Functional analysis of novel genetic markers of coronary artery disease identified by genome-wide association studies

Thesis submitted for the degree of Doctor of Philosophy
at the University of Leicester

by

Peter S. Braund MSc

Department of Cardiovascular Sciences
University of Leicester

2015

Functional analysis of novel genetic markers of coronary artery disease identified by genome-wide association studies

Peter S. Braund

Abstract

Coronary artery disease (CAD) is the commonest cause of death worldwide. It is a multifactorial disease caused by interplay between genes and environment. Defining genes that affect CAD risk and understanding their mechanisms may help to improve its prevention and treatment. Recent genome-wide association studies have identified several novel loci associated with CAD. In this thesis I examined two of these loci.

The first was a locus on chromosome 1p13.3. I showed that this locus was also associated with LDL-cholesterol. Fine-mapping of the locus and *in silico* analyses suggested that SNP rs12740374, where the risk allele disrupts binding of a liver-specific transcription factor, is probably the causal variant at the locus. Finally, I showed that the lipid-lowering effect of statin is independent of this locus, suggesting that targeting the mechanism by which this locus affects LDL-cholesterol may provide additional therapeutic benefit.

The second locus was on chromosome 13q34 located near the *COL4A1* and *COL4A2* genes which code for collagen Type 4, a key component of the arterial wall. *In silico* analyses identified an intronic 4-SNP haplotype within a bidirectional promoter that was likely to include the causal variant(s). A putative kidney expression QTL was identified with the CAD risk allele associated with lower expression of *COL4A1* and *COL4A2*. Functional experiments showed DNA-protein binding to the SNPs in the haploblock. However, there were no differences in level of DNA-binding between the alleles or a differential impact on gene expression. In a clinical study, there was no difference between the 13q34 genotype groups in restenosis rates in patients undergoing percutaneous coronary angioplasty.

In conclusion, the work on locus 1p13.3 shows that it is possible to identify a causal SNP from a GWA study signal, and elucidate its mechanism, but the studies on locus 13q34, indicates that this is a challenging process.

Acknowledgements

I would firstly like to thank my supervisors Prof. Nilesh J. Samani and Dr. Maciej Tomaszewski for all their help and support during this thesis.

I would secondly, like to thank all those in the Dept. of Cardiovascular Sciences that have helped in any way (either through advice or support) to progress the work in this thesis, especially Dr. Veryan Codd, Matthew Denniff, Dr Chris Nelson, Dr. Tom Webb, Dr. Clett Erridge, Prof. Alison Goodall and Dr Karl Herbert.

In particular, I would also like to dedicate this thesis to my wife Dr. Samantha Braund, for her tireless love, support, and encouragement, throughout my studies.

Experimental Acknowledgements

Chapter 4

Thank you to Matthew Denniff for performing the PKP study renal RNA extractions.

Thank you to Dr Tim McKeithan from the University of Nebraska for providing me with a luciferase reporter vector construct *BCL3* promoter and STAT3 (NF-κB) intronic enhancer, positive control, as a gift.

Publications and Abstracts

Publications (resulting from experimental work carried out in this thesis)

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

CARDIoGRAMplusC4D Consortium, Deloukas, P., Kanoni, S., Willenborg, C., Farrall, M., Assimes, T.L., et al. 2013. 'Large-scale association analysis identifies new risk loci for coronary artery disease'. Nature Genetics. 45 (1): 25-33.

Edmondson, A.C., Braund, P.S., Stylianou, I.M., Khera, A.V., Nelson, C.P., Wolfe, M.L., et al. 2011. 'Dense Genotyping of Candidate Gene Loci Identifies Variants Associated with High-Density Lipoprotein Cholesterol'. Circulation.Cardiovascular Genetics.

Guo, Y., Lanktree, M.B., Taylor, K.C., Hakonsarson, H., Lange, L.A., Keating, B.J., et al. 2012. 'Gene-centric meta-analyses of 108 912 individuals confirm known body mass index loci and reveal three novel signals'. Human Molecular Genetics.

IBC 50K CAD Consortium 2011. 'Large-scale gene-centric analysis identifies novel variants for coronary artery disease'. PLoS Genetics. 7 (9): e1002260.

Johnson, T., Gaunt, T.R., Newhouse, S.J., Padmanabhan, S., Tomaszewski, M., Kumari, M., et al. 2011. 'Blood pressure loci identified with a gene-centric array'. American Journal of Human Genetics. 89 (6): 688-700.

Lanktree, M.B., Guo, Y., Murtaza, M., Glessner, J.T., Bailey, S.D., Onland-Moret, N.C., et al. 2011. 'Meta-analysis of Dense Genecentric Association Studies Reveals Common and Uncommon Variants Associated with Height'. American Journal of Human Genetics. 88 (1): 6-18.

[#]Samani, N.J., Braund, P.S., Erdmann, J., Gotz, A., Tomaszewski, M., Linsel-Nitschke, P., et al. 2008. 'The novel genetic variant predisposing to coronary artery disease in the region of the PSRC1 and CELSR2 genes on chromosome 1 associates with serum cholesterol'. Journal of Molecular Medicine. 86 (11): 1233-1241.

Saxena, R., Elbers, C.C., Guo, Y., Peter, I., Gaunt, T.R., Mega, J.L., et al. 2012. 'Large-scale gene-centric meta-analysis across 39 studies identifies type 2 diabetes loci'. American Journal of Human Genetics. 90 (3): 410-425.

Shah, S., Nelson, C.P., Gaunt, T.R., van der Harst, P., Barnes, T., Braund, P.S., et al. 2011. 'Four Genetic Loci Influencing Electrocardiographic Indices of Left Ventricular Hypertrophy'. Circulation.Cardiovascular Genetics.

Tomaszewski, M., Debiec, R., Braund, P.S., Nelson, C.P., Hardwick, R., Christofidou, P., et al. 2010. 'Genetic architecture of ambulatory blood pressure in the general population: insights from cardiovascular gene-centric array'. Hypertension. 56 (6): 1069-1076.

