association with plaque and IMT, rs1035539 was not associated with vascular events. Considering the modest decrease in plaque odds (19% decrease with the G allele), detecting an association with event rate through the gene would not necessarily be expected, even if it were present.

rs1035539 was also not associated with cardiometabolic risk factors. If the effect on plaque was caused through an effect on a risk factor such as LDL cholesterol, we would expect to detect it. However, the lead SNP at the locus, rs4888378, was earlier associated with IMT independent of such risk factors (chapter 4.2.5). If the coding SNP is having an effect through the same gene, it may also be through a

pathway independent of these variables, such as the effect on cell remodelling in migration suggested in 7.3.6.

7.3.8 Sex-specific association with plaque

In chapter 4 it was seen that the *CFDP1-BCAR1-TMEM170A* locus's lead SNP rs4888378 was associated with IMT and vascular events only in women. In light of these differences produced by stratifying by sex, the relationships of rs1035539 with IMT and plaque were also stratified by sex. Association with plaque in IMPROVE and PLIC was here present in men but not women when the cohorts were stratified, but this association was true for men only under different genetic models in each cohort, so requires validation.

The associations with plaque were seen only under dominant or recessive models of the G allele, despite a trend for lower plaque with the G allele being present in each case. It may be the case that greater numbers of subjects are needed to detect associations with more confidence, and therefore determine whether associations are truly sex-specific, and whether the association is under a dominant or recessive model.

If it is indeed the case that rs1035539 is associated with plaque in men but not women, the relationship is surprising considering the association between rs48888378 and IMT was present only in women. It is possible that the SNPs affect different genes: the coding SNP rs1035539 on *BCAR1* and the regulatory rs4888378 on a different gene at the locus. Effects on the phenotype would therefore not be expected to be similar. However, the eQTL data implicated *BCAR1* in the association with rs4888378, and the fact that both SNPs are involved with very similar phenotypes (both of which plausibly involving the BCAR1 protein) means it is most likely that each involve this gene.

The difference may be due to the nature of the SNPs: a variant altering a protein and a variant altering expression might cause different downstream effects. For example, the presence of a harmful protein may have a different effect to a smaller amount of the wild-type, and these problems may be compensated by different systems. Further work will examine the consequences of changing the BCAR1 protein, and may provide some explanation as to the sex differences in association.

7.3.9 Conclusions and further work

In this chapter, the *CFDP1-BCAR1-TMEM170A* locus was examined with the aim of evaluating effects of coding variation on IMT phenotypes. One common coding SNP was present at the locus, altering an amino acid in the BCAR1 protein, and this was found to be associated with carotid plaque in two cohorts.

These findings further link *BCAR1* with the IMT phenotypes shown to be associated with the locus. Further work will look directly at the effect of the coding SNP's amino acid change on the protein and cell function.

8 Protein studies of BCAR1

8.1 Introduction

The previous chapter reported the investigation of nonsynonymous coding variants that might affect carotid IMT at the *CFDP1-BCAR1-TMEM170A* locus. One common SNP, rs1035539, was found in the gene *BCAR1* and genotyping revealed it to be associated with plaque and IMT in two cohorts.

In this chapter, the reason for this association was investigated by assaying the SNP's effect on BCAR1 and other proteins, and cell function, with the aim of further understanding the role of BCAR1 in atherosclerosis and intima-medial thickening. Finding out how changing BCAR1 disrupts cell signalling pathways or cell function may help to explain how genetic variants at the locus, particularly the coding SNP rs1035539, lead to changes in IMT.

As discussed in chapter 1.4, BCAR1, also known as p130cas, is an adaptor protein that acts as a docking protein for a number of proteins (Figure 12). Its involvement in cell adhesion, migration and remodelling make it a potentially important protein for initiation and growth of atherosclerotic plaque. Tyrosine phosphorylation of BCAR1 is key to its activation of downstream processes, and takes place primarily at 15 YxxP (Tyrosine-x-x-Proline) repeats within the substrate domain (Figure 79).

The SNP rs1035539 codes a proline to serine change in the proline-rich domain of the protein (Figure 79). In contrast to other domains such as the SH3 and substrate domains, the proline-rich region is not well characterised and no studies have yet characterised any known function. However, along with the rest of the protein, it is well-conserved between species[140], suggesting that alterations to an existing proline residue here may disrupt protein function.

In this chapter, site-directed mutagenesis was performed on a BCAR1 expression plasmid to cause the base change coded for by rs1035539, changing proline at position 76 of the protein to a serine. These plasmids were expressed in COS cells to investigate how the amino acid change affects levels and phosphorylation of BCAR1 and interacting proteins, localisation of the protein and cell movement. In order to carry out assays in a blood vessel primary cell line, an adenoviral expression vector was then created to express wild-type and mutant BCAR1 in HUVECs. Protein levels were again assayed, as was the ability of cells to migrate through a membrane.

8.2 Results

8.2.1 pEGFP-C2/BCAR1 plasmid

The BCAR1 gene had previously been cloned into the pEGFP-C2 vector at *EcoRI* and *BamHI* restriction sites, to create a plasmid encoding a GFP-BCAR1 fusion protein[185] (Figure 94). This plasmid was provided by the Cardiovascular Biology and Medicine group (CVB).

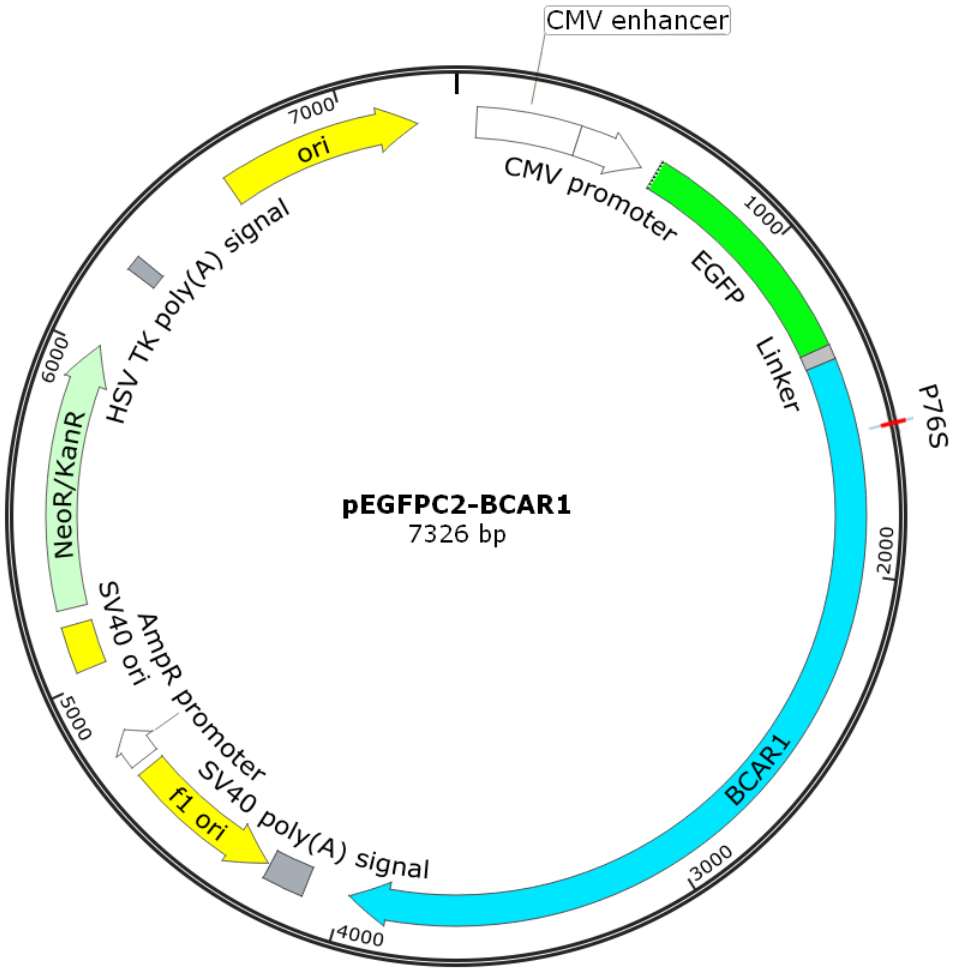

Figure 94: Map of pEGFPC2-BCAR1 plasmid. The plasmid was used as the wild-type BCAR1 expression vector, and was used for site-directed mutagenesis to create the mutant vector. The BCAR1 gene had previously been cloned into the plasmid. The position of the base change equivalent to rs1035539, changing a coded proline to serine, is marked by "P76S". Figure created using Snapgene[213].

8.2.2 Site-directed mutagenesis

Site-directed mutagenesis was performed to change C to T at the base corresponding to the position of rs1035539, coding for serine instead of proline at position 76 of the protein. Mutagenesis used Agilent Technologies' QuikChange Lightning mutagenesis kit. After transformation into OneShot TOP10

Chemically Competent cells, incubation, colony picking and plasmid purification, successful mutagenesis with no other introduced base changes was observed in one of four cultures grown. The C at the appropriate position in the *BCAR1* gene was successfully changed to a T (Figure 95).

Further sequencing across the gene confirmed that no additional bases had been modified. Cultures of transformed bacteria were therefore grown and maxi-prep plasmid purification carried out to purify sufficient stocks of the mutant plasmid.

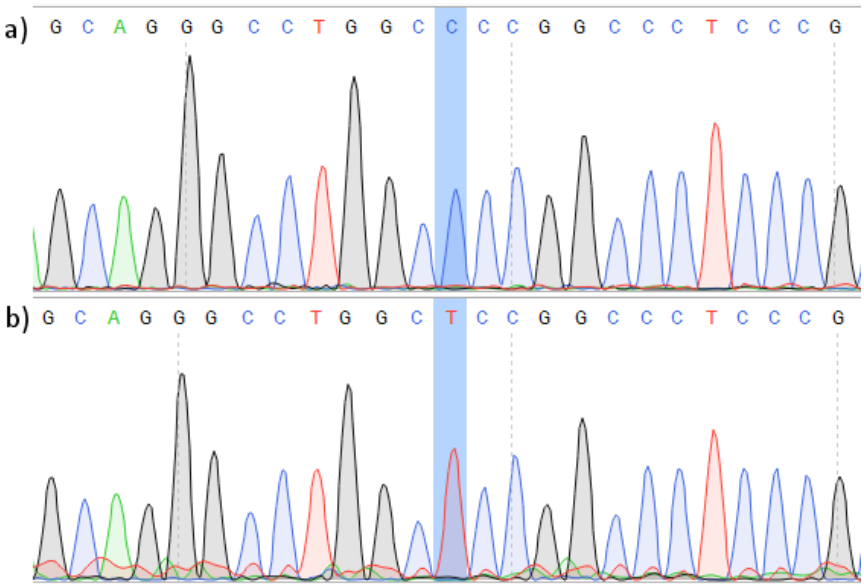

Figure 95: Sequence of the pEGFP-BCAR1 plasmid at the location of rs1035539 in the (a) original and (b) mutated plasmid. The C base was successfully mutated to T with no other introduced changes, altering proline to serine in the protein sequence.

8.2.3 Assays in COS and HEK293 cells

Four plasmids were expressed in cells for BCAR1 assays:

1. **pEGFP-C2:** basic GFP vector without *BCAR1* inserted. The vector was used as a negative control and for normalisation of transfection efficiency.
2. **pEGFP-C2/BCAR1(WT):** pEGFP-C2 vector with wild-type BCAR1/GFP fusion protein (proline at position 76).
3. **pEGFP-C2/BCAR1(M):** pEGFP-C2/BCAR1 with mutation coding serine at position 76, corresponding to the amino acid substitution caused by the SNP rs1035539.
4. **pEGFP-C2/BCAR1(15F):** pEGFP-C2/BCAR1 coding for an "unphosphorylatable" version of the BCAR1 gene. 15 key tyrosine residues in the substrate domain are mutated to phenylalanine and can no longer be phosphorylated[185].

The four plasmids are referred to in this chapter as “GFP”, “wild-type”, “mutant” and “15F” respectively.

8.2.3.1 Signalling assays

To investigate any differences between wild-type and mutant BCAR1 in total protein volume and protein phosphorylation, the four assay plasmids were transfected into COS (CV-1 in origin, carrying SV40) cells . Cells were incubated for 48 hours, and then treated with serum-free medium or one of three compounds to stimulate vascular pathways. Epidermal growth factor was used to stimulate cell growth and proliferation pathways, manganese chloride to stimulate integrin affinity (BCAR1 phosphorylation is integrin-dependent[333]), and nilotinib, a tyrosine-kinase inhibitor affecting Bcr-Abl, upstream of BCAR1.

Cell treatment had a significant effect on BCAR1 phosphorylation overall (Figure 96; ANOVA: $F = 6.4$, $p = 0.004$), with nilotinib resulting in greater phosphorylation than that seen in any of the other treatments (Tukey HSD; p values 0.008 to 0.01). However, there was no significant difference overall between wild-type and mutant BCAR1 ($F = 0.31$, $p = 0.58$). A trend for plasmid-treatment interaction appeared present: for example, phosphorylation appeared higher with mutant than wild-type BCAR1 when treated with nilotinib, but this difference was not significant ($p = 0.62$).

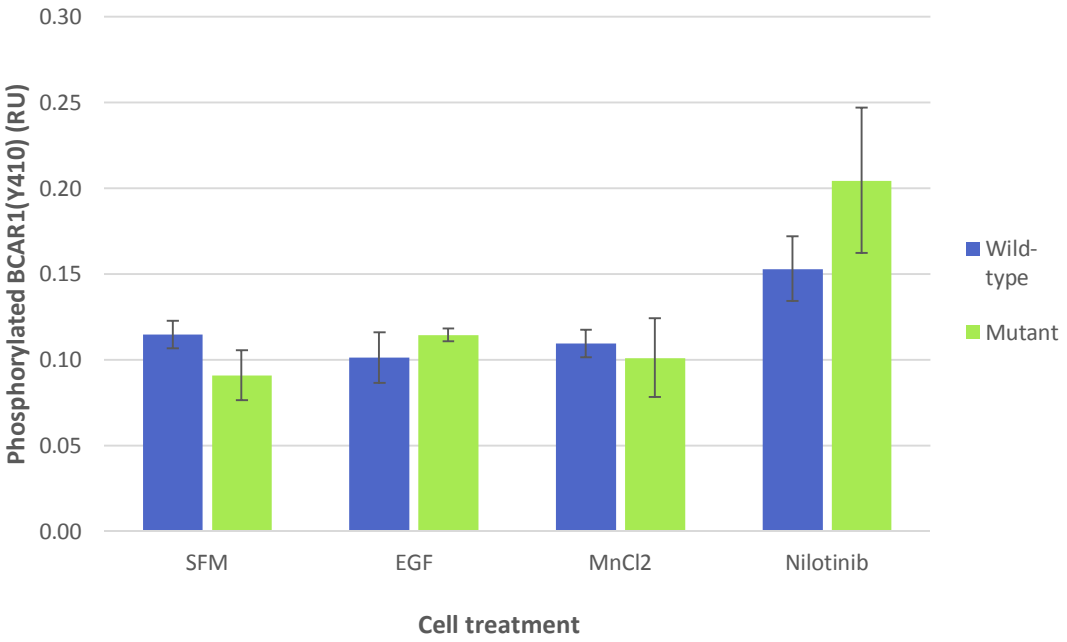

Figure 96: Phosphorylated BCAR1 by wild-type and mutant BCAR1 and cell treatment.COS cells were transfected with wild-type and mutant BCAR1 plasmids and incubated for 48 hours. After overnight incubation in medium with 0.5% FBS, cells were treated with serum-free medium (“SFM”) containing either nothing added, 2 ng/ml epidermal growth factor (“EGF”), 5 mM manganese chloride (“$MnCl_2$”) or 10 µM Nilotinib for 5, 5, 10 and

20 minutes respectively, then lysed. Proteins were detected via western blot. Data are averages from three independent experiments, and show mean±SE. Data is presented as phosphorylated BCAR1 at tyrosine 410 as a ratio of total BCAR1.

8.2.3.2 Protein localisation

As the pEGFPC2 vectors express BCAR1 in the form of a GFP fusion protein, detection of GFP signal could be used to locate the wild-type and mutant forms of the protein within different cellular compartments. Human embryonic kidney 293 (HEK293) cells were transfected with the pEGFPC2 expression vectors and fixed with formaldehyde 48 hours later. GFP signal was recorded under a fluorescent confocal microscope.

GFP signal was not present in all cells, showing that transfection was not 100% efficient. Cells transfected with the control GFP vector had the GFP protein distributed evenly throughout the cells. Wild-type and mutant BCAR1 was accumulated in points within the cytoplasm, and particularly at the cell membrane. No noticeable difference in localisation was observed between wild-type and mutant BCAR1 (Figure 97).

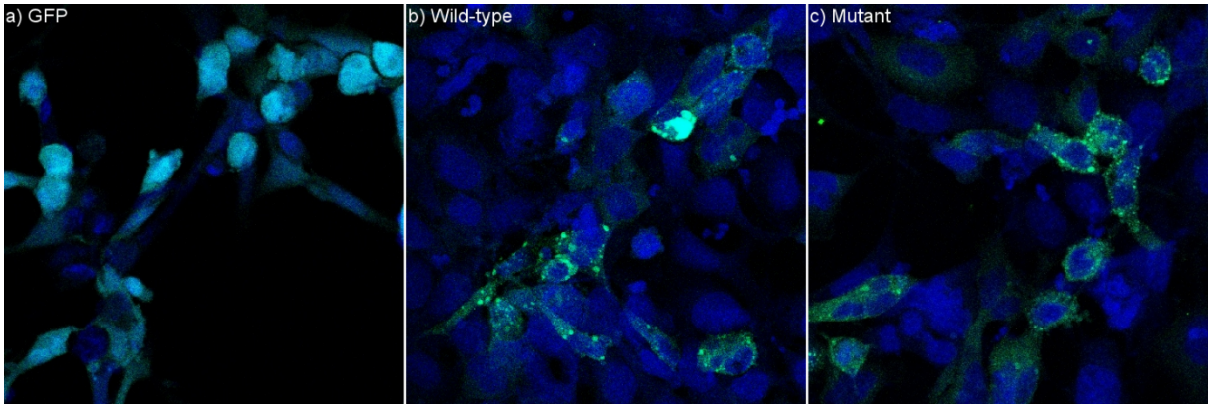

Figure 97: Location of GFP and GFP-BCAR1 proteins in COS cells.
COS cells were transfected with GFP, wild-type BCAR1 and mutant BCAR1 plasmids. 48 hours later, cells were fixed with formaldehyde and emitted GFP signal recorded under a fluorescent confocal microscope. Green signal shows GFP or GFP-BCAR1 protein; blue signal represents DAPI (staining cell nucleus). Images are representative of 2 replicates.
(a) pEGFP-C2: GFP control protein is distributed evenly throughout cells.
(b/c) Wild-type and mutant BCAR1 are accumulated at points within the cells, particularly at the cell membrane. No clear difference is visible between wild-type and mutant BCAR1.

8.2.3.3 Wound healing assay

Wound healing assays were performed on transfected COS cells to assess difference in cell movement between wild-type and mutant BCAR1. Cells were transfected with the four plasmids, and grown in a

96-well tissue culture plate. Uniform cell-free areas were created by scratching the cell surface. Cell confluence was monitored to observe the rate of cell movement to close the wound.

As shown in Figure 98, confluence after 62 hours (the last point at which all scratches had not yet closed) was higher in cells with wild-type BCAR1 than mutant (t-test: t = 3.4163, p = 1.3×10^{-3}). The overall speed of closure was higher in cells with wild-type BCAR1 than mutant BCAR1, but this was only of borderline statistical significance (t-test of regression coefficients: t = 1.93, p = 0.059).

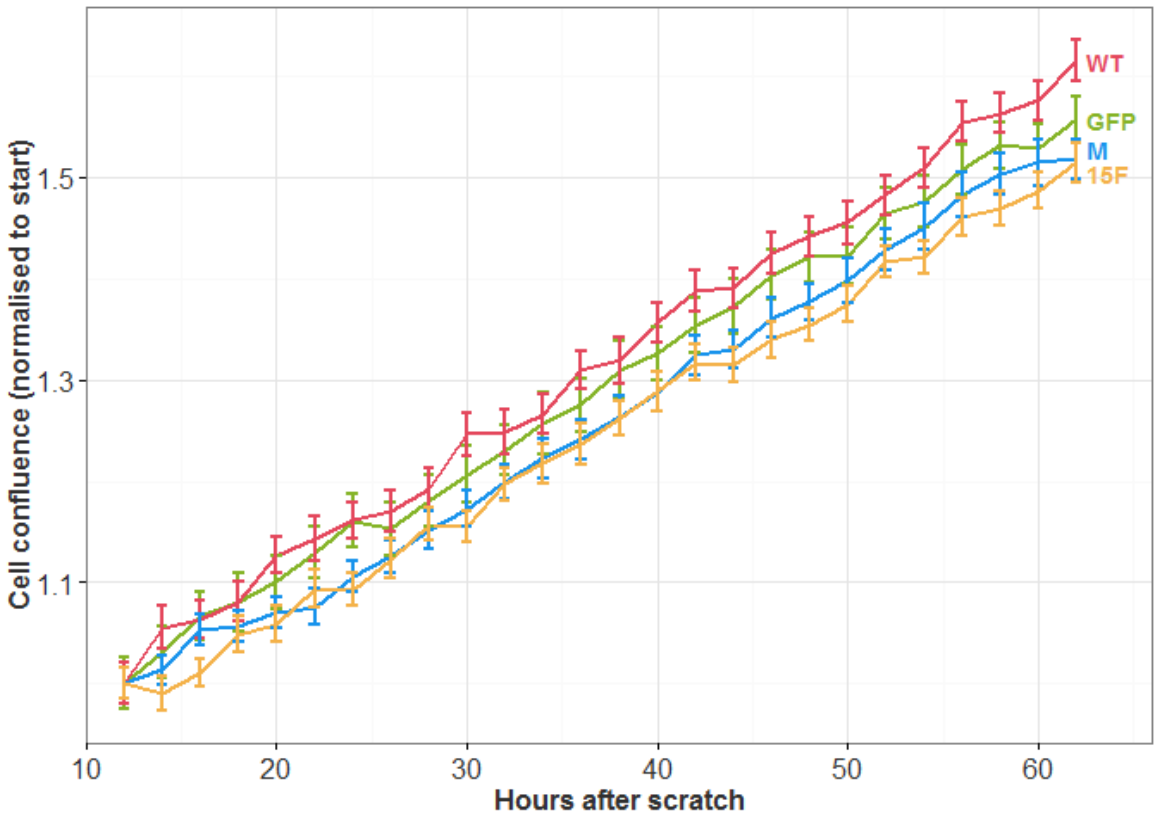

Figure 98: COS cell confluence by expression in scratched cell layers.
Confluent layers of cells were grown in a 96-well tissue culture plate. Uniform rectangular "wounds" of cell-free areas were creating by scratching each well with the Essen Bioscience WoundMaker, and cells were then incubated in standard culture conditions while being monitored by the Incucyte. Photos were taken at 2-hour intervals and confluence measured. Data was analysed from 12 hours after scratch, where cell confluence stabilised after wounding. Confluence is here presented as multiple of starting percentage. Results are average of 24 replicates per plasmid.
Cells with wild-type BCAR1 (WT) had higher confluence at the end of measurement than those with mutant BCAR1 (M) (p = 1.3×10^{-3}). Speed of closure was higher for WT than M cells, but this was of borderline statistical significance (p = 0.059).