# Joint first author.

Presentations

Braund, P., Erdmann, J., Tomaszewski, M., Götz, A., Hajat, C., Linsel-Nitschke, P., et al. A novel locus—PSRC11/CELSR2 gene locus on Chromosome 1p13.3, which affects serum cholesterol and is associated with increased risk of coronary artery disease *Heart* 2008; 94: A1-A4.

Braund, P.S., Erdmann, J., Tobin, M.D., Götz, A., Tomaszewski, M., Linsel-Nitschke, P., et al. on behalf of CARDIOGENICS. A recently identified genetic variant for coronary artery disease risk on chromosome 1p13.3 in the region of the PSRC1 and CELSR2 genes associates with serum cholesterol. *Eur Heart J* 2008; 29 (Suppl1): 733-900.

TABLE OF CONTENTS

List of tables

List of figures

List of abbreviations

>50%	greater than 50% luminal diameter stenosis
1958BC	1958 Birth Cohort
3' MGB	3 prime minor groove binder
3'UTR	3 prime untranslated region
A	adenosine
A549	adenocarcinomic human alveolar basal epithelial cells
ABC	ATP binding cassette transporter
ABCA1	ATP binding cassette transporter
ABCG5	ATP binding cassette transporter
ABCG8	ATP binding cassette transporter
ACE	Angiotensin Converting Enzyme
ACS	acute coronary syndromes
ADRB2	beta-2-adrenergic receptor
AIMs	Ancestry Informative Markers
ALOX5AP	arachidonate 5-lipoxygenase-activating protein
ANG-II	angiotensin 2
ANOVA	analysis of variance
AP	angina pectoris
AP1	activator protein 1
apo	apolipoprotein
APOB	apolipoprotein B
APOE	apolipoprotein E
ARH	autosomal recessive hypercholesterolaemia
AS	atherosclerosis
ASPs	affected sibling pairs
ASW	African ancestry in Southwest USA
B2M	beta-2-microglobulin
BACH2	BTB and CNC homology 1, basic leucine zipper transcription factor 2
BACs	bacterial artificial chromosomes
BCL3	B-cell lymphoma 3
BD	bipolar disorder
BHF-FHS	British Heart Foundation Family Heart Study
BM	basement membrane
BMI	body mass index
bp	base pair
BRCA1	breast cancer 1, early onset
BSA	bovine serum albumin
C	cytosine
CABG	coronary artery bypass graft surgery
CAC	coronary artery calciification
CAD	coronary artery disease

CARDIoGRAM	Coronary ARtery DIsease Genome-Wide Replication And Meta-Analysis
CARe	candidate-gene association resource
CBP	CREB binding protein
CD	Crohn’s disease
CD/CV	common disease / common variant
CD/RV	common disease / rare variant
CD36	**C**luster of **D**ifferentiation 36 - Scavenger receptor
CDKN2A	cyclin dependent kinase inhibitor 2A
CDKN2B	cyclin dependent kinase inhibitor 2B
CDKN2B-AS1	CDKN2B anti-sense 1, aka ANRIL
CEBP	CCAAT/enhancer binding protein
CEBPA	CCAAT/enhancer binding protein alpha
CEBPB	CCAAT/enhancer binding protein beta
CEC	circulatory endothelial cells
CELSR2	cadherin, EGF LAG seven-pass G-type receptor 2
CEU	Utah residents with Northern and Western European ancestry from the CEPH collection
CHD	coronary heart disease
CHF	Chronic (congestive) Heart Failure
ChIP	chromatin immunoprecipitation
ChIP-seq	chromatin immunoprecipitation with next generation sequencing
CHOD-PAP	cholesterol oxidase phenol 4-aminoantipyrine peroxidise
Chr.	chromosome
CI	Confidence Interval
cMYC	V-myc myelocytomatosis viral oncogene homolog (avian)
CNV	copy number variations
COL(I)	collagen type I
COL(III)	collagen type III
COL(IV)	collagen type IV
COL4A1	collagen type IV alpha 1
COL4A2	collagen type IV alpha 2
CRIC	chronic renal insufficiency cohort
CRISPR/Cas	clustered regularly interspaced short palindromic repeats/Cas gene
CRP	C-reactive protein
cSNPs	coding SNPs
Ct	cycle threshold
CTCBF	CTC box binding factor
CV	cardiovascular
CVD	cardiovascular disease
CXCL12	chemokine (C-X-C motif) ligand 12
D'	D prime
DBP	diastolic blood pressure
DCt	delta cycle threshold

DDCt	delta delta cycle threshold
DGF	digital genomic footprint
DIG	digoxigen
DLR	Dual-Luciferase assay
DMEM	Dulbecco's modified eagles medium
DMSO	dimethyl sulfoxide
DNA	deoxyribonucleic acid
DNase I	deoxyribonuclease 1
DNaseI HS	DNaseI hypersensitivity sites
DNase-seq	DNaseI hypersensitivity sites with next generation sequencing
dNTPs	deoxynucleotide triphosphates
ds	double stranded
DSBs	double-strand breaks
E	amplification efficiency
E2F6	E2F transcription factor 6
ECM	extracellular matrix
EDTA	ethylenediaminetetraacetic acid
EGFA	epidermal growth factor-like repeat A
EMSA	Electrophoresis mobility (gel) shift assay
ENCODE	ENCyclopaedia Of DNA Elements
eQTL	expression quantitative trait loci
ER	endoplasmic reticulum
ESCs	embryonic stem cells
ETS1	v-ets avian erythroblastosis virus E26 oncogene homolog 1
EVI-1	Aka MECOM - MDS1 and EVI1 complex locus
FAC1	bromodomain PHD finger transcription factor
FAIRE	Formaldehyde assisted isolation of regulatory elements
FAIRE-seq	FAIRE with next generation sequencing
FBS	foetal calf serum
FDB	familial ligand-defective apolipoprotein B-100
FEV	FEV (ETS oncogene family)
FGF	fibroblast growth factor
FH	familial hypercholesterolaemia
FLAP	5-lipoxygenase-activating protein
FOX	forkhead box
FOXO3	forkhead box O3
FOXP1	forkhead box P1
FPRP	false positive report probability
G	Guanine
GABPA	GA binding protein transcription factor, alpha subunit 60kDa
GAG	glycosaminoglycan
GAPDH	glyceraldehyde-3-phosphate dehydrogenase
GATA3	GATA binding protein 3