8.2.4 Assays in HUVECs

8.2.4.1 Transfection of HUVECs

After signalling and wound healing assays had been carried out, HUVECs were selected for further assays due to their suitability as a model for blood vessel phenotypes. Multiple methods were used to test efficacy of transfection into these cells, which are known to be difficult to transfect[284]. The GFP and WT plasmids were used for testing. As all the assay plasmids express GFP or a BCAR1-GFP fusion protein, effectiveness of transfection can easily be checked by detection of GFP signal. Transfection with these plasmids was tested using the lipid reagents Lipofectamine 3000 and jetPEI-HUVEC, and electroporation using the Amaxa Nucleofector I.

As shown in Figure 99, only electroporation achieved any transfection success, with about 15% of cells electroporated with the control GFP plasmid producing GFP signal. However, as seen in chapter 5.2.1.7, the GFP fusion plasmid was far more difficult to deliver to cells, with less than 1% of cells producing signal. Lysis of cells and blotting for BCAR1 showed a small amount of higher-weight BCAR1 (GFP-fusion protein) in these electroporated cells (Figure 100), but efficiency was not high enough to use this method for assays.

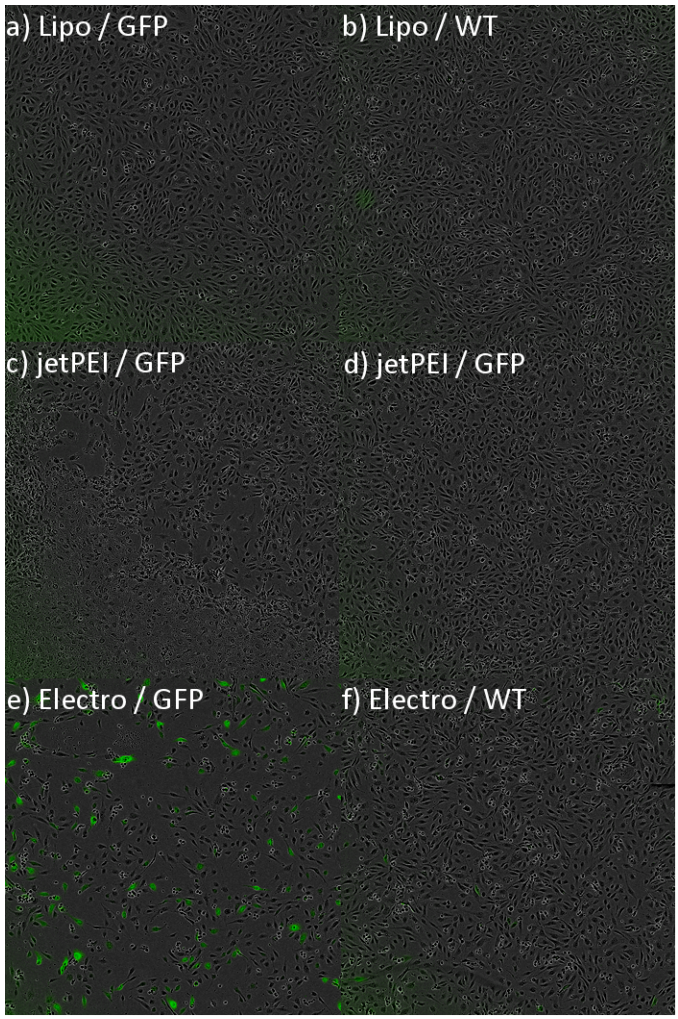

Figure 99: GFP signal shows transfection efficiency of pEGFPC2 (“GFP”) and pEGFP-C2/BCAR1-WT (“WT”) plasmids into HUVECS.
(a/b) Transfection using Lipofectamine 3000 was not successful: no cells produced GFP signal with either GFP or WT plasmid.
(c/d) Transfection using jetPEI-HUVEC also resulted in no cells producing GFP signal.
(e/f) Electroporation was successful with ~15% efficiency using the GFP plasmid. However, efficiency with the WT plasmid was much lower, with <1% cells producing GFP signal.

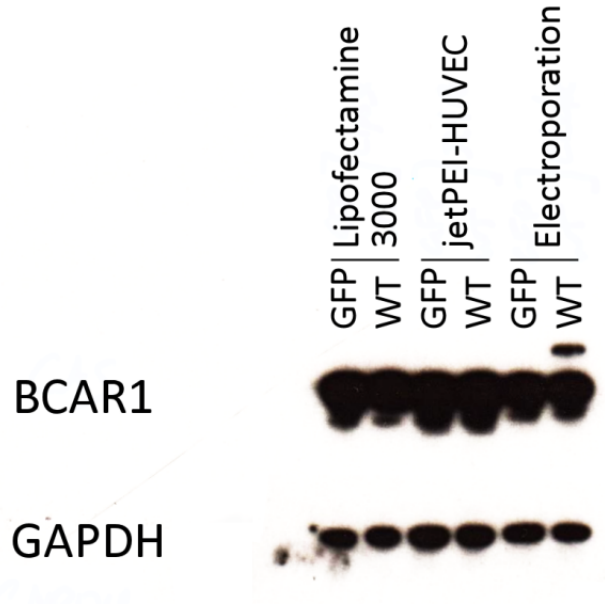

Figure 100: BCAR1 protein present in cells transfected using the three trial methods: Lipofectamine 3000, jetPEI-HUVEC and electroporation. Endogenous BCAR1 is similar between cells. GFP-fused BCAR1 is visible as a higher-weight protein. None is visible in cells transfected with Lipofectamine or jetPEI-HUVEC, where transfection was not successful.
Cells transfected with the GFP control plasmid showed no GFP-fused BCAR1, as only the GFP protein is expressed by the plasmid. A small amount of GFP-fused BCAR1 is present in cells electroporated with the WT plasmid, where a small number of cells successfully expressed the plasmid. GAPDH was used as a loading control.

8.2.4.2 Genotyping HUVECs for assays

After transfection tests showed that HUVECs could not be adequately transfected with BCAR1 expression vectors, genotyping of HUVECs for the rs1035539 variant was considered. As rs1035539 is a common variant with 32% MAF, the majority of HUVEC donors would be expected to have AA and GG genotypes, and their cells could be selected for use in protein assays to express the wild-type or mutant form of BCAR1.

This would require cells to come from individual donors, and several donors for each genotype would need to be tested to account for other genetic differences between donors: unlike with expression vectors expressed in cells from the same lot, the background genetic profile of individual donors differs. Individual-donor HUVECs were provided from the Queen Mary Cardiovascular Genomics and Stratified Medicine group for culture, genotyping and use in protein assays.

Chapter 7.3.3 outlined the difficulty of genotyping the rs1035539 polymorphism due to two retrogenes on chromosome 15, which have high similarity to the sequence around the SNP (Figure 93). Sequencing primers were therefore designed to amplify the sequence encompassing rs1035539 only. As shown in Figure 101, a 4 bp break in the sequence similarity to the retrogenes is present upstream

of the SNP. One primer was designed on this 4 bp sequence, with the other just outside the region of similarity (Figure 101). This allowed amplification and sequencing of the sequence specific to chromosome 16.

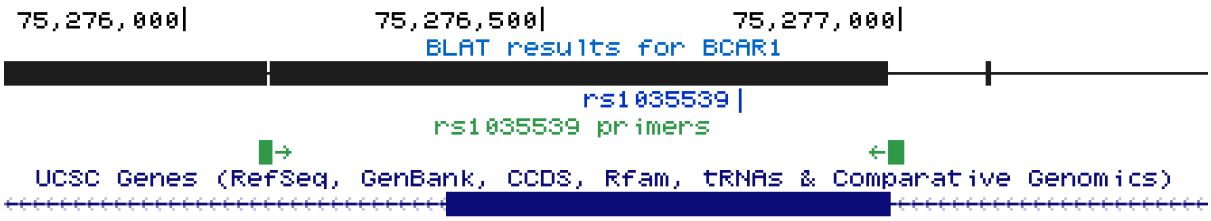

Figure 101: Location of rs1035539 PCR and sequencing primers. The left primer is positioned over the short sequence of non-similarity to the retrogenes, while the right primer is specific to this locus. This allows specific amplification of the SNP.

This sequencing method was successful in genotyping genomic DNA samples of known genotype. However, genotyping for rs1035539 was ultimately not used for selection of cells for protein assays. Primary HUVEC numbers that were obtained from the Queen Mary Cardiovascular Genomics and Stratified Medicine group were too low in number to acquire sufficient cells for assays without considerably exceeding passage 5; HUVECs have a limited lifespan before they differentiate or enter senescence[304], and passage 5 was here considered to be the maximum reliable passage number.

8.2.4.3 Production of BCAR1 adenovirus

As successful transfection of HUVEC cells with the pEGFP-C2/BCAR1 vectors could not be performed, an adenoviral vector was produced to infect HUVECs with wild-type or mutant vector for use in assays. Invitrogen's Gateway cloning system was used to produce the expression clones. The BCAR1 vector had previously been cloned into the Gateway pENTR-3C entry vector[185], and this pENTR-3C/BCAR1 vector was provided by the Cardiovascular Biology and Medicine group.

8.2.4.3.1 Site-directed mutagenesis

Site-directed mutagenesis was carried out on the pEGFP-C2/BCAR1 plasmid to change the base corresponding to rs1035539, changing Proline 76 to Serine, as in 8.2.2 (Figure 102). Mutagenesis was successful, with the C base mutated to T, and no other changes introduced into the sequence (Figure 103).

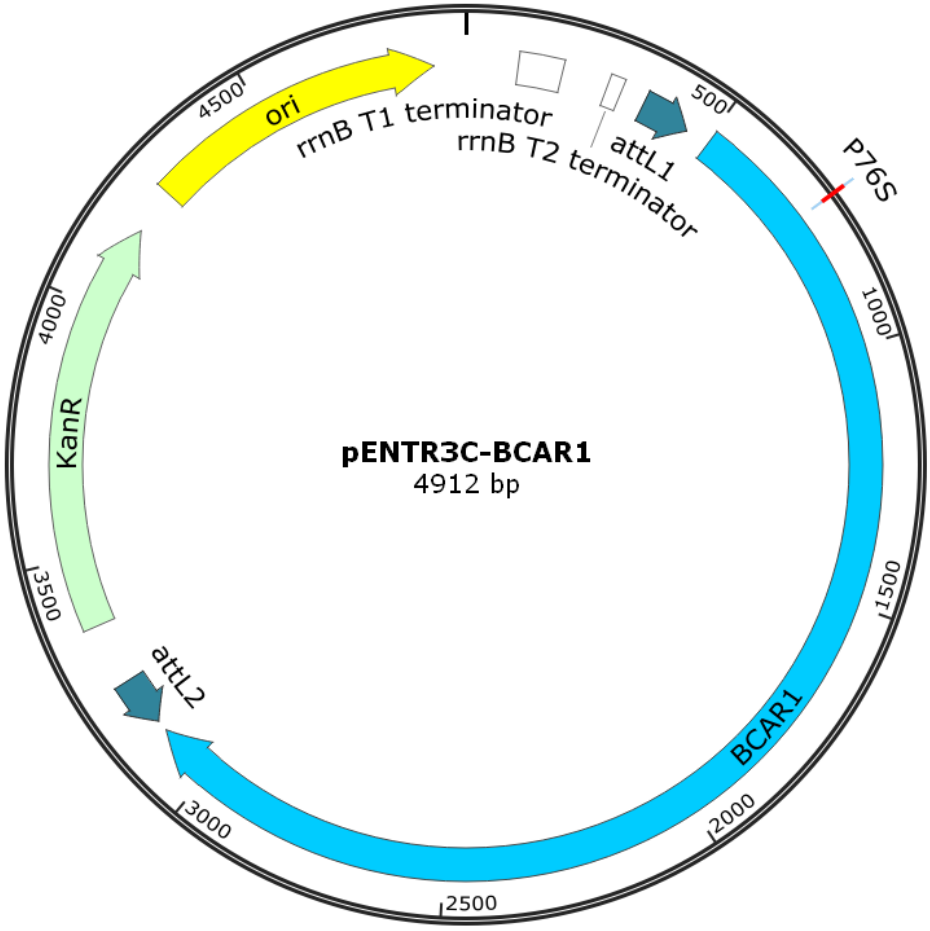

Figure 102: Map of pENTR3C-BCAR1 plasmid. The plasmid was used as the wild-type entry vector for production of adenovirus, and basis for site-directed mutagenesis to create the mutant vector. The BCAR1 gene had previously been cloned into the plasmid. The position of the base change equivalent to rs1035539, changing a coded proline to serine, is marked by "P76S". Figure created in Snapgene[213].

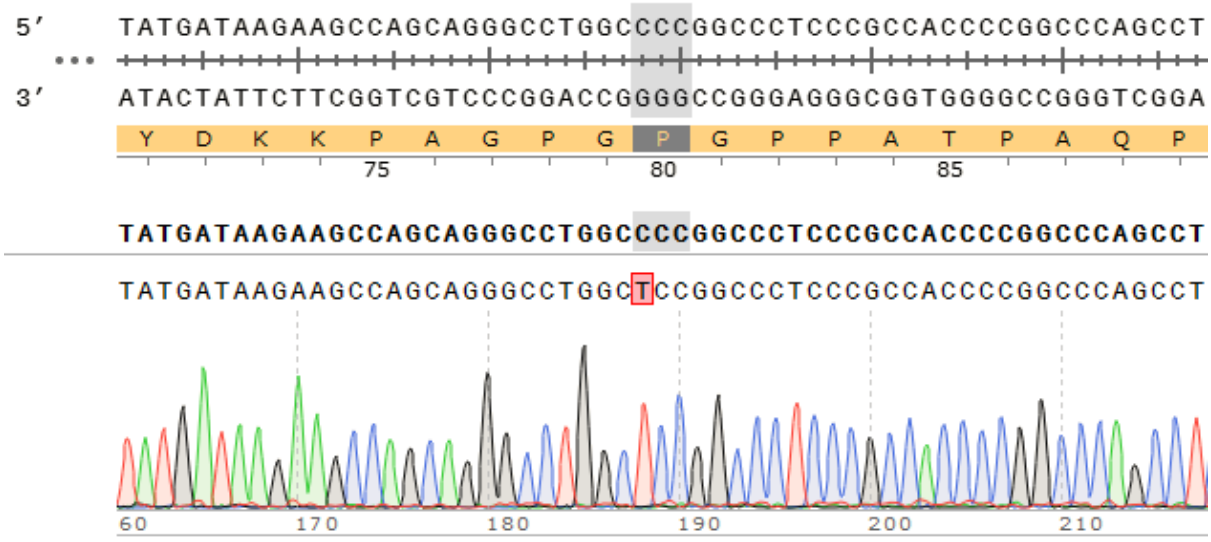

Figure 103: Sequence of the pENTR-3C/BCAR1 plasmid at the location of rs1035539 after site-directed mutagenesis. The C base was successfully mutated to T with no other introduced changes, altering proline to serine in the protein sequence. Picture from Snapgene[213].

The Gateway LR Recombination reaction was carried out to insert the wild-type and mutant *BCAR1* genes into the pAd/CMV/V5-DEST destination vector, creating the expression vector. Purified expression vector was used to infect HEK293 cells to allow production of virus particles, which were then purified using the Adenopure virus purification kit. Purified virus solutions were dialysed to replace elution buffer with a TE buffer containing glycerol, which is suitable for long-term storage. Viral titre in these solutions was then measured and calculated using the QuickTiter Adenovirus Quantitation Kit. Final wild-type and mutant viral titres were 4.3×10^{10} and 5.2×10^{10} VP/ml respectively.

8.2.4.4 Signalling assays

After successful production of BCAR1 adenovirus, the effect of BCAR1 type on protein levels and phosphorylation was tested by infecting HUVECs with control, wild-type and mutant viruses. Cells were incubated for 48 h after infection. They were then stimulated with 25 ng/ml VEGF for timepoints from 5-60 minutes, lysed and blotted for proteins.

8.2.4.5 Total BCAR1 does not differ between wild-type and mutant

The effect of expression vector and VEGF stimulation on total BCAR1 were compared using a two-way ANOVA (Figure 104). There was a significant effect of expression vector on total BCAR1 levels ($F = 63.8$, $p = 8.0 \times 10^{-14}$), but individually this was not seen between wild-type and mutant, or uninfected and control (each $p = 0.99$; Tukey HSD post-hoc test). There was also a significant effect of VEGF treatment time on total BCAR1 ($F = 7.4$, $p = 9.1 \times 10^{-5}$); with BCAR1 increasing with VEGF treatment time.

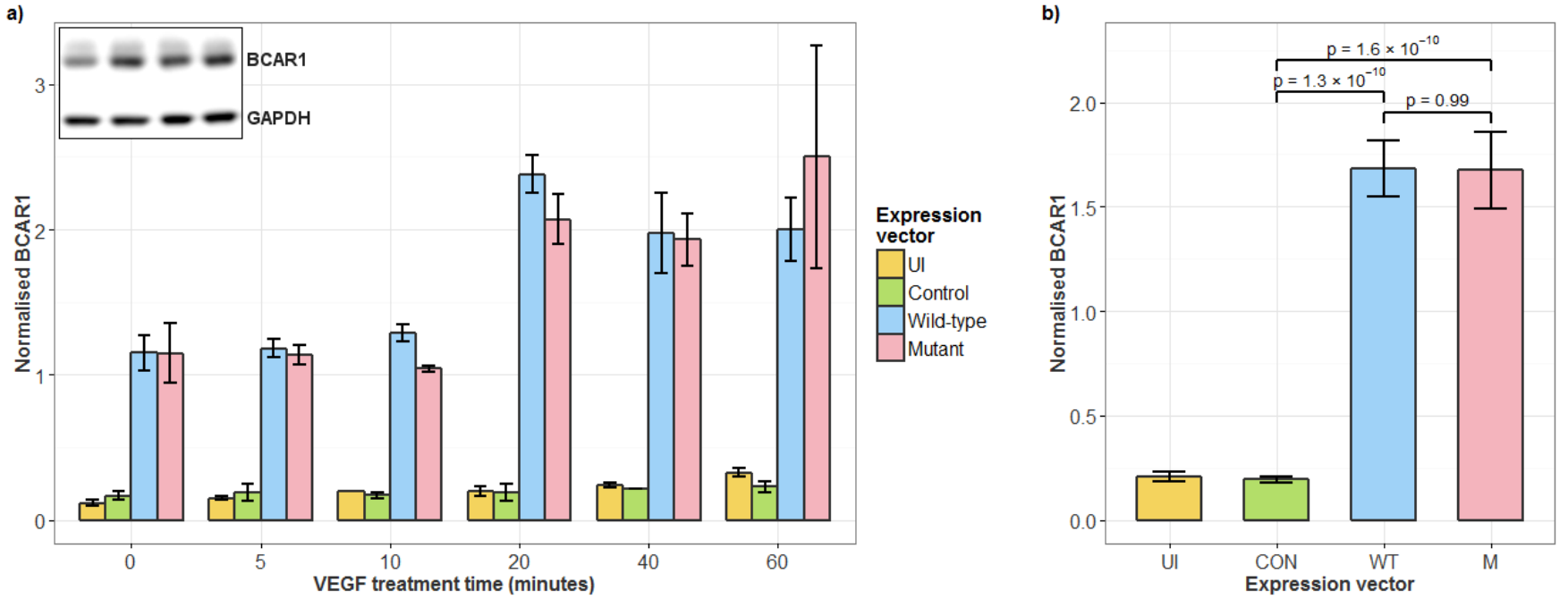

Figure 104: Effect of expression vector and VEGF treatment on total BCAR1.
HUVECs were left uninfected, or infected with control virus, wild-type and mutant BCAR1 virus ('UI', 'GFP', 'WT' and 'M'). Cells were incubated for 48 h, then stimulated with 25 ng/ml VEGF for 0 (no stimulation) to 60 minutes. Cells were then lysed and blotted for total BCAR1.
Figure shows data from two to three independent experiments (UI and control data from two experiments), presented as total BCAR1 in relative units (mean +/- SEM) normalised to total GAPDH. Representative blot is inset.
(a) BCAR1 increases with VEGF treatment time (ANOVA: $F = 7.7$, $p = 6.4 \times 10^{-5}$).
(b) BCAR1 levels were ~8× higher in cells with wild-type and mutant BCAR1 expression vectors than those with control ($p = 5.9 \times 10^{-12}$ and 7.1×10^{-12}; Tukey HSD post-hoc test). BCAR1 levels between WT and mutant were not different ($p = 0.99$)

8.2.4.6 Phosphorylated BCAR1

As phosphorylation of BCAR1 is integral to its role as an adapter protein, cell lysates were blotted for phosphorylated BCAR1. Antibodies were used to probe for phosphorylation at tyrosine-410 and tyrosine-249. These two tyrosines are located in two of the YxxP motifs in the BCAR1 substrate domain, and their phosphorylation is thought to be important for recruitment of proteins to trigger downstream signalling events[334].

Tyrosine-410 phosphorylation was affected by expression vector ($F = 18.8$, $p = 1.1 \times 10^{-8}$), and VEGF treatment time ($F = 8.6$, $p = 5.5 \times 10^{-6}$), but again there was no different between the two BCAR1 vectors (Figure 105). Tyr249 phosphorylation showed the same pattern with expression vector ($F = 7.7$, $p = 4.3 \times 10^{-4}$) but not VEGF treatment time ($F = 1.1$, $p = 0.38$) (Figure 106).

8.2.4.7 Phosphorylated paxillin

Phosphorylation of paxillin was tested, a focal adhesion protein involved that forms a complex with BCAR1[141]. Levels of paxillin phosphorylation were not affected by expression vector ($F = 0.84$, $p = 0.48$), but were affected by VEGF treatment time ($F = 6.2$, $p = 1.2 \times 10^{-4}$), with higher phosphorylation at 20 and 40 minutes than with no treatment (Figure 107).

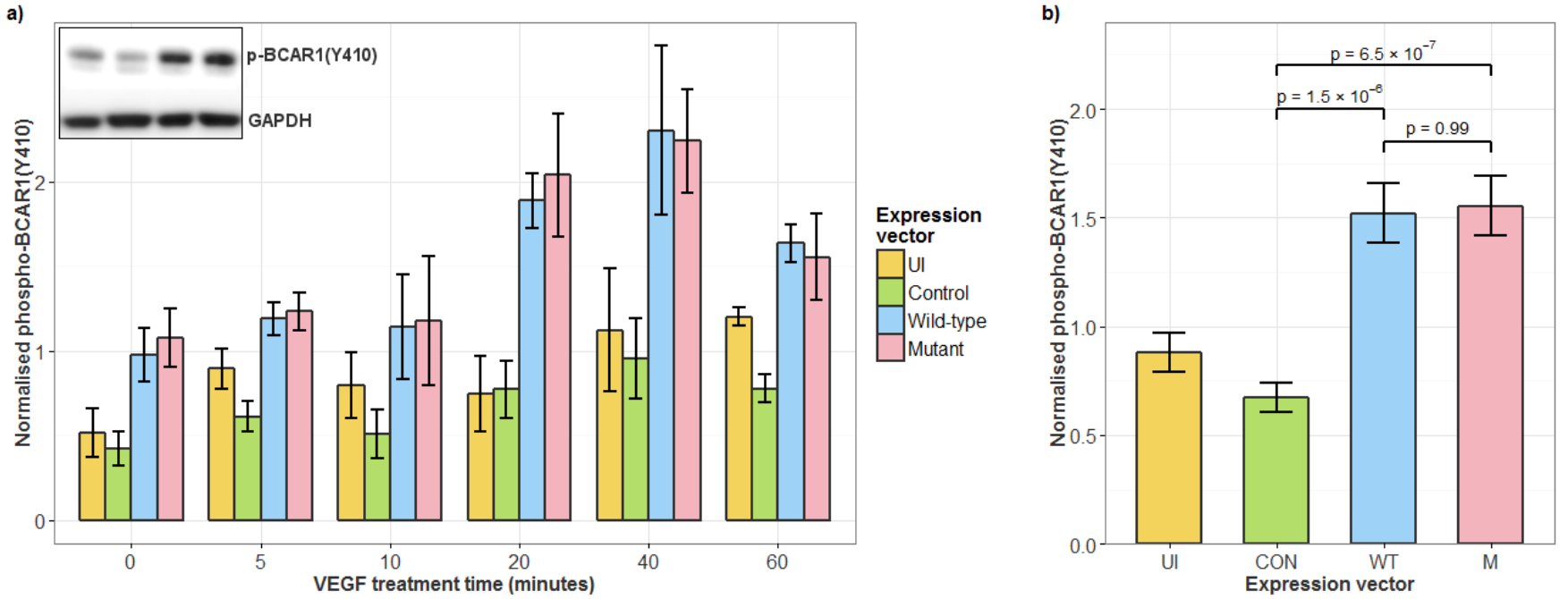

Figure 105: Effect of expression vector and VEGF treatment on phosphorylated BCAR1 (tyrosine 410).
HUVECs were left uninfected, or infected with control virus, wild-type and mutant BCAR1 virus ('UI', 'GFP', 'WT' and 'M'). Cells were incubated for 48 h, then stimulated with 25 ng/ml VEGF for 0 (no stimulation) to 60 minutes. Cells were then lysed and blotted for phosphorylated BCAR1(Y410).
Figure shows data from three to four independent experiments (UI and control data from three experiments), presented as total BCAR1 in relative units (mean +/- SEM) normalised to total GAPDH. Representative blot is inset.
(a) Phosphorylated BCAR1 increases with VEGF treatment time (ANOVA: $F = 8.6$, $p = 5.5 \times 10^{-6}$).
(b) Phosphorylated BCAR1 levels was higher in cells with wild-type and mutant BCAR1 than those with control ($p = 1.5 \times 10^{-6}$ and 6.5×10^{-7}; Tukey HSD post-hoc test). Levels between WT and mutant were not different ($p = 0.99$)

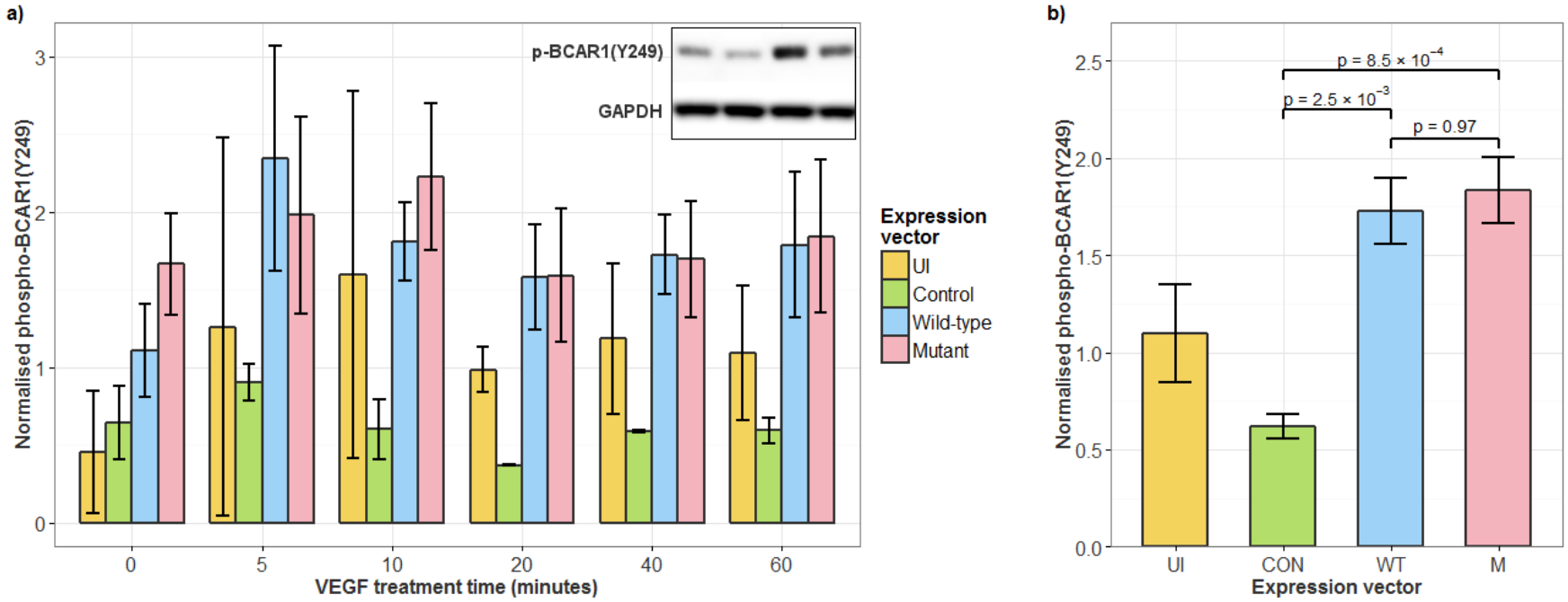

Figure 106: Effect of expression vector and VEGF treatment on phosphorylated BCAR1 (tyrosine 249).
HUVECs were left uninfected, or infected with control virus, wild-type and mutant BCAR1 virus ('UI', 'GFP', 'WT' and 'M'). Cells were incubated for 48 h, then stimulated with 25 ng/ml VEGF for 0 (no stimulation) to 60 minutes. Cells were then lysed and blotted for phosphorylated BCAR1(Y249).
Figure shows data from two to three independent experiments (UI and control data from two experiments), presented as total BCAR1 in relative units (mean +/- SEM) normalised to total GAPDH. Representative blot is inset.
(a) Phosphorylated BCAR1 increases with VEGF treatment time (ANOVA: $F = 7.7$, $p = 4.3 \times 10^{-4}$).
(b) Phosphorylated BCAR1 levels was higher in cells with wild-type and mutant BCAR1 than those with control ($p = 2.5 \times 10^{-3}$ and 8.5×10^{-4}; Tukey HSD post-hoc test). Levels between WT and mutant were not different ($p = 0.97$)

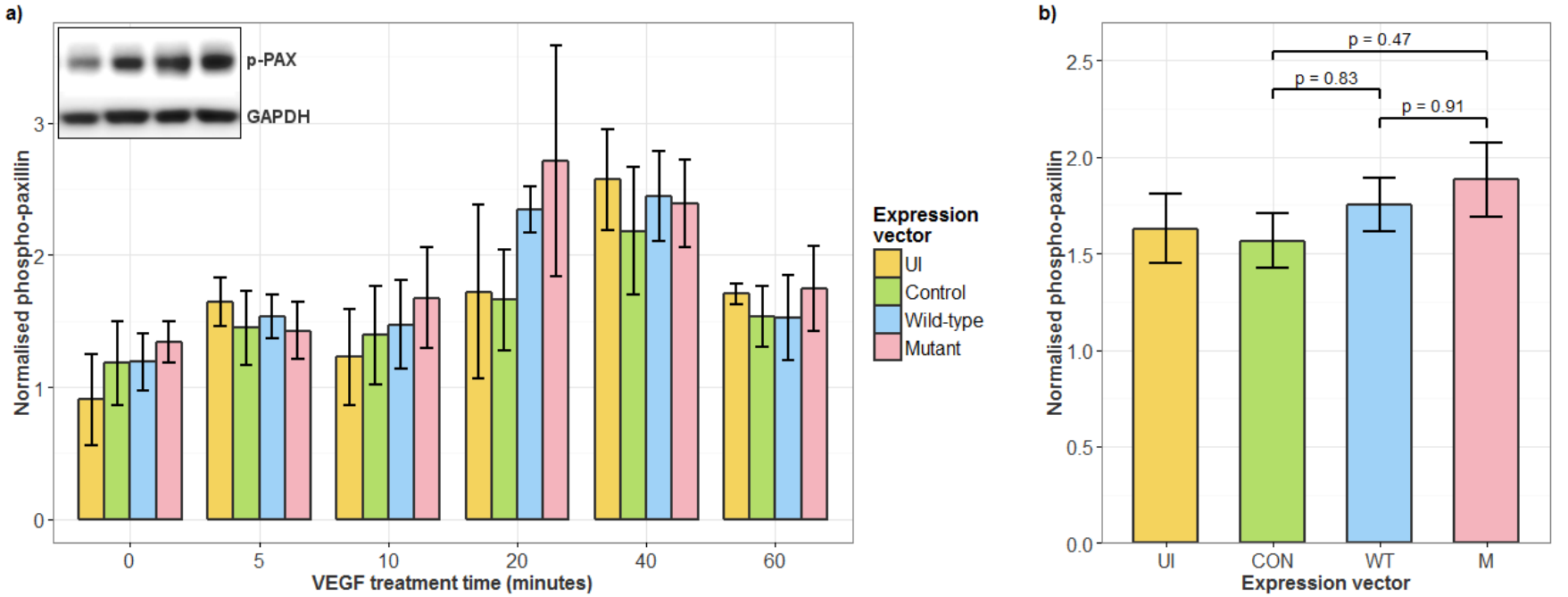

Figure 107: Effect of expression vector and VEGF treatment on phosphorylated paxillin.
HUVECs were left uninfected, or infected with control virus, wild-type and mutant BCAR1 virus ('UI', 'GFP', 'WT' and 'M'). Cells were incubated for 48 h, then stimulated with 25 ng/ml VEGF for 0 (no stimulation) to 60 minutes. Cells were then lysed and blotted for phosphorylated paxillin. Representative blot is inset.
Figure shows data from three to four independent experiments (UI and control data from three experiments), presented as total BCAR1 in relative units (mean +/- SEM) normalised to total GAPDH.
(a) Phosphorylated paxillin increased with VEGF treatment time (ANOVA: F = 6.2, p = 1.2 × 10-4).
(b) Phosphorylated paxillin did not differ between expression vector (ANOVA: F = 0.84, p = 0.48).

8.2.4.8 Well migration assays

After protein levels and phosphorylation had been shown not to differ between wild-type and mutant BCAR1, well migration assays were used to assess the effect of protein type on cell movement. HUVEC migration through a porous membrane was measured, using cells transfected with wild-type and mutant BCAR1, and GFP control virus. Cells were infected with each of the three viruses or no virus (uninfected), trypsinised and seeded into cell culture inserts with a porous membrane. Control wells contained medium with 10% serum while stimulated wells contained medium, 10% serum and 25 ng/ml VEGF.

There was a significant overall positive effect on VEGF stimulation on cell movement (Figure 108; two-way ANOVA; $F = 26.0$, $p = 3.3 \times 10^{-5}$), but only a near- significant effect of expression vector ($F = 2.9$, $p = 0.069$). Cells with mutant BCAR1 migrated more slowly than uninfected cells ($p = 0.04$, Tukey HSD post-hoc test), whereas those with wild-type BCAR1 did not ($p = 0.64$). Neither group showed a significant difference compared to GFP control ($p = 0.97$ and $p = 0.19$ for wild-type and mutant respectively). Cells with mutant BCAR1 appeared to migrate more slowly than those with wild-type BCAR1, although this difference was not significant ($p = 0.73$).

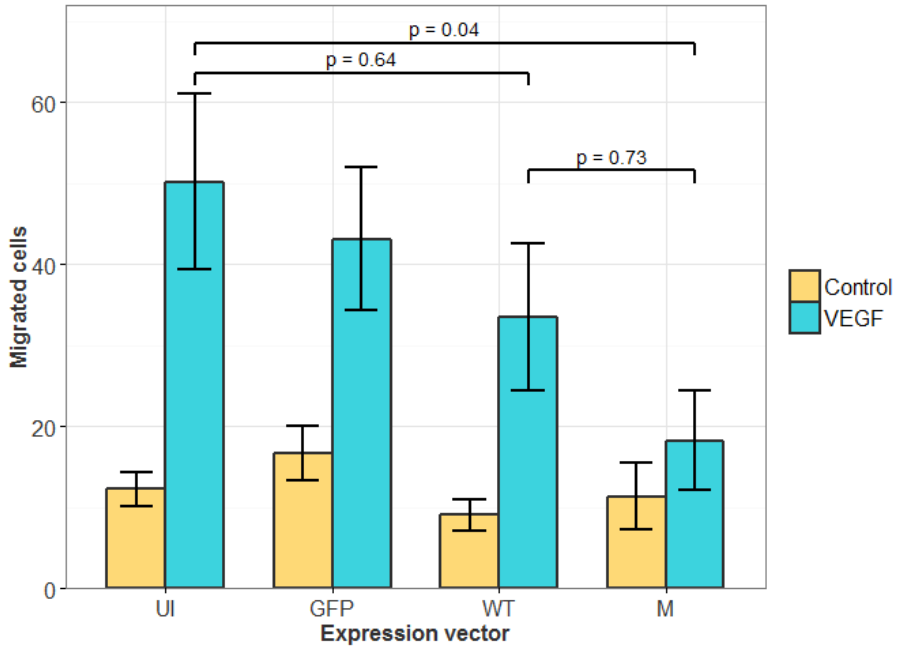

Figure 108: Average migrated cell number per expression vector for stimulated and unstimulated cells.
HUVECs were infected with no virus, GFP control virus, and wild-type and mutant BCAR1 virus ('UI', 'GFP', 'WT' and 'M'). 48 h later, wells were filled with medium with and without 25 ng/ml VEGF ('control' and 'VEGF' respectively) and cells seeded into well inserts containing a porous membrane. After 4 hours of incubation, cells on the membranes were fixed, stained and counted at 200× magnification to obtain the numbers of cells that had migrated through the membrane.
Bars represent mean number of cells of four independent experiments.
Cells with mutant BCAR1 migrate more slowly than uninfected cells ($p = 0.04$), but cells with wild-type BCAR1 do not ($p = 0.64$).
Error bars represent SEM.

8.3 Discussion

In this chapter, a proline/serine change at position 76 of the BCAR1 protein was assayed for its effects on protein function. Chapter 7 showed that this change, coded for by the SNP rs1035539, is associated with carotid plaque, so assays here looked at effects on signalling and protein function to understand how the SNP and the protein may be affecting plaque.

Assays in COS cells showed a potential difference in speed of cell migration, and HUVECs were then chosen for more detailed assays. In these cells, no differences between total BCAR1 and phosphorylation of BCAR1 and paxillin were found. However, a difference between cell migration was suggested between cells with wild-type and mutant BCAR1: cells with mutant BCAR1 migrated more slowly than uninfected cells, while wild-type and control cells did not.

8.3.1 Cell choice for expression assays

COS cells were selected for initial assays as they were known to be receptive to transfection. These cells are also responsive to epidermal growth factor[335], which is known to modulate proliferation and migration of endothelial and smooth muscle cells[336,337]. HEK293 cells used for protein localisation did not need to be responsive to growth factors and provided a human cell line in which to visualise the proteins.

However, a primary cell line suitable for the plaque phenotype was desired for further assays. Plaque as measured in the IMPROVE and PLIC cohorts is a derivative variable of intima-media thickness. rs1035539 may therefore be associated with changes in the intima (endothelium) or media (smooth muscle) layer. Endothelial cells were used here, but smooth muscle cells will also be used for further assays.

8.3.2 Signalling and wound healing assays (COS)

Initial signalling assays in COS cells suggested some differences in BCAR1 phosphorylation between wild-type and mutant, but this varied with cell treatment and was not statistically significant. It was clear that stimulation of cells had effects on pathways including BCAR1, and stimulation with growth factors was later used for assays on HUVECs.

A scratch assay was used to look at cell closure because of BCAR1's known role in cell motility. Cells with wild-type BCAR1 were shown to exhibit faster wound closure than those with mutant BCAR1. This was a preliminary experiment on cells that were an imperfect model, and would require more

repeats to draw robust conclusions from the results. However, it suggested that it would be sensible to examine cell movement in HUVECs.

8.3.3 Protein localisation (HEK293)

Fluorescent microscopy to visualise subcellular localisation of wild-type and mutant BCAR1 did not show a noticeable difference between the two. In each case the protein was largely clustered at points within the cytoplasm and at the plasma membrane, supporting known data about its subcellular location (see Human Protein Atlas[156]). BCAR1 is known to localise to the cytoplasm; here phosphorylation stimulates its movement to the plasma membrane[140], where it regulates cytoskeleton remodelling and cell adhesion[338]. However, the limitation should be noted that as a GFP fusion protein, the BCAR1 seen in the assays may not behave exactly as endogenous BCAR1 does.

A large difference in subcellular localisation between wild-type and mutant BCAR1 was not expected, due to the importance of the protein: any significant differences between the forms would be expected to have a large downstream effect, impeding the variant from becoming common in humans.

8.3.4 HUVECs: methods for assays

Genotyping HUVECs for rs1035539 was considered as a method of comparing wild-type and mutant BCAR1. This method was ultimately not used for assays, as too few single-donor HUVECs were available. Genotyping may be a method worth considering for future work, but the MAF of the variant determines its feasibility. As rs1035539 has a MAF of 32%, cells would be expected to be homozygous for the rare allele in only 10% of cases. As numerous differences in the rest of the genome exist between donors, multiple donors for each genotype would be required for assays to account for these differences. Even multiple donors would not separate out the effects of SNPs in LD; however, it was seen in chapter 7 that SNPs in LD with rs1035539 are few and less likely than this coding SNP to have a functional effect.

As with many primary cells, HUVECs are known to be difficult to transfect[284,285]. Here two lipid reagents and electroporation were trialled as a method of transfecting HUVECs for signalling and migration assays, but no method could achieve sufficient transfection for assays.

After other methods proved unsuitable for assays with HUVECs, adenoviral vectors were created to infect them. This method was not chosen initially as the production process is long and commonly

does not result in a successful virus; however, the infection efficiency with adenoviral vector was much higher than with transfection or electroporation. As adenoviruses do not integrate into the genome, they do not disrupt genomic DNA, allowing efficient transient expression of vectors.

8.3.5 HUVEC signalling assays

Production of the BCAR1 adenoviruses allowed infection of HUVECs for signalling assays. In these cells, total BCAR1 was found to be much higher in wild-type and mutant cells than in uninfected and control cells (Figure 104), as expected: expression vectors expressing BCAR1 in addition to endogenous production greatly increase the total. No difference was seen in total BCAR1 between wild-type and mutant, which was again anticipated, as the amino acid change is unlikely to affect total protein produced.

Phosphorylated BCAR1 on tyrosine residues 410 and 249 was measured. These two tyrosine residues make up two of the YxxP repeats in the substrate domain whose phosphorylation is important for downstream signalling[334], and are often used for BCAR1 phosphorylation assays. Phosphorylation was higher in cells transfected with wild-type or mutant BCAR1 vectors than those that weren't (Figure 105, Figure 106), due to the greater availability of the protein for phosphorylation. Here a difference in phosphorylation between wild-type and mutant BCAR1 might indicate a reason for the difference in plaque phenotypes, as phosphorylation of the protein is key in establishing its downstream effects[141]. However, no difference was seen between phosphorylation of these two forms of the protein, indicating the effect is occurring through a different method. The altered amino acid, proline at position 76, is not close in the primary sequence to these two tyrosine residues, nor is it in the substrate domain in which these residues are located, so would not necessarily be expected to alter their phosphorylation.

The change in the amino acid structure may have effects downstream in signalling pathways through another mechanism. Phosphorylation of paxillin was here tested as it is a focal adhesion protein that interacts with BCAR1[141]. However, again there was no significant difference between wild-type and mutant BCAR1 (Figure 107), and expression of either BCAR1 did not increase phosphorylation above endogenous amounts.

8.3.6 Well migration assays (HUVEC)

To investigate overall effect on cell function rather than individual proteins, well migration assays were used to measure the effect of the BCAR1 variant on cell movement. Stimulation of HUVECs

with VEGF resulted in significantly greater numbers of migrated cells, highlighting the importance of this molecule in endothelial cell migration[339].

Cells with mutant BCAR1 migrated more slowly than uninfected cells, whereas those with wild-type BCAR1 did not. Cells with mutant BCAR1 therefore appear to migrate more slowly than those with wild-type BCAR1, which matches with the results seen in HEK293 cells earlier in the wound healing assay. However, this difference was not significant: to draw strong conclusions, further biological replicates of this assay would have to be carried out.

If the difference in cell migration is indeed true, it raises the question of how the proline/serine change is causing the effect. It has been shown in this chapter that the variant does not cause an obvious change in phosphorylation in the substrate-binding domain. However, the adaptor protein role of BCAR1 might be important here. BCAR1 recruits and binds to many proteins (Figure 12), and as discussed in chapter 7.3.6, proline-rich regions often play a role in binding. If the variant is affecting binding of a protein or number of proteins to BCAR1, it may have effects on downstream proteins: further assays would test other proteins known to be affected by BCAR1 phosphorylation and protein recruitment.