GEE	generalised estimated equations
GerMIF	German myocardial infarction family (study)
GFR	Glomerular Filtration Rate
GLUT4	glucose transporter receptor 4
GM12878	a lymphoblastoid cell line
GRAPHIC	Genetic Regulation of Arterial Pressure of Humans In the Community
H1-hESC	human embryonic stem cells
H3K27Ac	histone 3 lysine 27 acetylation
H3K4Me1	histone 3 lysine 4 mono-methylation
H3K4Me3	histone 3 lysine 4 tri-methylation
HANAC	hereditary, angiopathy, nephropathy, aneurysms and muscle cramps
HASMC	human aorta smooth muscle cells
HAT	histone acetyltransferase
HCF	Human cardiac fibroblasts
HCHOLA3	autosomal dominant familial hypercholesterolemia
HCM	Human Cardiac Myocytes
HDAC	histone de-acetylase
HDL-C	high density lipoprotein cholesterol
HeLa	Henrietta Lacks cervical carcinoma epithelial-like cells
HepG2	hepatocellular carcinoma cells
HLC	human liver cohort
HLF	hepatic leukaemia factor
HMG-CoA	3-hydroxy-3-methyl-glutaryl-CoA
HR	homologous recombination
HRE	Human Renal Epithelial Cells
HSMM	Human Skeletal Muscle Myoblast
HTN	hypertension
HUVEC	human umbilical vein endothelial cells
HVMF	Human Villous Mesenchymal Fibroblasts
HWE	Hardy-Weinberg equilibrium
IBC	ITMAT-Broad-CARe
IBD	identical-by-descent
IBS	identical-by-state
ICAM-1	intracellular adhesion molecule 1
IDL	intermediate density lipoprotein
IL-1	interleukin 1
IL-1β	interleukin 1β
IL-8	interleukin 8
IMS	industrial methylated spirit
INFg	interferon gamma
iPSCs	induced pluripotent stem cells
ITMAT	Institute of Translational Medicine and Therapeutics
K562	chronic myelogenous erythroleukemia-type cells,

kb	kilobase
Kif2a	kinesin heavy chain member 2A
KORA	Cooperative Research in the Region of Augsburg
KpnI	restriction enzyme
L50%	greater than 50% loss of acute luminal gain
LB	luria broth
LCAT	lecithin cholesterol acyltransferase
LD	linkage disequilibrium
LDL-C	low density lipoprotein cholesterol
LDLR	LDL receptors
LGALS2	lectin, galactoside-binding, soluble, 2
lncRNA	long non-coding RNA
LOD	log-likelihood odds ratio
LOLIPOP	London Life Sciences Prospective Population Cohort
LPL	lipoprotein lipase
LTA	lymphotoxin alpha
LTA4H	leukotriene A4 hydrolase
Luc	luciferase
LVH	left ventricular hypertrophy
MACS	immunomagnetic cell sorting separation
MAF	minor allele frequency
MAFK	v-maf avian musculoaponeurotic fibrosarcoma oncogene homolog K
MAPRE3	microtubule-associated protein, RP/EB family, member 3
MAX	MYC associated factor X
MAZ	MYC-associated zinc finger protein (purine-binding transcription factor)
MCP-1	monocyte chemotactic protein 1
MDC-CC	Malmo¨ Diet and Cancer Study – Cardiovascular Cohort
MEF	mouse embryonic fibroblasts
MEF2A	myocyte enhancer factor 2A
MHC	major histocompatibility complex
MI	myocardial infarction
MIA3	melanoma inhibitory activity 3
MIF1	homocysteine-inducible, endoplasmic reticulum stress-inducible, ubiquitin-like domain member 1
miRNA	microRNA
miRNA-TS	MicroRNA target site
MMPs	Matrix metalloproteinases
MONICA	Monitoring of Trends and Determinations of Cardiovascular Disease
mRNA	messenger RNA
MSCs	mesenchymal stem cells
MTHFD1L	methylenetetrahydrofolate dehydrogenase (NADP+-dependent) 1-like
MULAN	multiple sequence local alignment and visualization tool
MXD1	MAX dimerization protein 1

MXI1	MAX interactor 1
MYBPHL	myosin binding protein H-like
MYT1	myelin transcription factor 1
NADPH	nicotinamide adenine dinucleotide phosphate
NCBI	National Center for Biotechnology Information
NFATC2	nuclear factor of activated T-cells, cytoplasmic, calcineurin-dependent 2
NF-κB	nuclear factor kappa-light-chain-enhancer of activated B cells
NHBLI	National Heart Lung Institute
NHEJ	non-homologous end-joining
NHEK	human epidermal keratinocytes
NHLF	human lung fibroblasts
NMR	nuclear magnetic resonance
NO	nitric oxide
nsSNPs	non-synonymous SNPs
nt	nucleotide
O4F	optimised for function
OR	odds ratio
oxLDL	oxidised LDL
P53	tumour suppressor
PAI	plasminogen activator inhibitor
PAI-1	plasminogen activator inhibitor 1
PARC	Pharmacogenomics and Risk of Cardiovascular Disease study
PAX3	paired box 3
PBS	phosphate-buffered saline
PCR	polymerase chain reaction
PCSK9	proprotein convertase subtilisin/kexin type 9
PDGF	platelet derived growth factor
PIC	pre-initiation complex
PKP	Polish kidney project
PLB	passive lysisbuffer
PNACL	Protein nucleotide acid chemistry laboratory
Pol2 / RNA Pol II	RNA Polymerase II
PP	pulse pressure
PPARG	peroxisome proliferator-activated receptor gamma
PROMIS	Pakistan Risk of Myocardial Infarction Study
PSRC1	proline-serine rich coiled-coil 1
PSSM	position specific score matrices
PTCA	percutaneous transluminal coronary angioplasty
PU box	purine box
P-value	probability value
PWM	position weight matrix
QC	quality control
qPCR	real time quantitative PCR