8.3.7 Conclusion

The proline/serine change in BCAR1 that corresponds to the SNP rs1035539 does not appear to affect total amounts or phosphorylation of the protein, nor of the associated protein paxillin. However, it may cause a change in speed of cell migration: to confirm this and to understand the mechanism of how it may contribute to plaque formation and remodelling, further work should look at other downstream proteins and their effects on phenotype.

8.3.8 Future work

Recent further work by the Cardiovascular Biology and Medicine Group has found an effect of wild-type and mutant BCAR1 on phosphorylation of Crk-like protein (CRKL), an adapter protein that binds to the substrate domain of BCAR1[340]. Further work will examine this relationship in greater depth and look at how phosphorylation of CRKL may affect cell phenotypes such as migration.

As discussed in 8.3.1, BCAR1 may be having an effect in the endothelium or smooth muscle cells. Immunohistochemistry assays could be used for further research here; if human plaque tissue could be obtained, the location of BCAR1 expression specifically within the plaque could be detected, giving further information about the role of the protein in plaque formation.

It would be desirable to assess the difference between wild-type and mutant BCAR1 in the absence of the endogenous protein, to increase power to detect differences between the wild-type and mutant protein. An siRNA knockdown was tested here to knock down endogenous expression of BCAR1, but successful knockdown could not be obtained, so this method was not used for assays. If knockdown could be successfully obtained, future work could carry out assays like those used here to remove endogenous BCAR1 and increase the relative effect of the P76S amino acid change on signalling and cell function.

Expressing vectors in cell lines to compare protein forms is a method with a degree of artificiality that does not truly represent the situation in the genome. One alternative to this was the idea of genotyping many primary cell samples, but as discussed, the other genomic differences between samples introduce complications.

An alternative that circumvents the problems of genotyping and expression vectors is the use of gene editing via clustered regularly-interspaced short palindromic repeats (CRISPR)[312]. Using CRISPR on primary cell lines would allow selective editing of a single base such as that of rs1035539, changing only the amino acid of interest and keeping all other factors the same. As CRISPR technology improves and can be conducted more routinely, it is likely to be of great value for protein assays such as those involving BCAR1.

9 Discussion

9.1 Overview

Studying the genetic basis of carotid intima-media thickness allows the discovery of loci involved in the development of atherosclerosis. The hypothesis of this thesis was that genetic variants which contribute to the formation and progression of atherosclerosis can be investigated by studying the surrogate variable intima-media thickness, and that by verifying the genetic associations with carotid IMT and identifying a potential molecular mechanism from variant to disease, novel pathways leading to atherosclerosis can be identified.

To test this, the *CFDP1-BCAR1-TMEM170A* locus on chromosome 16, previously implicated in IMT and CAD, was investigated in order to uncover the functional variation causing changes in IMT to identify and the gene or genes involved. Analysis of eQTL data identified *BCAR1* as the gene likely to be involved in the phenotype. Functional analyses identified a variant likely to be causing the observed association through gene regulation, and it was explored how this variant might be mediating this effect. Additional genotyping and meta-analysis suggested that the effect of the functional variant on common-carotid IMT and vascular events was present only (or at least most strongly) in women, suggesting there may be an interaction between BCAR1 pathways and sex-specific physiology, such as the relative amounts of sex hormones.

The complexity of genetic regulation at loci such as *CFDP1-BCAR1-TMEM170A* shows the necessity of investigating long-range genetic regulation when considering the effect of genetic loci on phenotypic traits and disease. The use of circular chromosome conformation capture to look for long-range interactions was explored and a protocol developed to study the *CFDP1-BCAR1-TMEM170A* locus, though ultimately final sequencing data could not be obtained. Examination of sequencing data prompted the investigation of a coding variant in *BCAR1*, which showed some associations with carotid plaque or IMT in two cohorts, though these were under different models and would require validation. Further work examined the effect of the variant on the protein itself, particularly in pathways related to atherosclerosis. Results from protein assays suggested that the risk allele may cause slower migration of cells, potentially affecting the remodelling of the vascular wall during the process of atherosclerosis.

9.2 Regulation of gene expression at the *CFDP1-BCAR1-TMEM170A* locus

The *CFDP1-BCAR1-TMEM170A* locus formed the basis of study for much of this thesis. After identification of risk loci by GWAS, the importance of functional characterisation is increasingly being recognised so as to understand the biological mechanisms behind a locus's association[341]. Only a small number of GWAS loci have been studied to successfully characterise the functional variation. For example, Musunuru and colleagues studied a SNP associated with LDL-cholesterol and MI, present at a gene-dense locus between two genes *PSRC1* and *CELSR2*. Experiments showed the functional SNP to alter a binding site for the transcription factor C/EBP, resulting in altered *SORT1* expression, and modulating hepatic VLDL secretion[220]. Zhou and colleagues studied a locus near the gene *HHIP*, identifying SNPs that lie in an enhancer upstream of *HHIP* which modifies the gene's expression through chromatin looping. One SNP was found to affect binding of the Sp3 transcription factor, lowering gene expression[342].

The *CFDP1-BCAR1-TMEM170A* locus, identified by Gertow and colleagues as being associated with carotid IMT and CAD, was studied here in the same way, with the aim of identifying potentially novel genes in the pathogenesis of atherosclerosis.

Bioinformatics tools were used to find SNPs in LD with the lead SNP, and to annotate these for regulatory marks that signify areas involved in gene expression. While genomic annotation projects like ENCODE are making increasing amounts of information available to assess variants for regulatory potential, the lack of definitive tools to consolidate this information can leave uncertainties in what weighting is given to information to rank variants. A degree of subjectivity was therefore required in compiling a shortlist of six candidate SNPs for functional analysis. Two newer tools for variant ranking include GWAVA and CADD, which use machine learning methods to prioritise non-coding (or coding and non-coding) variations with genomic and epigenomic annotations[135,343]. A more conclusive tool would need to bring together all relevant annotation resources and be continuously updated.

The candidate SNPs were first tested for allele-specific protein binding using EMSAs, in which only the lead SNP rs4888378 was identified as causing allele-specific protein binding. Multiplex competitor assays further probed this relationship using binding sequences of known transcription factors, indicating that the protein binding was a member of the FOXA family, a subfamily of hepatocyte nuclear factors involved in regulation of metabolism and glucose homeostasis[277,280]. FOXA1 and FOXA2 are thought to bind enhancers with FOXA motifs and open compacted chromatin

through DNA demethylation and H3K4 methylation[279,344]. It is not immediately clear how FOXA proteins might be involved in the phenotype considering the later implication of BCAR1, but it should be noted that a risk locus near the *BCAR1* gene is associated with type 2 diabetes and β-cell function in glucose metabolism[345,346]. It would next be ideal to verify the identity of the binding protein using a supershift EMSA assay, but a suitable EMSA-verified antibody that bound to control bands could not be identified, leaving this transcription factor yet to be verified, and precluded the use of ChIP to verify *in vivo* binding.

After the effect of this SNP on DNA-protein interactions had been characterised, luciferase assays were used to examine its effect on gene expression. The introduction of a single base change allows assessment of the effect of a SNP in isolation from other SNPs in LD, and these assays have previously been used to find functional SNPs in cardiovascular phenotypes[220,347]. Expression was found to vary according to the length of fragment around the SNP that was inserted into the reporter vector, suggesting the possible effect of a binding repressor. When the sequence was cropped to avoid predicted repressor binding elements, the protective A allele produced higher expression than the risk G. This fits with the eQTL data, which showed higher expression of *BCAR1* in the presence of the G allele, and suggests the SNP could be directly causing this effect. At the time of analysis *BCAR1* had not yet been implicated as the gene of interest, so the reporter vector contained the SV40 promoter, as other published studies have used when assaying the effect of a SNP on expression[220]. Further work would clone the *BCAR1* promoter into the reporter vector to assess the effect directly on this promoter.

The effect of fragment length on expression highlights the importance of all the interacting sequence elements around the SNP. An *in vitro* assay like the luciferase reporter cannot perfectly reflect the true state of gene expression *in vivo*, especially as promoters may be interacting with distant regulatory elements[287]. It is here that genome-editing techniques such as CRISPR, discussed in section 9.4, may be useful in measuring the true effect of a variant on gene expression.

In addition to other bioinformatics sources, eQTL data is also extremely valuable in the analysis of loci like *CFDP1-BCAR1-TMEM170A*. Risk-associated variants that do not alter the coding region of a gene are generally thought to exert their effects through regulation of expression[348], and gene expression itself is useful for the study of regulatory variation as it is the first step by which regulatory variants are likely to perturb molecular pathways. Gertow and colleagues found the lead SNP to correlate with *TMEM170A*, *BCAR1* and *LDHD* (the latter two with only nominal significance)

using expression data from a biobank with relevant tissues (aortic intima-media, aortic adventitia and carotid plaque).

Further analysis here used the gene expression dataset GTEx to look at SNP-expression relationships in artery tissues, and in this dataset only the *BCAR1* association was convincing, being present in more than one tissue and remaining after correction for multiple testing. The larger sample sizes and small p-values ($p < 10^{-10}$ in tibial and aortic artery) make a convincing case for *BCAR1* being the gene upon which the functional variant is acting, although there have also been cases of enhancers acting upon multiple genes[349].

In another approach to examine the effect of variants on chromatin structure, using the Gilad/Pritchard eQTL browser, two dsQTLs at the locus were found that affect local DNase-I sensitivity. EMSAs on these two SNPs did not show any protein-binding bands for the alleles of both SNPs, indicating this effect may not be directly through altering the binding affinity of transcription factors.

To look further at the effect of the locus on IMT phenotypes, the additional cohort PLIC was genotyped in order to study IMT and the related phenotype of IMT progression, which may provide more information than a single-timepoint variable. IMT progression has been shown to be a predictor of stroke[245], but recently published results from the PROG-IMT collaboration did not find an association between IMT progression and vascular risk[350].

The association between the lead SNP and baseline was not replicated, perhaps due to a lack of power brought about by smaller study size and inadequate assay call rate, or to the different background characteristics of the IMPROVE and PLIC cohorts. IMT progression was also not seen to have a significant association, but stratification by sex showed that women with the lower-risk allele had slower progression of IMT. To investigate this possible effect of sex on IMT phenotypes, a meta-analysis was carried out on common-carotid IMT in five cohorts, in which it was found that the effect of the A allele on thinner IMT was seen in women but not men. This effect of sex was also seen on vascular events.

This raised the question of how difference between the sexes interacts with the effect of the variant on IMT. Oestrogen was implicated as an involved molecule, because of the cardiovascular risk factors, only female hormonal status is known to be sex-specific[248]. Oestrogen is one of the many components of BCAR1 signalling pathways, with oestrogen treatment causing BCAR1 to transiently

associate with oestrogen receptor alpha (ERα)[351]. It also has a potential role in atherosclerosis. Oestrogen is thought to have a protective role in regards to blood lipids[256], but it also increases fibrinolytic potential[257], which is a marker of lower CVD risk[258].

Previous studies have also directly looked at the effect of oestrogen on carotid IMT. Discrepancies exist for the effect of oestrogen therapy on IMT, but this may be due to study heterogeneity. For example, in a study in 2006 on two post-menopausal oestrogen therapies, it was found that both therapies were associated with increased progression of CC-IMT[352], but a later study found longer duration of oestrogen therapy to be associated with slower IMT-progression[353]. Recent research indicates that oestrogen replacement therapy does decrease the progression of IMT, but in a time-dependent manner, being true only when it is initiated within 6 years after menopause[354].

It is also interesting to note that genetic variants in oestrogen receptors have been implicated in CVD risk in men: a dinucleotide repeat polymorphism in an ER regulatory region is associated with severity of CAD in men[355], with the longer repeat being associated with increased narrowing and formation of plaque. A male patient with a non-functional ER protein caused by a premature stop codon experienced early atherosclerosis despite low LDL levels[356].

The *CFDP1-BCAR1-TMEM170A* locus is a strong example of an association signal that spans a region without a clear candidate gene. Many genes are present close to the lead SNP, and even so it is known that regulatory elements can be far from their target genes[357]. It is therefore important to use methods to study long-range interactions important for genetic regulation. Circular chromosome conformation capture (4C) was planned for this thesis because it allows loci to be studied without bias to find out what regions they interact with.

The two *BCAR1* promoters and the proposed functional SNP were chosen for analysis; one plausible hypothesis is that the SNP is present in an enhancer which physically interacts with a promoter, with the SNP disrupting the enhancer's effect on gene expression, and having a subsequent effect on atherosclerosis development. Although all the stages of 4C design and development were completed, sequencing results could unfortunately not be obtained within the time limitations of this thesis, so data could not be produced regarding the long-rang interactions of these genetic locations. With the strategy, viewpoints and primers designed for this locus and 4C libraries prepared, the foundations are present for the method to be taken forward in future.

9.3 Coding variation at the *CFDP1-BCAR1-TMEM170A* locus

A large section of this thesis focused on the regulatory aspects of genetic variation in IMT, but variants that lie in coding regions can also provide a great deal of information about the impact of genes on a phenotype. These variants may affect the protein structure through substitution of amino acids, frame shift or termination of the protein, or through alteration of splice site. Variants occur less frequently in coding regions than non-coding regions[324], and occur at a lower MAF on average[358], so the number of common coding variants at a locus would be expected to be small. Nevertheless, searching exome sequence data here showed a common SNP in *BCAR1*, rs1035539, which codes for a proline/serine change in the proline-rich domain of the protein.

It was first necessary to consider whether the SNP was responsible for the association with IMT found by Gertow and colleagues, since as a SNP altering protein structure, it is more likely to alter phenotype than other SNPs[321]. Its absence from the 1000 Genomes panels used for LD calculations meant it had to be genotyped in IMPROVE to calculate LD with the lead SNP, which showed there to be no correlation between the two ($r^2 = 0.005$).

However, the SNP was found to be associated with carotid plaque under a dominant model. If a true association, the suggestion of independent signal associated with a similar carotid phenotype backs up the original association, and also indicates the possibility that disturbance of *BCAR1* affects plaque and IMT. This reinforces the hypothesis that the functional regulatory SNP is affecting *BCAR1*.

Further research on this SNP involved genotyping in PLIC, where an association with IMT was observed, but not with plaque. These results would require additional validation to confirm whether an association is truly present. If true, these slightly different results may be influenced by different phenotype definitions, and demonstrate one of the issues with IMT as a phenotype. As measurements can be taken at multiple segments of the carotid tree, and presented in different ways (such as average or maximum values), a large number of variables can be produced, causing problems of multiple testing if using all for analysis. There is no one definitive variable, although measures of common-carotid IMT are often used and have been suggested to be of equal use to measuring all the segments[83].

A surprising result found for the coding SNP was that stratifying the two cohorts by sex showed the observed associations to be present in men but not women, which was opposite to the relationship seen with the regulatory lead SNP, seen only in women. As these findings were true only under

different genetic models in men (additive and recessive), further genotyping and association analysis would be required to determine whether there is truly a sex-specific effect. Nevertheless, if the effect seen is true it hints at different mechanisms of effect of the regulatory and coding SNPs. To alter a protein's function and to alter its total expression may result in discrete effects; a protein with a damaging effect may interact with molecular pathways in a different way to a different amount of the wild-type protein.

To investigate how the SNP affects the protein itself, assays were carried out to look at the effect of the SNP on BCAR1 function. An expression vector was created for the "mutant" rs1035539 protein, and wild-type and mutant vectors were transfected into COS and HEK293 cells. The same was done to create viral expression vectors for infection of HUVECs. Assays were designed to examine phosphorylation and cell behaviour, since phosphorylation is vital to the function of BCAR1. No difference was seen between wild-type and mutant proteins; however, a small difference was observed in cell migration, with cells with the mutant (risk) form migrating more slowly.

How might the amino acid cause this change in migration? One theory is that the effect is mediated through the adaptor role of BCAR1. Proline-rich domains in proteins often play a role in binding, particularly in protein complexes[331]. The proline/serine change may modulate recruitment of other proteins. Mutations in proline-rich regions have been shown to affect binding; for example, proline/glutamine mutations in the proline-rich region of $p22^{phox}$, a NADPH oxidase subunit, impair binding of Nox enzymes and inhibit reactive oxygen production[359].

Further work by the UCL Cardiovascular Biology and Medicine group has suggested an effect on phosphorylation of CRKL, an adapter protein that binds to BCAR1[340]. This is a promising route for further study to investigate whether this is a true effect, and whether it contributes to changes in atherosclerosis and IMT. Further work would also look at the effect of the mutation in smooth muscle cells in addition to endothelial cells, as these are also a principal component of plaque in which BCAR1 plays a role: its presence and phosphorylation status are important for the migratory response in VSMCs[141].

9.4 Conclusions and future directions

Using several approaches to explore the *CFDP1-BCAR1-TMEM170A* locus, the likely gene through which the association is acting has been identified, supported by evidence from regulatory and coding studies. *In vitro* studies have enabled a functional SNP to be proposed, which may exert its

effect through differential binding of a FOXA protein and consequent differential expression of *BCAR1*. Meta-analysis revealed a female-specific association with common-carotid IMT, suggesting possible involvement of oestrogen, a key mediator in BCAR1 pathways. A coding variant was found to correlate with related IMT phenotypes under certain inheritance models, but the results differed between cohorts and require validation. Protein studies on the corresponding wild-type and mutant forms of the variant indicated differences in endothelial migration.

Results in this thesis point towards *BCAR1* as the gene most likely to be involved in the IMT phenotype. In addition to this gene's suitability as a candidate for atherosclerosis-related phenotypes, the suggested presence of an independent signal within the *BCAR1* coding region gives more power to its proposal as the causal gene. The coding variant studied for its relationship with plaque appears to show interesting effects on migration in cell-based assays. Validation of the association and further research on the variant's functional effects will demonstrate whether there is strong evidence for *BCAR1* as the causal gene, and whether the mechanism is indeed through cell migration.

However, certain considerations should be taken into account when looking at other studies studying these phenotypes. The *CFDP1-BCAR1-TMEM170A* locus did not achieve genome-wide significance for association with CAD in the CARDIoGRAM consortium, although significance on a non-genome-wide level was observed (OR = 1.03, p = 0.011)[117]. As the *a priori* premise was to examine this SNP only, this suggests a true association is present, but raises the question of why a stronger association was not seen.

The CARDIoGRAM consortium is a meta-analysis of 22 GWAS studies (total sample size 86,995); it is possible that inter-study heterogeneity weakened the strength of the association. It should also be noted that the association tested in CARDIoGRAM is with CAD, the endpoint of the proposed effect pathway of SNP affecting eQTL, which in turn affects the intermediate trait (IMT), which influences risk of CAD. At each stage power is lost to detect a true association.The rs4888378 locus did also not come up as a genome-wide significant hit in the CHARGE consortium (total sample size 31,211), looking at associations with IMT[118]. Results may have been affected by the differences in baseline characteristics: subjects in CHARGE were taken from the general population and ranged from 44-76 years old, whereas those in IMPROVE were of high cardiovascular risk, and 54-79 years of age. There is also the possibility that a less significant association is present that did not reach genome-wide significance. The locus studied in this thesis may be an example of the winner's curse, in which initial

studies tend to overestimate the genetic effect size of a variant[360]. It could be the case that the BCAR1 protein is indeed involved in atherosclerosis-related pathways, but that selection pressure has not allowed variants with a strong effect on phenotype to become common. In this case, a strong GWAS signal would not be easily detectable within the gene.

To establish whether this is the case, further mechanistic studies should be carried out to determine the effect of BCAR1 mutants on proteins and cell function. For example, more work on wild-type and mutant BCAR1 protein will be of particular interest, especially on downstream proteins such as CRKL. Greater replicate numbers for experiments such as the migration assay will allow increased confidence in whether the variant truly affects these phenotypes. Further plans also involve genotyping multiple samples of primary HUVECs and other cells for the coding variant, allowing selection of AA and GG homozygotes for use in assays. This presents a method of assaying variant effect without disturbing cell behaviour with viral infection, but many samples of similar background will be required for each assay to reduce the effect of other background genetic factors.

Further assays could aim to elucidate further the role of the BCAR1 protein in atherosclerosis and vascular tissues. Immunohistochemistry could be performed on human vascular tissue samples: this could determine the location of the protein within vascular tissues, particularly whether it is more prominent in endothelial or smooth muscle cells. It could be particularly interesting to compare protein quantity and location between samples with and without the presence of plaque, in addition to comparing male and female samples, and those of different rs1035539 genotype.

Exploration of effects on BCAR1 pathways should also focus on the possible involvement of oestrogen. Protein assays could examine the effect of oestrogen treatment on the behaviour of cells with wild-type and mutant protein, and if there is a difference, whether the oestrogen receptor itself is involved in this process. Cells from both male and female donors should be used and compared. This is particularly valuable in light of recent studies, which have shown that transcriptional differences exist between male and female endothelial cells. One study found migration and proliferation were higher in female cells[361], while another showed female HUVECs to display a stronger transcriptional response to shear stress[362].

The strong linkage disequilibrium at the locus presented several limitations for the conclusive identification of the functional variant. Blocks of LD in the genome present obstacles in association

studies, as variants in complete LD will show identical correlations with traits. The large number of SNPs in strong LD with the lead SNP at the *CFDP1-BCAR1-TMEM170A* locus made this signal unusually challenging in the selection of candidate regulatory SNPs. Where loci are well-covered by genotyping arrays, genotype imputation can be carried out to predict genotypes that have not been directly assayed[363]. Imputation uses genotyped variants and known haplotype patterns to predict the most likely genotypes at untyped variants, allowing fine-mapping of the signal. Denser genotyping of the locus may allow imputation to refine the signal at this locus. It has already been shown here that at least two independent signals associated with carotid plaque and IMT exist. Others may exist, which would be expected to affect *BCAR1*, perhaps through promoter or other enhancer activity. Genotyping cohorts of ethnic groups other than Europeans should be carried out to take advantage of different patterns of LD, producing different association signals surrounding functional variants. African populations would be of particular value due to their shorter haplotypic blocks[236]. LD patterns in central Europeans and Yoruban Africans at part of the *CFDP1-BCAR1-TMEM170A* locus are shown in Figure 109, where it can be seen that haplotype blocks are less extensive in the Yoruban population.