R	fold change
r^2	square of the correlation coefficient
R2	coefficient of determination
RA	rheumatoid arthritis
rAAV	recombination (vector) adeno-associated virus
RAS	renin angiotensinogen system
RER	rough endoplasmic reticulum
RMEC	renal microvascular endothelial cells
RNA	ribonucleic acid
RORA_2	RAR-related orphan receptor A 2
rs	reference SNP ID
RT-PCR	reverse transcription polymerase chain reaction
SacI	restriction enzyme
SARS	serine-tRNA ligase, cytoplasmic
SBP	systolic blood pressure
SCD	sudden cardiac death
SD	standard deviation
SE	standard error
SEM	standard error of the mean
SHARP	Subcutaneous Heparin and Angioplasty Restenosis Prevention Trial
SHS	Silesian hypertension study
SIN3A	SIN3 transcription regulator family member A
SMAD3	SMAD family member 3
SMGS	smooth muscle growth supplement
SNAP	SNP Association and Proxy (search tool)
SNP	single nucleotide polymorphism
SORT1	sortilin 1
SOX10	SRY (sex determining region Y)-box 10
SP1	specificity protein 1
SPI1	spleen focus forming virus (SFFV) proviral integration oncogene
SPIB	Spi-B transcription factor (Spi-1/PU.1 related)
SRTB	Silesian Renal Tissue Bank
SSC	saline-sodium citrate
STAT1	signal transducer and activator of transcription 1
STAT3	signal transducer and activator of transcription 3
STAT6	signal transducer and activator of transcription 6
T	Thymine
T1D	type 1 diabetes
T2D	type 2 diabetes
TAD	transcriptional activation domain
TAE	Tris-Acetate-EDTA buffer
TALENs	transcription activator-like effector nucleases
TBE	tris boric acid EDTA

TBP	TATA box protein
TC	total cholesterol
TDT	Transmission disequilibrium test
TEAD1	TEA domain family member 1 (SV40 transcriptional enhancer factor)
TEN	Tris EDTA NaCl buffer
TF	transcription factor
TFBS	transcription factor binding site
TG	triglycerides
TGF-β	transforming growth factor beta
TGN	trans-Golgi network
THBS	thrombospondin
TIMPs	tissue inhibitors of metalloproteinases
TNF	tumour necrosis factor
TNFα	tumour necrosis factor alpha
TSS	transcription start site
UCSC	University of California Santa Cruz
v/v	volume/volume
VCAM-1	vascular adhesion molecule 1
VLDL	very low density lipoprotein
VSMC	vascular smooth muscle VSMC
w/v	weight/volume
WHO	world health organisation
WHR	waist hip ratio
WOSCOPS	West Of Scotland Coronary Prevention Study
WTCCC	Wellcome Trust Case Control Consortium
YRI	Yoruba in Ibadan, Nigeria
YY1	Yin Yang 1
ZEB1	zinc finger E-box binding homeobox 1
ZFNs	zinc finger nucleases

1. Chapter 1

Introduction

1.1 Epidemiology of coronary artery disease

Coronary artery disease (CAD) is the single largest killer of men and women in the world. In 2008 it was responsible for 7.3 million deaths world-wide accounting for 1 in 8 of all deaths for that year (World Health Organisation (WHO) 2011). This is in spite of the fact that premature CAD death rates have dropped dramatically over the past 40 years. For instance, in the United States (US) there was a 40% decrease in death rate between 1968 (peak year) to 1984 (Kumar, Cotran & Robbins 1992), and in the United Kingdom (UK) there was a 65% decrease in death rate from CAD between 1980 and 2007 (Scarborough et al. 2010). In 2009, CAD accounted for 82,528 (1 in 7) deaths in the UK. More importantly as regards productive life years, premature deaths from CAD (those below the age of 75) accounted for 1 in 3 deaths (British Heart Foundation (BHF) 2010). In terms of morbidity, the latest UK prevalence estimates indicate that approximately one million men and 500,000 women in the UK have suffered a myocardial infarction (MI) and nearly 2.7 million are registered as having CAD. In addition, the overall cost to the NHS in 2006 was estimated to be £3.3 billion, approximately £54 per capita (Scarborough et al. 2010). For these reasons it is important that we continue to enhance our knowledge of the biological mechanisms that result in CAD, as this will enable further therapeutic advancement, and help to diminish the social and economic burden caused by CAD.

1.2 Definition of CAD

CAD is defined as atherosclerosis (AS) of the coronary arteries (blood vessels that supply the heart muscle). This is a gradual decades-long process resulting in slowly progressive narrowing of coronary artery blood vessel walls, by lipid-laden plaques that eventually lead to symptoms of ischaemia, a state where blood vessel narrowing, or atherothrombotic occlusion causes a deficiency of oxygen and nutrient supply to the myocardium (heart muscle) – (**Figure 1.1**). CAD develops silently, and can be asymptomatic for many years, or throughout an individual's life.

When clinically apparent it usually presents as one or more of four syndromes: 1) angina pectoris; 2) myocardial infarction; 3) chronic CAD with heart failure and 4) sudden cardiac death; (Regieli et al. 2007).