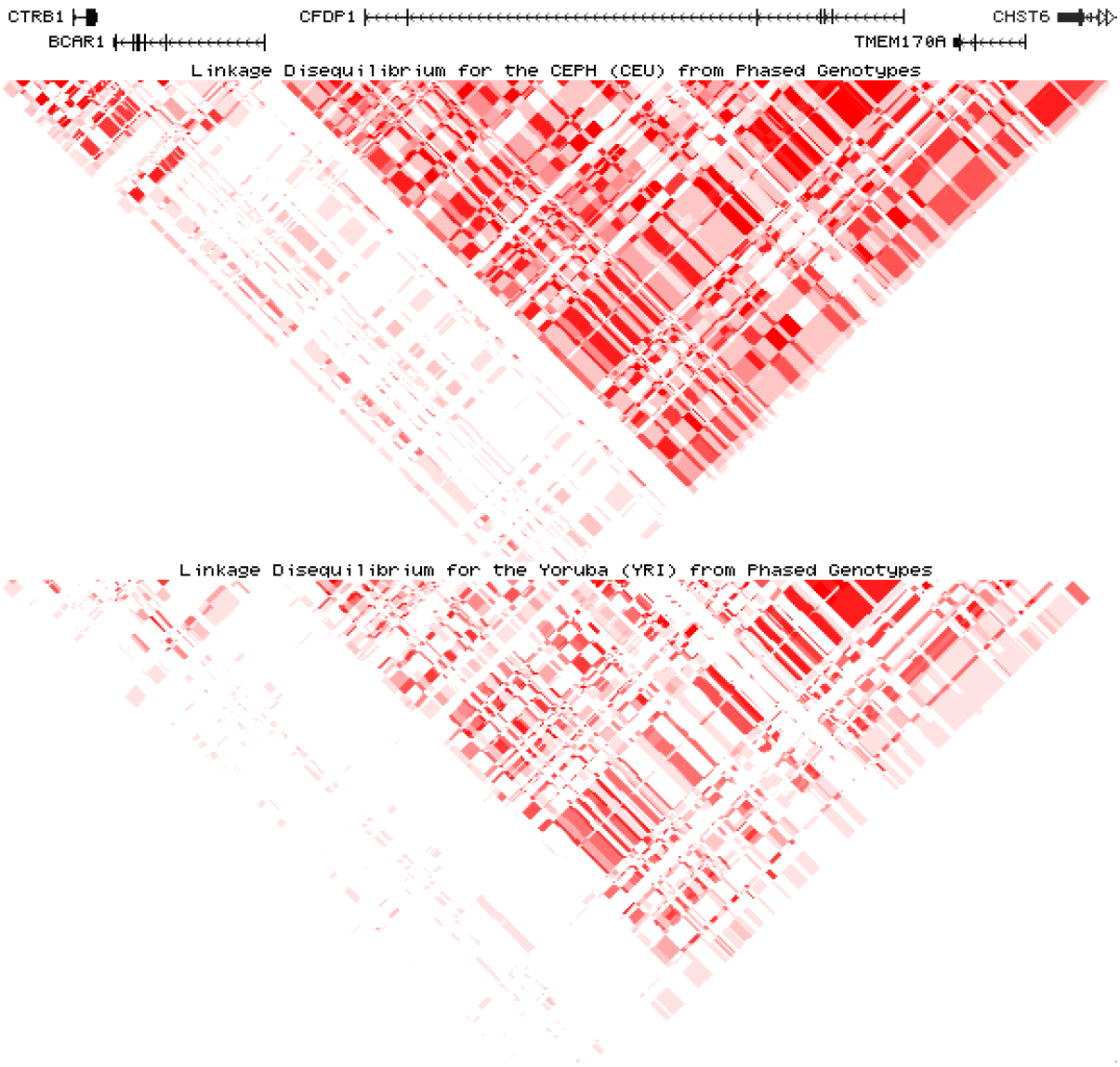

Figure 109: LD structure at the *CFDP1-BCAR1-TMEM170A* locus in central Europeans (CEU, above) and Yoruban Africans (YRI). Data from HapMap[96], visualised on UCSC Genome Browser[127]. LD at the locus can be seen to be less extensive in the Yoruban population than the European population, making these populations particularly valuable for fine-mapping of association signals.

Functional assays on the regulatory effects of SNPs at the locus suggested that a FOXA protein was binding to rs4888378, affecting the action of an enhancer and therefore impacting on expression of a gene, proposed to be *BCAR1*. To test this hypothesis, the proteins in the FOXA family could be knocked down using siRNA, and the impact on *BCAR1* expression, or on protein binding via EMSA, observed. Alternatively, a plasmid containing *FOXA* could be overexpressed and the same outcomes measured.

As discussed during this thesis, the selection of relevant cells is important for experiments using cell culture. For example, here HUVECs were used to study vascular phenotypes. However, it is now

becoming more feasible to use alternative methods of obtaining suitable cells, such as induced pluripotent stem cells (iPSCs)[364]. Using retroviral expression of pluripotency-associated transcription factors, such as the Oct-3/4 and Sox families, Nanog and Lin28, human somatic calls can be reprogrammed to become pluripotent, providing the possibility to generate any type of cell in the body.

In additional to providing a source of pluripotent stem cells without using human embryonic stem cells, which require destruction of human embyros[365], iPSCs are particularly useful as they can be generated from cells of a particular genotype. They could therefore be used to provide genotype-matched cells of different types. For example, the signalling and wound healing assays in chapter 8 could be carried out in various vascular cells with the same genotype, to remove the possibility of background genetic variation affecting results. iPSCs have been used to generate human vascular smooth muscle cells and endothelial for functional assays and disease modelling[366,367]. iPSCs would also allow for the generation of subject-specific cell lines; for example, assays could be carried out on cells generated from groups of patients with particularly high or low IMT measures.

Functional assays like EMSA and luciferase reporter assays are valuable for detecting allele-specific effects in isolation from SNPs in LD, but they cannot take into account interactions with the rest of the genome or chromatin state. To study these interactions work on chromosome conformation capture should be continued, using the *CFDP1-BCAR1-TMEM170A* locus and other regions of interest, such as the IMT/plaque loci identified by the CHARGE consortium.

To directly assay the effect of proposed functional SNPs, genome editing techniques could be used. Designer zinc finger nucleases (ZFNs)[368] and transcription activator-like effector nucleases (TALENs)[369] are two methods that have been used for targeted genome editing, but particular promise is shown by the clustered regularly-interspaced short palindromic repeats (CRISPR) system[312], which, compared to previous systems, is affordable and relatively easy to engineer.

The CRISPR/Cas9 genome editing system (CRISPR combined with CRISPR-associated nuclease 9) is based on a prokaryotic immune system protecting against foreign genetic elements, in which CRISPR elements recognise and cut offending DNA sequences[370]. Genome editing using CRISPR uses a guide RNA that binds to the sequence at the locus being targeted. Upon binding, the Cas9 nuclease cuts the DNA strands, and non-homologous end joining occurs, resulting in deletion of the target sequence. Alternatively, a DNA template can be introduced, containing sequences homologous to

the region surrounding the target site. After cutting of the double-stranded DNA, homologous recombination can occur, inserting the target sequence into the gap[312] (Figure 110).

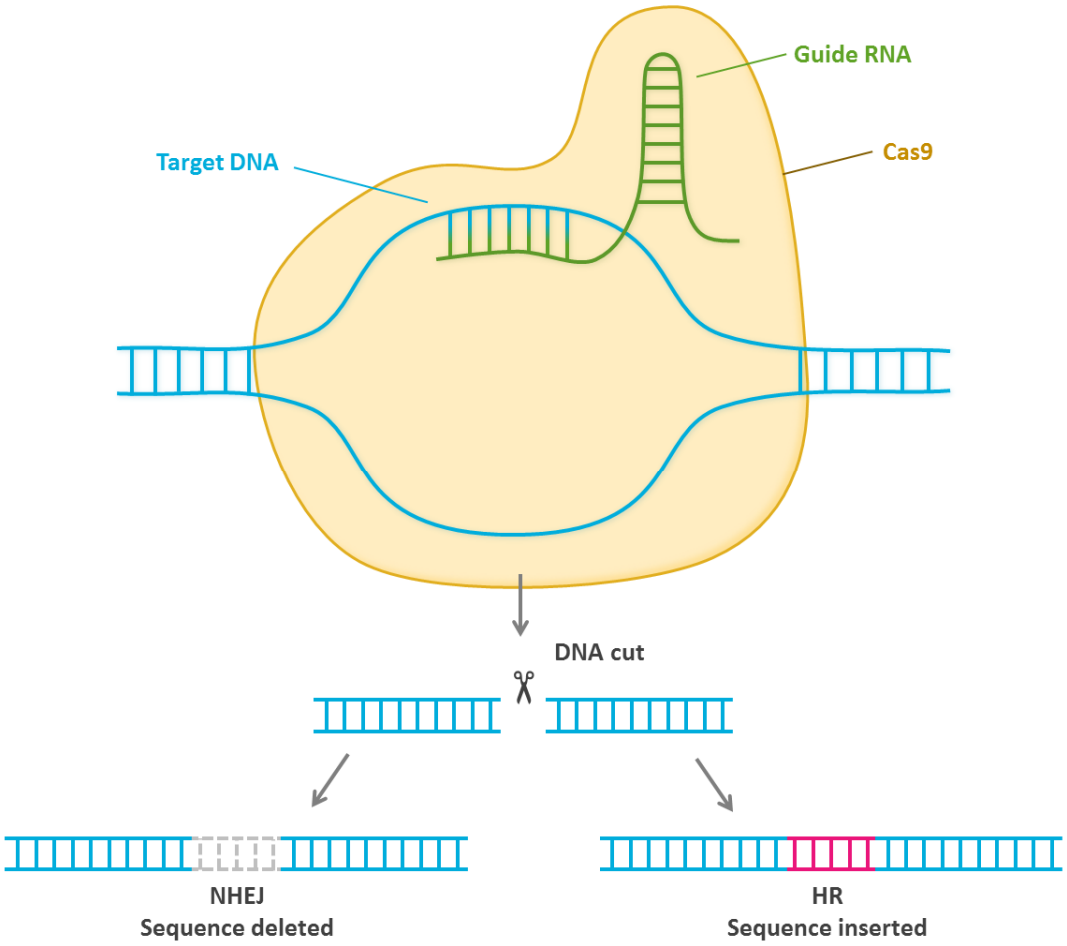

Figure 110: CRISPR for targeted genome editing. A guide RNA is generated that is complementary to the target genomic sequence. This RNA guides the Cas9 enzyme (CRISPR-associated nuclease 9) to the target sequence. Cas9 cuts both strands of the DNA sequence on either side of the target sequence. The ends may join by non-homologous end joining (NHEJ), resulting in a deletion of the target sequence. Alternatively, if a DNA template with flanking homologous sequences has been introduced, this template will be inserted into the gap with homologous recombination (HR).

To directly assay the effect of proposed functional SNPs, the CRISPR system could be used to edit only the SNP of interest in the genome. This would allow cells to be assayed without using artificial expression vectors or otherwise modify the cells' natural state[312]. CRISPR techniques are still new, but have already been used to study the effects of SNPs and cell culture systems in a similar method to that proposed. Claussnitzer et al examined the obesity-linked *FTO* locus, using modelling methods to propose a causal SNP, and testing its effect on protein-binding and expression using EMSA and luciferase reporter assays[371]. They then used CRISPR to validate this SNP in preadipocytes, finding it to affect *IRX3* and *IRX5* expression with a downstream effect on mitochondrial thermogenesis. Such a method would be useful in analysis of both regulatory and coding SNPs such as those studied here,

preventing the need for transfection and infection protocols that may itself modify cell behaviour. For example, the assay proposed above involving selection of cells with AA and GG coding variant genotypes, could instead use CRISPR to directly edit the SNP of interest to produce the mutant protein, eliminating the need to control for background factors.

Genome editing could also be carried out to knock out enhancer regions such as the proposed enhancer at the site of rs4888378. Expression of *BCAR1* and other nearby genes could then be measured in order to evaluate the importance of the enhancer in gene expression. Similar methods have previously been used with success; for example, Canver and colleagues used CRISPR to investigate an enhancer for the gene *BCL11A*, containing variation associated with foetal haemoglobin levels. Deletion of the enhancer was found to greatly reduce *BCL11A* expression, and more finely targeted high-throughput mutagenesis revealed critical features of the enhancer[372]. Editing of enhancers to examine expression could be expanded by using a microarray to analyse gene expression on a global level[373], to assess whether this has effects on gene expression other pathways.

The methods above can be carried out in cell culture, but further work could use animal models to examine the effect of enhancer disruption on a larger scale. For example, Sur and colleagues examined a putative *MYC* enhancer containing a SNP strongly associated with colorectal cancer mortality. They deleted the enhancer in mice using Cre-Lox recombination and found a resultant reduction in *MYC* expression and a marked resistance to intestinal tumourigenesis[374].

BCAR1 is involved in numerous processes in the body. In drug development, drugs with many roles are more likely to have off-target effects, a major problem in the development process[375]. Future therapeutic agents to target BCAR1 with aim of reducing atherosclerosis are therefore likely to be unviable. However, if further work could identify downstream proteins affected by the amino acid change and involved in atherosclerosis, these molecules could potentially be targets for reduction of risk. This study has shown how the use of multiple complementary tools can be used to identify novel pathways for complex diseases such as atherosclerosis from GWAS findings. Future use of such analyses on a larger scale is likely to help to realise the potential of genetic associations to identify novel therapeutic targets.

Bibliography

1. Huxley A. Along the Road: Notes and Essays of a Tourist. George H. Doran; 1925.

2. WHO. Global status report on noncommunicable diseases 2014. *World Health*. 2014;176.

3. World Health Organisation. Cardiovascular diseases (CVDs) fact sheet No 317 [Internet]. 2015 [cited 2016 Apr 1];Available from: http://www.who.int/mediacentre/factsheets/fs317/en/#.Vv5GGuFVbcE.mendeley

4. Roth GA, Forouzanfar MH, Moran AE, Barber R, Nguyen G, Feigin VL, et al. Demographic and epidemiologic drivers of global cardiovascular mortality. *New England Journal of Medicine*. 2015;372:1333–41.

5. Kathiresan S, Srivastava D. Genetics of human cardiovascular disease. *Cell*. 2012;148:1242–1257.

6. Deloukas P, Kanoni S, Willenborg C, Farrall M, Assimes TL, Thompson JR, et al. Large-scale association analysis identifies new risk loci for coronary artery disease. *Nature Genetics*. 2013;45:25–33.

7. Weisfeldt ML, Zieman SJ. Advances in the prevention and treatment of cardiovascular disease. *Health Affairs*. 2007;26:25–37.

8. Lusis AJ. Atherosclerosis. *Nature*. 2000;407:233–241.

9. Ouriel K. Peripheral arterial disease. *Lancet*. 2001;358:1257–1264.

10. Hansson GK, Hermansson A. The immune system in atherosclerosis. *Nature immunology*. 2011;12:204–212.

11. Tabas I, Williams KJ, Borén J. Subendothelial Lipoprotein Retention as the Initiating Process in Atherosclerosis: Update and Therapeutic Implications. *Circulation*. 2007;116:1832–1844.

12. Deanfield JE, Halcox JP, Rabelink TJ. Endothelial function and dysfunction: Testing and clinical relevance. *Circulation*. 2007;115:1285–1295.

13. Dharmashankar K, Widlansky ME. Vascular Endothelial Function and Hypertension: Insights and Directions. *Current Hypertension Reports*. 2010;12:448–455.

14. Messner B, Bernhard D. Smoking and Cardiovascular Disease: Mechanisms of Endothelial Dysfunction and Early Atherogenesis . *Arteriosclerosis, Thrombosis, and Vascular Biology* . 2014;34:509–515.

15. Sena CM, Pereira AM, Seiça R. Endothelial dysfunction — A major mediator of diabetic vascular disease. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*. 2013;1832:2216–2231.

16. Greaves DR, Gordon S. The macrophage scavenger receptor at 30 years of age: current knowledge and future challenges. *Journal of Lipid Research*. 2009;50:S282–S286.

17. Louis SF, Zahradka P. Vascular smooth muscle cell motility: From migration to invasion. *Experimental and Clinical Cardiology*. 2010;15.

18. Allahverdian S, Pannu PS, Francis GA. Contribution of monocyte-derived macrophages and smooth muscle cells to arterial foam cell formation. *Cardiovascular Research*. 2012;95:165–

172.

19. Newby AC, Zaltsman AB. Fibrous cap formation or destruction--the critical importance of vascular smooth muscle cell proliferation, migration and matrix formation. *Cardiovascular research*. 1999;41:345–360.

20. Webb NR. Getting to the core of atherosclerosis. *Nature medicine*. 2008;14:1015–6.

21. Hermiller JB, Tenaglia AN, Kisslo KB, Phillips HR, Bashore TM, Stack RS, et al. In vivo validation of compensatory enlargement of atherosclerotic coronary arteries. *The American Journal of Cardiology*. 1993;71:665–668.

22. Castelli WP. Epidemiology of coronary heart disease: The Framingham study. *The American Journal of Medicine*. 1984;76:4–12.

23. Lloyd-Jones DM, Larson MG, Beiser A, Levy D. Lifetime risk of developing coronary heart disease. *The Lancet*. 1999;353:89–92.

24. Jousilahti P, Vartiainen E, Tuomilehto J, Puska P. 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*. 1999;99:1165–1172.

25. Mensah GA, Mendis S, Greenland K, MacKay J. The atlas of heart disease and stroke. World Health Organization; 2004.

26. Balarajan R. Ethnic differences in mortality from ischaemic heart disease and cerebrovascular disease in England and Wales. *BMJ (Clinical research ed)*. 1991;302:560–564.

27. Lloyd-Jones D, Adams R, Carnethon M, De Simone G, Ferguson TB, Flegal K, et al. Heart Disease and Stroke Statistics—2009 Update: A Report From the American Heart Association Statistics Committee and Stroke Statistics Subcommittee. *Circulation*. 2009;119:480–486.

28. Hawe E, Talmud PJ, Miller GJ, Humphries SE. Family History is a Coronary Heart Disease Risk Factor in the Second Northwick Park Heart Study. *Annals of Human Genetics*. 2003;67:97–106.

29. Lloyd-Jones DM, Nam B-H, D'Agostino, Sr RB, Levy D, Murabito JM, Wang TJ, et al. Parental Cardiovascular Disease as a Risk Factor for Cardiovascular Disease in Middle-aged Adults. *JAMA*. 2004;291:2204.

30. Alwan A. Global status report on noncommunicable diseases 2010. World Health Organization; 2011.

31. Ockene IS, Miller NH. Cigarette Smoking, Cardiovascular Disease, and Stroke: A Statement for Healthcare Professionals From the American Heart Association. *Circulation*. 1997;96:3243–3247.

32. Yamaguchi Y, Matsuno S, Kagota S, Haginaka J, Kunitomo M. Oxidants in cigarette smoke extract modify low-density lipoprotein in the plasma and facilitate atherogenesis in the aorta of Watanabe heritable hyperlipidemic rabbits. *Atherosclerosis.* 2001;156:109–117.

33. Meade TW, Imeson J, Stirling Y. Effects of changes in smoking and other characteristics on clotting factors and the risk of ischaemic heart disease. *The Lancet*. 1987;330:986–988.

34. Benowitz NL. The Role of Nicotine in Smoking-Related Cardiovascular Disease. *Preventive Medicine*. 1997;26:412–417.

35. Shiroma EJ, Lee I-M. Physical activity and cardiovascular health: lessons learned from epidemiological studies across age, gender, and race/ethnicity. *Circulation*. 2010;122:743–752.

36. Marmot M, Brunner E. Alcohol and cardiovascular disease: the status of the U shaped curve. *BMJ (Clinical research ed)*. 1991;303:565–568.

37. Lucas DL, Brown RA, Wassef M, Giles TD. Alcohol and the cardiovascular system: Research challenges and opportunities. *Journal of the American College of Cardiology*. 2005;45:1916–1924.

38. Gaziano JM, Buring JE, Breslow JL, Goldhaber SZ, Rosner B, VanDenburgh M, et al. Moderate alcohol intake, increased levels of high-density lipoprotein and its subfractions, and decreased risk of myocardial infarction. *The New England Journal of Medicine*. 1993;329:1829–34.

39. Mukamal KJ, Rimm EB. Alcohol's Effects on the Risk for Coronary Heart Disease. *Alcohol Research & Health*. 2001;25:255–261.

40. Holmes M V, Dale CE, Zuccolo L, Silverwood RJ, Guo Y, Ye Z, et al. Association between alcohol and cardiovascular disease: Mendelian randomisation analysis based on individual participant data. *BMJ*. 2014;349:g4164–g4164.

41. Whitworth JA. 2003 World Health Organization (WHO)/International Society of Hypertension (ISH) statement on management of hypertension. *Journal of Hypertension*. 2003;21:1983–1992.

42. Alexander RW. Hypertension and the Pathogenesis of Atherosclerosis: Oxidative Stress and the Mediation of Arterial Inflammatory Response: A New Perspective. *Hypertension*. 1995;25:155–161.

43. World Health Organisation. A global brief on hypertension: silent killer, global public health crisis. 2013.

44. Di Cesare M, Bennett JE, Best N, Ezzati M, Stevens GA, Danaei G. The contributions of risk factor trends to cardiometabolic mortality decline in 26 industrialized countries. *International Journal of Epidemiology*. 2013;42:838–848.

45. Blood Pressure Lowering Treatment Trialists' Collaboration. Effects of ACE inhibitors, calcium antagonists, and other blood-pressure-lowering drugs: results of prospectively designed overviews of randomised trials. *The Lancet*. 2000;356:1955–1964.

46. Hubert HB, Feinleib M, McNamara PM, Castelli WP. Obesity as an independent risk factor for cardiovascular disease: a 26-year follow-up of participants in the Framingham Heart Study. *Circulation*. 1983;67:968–977.

47. Wang YC, McPherson K, Marsh T, Gortmaker SL, Brown M. Health and economic burden of the projected obesity trends in the USA and the UK. *The Lancet*. 2011;378:815–825.

48. Gaziano TA. Reducing The Growing Burden Of Cardiovascular Disease In The Developing World. *Health Affairs*. 2007;26:13–24.

49. Kotsis V, Stabouli S, Bouldin M, Low A, Toumanidis S, Zakopoulos N. Impact of obesity on 24-hour ambulatory blood pressure and hypertension. *Hypertension*. 2005;45:602–607.

50. Kahn SE, Hull RL, Utzschneider KM. Mechanisms linking obesity to insulin resistance and type 2 diabetes. *Nature*. 2006;444:840–846.

51. Kannel WB, McGee DL. Diabetes and Cardiovascular Disease. *JAMA*. 1979;241:2035.

52. Aronson D, Rayfield EJ. How hyperglycemia promotes atherosclerosis: molecular mechanisms. *Cardiovascular Diabetology*. 2002;1:1.

53. Brown MS, Goldstein JL. A receptor-mediated pathway for cholesterol homeostasis. *Science*. 1986;232:34–47.

54. Hokanson JE, Austin MA. Plasma Triglyceride Level is a Risk Factor for Cardiovascular Disease Independent of High-Density Lipoprotein Cholesterol Level: A Metaanalysis of Population-Based Prospective Studies. *Journal of Cardiovascular Risk*. 1996;3:213–219.

55. Byrne CD. Triglyceride-rich lipoproteins: Are links with atherosclerosis mediated by a procoagulant and proinflammatory phenotype? *Atherosclerosis*. 1999;145:1–15.

56. Després J-P, Lemieux I, Dagenais G-R, Cantin B, Lamarche B. HDL-cholesterol as a marker of coronary heart disease risk: the Québec cardiovascular study. *Atherosclerosis*. 2000;153:263–272.

57. Voight BF, Peloso GM, Orho-Melander M, Frikke-Schmidt R, Barbalic M, Jensen MK, et al. Plasma HDL cholesterol and risk of myocardial infarction: A mendelian randomisation study. *The Lancet*. 2012;380:572–580.

58. Manson JE, Hu FB, Rich-Edwards JW, Colditz GA, Stampfer MJ, Willett WC, et al. A Prospective Study of Walking as Compared with Vigorous Exercise in the Prevention of Coronary Heart Disease in Women. *New England Journal of Medicine*. 1999;341:650–658.

59. Tanasescu M, MF L, EB R, WC W, MJ S, FB H. Exercise type and intensity in relation to coronary heart disease in men. *JAMA*. 2002;288:1994–2000.

60. Teslovich TM, Musunuru K, Smith A V, Edmondson AC, Stylianou IM, Koseki M, et al. Biological, clinical and population relevance of 95 loci for blood lipids. *Nature*. 2010;466:707–13.

61. Willerson JT, Ridker PM. Inflammation as a cardiovascular risk factor. *Circulation*. 2004;109:II2-I10.

62. Buckley DI, Fu R, Freeman M, Rogers K, Helfand M. C-Reactive Protein as a Risk Factor for Coronary Heart Disease: A Systematic Review and Meta-analyses for the U.S. Preventive Services Task Force. *Annals of Internal Medicine*. 2009;151:483.

63. Oh J, Teoh H, Leiter LA. Should C-Reactive Protein Be a Target of Therapy? *Diabetes Care*. 2011;34:S155–S160.

64. C Reactive Protein Coronary Heart Disease Genetics Collaboration (CCGC), Wensley F, Gao P, Burgess S, Kaptoge S, Di Angelantonio E, et al. Association between C reactive protein and coronary heart disease: mendelian randomisation analysis based on individual participant data. *BMJ (Clinical research ed)*. 2011;342:d548.

65. Danesh J, Kaptoge S, Mann AG, Sarwar N, Wood A, Angleman SB, et al. Long-term interleukin-6 levels and subsequent risk of coronary heart disease: Two new prospective studies and a systematic review. *PLoS Medicine*. 2008;5:0600–0610.

66. Interleukin-6 Receptor Mendelian Randomisation Analysis (IL6R MR) Consortium. The interleukin-6 receptor as a target for prevention of coronary heart disease: a mendelian randomisation analysis. *Lancet*. 2012;379:1214–1224.

67. Kromhout D, Menotti A, Kesteloot H, Sans S. Prevention of Coronary Heart Disease by Diet and Lifestyle: Evidence From Prospective Cross-Cultural, Cohort, and Intervention Studies. *Circulation*. 2002;105:893–898.

68. Villareal DT, Miller B V, Banks M, Fontana L, Sinacore DR, Klein S. Effect of lifestyle intervention on metabolic coronary heart disease risk factors in obese older adults. *The American Journal of Clinical Nutrition*. 2006;84:1317–1323.

69. Cholesterol Treatment Trialists' (CTT) Collaborators. Efficacy and safety of cholesterol-lowering treatment: prospective meta-analysis of data from 90 056 participants in 14 randomised trials of statins. *The Lancet*. 2005;366:1267–1278.

70. Naci H, Brugts J, Ades T. Comparative Tolerability and Harms of Individual Statins: A Study-Level Network Meta-Analysis of 246 955 Participants From 135 Randomized, Controlled Trials. *Circulation: Cardiovascular Quality and Outcomes*. 2013;6:390–399.

71. Ellis JJ, Erickson SR, Stevenson JG, Bernstein SJ, Stiles RA, Fendrick AM. Suboptimal Statin Adherence and Discontinuation in Primary and Secondary Prevention Populations. *Journal of General Internal Medicine*. 2004;19:638–645.

72. Simon A, Megnien J-L, Chironi G. The Value of Carotid Intima-Media Thickness for Predicting Cardiovascular Risk. *Arteriosclerosis, Thrombosis, and Vascular Biology*. 2010;30:182–185.

73. Stein JH, Korcarz CE, Hurst RT, Lonn E, Kendall CB, Mohler ER, et al. Use of Carotid Ultrasound to Identify Subclinical Vascular Disease and Evaluate Cardiovascular Disease Risk: A Consensus Statement from the American Society of Echocardiography Carotid Intima-Media Thickness Task Force Endorsed by the Society for Vascular. *Journal of the American Society of Echocardiography*. 2008;21:93–111.

74. Nikic P, Savic M, Jakovljevic V, Djuric D. Carotid atherosclerosis, coronary atherosclerosis and carotid intima-media thickness in patients with ischemic cerebral disease: Is there any link? *Experimental and Clinical Cardiology*. 2006;11:102–106.

75. Geroulakos G, O'Gorman DJ, Kalodiki E, Sheridan DJ, Nicolaides AN. The carotid intima-media thickness as a marker of the presence of severe symptomatic coronary artery disease. *European Heart Journal*. 1994;15:781–785.

76. Amato M, Montorsi P, Ravani A, Oldani E, Galli S, Ravagnani PM, et al. Carotid intima-media thickness by B-mode ultrasound as surrogate of coronary atherosclerosis: Correlation with quantitative coronary angiography and coronary intravascular ultrasound findings. *European Heart Journal*. 2007;28:2094–2101.

77. van der Meer IM, Bots ML, Hofman A, del Sol AI, van der Kuip DAM, Witteman JCM. Predictive value of noninvasive measures of atherosclerosis for incident myocardial infarction: the Rotterdam Study. *Circulation*. 2004;109:1089–1094.

78. Polak JF, Pencina MJ, Pencina KM, O'Donnell CJ, Wolf PA, D'Agostino RB. Carotid-Wall Intima–Media Thickness and Cardiovascular Events. *New England Journal of Medicine*. 2011;365:213–221.

79. Bots ML, Hoes AW, Koudstaal PJ, Hofman A, Grobbee DE. Common Carotid Intima-Media

Thickness and Risk of Stroke and Myocardial Infarction: The Rotterdam Study. *Circulation*. 1997;96:1432–1437.

80. Lorenz MW, Markus HS, Bots ML, Rosvall M, Sitzer M. Prediction of Clinical Cardiovascular Events With Carotid Intima-Media Thickness: A Systematic Review and Meta-Analysis. *Circulation*. 2007;115:459–467.

81. Lorenz MW, Schaefer C, Steinmetz H, Sitzer M. Is Carotid intima media thickness useful for individual prediction of cardiovascular risk? Ten-year results from the Carotid Atherosclerosis Progression Study (CAPS). *European Heart Journal*. 2010;31:2041–2048.

82. Peters S a. E, den Ruijter HM, Bots ML, Moons KGM. Improvements in risk stratification for the occurrence of cardiovascular disease by imaging subclinical atherosclerosis: a systematic review. *Heart*. 2012;98:177–184.

83. Nambi V, Chambless L, He M, Folsom AR, Mosley T, Boerwinkle E, et al. Common carotid artery intima-media thickness is as good as carotid intima-media thickness of all carotid artery segments in improving prediction of coronary heart disease risk in the Atherosclerosis Risk in Communities (ARIC) study. *European Heart Journal*. 2012;33:183–190.

84. Plichart M, Celermajer DS, Zureik M, Helmer C, Jouven X, Ritchie K, et al. Carotid intima-media thickness in plaque-free site, carotid plaques and coronary heart disease risk prediction in older adults. The Three-City Study. *Atherosclerosis*. 2011;219:917–924.

85. Bots ML, den Ruijter HM. Should We Indeed Measure Carotid Intima-Media Thickness for Improving Prediction of Cardiovascular Events After IMPROVE? *Journal of the American College of Cardiology*. 2012;60:1500–1502.

86. Friedman MH, Bargeron CB, Deters OJ, Hutchins GM, Mark FF. Correlation between wall shear and intimal thickness at a coronary artery branch. *Atherosclerosis*. 1987;68:27–33.

87. Marks D, Thorogood M, Neil HAW, Humphries SE. A review on the diagnosis, natural history, and treatment of familial hypercholesterolaemia. *Atherosclerosis*. 2003;168:1–14.

88. Soutar AK, Naoumova RP. Mechanisms of disease: genetic causes of familial hypercholesterolemia. *Nature Clinical Practice Cardiovascular medicine*. 2007;4:214–225.

89. Lehrman MA, Schneider WJ, Südhof TC, Brown MS, Goldstein JL, Russell DW. Mutation in LDL receptor: Alu-Alu recombination deletes exons encoding transmembrane and cytoplasmic domains. *Science*. 1985;227:140–6.

90. Jarcho JA, McKenna W, Pare JAP, Solomon SD, Holcombe RF, Dickie S, et al. Mapping a Gene for Familial Hypertrophic Cardiomyopathy to Chromosome 14q1. *New England Journal of Medicine*. 1989;321:1372–1378.

91. Geisterfer-Lowrance AAT, Kass S, Tanigawa G, Vosberg H-P, McKenna W, Seidman CE, et al. A molecular basis for familial hypertrophic cardiomyopathy: a β cardiac myosin heavy chain gene missense mutation. *Cell*. 1990;62:999–1006.

92. Reich DE, Lander ES. On the allelic spectrum of human disease. *Trends in Genetics*. 2016;17:502–510.

93. Bush WS, Moore JH. Chapter 11: Genome-Wide Association Studies. *PLoS Computational Biology*. 2012;8.

94. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. *Nature*. 2009;461:747–53.

95. Human Genome Sequencing Consortium I. Finishing the euchromatic sequence of the human genome. *Nature*. 2004;431:931–945.

96. The International HapMap Consortium. A haplotype map of the human genome. *Nature*. 2005;437:1299–1320.

97. Manolio TA. Genomewide Association Studies and Assessment of the Risk of Disease. *New England Journal of Medicine*. 2010;363:166–176.

98. Cox MM, Doudna JA, O'Donnell M. Molecular Biology. New York: W H Freeman; 2011.

99. Kruglyak L. The road to genome-wide association studies. *Nature Reviews Genetics*. 2008;9:314–318.

100. Padmanabhan S, Hastie C, Prabhakaran D, Dominczak AF. Genomic approaches to coronary artery disease. *The Indian Journal of Medical Research*. 2010;132:567–78.

101. Jannot AS, Ehret G, Perneger T. P < 5 x 10-8 has emerged as a standard of statistical significance for genome-wide association studies. *Journal of Clinical Epidemiology*. 2015;68:460–465.

102. Hochberg Y, Benjamini Y. More powerful procedures for multiple significance testing. *Statistics in Medicine*. 1990;9:811–818.

103. Klein RJ, Zeiss C, Chew EY, Tsai J-Y, Sackler RS, Haynes C, et al. Complement factor H polymorphism in age-related macular degeneration. *Science*. 2005;308:385–9.

104. Skerka C, Lauer N, Weinberger AAWA, Keilhauer CN, Sühnel J, Smith R, et al. Defective complement control of Factor H (Y402H) and FHL-1 in age-related macular degeneration. *Molecular Immunology*. 2007;44:3398–3406.

105. The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*. 2007;447:661–78.

106. Samani NJ, Erdmann J, Hall AS, Hengstenberg C, Mangino M, Mayer B, et al. Genomewide Association Analysis of Coronary Artery Disease. *New England Journal of Medicine*. 2007;357:443–453.

107. Helgadottir A, Thorleifsson G, Manolescu A, Gretarsdottir S, Blondal T, Jonasdottir A, et al. A Common Variant on Chromosome 9p21 Affects the Risk of Myocardial Infarction. *Science*. 2007;316:1491–1493.

108. McPherson R, Pertsemlidis A, Kavaslar N, Stewart A, Roberts R, Cox DR, et al. A Common Allele on Chromosome 9 Associated with Coronary Heart Disease. *Science*. 2007;316:1488–1491.

109. Rafiq S, Anand S, Roberts R. Genome-wide association studies of hypertension: Have they been fruitful? *Journal of Cardiovascular Translational Research*. 2010;3:189–196.

110. Ganesh SK, Tragante V, Guo W, Guo Y, Lanktree MB, Smith EN, et al. Loci influencing blood pressure identified using a cardiovascular gene-centric array. *Human Molecular Genetics*. 2013;22:1663–1678.

111. Ganesh SK, Chasman DI, Larson MG, Guo X, Verwoert G, Bis JC, et al. Effects of long-term averaging of quantitative blood pressure traits on the detection of genetic associations. *American Journal of Human Genetics*. 2014;95:49–65.

112. Michailidou K, Hall P, Gonzalez-Neira A, Ghoussaini M, Dennis J, Milne RL, et al. Large-scale genotyping identifies 41 new loci associated with breast cancer risk. *Nature Genetics*. 2013;45:353–61, 361–2.

113. Keating BJ, Tischfield S, Murray SS, Bhangale T, Price TS, Glessner JT, et al. Concept, design and implementation of a cardiovascular gene-centric 50 K SNP array for large-scale genomic association studies. *PLoS ONE*. 2008;3.

114. Voight BF, Kang HM, Ding J, Palmer CD, Sidore C, Chines PS, et al. The metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. *PLoS Genetics*. 2012;8:e1002793.

115. Yoneyama S, Guo Y, Lanktree MB, Barnes MR, Elbers CC, Karczewski KJ, et al. Gene-centric meta-analyses for central adiposity traits in up to 57 412 individuals of european descent confirm known loci and reveal several novel associations. *Human Molecular Genetics*. 2014;23:2498–2510.

116. Willer CJ, Schmidt EM, Sengupta S, Peloso GM, Gustafsson S, Kanoni S, et al. Discovery and refinement of loci associated with lipid levels. *Nature genetics*. 2013;45:1274–83.

117. Schunkert H, König IR, Kathiresan S, Reilly MP, Assimes TL, Holm H, et al. Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. *Nature genetics*. 2011;43:333–8.

118. Bis JC, Kavousi M, Franceschini N, Isaacs A, Abecasis GR, Schminke U, et al. Meta-analysis of genome-wide association studies from the CHARGE consortium identifies common variants associated with carotid intima media thickness and plaque. *Nature Genetics*. 2011;43:940–7.

119. Gertow K, Sennblad B, Strawbridge RJ, Ohrvik J, Zabaneh D, Shah S, et al. Identification of the BCAR1-CFDP1-TMEM170A locus as a determinant of carotid intima-media thickness and coronary artery disease risk. *Circulation Cardiovascular genetics*. 2012;5:656–665.

120. Diekwisch TGH, Luan X, McIntosh JE. CP27 Localization in the Dental Lamina Basement Membrane and in the Stellate Reticulum of Developing Teeth. *Journal of Histochemistry & Cytochemistry*. 2002;50:583–586.

121. Christodoulou A, Santarella-Mellwig R, Santama N, Mattaj IW. Transmembrane protein TMEM170A is a novel regulator of ER and NE morphogenesis in human cells. *Journal of Cell Science*. 2016;1552–1565.

122. Akama TO, Misra AK, Hindsgaul O, Fukuda MN, Institute B. Enzymatic synthesis in vitro of the disulfated disaccharide unit of corneal keratan sulfate. *Journal of Biological Chemistry*. 2002;277:42505–42513.

123. Akama TO, Nishida K, Nakayama J, Watanabe H, Ozaki K, Nakamura T, et al. Macular corneal dystrophy type I and type II are caused by distinct mutations in a new sulphotransferase gene. *Nature Genetics*. 2000;26:237–241.

124. Szabó A, Sahin-Tóth M. Determinants of chymotrypsin C cleavage specificity in the calcium-binding loop of human cationic trypsinogen. *FEBS Journal*. 2012;279:4283–4292.

125. Chowdhury K, Dietrich S, Balling R, Guenet J-L, Gruss P. Structure, expression and chromosomal localization of Zfp-1, a murine zinc finger protein gene. *Nucleic Acids Research*. 1989;17:10427–10438.

126. Flick MJ, Konieczny SF. Identification of putative mammalian. *Biochemical and Biophysical Research Communications*. 2002;295:910–916.

127. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The Human Genome Browser at UCSC. *Genome Research*. 2002;12:996–1006.

128. Manolio TA. Bringing genome-wide association findings into clinical use. *Nature Reviews Genetics*. 2013;14:549–558.

129. Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis C a, Doyle F, et al. An integrated encyclopedia of DNA elements in the human genome. *Nature*. 2012;489:57–74.

130. Graur D, Zheng Y, Price N, Azevedo RBR, Zufall RA, Elhaik E. On the immortality of television sets: "Function" in the human genome according to the evolution-free gospel of encode. *Genome Biology and Evolution*. 2013;5:578–590.

131. Niu DK, Jiang L. Can ENCODE tell us how much junk DNA we carry in our genome? *Biochemical and Biophysical Research Communications*. 2013;430:1340–1343.

132. Bernstein BE, Stamatoyannopoulos JA, Costello JF, Ren B, Milosavljevic A, Meissner A, et al. The NIH roadmap epigenomics mapping consortium. *Nature Biotechnology*. 2010;28:1045–1048.

133. Ward LD, Kellis M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Research*. 2012;40:D930-934.

134. Boyle AP, Hong EL, Hariharan M, Cheng Y, Schaub MA, Kasowski M, et al. Annotation of functional variation in personal genomes using RegulomeDB. *Genome Research*. 2012;22:1790–1797.

135. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics*. 2014;46:310–315.

136. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature Protocols*. 2009;4:1073–81.

137. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. *Nature Methods*. 2010;7:248–9.

138. Reynolds AB, Kanner SB, Wang HC, Parsons JT. Stable association of activated pp60src with two tyrosine-phosphorylated cellular proteins. *Molecular and Cellular Biology*. 1989;9:3951–8.

139. Honda H, Oda H, Nakamoto T, Honda Z, Sakai R, Suzuki T, et al. Cardiovascular anomaly , impaired actin bundling and resistance to Src-induced transformation in mice lacking p130 Cas. *Nature Genetics*. 1998;19:361–365.

140. Sakai R, Iwamatsu A, Hirano N, Ogawa S, Tanaka T, Mano H, et al. A novel signaling molecule, p130, forms stable complexes in vivo with v-Crk and v-Src in a tyrosine phosphorylation-

dependent manner. *The EMBO Journal*. 1994;13:3748–3756.

141. Barrett A, Pellet-Many C, Zachary IC, Evans IM, Frankel P. p130Cas: a key signalling node in health and disease. *Cellular Signalling*. 2013;25:766–77.

142. Defilippi P, Di Stefano P, Cabodi S. p130Cas: a versatile scaffold in signaling networks. *Trends in Cell Biology*. 2006;16:257–63.

143. Harte MT, Hildebrand JD, Burnham R, Bouton AH, Parsons T, Burnham MR, et al. p130 Cas, a Substrate Associated with v-Src and v-Crk, Localizes to Focal Adhesions and Binds to Focal Adhesion Kinase. *Journal of Biological Chemistry*. 1996;271:13649–13655.

144. Astier A, Avraham H, Manie SN, Groopman J, Canty T, Avraham S, et al. The Related Adhesion Focal Tyrosine Kinase Is Tyrosine-phosphorylated after β1-Integrin Stimulation in B Cells and Binds to p130cas. *Journal of Biological Chemistry*. 1997;272:228–232.

145. Kirsch KH, Georgescu M-M, Hanafusa H. Direct binding of p130(Cas) to the guanine nucleotide exchange factor C3G. *Journal of Biological Chemistry*. 1998;273:25673–25679.

146. Nakamoto T, Sakai R, Honda H. Requirements for localization of p130cas to focal adhesions. *Molecular and Cellular Biology*. 1997;17:3884–3897.

147. Meenderink LM, Ryzhova LM, Donato DM, Gochberg DF, Kaverina I, Hanks SK. P130Cas Src-binding and substrate domains have distinct roles in sustaining focal adhesion disassembly and promoting cell migration. *PLoS ONE*. 2010;5.

148. Cantley LC, Songyang Z. Specificity in recognition of phosphopeptides by src-homology 2 domains. *Journal of Ce*. 1994;18:121–6.

149. Briknarová K, Nasertorabi F, Havert ML, Eggleston E, Hoyt DW, Li C, et al. The serine-rich domain from Crk-associated substrate (p130cas) is a four-helix bundle. *Journal of Biological Chemistry*. 2005;280:21908–21914.

150. Nakamoto T, Sakai R, Ozawa K, Yazaki Y, Hirai H. Direct Binding of C-terminal Region of p130 to SH2 and SH3 Domains of Src Kinase. *Journal of Biological Chemistry*. 1996;271:8959–8965.

151. Pugacheva EN, Golemis EA. The focal adhesion scaffolding protein HEF1 regulates activation of the Aurora-A and Nek2 kinases at the centrosome. *Nature Cell Biology*. 2005;7:937–946.

152. Ishino M, Ohba T, Sasaki H, Sasaki T. Molecular cloning of a cDNA encoding a phosphoprotein, Efs, which contains a Src homology 3 domain and associates with Fyn. *Oncogene*. 1995;11:2331–2338.

153. Donlin LT, Danzl NM, Wanjalla C, Alexandropoulos K. Deficiency in expression of the signaling protein Sin/Efs leads to T-lymphocyte activation and mucosal inflammation. *Molecular and Cellular Biology*. 2005;25:11035–11046.

154. Singh MK, Dadke D, Nicolas E, Serebriiskii IG, Apostolou S, Canutescu A, et al. A novel Cas family member, HEPL, regulates FAK and cell spreading. *Molecular Biology of the Cell*. 2008;19:1627–36.

155. Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, et al. The Genotype-Tissue Expression (GTEx) project. *Nature Genetics*. 2013;45:580–585.

156. Uhlen M, Fagerberg L, Hallstrom BM, Lindskog C, Oksvold P, Mardinoglu A, et al. Tissue-based map of the human proteome. *Science*. 2015;347:1260419–1260419.

157. Petch LA, Bockholt SM, Bouton A, Parsons JT, Burridge K. Adhesion-induced tyrosine phosphorylation of the p130 src substrate. *Journal of Cell Science*. 1995;108:1371–9.

158. Webb DJ, Donais K, Whitmore LA, Thomas SM, Turner CE, Parsons JT, et al. FAK-Src signalling through paxillin, ERK and MLCK regulates adhesion disassembly. *Nature Cell Biology*. 2004;6:154–161.