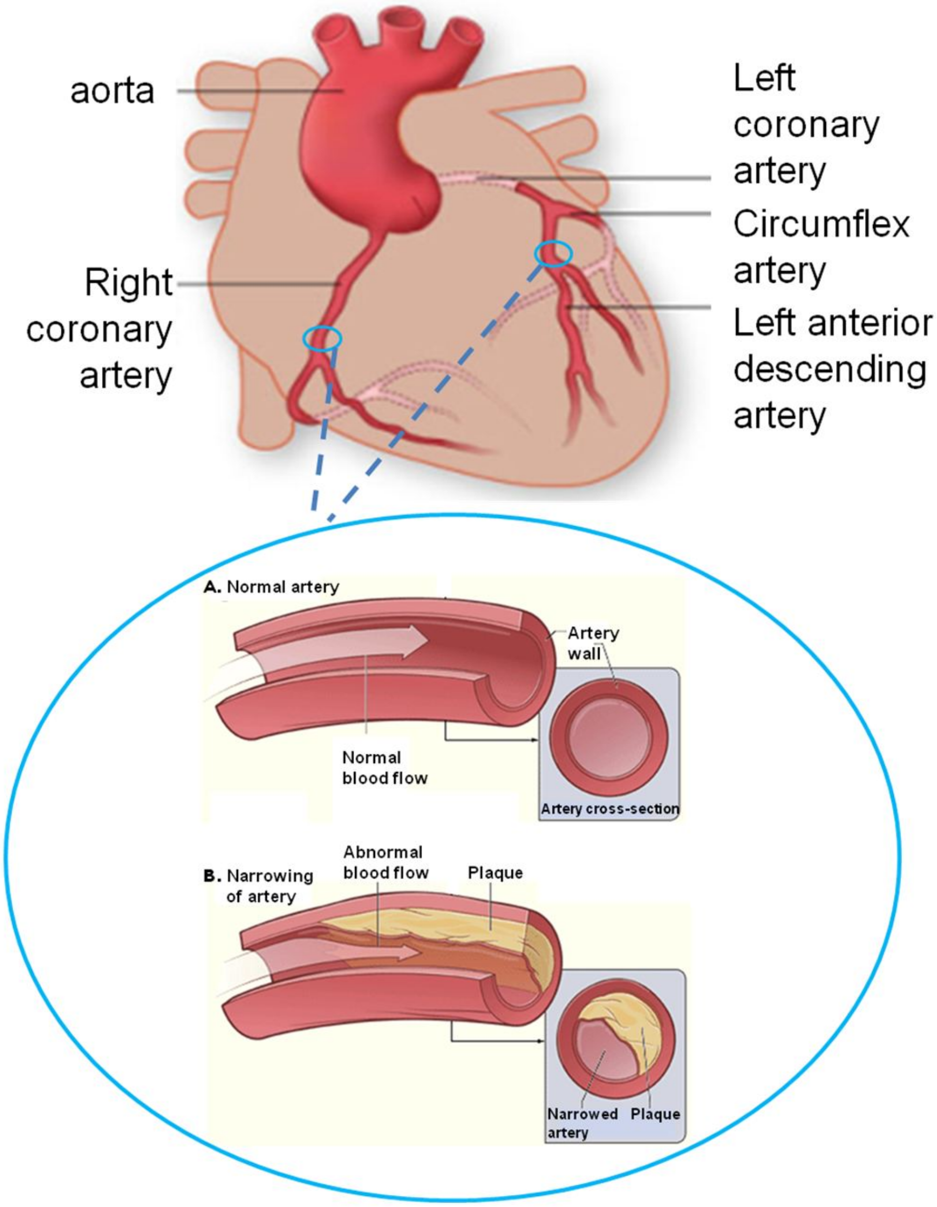


Figure 1.1. Simple schematic to show the position of the coronary arteries within the heart and a cross-section of atherosclerosis within the artery wall (National Heart Lung and Blood Institute).

1.3 Clinical Manifestations of CAD

1.3.1 Angina pectoris

Angina pectoris (AP) in Greek literally means “a strangling of the chest”, and is generally described by patients as a tight, heavy or gripping chest pain on exertion sometimes radiating to the jaw, neck, shoulder or arms and settling with rest within a few minutes. Angina typically occurs when the narrowing caused by AS is sufficiently severe (usually ≥75% stenosis of one or more coronary arteries) enough, to restrict the required blood flow for the needs of the myocardium – hence its typical relation to exertion. For the same reason it can be exacerbated by emotion, cold weather or a full stomach. Other symptoms may include nausea, sweating and breathlessness (Kumar & Clark 1998, 4th Edn Clinical Medicine).

1.3.2 Myocardial infarction

Myocardial infarction (MI) in Greek literally means - “death of the heart muscle”. It is caused by an ischaemic episode brought about by the complete occlusion of one or more coronary arteries, with sufficient duration to result in necrosis of cardiomyocytes.

The symptoms for MI match that of angina, only they are more pronounced and occur at rest, i.e. sudden severe crushing chest pain (often with referred pain to the neck, jaw and arm), breathlessness, nausea, weakness, anxiety and sweating.

There are in fact two forms of MI - the more common and life threatening ***transmural*** infarct that involves the whole thickness of the ventricular wall, and the ***subendocardial*** infarct that affects the inner ventricular wall. In fact, the transmural MI starts off life as a subendocardial infarct (as this is most distant from the epicardially located coronary arteries), but may progress to a transmural infarct depending upon: 1) the severity and duration of the ischaemia; 2) the presence and extent of the collateral flow; and 3) myocardial demand. In approximately 90% of cases a (transmural) MI occurs, due to an acute thrombotic event over an atherosclerotic plaque in one of the major coronary arteries, which has ruptured exposing its pro-thrombotic lipid core to the flowing blood. The process involves platelet aggregation and vasospasm. This results in either complete occlusion at its point of origin, or creates thromboemboli large enough to occlude a downstream coronary artery branch. Interestingly, although the majority of patients who present with an MI have severe stenosis, plaque rupture, and thrombosis leading to coronary occlusion - this can occur in coronary arteries where the degree of luminal narrowing is much lower, i.e. ~25%-50%. The reason for this is that there is no direct relationship between severity of plaque stenosis, and the vulnerability of plaques to rupture and cause coronary occlusion. Vulnerable plaques tend to have a large pro-atherogenic necrotic core, with a thin fibrous cap, and a compensatory enlargement in the subendothelial intima. However, many highly stenosed lesions are likely to be the result of previously disrupted unstable plaques, which have undergone a thrombotic response that did not result in a MI, but did lead to an infiltration of collagen, fibrin, and calcium, which resulted in substantial hardening and thickening of the plaque cap, so that it is less prone to rupture.