159. Ruoslahti E, Pierschbacher MD. New perspectives in cell adhesion: RGD and integrins. *Science*. 1987;238:491–497.

160. Giancotti FG. Integrin Signaling. *Science*. 1999;285:1028–1033.

161. Ngalim SH, Magenau A, Le Saux G, Gooding JJ, Gaus K. How do cells make decisions: Engineering micro- and nanoenvironments for cell migration. *Journal of Oncology*. 2010;1–7.

162. Zhao Z, Tan SH, Machiyama H, Kawauchi K, Araki K, Hirata H, et al. Association between tensin 1 and p130Cas at focal adhesions links actin inward flux to cell migration. *Biology Open*. 2016;1–8.

163. Nojima Y, Morino N, Mimura T, Hamasaki K, Furuya H, Sakai R, et al. Integrin-mediated cell adhesion promotes tyrosine phosphorylation of p130Cas, a Src homology 3-containing molecule having multiple Src homology 2-binding motifs. *Journal of Biological Chemistry*. 1995;270:15398–15402.

164. Gutkind JS, Robbins KC. Activation of transforming G protein-coupled receptors induces rapid tyrosine phosphorylation of cellular proteins, including p125 FAK and the p130 v-src substrate. *Biochemical and Biophysical Research Communications*. 1992;188:155–161.

165. Sharma A, Mayer BJ. Phosphorylation of p130Cas initiates Rac activation and membrane ruffling. *BMC cell biology*. 2008;9:50.

166. Avraham HK, Lee T-H, Koh Y, Kim T-A, Jiang S, Sussman M, et al. Vascular Endothelial Growth Factor Regulates Focal Adhesion Assembly in Human Brain Microvascular Endothelial Cells through Activation of the Focal Adhesion Kinase and Related Adhesion Focal Tyrosine Kinase. *Journal of Biological Chemistry*. 2003;278:36661–36668.

167. Cross MJ, Dixelius J, Matsumoto T, Claesson-Welsh L. VEGF-receptor signal transduction. *Trends in Biochemical Sciences*. 2003;28:488–494.

168. Korah R, Choi L, Barrios J, Wieder R. Expression of FGF-2 alters focal adhesion dynamics in migration- restricted MDA-MB-231 breast cancer cells. *Breast Cancer Research and Treatment*. 2004;88:17–28.

169. Rankin S, Rozengurt E. Platelet-derived growth factor modulation of focal adhesion kinase (p125FAK) and paxillin tyrosine phosphorylation in Swiss 3T3 cells. Bell-shaped dose response and cross-talk with bombesin. *Journal of Biological Chemistry*. 1994;269:704–710.

170. Ojaniemi M, Vuori K. Epidermal Growth Factor Modulates Tyrosine Phosphorylation of p130Cas : involvement of phosphatidylinositol 3'-kinase and actin cytoskeleton. *Journal of Biological Chemistry*. 1997;272:25993–25998.

171. Casamassima A, Rozengurt E. Insulin-like Growth Factor I Stimulates Tyrosine Phosphorylation of p130Cas, Focal Adhesion Kinase, and Paxillin. *Journal of Biological Chemistry*. 1998;273:26149–26156.

172. Fonseca PM, Shin NY, Brábek J, Ryzhova L, Wu J, Hanks SK. Regulation and localization of CAS substrate domain tyrosine phosphorylation. *Cellular Signalling*. 2004;16:621–629.

173. Burridge K, Turner CE, Romer LH. Tyrosine phosphorylation of paxillin and pp125FAK accompanies cell adhesion to extracellular matrix: a role in cytoskeletal assembly. *The Journal of Cell Biology*. 1992;119:893–903.

174. Bockholt SM, Burridge K. Cell spreading on extracellular matrix proteins induces tyrosine phosphorylation of tensin. *Journal of Biological Chemistry*. 1993;268:14565–14567.

175. Hamasaki K, Mimura T, Morino N, Furuya H, Nakamoto T, Aizawa S, et al. Src kinase plays an essential role in integrin-mediated tyrosine phosphorylation of Crk-associated substrate p130Cas. *Biochemical and Biophysical Research Communications*. 1996;222:338–343.

176. Romashko AA, Young MRI. Protein phosphatase-2A maintains focal adhesion complexes in keratinocytes and the loss of this regulation in squamous cell carcinomas. *Clinical & Experimental Metastasis*. 2004;21:371–9.

177. Yamakita Y, Totsukawa G, Yamashiro S, Fry D, Zhang X, Hanks SK, et al. Dissociation of FAK/p130CAS/c-Src Complex during Mitosis: Role of Mitosis-specific Serine Phosphorylation of FAK. *The Journal of Cell Biology*. 1999;144:315–324.

178. Polte TR, Hanks SK. Interaction between focal adhesion kinase and Crk-associated tyrosine kinase substrate p130Cas. *Proceedings of the National Academy of Sciences of the United States of America*. 1995;92:10678–82.

179. Chodniewicz D, Klemke RL. Regulation of integrin-mediated cellular responses through assembly of a CAS/Crk scaffold. *Biochimica et Biophysica Acta - Molecular Cell Research*. 2004;1692:63–76.

180. Provenzano PP, Keely PJ. Mechanical signaling through the cytoskeleton regulates cell proliferation by coordinated focal adhesion and Rho GTPase signaling. *Journal of Cell Science*. 2011;124:1195–1205.

181. Rivera GM, Antoku S, Gelkop S, Shin NY, Hanks SK, Pawson T, et al. Requirement of Nck adaptors for actin dynamics and cell migration stimulated by platelet-derived growth factor B. *Proceedings of the National Academy of Sciences of the United States of America*. 2006;103:9536–41.

182. Chatzizisis Y, Coskun A, Jonas M, Edelman E, Feldman C, Stone P. Role of Endothelial Shear Stress in the Natural History of Coronary Atherosclerosis and Vascular Remodeling. *Journal of the American College of Cardiology*. 2007;49:2379–2393.

183. Okuda M, Takahashi M, Suero J, Murry CE, Traub O, Kawakatsu H, et al. Shear Stress Stimulation of p130 cas Tyrosine Phosphorylation Requires Calcium-dependent c-Src Activation. *Journal of Biological Chemistry*. 1999;274:26803–26809.

184. Sawada Y, Tamada M, Dubin-Thaler BJ, Cherniavskaya O, Sakai R, Tanaka S, et al. Force Sensing by Mechanical Extension of the Src Family Kinase Substrate p130Cas. *Cell*. 2006;127:1015–1026.

185. Evans IM, Yamaji M, Britton G, Pellet-Many C, Lockie C, Zachary IC, et al. Neuropilin-1 signaling through p130Cas tyrosine phosphorylation is essential for growth factor-dependent migration of glioma and endothelial cells. *Molecular and Cellular Biology*. 2011;31:1174–85.

186. Tang DD. p130 Crk-Associated Substrate (CAS) in Vascular Smooth Muscle. *Journal of Cardiovascular Pharmacology and Therapeutics*. 2009;14:89–98.

187. Chen C-H, Ho Y-C, Ho H-H, Chang I-C, Kirsch KH, Chuang Y-J, et al. Cysteine-rich protein 2 alters p130Cas localization and inhibits vascular smooth muscle cell migration. *Cardiovascular Research*. 2013;100:461–471.

188. Kovacic-Milivojević B, Roediger F, Almeida E a, Damsky CH, Gardner DG, Ilić D. Focal adhesion kinase and p130Cas mediate both sarcomeric organization and activation of genes associated with cardiac myocyte hypertrophy. *Molecular Biology of the Cell*. 2001;12:2290–307.

189. Tu L, De Man FS, Girerd B, Huertas A, Chaumais M-C, Lecerf F, et al. A Critical Role for p130Cas in the Progression of Pulmonary Hypertension in Humans and Rodents. *American Journal of Respiratory and Critical Care Medicine*. 2012;186:666–676.

190. Brinkman A, van der Flier S, Kok EM, Dorssers LC. BCAR1, a human homologue of the adapter protein p130Cas, and antiestrogen resistance in breast cancer cells. *Journal of the National Cancer Institute*. 2000;92:112–120.

191. Flier S van der, Brinkman A, Look MP, Kok EM, Gelder MEM, Klijn JGM, et al. Bcar1/p130Cas Protein and Primary Breast Cancer: Prognosis and Response to Tamoxifen Treatment. *Journal of the National Cancer Institute*. 2000;92:120–127.

192. Konstantinovsky S, Davidson B, Reich R. Ezrin and BCAR1/p130Cas mediate breast cancer growth as 3-D spheroids. *Clinical and Experimental Metastasis*. 2012;29:527–540.

193. Brábek J, Constancio SS, Shin N-Y, Pozzi A, Weaver AM, Hanks SK. CAS promotes invasiveness of Src-transformed cells. *Oncogene*. 2004;23:7406–15.

194. Wendt MK, Smith JA, Schiemann WP. p130Cas Is Required for Mammary Tumor Growth and Transforming Growth Factor-beta-mediated Metastasis through Regulation of Smad2/3 Activity. *Journal of Biological Chemistry*. 2009;284:34145–34156.

195. Kook S, Shim SR, Choi SJ, Ahnn J, Kim J Il, Eom SH, et al. Caspase-mediated Cleavage of p130cas in Etoposide-induced Apoptotic Rat-1 Cells. *Molecular Biology of the Cell*. 2000;11:929–939.

196. The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature*. 2012;491:56–65.

197. Johnson AD, Handsaker RE, Pulit SL, Nizzari MM, O'Donnell CJ, de Bakker PIW. SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. *Bioinformatics*. 2008;24:2938–2939.

198. Lewis BP, Burge CB, Bartel DP. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*. 2005;120:15–20.

199. Machiela MJ, Chanock SJ. LDlink: A web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. *Bioinformatics*. 2015;31:3555–3556.

200. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics*. 2007;81:559–575.

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

202. Baldassarre D, Hamsten A, Veglia F, de Faire U, Humphries SE, Smit AJ, et al. Measurements of Carotid Intima-Media Thickness and of Interadventitia Common Carotid Diameter Improve Prediction of Cardiovascular Events. *Journal of the American College of Cardiology*. 2012;60:1489–1499.

203. Baragetti A, Palmen J, Garlaschelli K, Grigore L, Pellegatta F, Tragni E, et al. Telomere shortening over 6 years is associated with increased subclinical carotid vascular damage and worse cardiovascular prognosis in the general population. *Journal of Internal Medicine*. 2015;277:478–487.

204. Lee AJ, Mowbray PI, Lowe GDO, Rumley A, Fowkes FGR, Allan PL. Blood Viscosity and Elevated Carotid Intima-Media Thickness in Men and Women: The Edinburgh Artery Study. *Circulation*. 1998;97:1467–1473.

205. Marmot M, Brunner E. Cohort Profile: The Whitehall II study. *International Journal of Epidemiology*. 2005;34:251–256.

206. Berglund G, Elmståhl S, Janzon L, Larsson SA. Design and feasibility. *Journal of Internal Medicine*. 1993;233:45–51.

207. R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2013.

208. NHLBI GO Exome Sequencing Project. Exome Variant Server [Internet]. Available from: http://evs.gs.washington.edu/EVS/

209. Smith AJP, Humphries SE. Characterization of DNA-binding proteins using multiplexed competitor EMSA. *Journal of Molecular Biology*. 2009;385:714–717.

210. Clontech. In-Fusion Primer Design Tool [Internet]. Available from: http://bioinfo.clontech.com/infusion/

211. Genomatix Software Suite [Internet]. Available from: https://www.genomatix.de/

212. Untergasser A, Nijveen H, Rao X, Bisseling T, Geurts R, Leunissen JAM. Primer3Plus, an enhanced web interface to Primer3. *Nucleic Acids Research*. 2007;35:W71–W74.

213. Biotech G. SnapGene® software.

214. Technologies A. QuikChange Primer Design tool [Internet]. Available from: https://www.genomics.agilent.com/primerDesignProgram.jsp

215. de Martin R, Raidl M, Hofer E, Binder BR. Adenovirus-mediated expression of green fluorescent protein. *Gene Therapy*. 1997;4:493–495.

216. Steinberg MH. Sickle cell anemia, the first molecular disease: overview of molecular etiology, pathophysiology, and therapeutic approaches. *The Scientific World Journal*. 2008;8:1295–1324.

217. Freedman ML, Monteiro ANA, Gayther SA, Coetzee GA, Risch A, Plass C, et al. Principles for the post-GWAS functional characterization of cancer risk loci. *Nature Genetics*. 2011;43:513–

518.

218. Smith AJP, Humphries SE, Talmud PJ. Identifying functional noncoding variants from genome-wide association studies for cardiovascular disease and related traits. *Current Opinion in Lipidology*. 2015;26:120–126.

219. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*. 2012;489:57–74.

220. Musunuru K, Strong A, Frank-Kamenetsky M, Lee NE, Ahfeldt T, Sachs K V, et al. From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature*. 2010;466:714–719.

221. Miller CL, Anderson DR, Kundu RK, Raiesdana A, Nürnberg ST, Diaz R, et al. Disease-Related Growth Factor and Embryonic Signaling Pathways Modulate an Enhancer of TCF21 Expression at the 6q23.2 Coronary Heart Disease Locus. *PLoS Genetics*. 2013;9:1–17.

222. Richardson K, Nettleton JA, Rotllan N, Tanaka T, Smith CE, Lai CQ, et al. Gain-of-function lipoprotein lipase variant rs13702 modulates lipid traits through disruption of a MicroRNA-410 seed site. *American Journal of Human Genetics*. 2013;92:5–14.

223. Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, et al. The accessible chromatin landscape of the human genome. *Nature*. 2012;489:75–82.

224. Sims RJ, Nishioka K, Reinberg D. Histone lysine methylation: a signature for chromatin function. *Trends in Genetics*. 2003;19:629–639.

225. Gross DS, Garrard WT. Nuclease hypersensitive sites in chromatin. *Annual Review of Biochemistry*. 1988;57:159–197.

226. Rockman M V., Kruglyak L. Genetics of global gene expression. *Nature Reviews Genetics*. 2006;7:862–872.

227. eQTL resources from the Gilad/Pritchard group [Internet]. Available from: http://eqtl.uchicago.edu/cgi-bin/gbrowse/eqtl/

228. Meyer KB, Maia AT, O'Reilly M, Teschendorff AE, Chin SF, Caldas C, et al. Allele-specific up-regulation of FGFR2 increases susceptibility to breast cancer. *PLoS Biology*. 2008;6:1098–1103.

229. Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, et al. Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. *Science*. 2012;337:1190–1195.

230. Hayes JE, Trynka G, Vijai J, Offit K, Raychaudhuri S, Klein RJ. Tissue-specific enrichment of lymphoma risk loci in regulatory elements. *PLoS ONE*. 2015;10:1–14.

231. Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, et al. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*. 2011;473:43–9.

232. Harismendy O, Notani D, Song X, Rahim NG, Tanasa B, Heintzman N, et al. 9p21 DNA variants associated with coronary artery disease impair interferon-γ signalling response. *Nature*. 2011;470:264–8.

233. Oldoni F, Palmen J, Giambartolomei C, Howard P, Drenos F, Plagnol V, et al. Post-GWAS methodologies for localisation of functional non-coding variants: ANGPTL3. *Atherosclerosis*.

2015;246:193–201.

234. Genomics England. 100,000 Genomes Project [Internet]. Available from: http://www.genomicsengland.co.uk/the-100000-genomes-project/

235. Ardlie KG, Kruglyak L, Seielstad M. Patterns of Linkage Disequilibrium in the Human Genome. *Nature Reviews Genetics*. 2002;3:299–309.

236. Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, et al. The Structure of Haplotype Blocks in the Human Genome. *Science*. 2002;296:2225–2229.

237. Wu Y, Waite LL, Jackson AU, Sheu WHH, Buyske S, Absher D, et al. Trans-Ethnic Fine-Mapping of Lipid Loci Identifies Population-Specific Signals and Allelic Heterogeneity That Increases the Trait Variance Explained. *PLoS Genetics*. 2013;9.

238. Peprah E, Xu H, Tekola-Ayele F, Royal CD. Genome-wide association studies in Africans and African Americans: expanding the framework of the genomics of human traits and disease. *Public Health Genomics*. 2015;18:40–51.

239. Degner JF, Pai A a, Pique-Regi R, Veyrieras J-B, Gaffney DJ, Pickrell JK, et al. DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature*. 2012;482:390–4.

240. Miller WL, Hartman KA, Burritt MF, Grill DE, Rodeheffer RJ, Burnett JC, et al. Serial Biomarker Measurements in Ambulatory Patients With Chronic Heart Failure: The Importance of Change Over Time. *Circulation*. 2007;116:249–257.

241. Mackinnon AD, Jerrard-Dunne P, Sitzer M, Buehler A, von Kegler S, Markus HS. Rates and Determinants of Site-Specific Progression of Carotid Artery Intima-Media Thickness: The Carotid Atherosclerosis Progression Study. *Stroke*. 2004;35:2150–2154.

242. Tanaka H, Nishino M, Ishida M, Fukunaga R, Sueyoshi K. Progression of carotid atherosclerosis in Japanese patients with coronary artery disease. *Stroke; a journal of cerebral circulation*. 1992;23:946–51.

243. Hodis HN, Mack WJ, LaBree L, Selzer RH, Liu C, Liu C, et al. The role of carotid arterial intima-media thickness in predicting clinical coronary events. *Annals of Internal Medicine*. 1998;128:262–269.

244. Hirano M, Nakamura T, Kitta Y, Takishima I, Deyama J, Kobayashi T, et al. Short-term progression of maximum intima-media thickness of carotid plaque is associated with future coronary events in patients with coronary artery disease. *Atherosclerosis*. 2011;215:507–12.

245. Polak JF, Pencina MJ, O'Leary DH, D'Agostino RB. Common carotid artery intima-media thickness progression as a predictor of stroke in multi-ethnic study of atherosclerosis. *Stroke*. 2011;42:3017–21.

246. Lorenz MW, Polak JF, Kavousi M, Mathiesen EB, Völzke H, Tuomainen T-P, et al. Carotid intima-media thickness progression to predict cardiovascular events in the general population (the PROG-IMT collaborative project): a meta-analysis of individual participant data. *The Lancet*. 2012;379:2053–2062.

247. Baldassarre D, Veglia F, Hamsten A, Humphries SE, Rauramaa R, de Faire U, et al. Progression of Carotid Intima-Media Thickness as Predictor of Vascular Events: Results from the IMPROVE Study. *Arteriosclerosis, Thrombosis, and Vascular Biology*. 2013;33:2273–2279.

248. Roeters van Lennep JE, Westerveld HT, Erkelens DW, van der Wall EE. Risk factors for coronary heart disease: implications of gender. *Cardiovascular Research*. 2002;53:538–549.

249. Johnson A. Sex differentials in coronary heart disease: the explanatory role of primary risk factors. *Journal of Health and Social Behavior*. 1977;18:46–54.

250. Kofler BM, Miles E a, Curtis P, Armah CK, Tricon S, Grew J, et al. Apolipoprotein E genotype and the cardiovascular disease risk phenotype: impact of sex and adiposity (the FINGEN study). *Atherosclerosis*. 2012;221:467–70.

251. Döring A, Gieger C, Mehta D, Gohlke H, Prokisch H, Coassin S, et al. SLC2A9 influences uric acid concentrations with pronounced sex-specific effects. *Nature Genetics*. 2008;40:430–436.

252. Zhou J, Huang Y, Huang RS, Wang F, Xu L, Le Y, et al. A case-control study provides evidence of association for a common SNP rs974819 in PDGFD to coronary heart disease and suggests a sex-dependent effect. *Thrombosis Research*. 2012;130:602–606.

253. Ober C, Loisel DA, Gilad Y. Sex-specific genetic architecture of human disease. *Nature Reviews Genetics*. 2008;9:911–922.

254. Juonala M, Viikari JSA, Laitinen T, Marniemi J, Helenius H, Rönnemaa T, et al. Interrelations between brachial endothelial function and carotid intima-media thickness in young adults: The Cardiovascular Risk in Young Finns Study. *Circulation*. 2004;110:2918–2923.

255. Dawson JD, Sonka M, Blecha MB, Lin W, Davis PH. Risk Factors Associated with Aortic and Carotid Intimal Medial Thickness in Adolescents and Young Adults: the Muscatine Offspring Study. *Journal of the American College of Cardiology*. 2009;53:2273–2279.

256. Bonithon-Kopp C, Scarabin P-Y, Darne B, Malmejac A, Guize L. Menopause-related changes in lipoproteins and some other cardiovascular risk factors. *International Journal of Epidemiology*. 1990;19:42–48.

257. Shahar E, Folsom AR, Salomaa V V, Stinson VL, McGovern PG, Shimakawa T, et al. Relation of Hormone-Replacement Therapy to Measures of Plasma Fibrinolytic Activity. *Circulation*. 1996;93:1970–1975.

258. Meade TW, Ruddock V, Stirling Y, Chakrabarti R, Miller GJ. Fibrinolytic activity, clotting factors, and long-term incidence of ischaemic heart disease in the Northwick Park Heart Study. *The Lancet*. 1993;342:1076–1079.

259. Ku DN, Giddens DP, Zarins CK, Glagov S. Pulsatile flow and atherosclerosis in the human carotid bifurcation. Positive correlation between plaque location and low oscillating shear stress. *Arteriosclerosis, Thrombosis, and Vascular Biology*. 1985;5:293–302.

260. Gnasso A, Carallo C, Irace C, Spagnuolo V, De Novara G, Mattioli PL, et al. Association Between Intima-Media Thickness and Wall Shear Stress in Common Carotid Arteries in Healthy Male Subjects. *Circulation*. 1996;94:3257–3262.

261. Malek AM, Alper SL, Izumo S. Hemodynamic shear stress and its role in atherosclerosis. *JAMA*. 1999;282:2035–2042.

262. Levesque MJ, Nerem RM, Sprague E a. Vascular endothelial cell proliferation in culture and the influence of flow. *Biomaterials*. 1990;11:702–707.

263. Zaidel-Bar R, Kam Z, Geiger B. Polarized downregulation of the paxillin-p130CAS-Rac1

pathway induced by shear flow. *Journal of Cell Science*. 2005;118:3997–4007.

264. Hedblad B, Nilsson P, Janzon L, Berglund G. Relation between insulin resistance and carotid intima-media thickness and stenosis in non-diabetic subjects. Results from a cross-sectional study in Malmö, Sweden. *Diabetic Medicine*. 2000;17:299–307.

265. Rimm EB, Stampfer MJ, Giovannucci E, Ascherio A, Spiegelman D, Colditz GA, et al. Body Size and Fat Distribution as Predictors of Coronary Heart Disease among Middle-aged and Older US Men. *American Journal of Epidemiology*. 1995;141:1117–1127.