Post MI, the heart is left with lasting scar tissue and becomes functionally impaired; this is due to a limited inherent ability of the heart to repair, and heal following ischaemic tissue damage. Also, the heart wall may end up much thinner and prone to forming aneurysms (balloon-like dilation of blood vessels), following an acute MI event.

1.3.3 Chronic CAD with heart failure

Congestive heart failure (CHF) is a common progressive disorder with a high mortality rate. CHF is a state where the myocardium is unable to pump enough blood to meet metabolic demand, or it can only do so at a raised filling pressure. The onset of CHF occurs during the end-stage of chronic heart disease and is often due to: 1) chronic work overload (such as hypertension or valve disease); or 2) chronic CAD following a MI and a significant level of ischaemic damage, which undergoes cardiac hypertrophy of the non-infarcted myocardium, a compensatory response to increased mechanical work. However, if the severity of CAD is bad enough, CHF can sometimes occur due to CAD even without a previous infarction; and even without prior ischaemic symptoms. This is a situation where the first presentation of CAD is heart failure.

1.3.4 Sudden cardiac death

Sudden cardiac death (SCD) is an unexpected death of cardiac cause, in subjects that either had no previous symptoms of heart disease, or very soon after symptoms have first become apparent (in most cases this is within 1 hour of onset). The commonest reason for SCD is a fatal arrhythmia (this could be a ventricular fibrillation or asystole) set off by an acute myocardial ischaemic event. Indeed the most commonly observed morphology in 80% to 90% of SCD cases is ≥75% (often >90%) stenosis in one or more of the three major coronaries. Indeed, about 50% of cases will have had acute plaque disruption, and 25% have diagnostic features of acute MI (Farb et al. 1995). It is this manifestation of CAD that causes the greatest public concern, and raises the profile of CAD as the number one killer – sudden cardiac death without any prior symptoms or warning.

1.4 Pathophysiology of CAD

The principle cause of CAD is the most important form of arteriosclerosis (which from Greek literally means "hardening of the arteries") called atherosclerosis. This term is taken from two Greek words *athera* meaning "gruel" and *sclerosis* meaning "hardening", which is a morphological description of arterial wall thickening with an encased lump of gruel-like (porridge) material. AS is a chronic progressive condition, where atherosclerotic plaques develop in the inner lining of the coronary arteries, which run over the surface of the heart. Each coronary artery is essential for supplying the myocardium (heart muscle) with nutrients and oxygenated blood.

A normal artery consists of three layers. The first is the (tunica) intima, the innermost layer that consists of a single layer of endothelial cells that is in contact with blood, which is structurally supported in normal arteries by a sheet-like layer of specialised extracellular matrix (ECM), called the basement membrane (BM). The basement membrane is made up of four major constituents: collagen type IV, laminin, nidogen/entactin and perlecan. Both collagen type IV and laminin self-assemble into suprastructures that form a scaffold of enmeshed networks to form the basic framework for the BM. Nidogen/entactin and the multidomain core protein of perlecan stabilise this BM scaffold by bridging the collagen/laminin networks, and increasing BM stability and structural integrity. Of note, perlecan consists not only of a multidomain core protein, but also three long chains of glycosaminoglycans, usually heparan sulphate (proteoglycans) (Paulsson 1992; LeBleu, Macdonald & Kalluri 2007). The second is the (tunica) media the middle layer, which is made up of a strata of smooth muscle cells (SMCs) (capable of contraction in response to vasoconstrictors: endothelin, catecholamines and angiotensin II; or relaxing in response to vasodilators: kinins, nitric oxide (NO) and prostacyclin) - surrounded by an internal and external elastic lamina. The third is the (tunica) externa (also called adventitia) the outermost layer consisting of connective tissue, mainly collagen and elastic fibres, plus fibroblasts (that make connective tissue), mast cells, nerve endings and arterioles (the vasa vasorum) (Bhatia 2010; Libby, Ridker & Hansson 2011).

1.4.1 Overview of Atherosclerotic plaque development and progression:

AS is a chronic inflammatory process. Each stage of its development - initiation, progression and thrombosis, are controlled to a large extent by inflammatory cellular and molecular mediators of the immune response (see **Figure 1.2** and **Figure 1.3** for two different cross-sections of the atheromatic plaque as it progresses).

AS is initiated when the arterial endothelium is activated at focal areas such as, blood vessel openings and branch points by a combination of laminar shear stress, turbulent flow haemodynamic disturbances and other risk factors that include: dyslipidaemia, hypertension, hyperglycaemia end-products, homocysteine, bacterial products, viruses, cigarette smoke toxins, hypoxia, pro-inflammatory cytokines and complement. These cause endothelial cell expression of adhesion molecules such as, intracellular adhesion molecule 1 (ICAM-1), vascular adhesion molecule 1 (VCAM-1), E- and P- selectin and chemokines interleukin 8 (IL-8) and monocyte chemotactic protein 1 (MCP-1), which triggers the inflammatory innate immune response and leads to recruitment, rolling and tight adhesion followed by transmigration of blood leukocytes into the sub-endothelial intimal space via a process called diapedesis (Businaro et al. 2012).

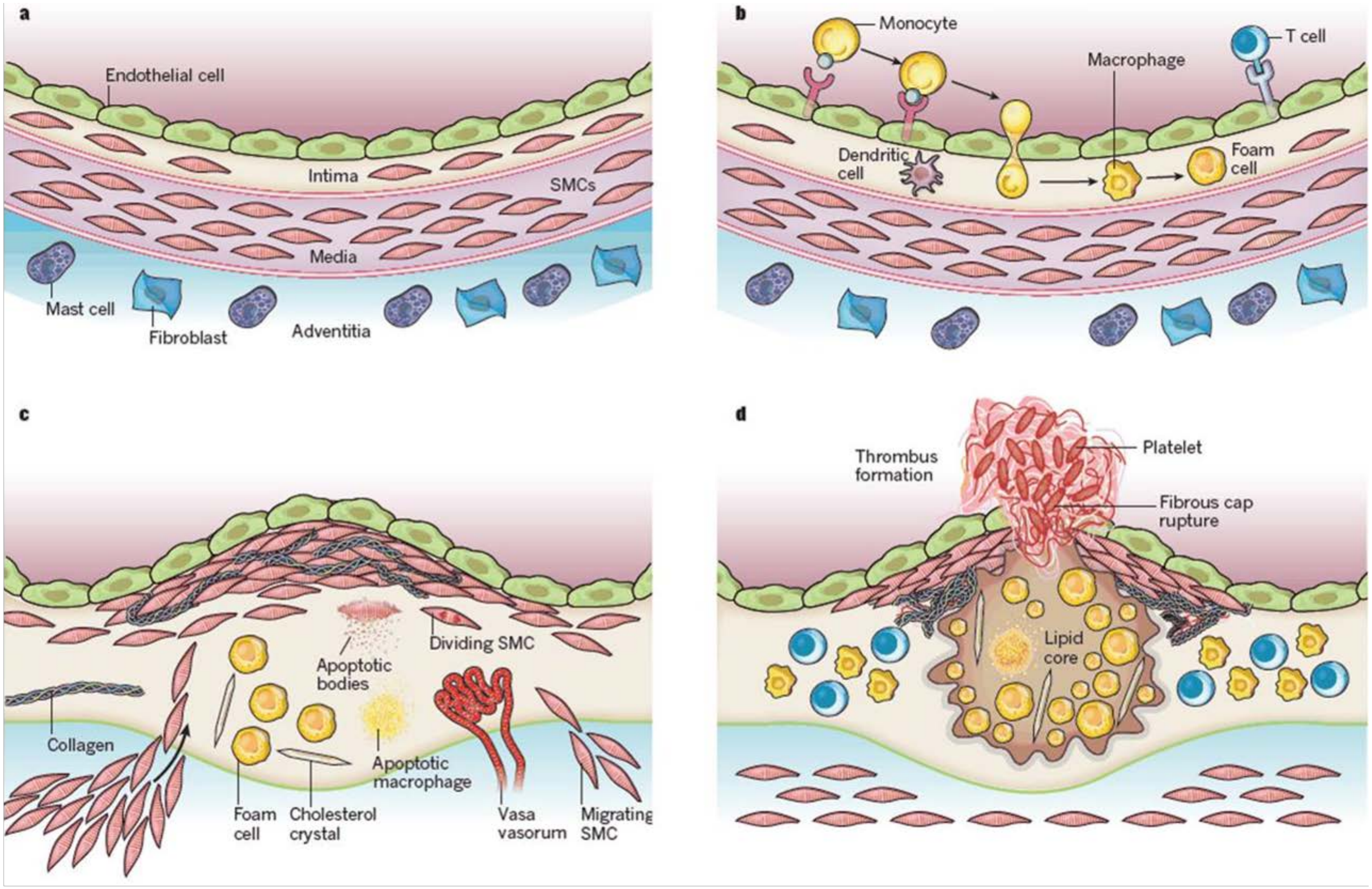


Figure 1.2. Stages in the development of atherosclerotic lesions. Taken from Libby, Ridker & Hansson 2011.

The main leukocyte infiltration includes mononuclear phagocytes (monocytes) and lymphocytes (T cells) and to a lesser degree polymorphonuclear granulocytes (mainly Neutrophils). Signals are immediately sent to the endogenous cells of the arterial wall – endothelial cells and SMCs by mediators of inflammatory and immunity pathways. This results in an almost simultaneous increase in permeability of the endothelium and the underlying basement membrane, by inflammatory mediator autacoids, such as histamine and leukotrienes (eicosanoids), which enablie recruitment and retention of cholesterol carrying low density lipoproteins (LDL) (Libby & Theroux 2005).

ECM secreted inflammatory mediators such as proteoglycans can trap LDL, particularly oxidised LDL (oxLDL), within the intima via their hyper-elongated glycosaminoglycan (GAG) chains forming complexes of lipid droplets that accumulate initially as an early grade fatty streak (as seen under the microscope) and a diffuse intimal thickening. As the lipid level accumulates still further, the fatty streak as visualised under the microscope,