266. Wang Z, Hoy WE. Waist circumference, body mass index, hip circumference and waist-to-hip ratio as predictors of cardiovascular disease in Aboriginal people. *European Journal of Clinical Nutrition*. 2004;58:888–893.

267. Lissner L, Björkelund C, Heitmann BL, Seidell JC, Bengtsson C. Larger hip circumference independently predicts health and longevity in a Swedish female cohort. *Obesity Research*. 2001;9:644–6.

268. Seidell JC, Pérusse L, Després JP, Bouchard C. Waist and hip circumferences have independent and opposite effects on cardiovascular disease risk factors: The Quebec Family Study. *American Journal of Clinical Nutrition*. 2001;74:315–321.

269. Bonithon-Kopp C, Touboul P-J, Berr C, Leroux C, Mainard F, Courbon D, et al. Relation of Intima-Media Thickness to Atherosclerotic Plaques in Carotid Arteries: The Vascular Aging (EVA) Study . *Arteriosclerosis, Thrombosis, and Vascular Biology* . 1996;16:310–316.

270. Speliotes EK, Willer CJ, Berndt SI, Monda KL, Thorleifsson G, Jackson AU, et al. Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nature Genetics*. 2010;42:937–948.

271. Pruim RJ, Welch RP, Sanna S, Teslovich TM, Chines PS, Gliedt TP, et al. LocusZoom: Regional visualization of genome-wide association scan results. *Bioinformatics*. 2011;27:2336–2337.

272. Bulun SE, Simpson ER. Competitive reverse transcription-polymerase chain reaction analysis indicates that levels of aromatase cytochrome P450 transcripts in adipose tissue of buttocks, thighs, and abdomen of women increase with advancing age. *The Journal of Clinical Endocrinology & Metabolism*. 1994;78:428–432.

273. Weatherman R V. Untangling the estrogen receptor web. *Nature Chemical Biology*. 2006;2:175–176.

274. Deroo BJ, Korach KS. Review series estrogen receptors and human disease. *The Journal of Clinical Investigation*. 2006;116:561–570.

275. Corporation P. pGL3 Luciferase Reporter pGL3 Luciferase Reporter Vectors. 2008;1–30.

276. Cartharius K, Frech K, Grote K, Klocke B, Haltmeier M, Klingenhoff A, et al. MatInspector and beyond: promoter analysis based on transcription factor binding sites. *Bioinformatics*. 2005;21:2933–2942.

277. Overdier DG, Porcella A, Costa RH. The DNA-binding specificity of the hepatocyte nuclear factor 3 / forkhead domain is influenced by amino-acid residues adjacent to the recognition helix. *Molecular and Cellular Biology*. 1994;14:2755–2766.

278. Friedman JR, Kaestner KH. The Foxa family of transcription factors in development and

metabolism. *Cellular and Molecular Life Sciences*. 2006;63:2317–2328.

279. Cirillo LA, Lin FR, Cuesta I, Friedman D, Jarnik M, Zaret KS. Opening of Compacted Chromatin by Early Developmental Transcription Factors HNF3 (FoxA) and GATA-4. *Molecular Cell*. 2002;9:279–289.

280. Wolfrum C, Asilmaz E, Luca E, Friedman JM, Stoffel M. Foxa2 regulates lipid metabolism and ketogenesis in the liver during fasting and in diabetes. *Nature*. 2004;432:1027–1032.

281. Kheradpour P, Kellis M. Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. *Nucleic Acids Research*. 2013;42:2976–87.

282. Dominguez R, Micevych P. Estradiol Rapidly Regulates Membrane Estrogen Receptor α Levels in Hypothalamic Neurons. *The Journal of Neuroscience*. 2010;30:12589–12596.

283. Cowell IG, Hurst HC. Transcriptional repression by the human bZIP factor E4BP4: definition of a minimal repression domain. *Nucleic Acids Research*. 1994;22:59–65.

284. Colombo MG, Citti L, Basta G, De Caterina R, Biagini A, Rainaldi G. Differential ability of human endothelial cells to internalize and express exogenous DNA. *Cardiovascular Drugs and Therapy*. 2001;15:25–29.

285. Hernández JL, Coll T, Ciudad CJ. A highly efficient electroporation method for the transfection of endothelial cells. *Angiogenesis*. 2004;7:235–241.

286. Kleinjan DA, van Heyningen V. Long-range control of gene expression: emerging mechanisms and disruption in disease. *American Journal of Human Genetics*. 2005;76:8–32.

287. Sanyal A, Lajoie BR, Jain G, Dekker J. The long-range interaction landscape of gene promoters. *Nature*. 2012;489:109–113.

288. Tolhuis B, Palstra R-J, Splinter E, Grosveld F, de Laat W. Looping and Interaction between Hypersensitive Sites in the Active β-globin Locus. *Molecular Cell*. 2002;10:1453–1465.

289. Dekker J, Rippe K, Dekker M, Kleckner N. Capturing Chromosome Conformation. *Science*. 2002;295:1306–1311.

290. Venter JC, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Adams MD, et al. The Sequence of the Human Genome. *Science*. 2001;291:1304–1351.

291. Samani NJ, Erdmann J, Hall AS, Hengstenberg C, Mangino M, Mayer B, et al. Genomewide association analysis of coronary artery disease. *The New England Journal of Medicine*. 2007;357:443–53.

292. Hoffman EA, Frey BL, Smith LM, Auble DT. Formaldehyde crosslinking: A tool for the study of chromatin complexes. *Journal of Biological Chemistry*. 2015;290:26404–26411.

293. Zhao Z, Tavoosidana G, Sjölinder M, Göndör A, Mariano P, Wang S, et al. Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nature genetics*. 2006;38:1341–1347.

294. Dostie J, Richmond TA, Arnaout RA, Selzer RR, Lee WL, Honan TA, et al. Chromosome Conformation Capture Carbon Copy (5C): A massively parallel solution for mapping interactions between genomic elements. *Genome Research*. 2006;16:1299–1309.

295. Fullwood MJ, Liu MH, Pan YF, Liu J, Xu H, Mohamed Y Bin, et al. An oestrogen-receptor-alpha-

bound human chromatin interactome. *Nature*. 2009;462:58–64.

296. Belton JM, McCord RP, Gibcus JH, Naumova N, Zhan Y, Dekker J. Hi-C: A comprehensive technique to capture the conformation of genomes. *Methods*. 2012;58:268–276.

297. Dostie J, Dekker J. Mapping networks of physical interactions between genomic elements using 5C technology. *Nature Protocols*. 2007;2:988–1002.

298. Zhang J, Poh HM, Peh SQ, Sia YY, Li G, Mulawadi FH, et al. ChIA-PET analysis of transcriptional chromatin interactions. *Methods*. 2012;58:289–299.

299. de Wit E, de Laat W. A decade of 3C technologies: Insights into nuclear organization. *Genes and Development*. 2012;26:11–24.

300. Flicek P, Amode MR, Barrell D, Beal K, Billis K, Brent S, et al. Ensembl 2014. *Nucleic Acids Research* . 2014;42:D749–D755.

301. Kumbrink J, Kirsch KH. Regulation of p130(Cas)/BCAR1 expression in tamoxifen-sensitive and tamoxifen-resistant breast cancer cells by EGR1 and NAB2. *Neoplasia*. 2012;14:108–20.

302. Van De Werken HJG, De Vree PJP, Splinter E, Holwerda SJB, Klous P, De Wit E, et al. 4C technology: Protocols and data analysis. 1st ed. Elsevier Inc.; 2012.

303. Stadhouders R, Kolovos P, Brouwer R, Zuin J, van den Heuvel A, Kockx C, et al. Multiplexed chromosome conformation capture sequencing for rapid genome-scale high-resolution detection of long-range chromatin interactions. *Nat Protocols*. 2013;8:509–524.

304. Bouïs D, Hospers GAP, Meijer C, Molema G, Mulder NH. Endothelium in vitro: A review of human vascular endothelial cell lines for blood vessel-related research. *Angiogenesis*. 2001;4:91–102.

305. Splinter E, de Wit E, van de Werken HJG, Klous P, de Laat W. Determining long-range chromatin interactions for selected genomic sites using 4C-seq technology: From fixation to computation. *Methods*. 2012;58:221–230.

306. Gondor A, Rougier C, Ohlsson R. High-resolution circular chromosome conformation capture assay. *Nature Protocols*. 2008;3:303–313.

307. Naumova N, Smith EM, Zhan Y, Dekker J. Analysis of long-range chromatin interactions using Chromosome Conformation Capture. *Methods*. 2012;58:192–203.

308. Mifsud B, Tavares-Cadete F, Young AN, Sugar R, Schoenfelder S, Ferreira L, et al. Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nature Genetics*. 2015;47:598–606.

309. Würtele H, Chartrand P. Genome-wide scanning of HoxB1-associated loci in mouse ES cells using an open-ended Chromosome Conformation Capture methodology. *Chromosome Research*. 2006;14:477–495.

310. Tan-Wong SM, French JD, Proudfoot NJ, Brown M a. Dynamic interactions between the promoter and terminator regions of the mammalian BRCA1 gene. *Proceedings of the National Academy of Sciences of the United States of America*. 2008;105:5160–5165.

311. Visser M, Kayser M, Palstra RJ. HERC2 rs12913832 modulates human pigmentation by attenuating chromatin-loop formation between a long-range enhancer and the OCA2 promoter. *Genome Research*. 2012;22:446–455.

312. Cong L, Ran FA, Cox D, Lin S, Barretto R, Habib N, et al. Multiplex Genome Engineering Using CRISPR/Cas Systems. *Science*. 2013;339:819–823.

313. Geyer PK, Corces VG. Dna Position-Specific Repression of Transcription By a Drosophila Zinc Finger Protein. *Genes & Development*. 1992;6:1865–1873.

314. Sun F, Elgin SC. Putting Boundaries on Silence. *Cell*. 1999;99:459–462.

315. Dean A. In the loop: Long range chromatin interactions and gene regulation. *Briefings in Functional Genomics*. 2011;10:3–10.

316. Phillips JE, Corces VG. CTCF: Master Weaver of the Genome. *Cell*. 2009;137:1194–1211.

317. Ong C, Corces V. Enhancer function: new insights into the regulation of tissue-specific gene expression. *Nature Reviews Genetics*. 2011;12:283–93.

318. Maron BJ, Maron MS, Semsarian C. Genetics of hypertrophic cardiomyopathy after 20 years: Clinical perspectives. *Journal of the American College of Cardiology*. 2012;60:705–715.

319. Cargill M, Altshuler D, Ireland J, Sklar P, Ardlie K, Patil N, et al. Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nature Genetics*. 1999;22:231–238.

320. Kryukov G V, Pennacchio LA, Sunyaev SR. Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. *American Journal of Human Genetics*. 2007;80:727–39.

321. Ng PC, Henikoff S. Predicting the effects of amino acid substitutions on protein function. *Annual Review of Genomics and Human Genetics*. 2006;7:61–80.

322. Macias MJ, Wiesner S, Sudol M. WW and SH3 domains, two different scaffolds to recognize proline-rich ligands. *FEBS Letters*. 2002;513:30–37.

323. Betts MJ, Russell RB. Amino acid properties and consequences of substitutions. In: Barnes MR, Gray IC, editors. Bioinformatics for Geneticists. Wiley; 2003.

324. Zhao Z, Fu YX, Hewett-Emmett D, Boerwinkle E. Investigating single nucleotide polymorphism (SNP) density in the human genome and its implications for molecular evolution. *Gene*. 2003;312:207–213.

325. Li B, Leal S. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *The American Journal of Human Genetics*. 2008;83:311–321.

326. Morris AP, Zeggini E. An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genetic Epidemiology*. 2010;34:188–193.

327. Lee S, Abecasis GR, Boehnke M, Lin X. Rare-Variant Association Analysis: Study Designs and Statistical Tests. *American Journal of Human Genetics*. 2014;95:5–23.

328. Choi Y, Sims GE, Murphy S, Miller JR, Chan AP. Predicting the Functional Effect of Amino Acid Substitutions and Indels. *PLoS ONE*. 2012;7.

329. Kent WJ. BLAT—The BLAST-Like Alignment Tool. *Genome Research*. 2002;12:656–664.

330. Zhang Z, Gerstein M. Patterns of nucleotide substitution, insertion and deletion in the human

genome inferred from pseudogenes. *Nucleic Acids Research*. 2003;31:5338–5348.

331. Williamson MP. The structure and function of proline-rich regions in proteins. *The Biochemical Journal*. 1994;297:249–60.

332. Kay BK, Williamson MP, Sudol M. The importance of being proline: the interaction of proline-rich motifs in signaling proteins with their cognate domains. *The FASEB Journal*. 2000;14:231–241.

333. Nikonova AS, Gaponova A V., Kudinov AE, Golemis EA. CAS proteins in health and disease: an update. *IUBMB Life*. 2014;66:387–395.

334. Cunningham-Edmonson A, Hanks S. p130Cas substrate domain signaling promotes migration, invasion, and survival of estrogen receptor-negative breast cancer cells. *Breast Cancer: Targets and Therapy*. 2009;39–52.

335. Sawano A, Takayama S, Matsuda M, Miyawaki A. Lateral propagation of EGF signaling after local stimulation is dependent on receptor density. *Developmental Cell*. 2002;3:245–257.

336. Berk BC, Brock TA, Webb RC, Taubman MB, Atkinson WJ, Gimbrone MA, et al. Epidermal growth factor, a vascular smooth muscle mitogen, induces rat aortic contraction. *The Journal of Clinical Investigation*. 1985;75:1083–6.

337. Maretzky T, Evers A, Zhou W, Swendeman SL, Wong P-M, Rafii S, et al. Migration of growth factor-stimulated epithelial and endothelial cells depends on EGFR transactivation by ADAM17. *Nature Communications*. 2011;2:229.

338. Bouton a H, Riggins RB, Bruce-Staskal PJ. Functions of the adapter protein Cas: signal convergence and the determination of cellular responses. *Oncogene*. 2001;20:6448–58.

339. Senger DR, Ledbetter SR, Claffey KP, Papadopoulos-Sergiou A, Peruzzi CA, Detmar M. Stimulation of endothelial cell migration by vascular permeability factor/vascular endothelial growth factor through cooperative mechanisms involving the alphavbeta3 integrin, osteopontin, and thrombin. *The American Journal of Pathology*. 1996;149:293–305.

340. Salgia R, Pisick E, Sattler M, Li J, Uemura N, Wong W, et al. p130 CAS Forms a Signaling Complex with the Adapter Protein CRKL in Hematopoietic Cells Transformed by the BCR / ABL Oncogene. *Journal of Biological Chemistry*. 1996;271:25198–25203.

341. Editorial. On beyond GWAS. *Nature Genetics*. 2010;42:551–551.

342. Zhou X, Baron RM, Hardin M, Cho MH, Zielinski J, Hawrylkiewicz I, et al. Identification of a chronic obstructive pulmonary disease genetic determinant that regulates HHIP. *Human Molecular Genetics*. 2012;21:1325–1335.

343. Ritchie GRS, Dunham I, Zeggini E, Flicek P. Functional annotation of noncoding sequence variants. *Nature Methods*. 2014;11:294–6.

344. Serandour AA, Avner S, Percevault F, Demay F, Bizot M, Lucchetti-Miganeh C, et al. Epigenetic switch involved in activation of pioneer factor FOXA1-dependent enhancers. *Genome Research*. 2011;21:555–565.

345. Morris AP, Voight BF, Teslovich TM, Ferreira T, Segrè A V, Steinthorsdottir V, et al. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nature Genetics*. 2012;44:981–990.

346. Harder MN, Ribel-Madsen R, Justesen JM, Sparsø T, Andersson E a, Grarup N, et al. Type 2 diabetes risk alleles near BCAR1 and in ANK1 associate with decreased β-cell function whereas risk alleles near ANKRD55 and GRB14 associate with decreased insulin sensitivity in the Danish Inter99 cohort. *The Journal of Clinical Endocrinology & Metabolism*. 2013;98:E801-6.

347. Smith AJP, Palmen J, Putt W, Talmud PJ, Humphries SE, Drenos F. Application of statistical and functional methodologies for the investigation of genetic determinants of coronary heart disease biomarkers: Lipoprotein lipase genotype and plasma triglycerides as an exemplar. *Human Molecular Genetics*. 2010;19:3936–3947.

348. Gyoung Tak Y, Farnham PJ. Making sense of GWAS: using epigenomics and genome engineering to understand the functional relevance of SNPs in non-coding regions of the human genome. *Epigenetics & Chromatin*. 2015;8:57.

349. Mohrs M, Blankespoor CM, Wang ZE, Loots GG, Afzal V, Hadeiba H, et al. Deletion of a coordinate regulator of type 2 cytokine expression in mice. *Nature Immunology*. 2001;2:842–847.

350. Lorenz MW, Price JF, Robertson C, Bots ML, Polak JF, Poppert H, et al. Carotid Intima-Media Thickness Progression and Risk of Vascular Events in People With Diabetes: Results From the PROG-IMT Collaboration. *Diabetes Care*. 2015;38:1921–1929.

351. Cabodi S, Moro L, Baj G, Smeriglio M, Di Stefano P, Gippone S, et al. p130Cas interacts with estrogen receptor α and modulates non-genomic estrogen signaling in breast cancer cells. *Journal of Cell Science*. 2004;117:1603–1611.

352. Bots ML, Evans GW, Riley W, McBride KH, Paskett ED, Helmond FA, et al. The effect of tibolone and continuous combined conjugated equine oestrogens plus medroxyprogesterone acetate on progression of carotid intima-media thickness: The Osteoporosis Prevention and Arterial effects of tiboLone (OPAL) study. *European Heart Journal*. 2006;27:746–755.

353. Mortensen KH, Andersen NH, Hjerrild BE, Hørlyck A, Stochholm K, Højbjerg Gravholt C. Carotid intima-media thickness is increased in Turner syndrome: Multifactorial pathogenesis depending on age, blood pressure, cholesterol and oestrogen treatment. *Clinical Endocrinology*. 2012;77:844–851.

354. Hodis HN, Mack WJ, Henderson VW, Shoupe D, Budoff MJ, Hwang-Levine J, et al. Vascular Effects of Early versus Late Postmenopausal Treatment with Estradiol. *New England Journal of Medicine*. 2016;374:1221–1231.

355. Kunnas TA, Laippala P, Penttilä A, Lehtimäki T, Karhunen PJ. Association of polymorphism of human alpha oestrogen receptor gene with coronary artery disease in men: a necropsy study. *British Medical Journal*. 2000;321:273–274.

356. Sudhir K, Chou TM, Chatterjee K, Smith EP, Williams TC, Kane JP, et al. Premature Coronary Artery Disease Associated With a Disruptive Mutation in the Estrogen Receptor Gene in a Man. *Circulation*. 1997;96:3774–3777.

357. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, et al. Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science*. 2009;326:289–293.

358. Barreiro LB, Laval G, Quach H, Patin E, Quintana-Murci L. Natural selection has driven population differentiation in modern humans. *Nature Genetics*. 2008;40:340–345.

359. Kawahara T, Ritsick D, Cheng G, Lambeth JD. Point mutations in the proline-rich region of p22phox are dominant inhibitors of Nox1- and Nox2-dependent reactive oxygen generation. *Journal of Biological Chemistry*. 2005;280:31859–31869.

360. Xiao R, Boehnke M. Quantifying and correcting for the winner's curse in quantitative-trait association studies. *Genetic Epidemiology*. 2011;35:133–138.

361. Addis R, Campesi I, Fois M, Capobianco G, Dessole S, Fenu G, et al. Human umbilical endothelial cells (HUVECs) have a sex: characterisation of the phenotype of male and female cells. *Biology of Sex Differences*. 2014;5:18.

362. Lorenz M, Koschate J, Kaufmann K, Kreye C, Mertens M, Kuebler WM, et al. Does cellular sex matter? Dimorphic transcriptional differences between female and male endothelial cells. *Atherosclerosis*. 2015;240:61–72.

363. Marchini J, Howie B. Genotype imputation for genome-wide association studies. *Nature Reviews Genetics*. 2010;11:499–511.

364. Yu J, Vodyanik MA, Smuga-Otto K, Antosiewicz-Bourget J, Frane JL, Tian S, et al. Induced pluripotent stem cell lines derived from human somatic cells. *Science*. 2007;318:1917–1920.

365. Thomson JA, Itskovitz-Eldor J, Shapiro SS, Waknitz MA, Swiergiel JJ, Marshall VS, et al. Embryonic Stem Cell Lines Derived from Human Blastocysts. *Science*. 1998;282:1145 LP-1147.

366. Belair DG, Whisler JA, Valdez J, Velazquez J, Molenda JA, Vickerman V, et al. Human Vascular Tissue Models Formed from Human Induced Pluripotent Stem Cell Derived Endothelial Cells. *Stem Cell Reviews and Reports*. 2015;11:511–525.

367. Dash BC, Jiang Z, Suh C, Qyang Y. Induced pluripotent stem cell-derived vascular smooth muscle cells: methods and application. *The Biochemical Journal*. 2015;465:185–94.

368. Miller JC, Holmes MC, Wang J, Guschin DY, Lee Y-L, Rupniewski I, et al. An improved zinc-finger nuclease architecture for highly specific genome editing. *Nature Bi4otechnology*. 2007;25:778–85.

369. Joung JK, Sander JD. TALENs: a widely applicable technology for targeted genome editing. *Nature Reviews Molecular Cell Biology*. 2013;14:49–55.

370. Marraffini LA, Sontheimer EJ. CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea. *Nature Reviews Genetics*. 2010;11:181–190.

371. Claussnitzer M, Dankel SN, Kim K-H, Quon G, Meuleman W, Haugen C, et al. FTO Obesity Variant Circuitry and Adipocyte Browning in Humans. *The New England journal of medicine*. 2015;373:895–907.

372. Canver MC, Smith EC, Sher F, Pinello L, Sanjana NE, Shalem O, et al. BCL11A enhancer dissection by Cas9-mediated in situ saturating mutagenesis. *Nature*. 2015;527:192–7.

373. Schulze A, Downward J. Navigating gene expression using microarrays — a technology review. *Nature Cell Biology*. 2001;3:E190–E195.

374. Sur IK, Hallikas O, Vaharautio A, Yan J, Turunen M, Enge M, et al. Mice lacking a Myc enhancer that includes human SNP rs6983267 are resistant to intestinal tumors. *Science*. 2012;338:1360–1363.

375. MacDonald ML, Lamerdin J, Owens S, Keon BH, Bilter GK, Shang Z, et al. Identifying off-target

effects and hidden phenotypes of drugs in human cells. *Nature Chemical Biology*. 2006;2:329–337.