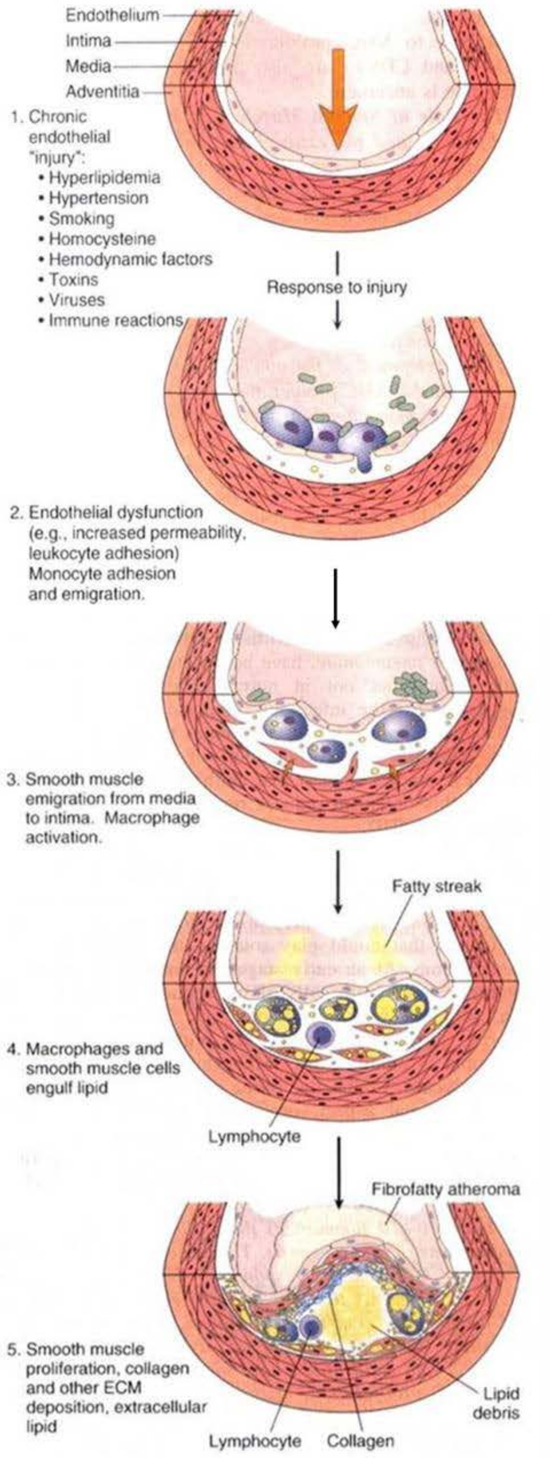

Endothelium
Intima
Media
Adventitia
1. Chronic endothelial "injury":
• Hyperlipidemia
• Hypertension
• Smoking
• Homocysteine
• Hemodynamic factors
• Toxins
• Viruses
• Immune reactions
Response to injury
2. Endothelial dysfunction (e.g., increased permeability, leukocyte adhesion) Monocyte adhesion and emigration.
3. Smooth muscle emigration from media to intima. Macrophage activation.
Fatty streak
4. Macrophages and smooth muscle cells engulf lipid
Lymphocyte
Fibrofatty atheroma
5. Smooth muscle proliferation, collagen and other ECM deposition, extracellular lipid
Lipid debris
Lymphocyte
Collagen

Figure 1.3. Evolution of artery wall changes in the response to injury hypothesis. 1. Normal; 2. Endothelial injury with adhesion of monocytes and platelets (the latter to sites where endothelial cells have been lost); 3. Migration of monocytes and SMCs into the intima; 4. SMC proliferation in the intima with ECM production; 5. Well-developed plaque. Taken from (Kumar et al. 2009d).

becomes a more substantial higher grade fatty streak that is then considered to be a pathological intimal thickening (Businaro et al. 2012).

Once within the intima LDL chemoattract and stimulate monocytes (the most abundant of leukocytes in the plaque) to differentiate into macrophages that locally along with endothelial cells expose LDL to: 1) non-enzymatic oxidation, e.g. copper II cations (Cu^{2+}) (Steinbrecher et al. 1984), haem (Miller, Altamentova & Shaklai 1997) and glycation (Libby & Theroux 2005); 2) enzymatic oxidative modification e.g. lipoxygenases (Parthasarathy, Wieland & Steinberg 1989), myeloperoxidase (Savenkova, Mueller & Heinecke 1994), nicotinamide adenine dinucleotide phosphate (NADPH) oxidase (McNally et al. 1990), and other peroxidases that leads to further recruitment of leukocytes and SMCs into the intima, thus continuing the immune response. The presence of oxLDL is cytotoxic to endothelial and SMCs alike and leads to endothelial dysfunction. As the fatty streak oxLDL lipids accumulate, macrophages engulf them via surface scavenger receptor (CD36) phagocytosis, and they subsequently become ‘lipid-laden’ foam cells, (so-called because of their microscopic appearance). One would think that the removal of dangerous oxLDL by macrophages is a protective immune response, but resident macrophages tend to have a commonly described ‘M1’ pro-inflammatory phenotype, whereby upon binding to oxLDL via surface CD36 (integral membrane proteins) they trigger the adaptive immune response. This causes the expression of pro-inflammatory cytokines such as interleukin 1β (IL-1β) and tumour necrosis factor (TNF), as well as, mimicry of dendritic cells by presenting surface antigens via the major histocompatibility complex (MHC) class II molecules to attract more $CD4^+$ T-helper cells (lymphocytes). These cytokines activate endothelial cells, SMCs, macrophages and lymphocytes to generate further pro-inflammatory cytokines. This results in a low level chronic self-perpetuating inflammation that no longer requires oxLDL to continue and is represented by increased systemic levels of not only cytokines, but also C-reactive protein (CRP) (Little, Chait & Bobik 2011).

The atherosclerotic lesion progresses from the fatty streak to a mature atheroma by recruiting SMCs from the media into the intima. The intimal space SMC migration and proliferation are regulated by growth promoter mediators, such as platelet derived growth factor (PDGF) (secreted by platelets and also macrophages, endothelial cells and SMCs), fibroblast growth factor (FGF), endothelin-1, thrombin, interferon-gamma (INF-γ) and interleukin 1 (IL-1); or growth inhibitors such as NO, TGF-β and heparan sulphates (member of the GAG family). SMCs secrete extracellular matrix components that include collagen, elastin and proteoglycans to form a superficial protective fibrous cap that contains SMCs and dense collagen, whereas to the sides and beneath the cap ‘the shoulder’ is an area that consists primarily of SMCs, macrophages and T cells.

Encapsulated below the fibrous cap are plenty of macrophage-derived foam cells that with the passing of time, along with other cells within the plaque (e.g. SMCs) are prone to die by apoptosis. This causes the release of lipids and tissue factor extracellularly. A failed process of efferocytosis (the inefficient clearance of dead cells) results in the accumulation of cell debris and extracellular cholesterol crystals, to form a lipid-rich pool known as the necrotic core.

The fibrous cap tends to encapsulate one of two plaque forms, one that is considered to be stable, and another that is said to be vulnerable, and more prone to rupture leading to acute coronary syndromes (ACS). The stable plaque tends to contain very little cholesterol, and an excess of SMCs and collagen to create a thick fibrous cap; whereas the unstable plaque has fewer SMCs and collagen, resulting in a thinner fibrous cap with a large lipid core.

The stable plaque is achieved in one of two ways, either the plaque was always stable due to its formation of a thick fibrous cap, to encapsulate a lipid-laden core that has been able to avoid plaque disruption; or it is a formerly unstable plaque that has undergone efficient thrombotic and fibrotic remodelling post plaque disruption, such that it had remained asymptomatic. The latter of these stable plaques probably now exists as a constricted, or fixed plaque due to the recruitment of collagen via PDGF and fibrin, via fibrinogen from exposure to blood that causes a hardening and stenosis of the coronary artery wall. In time this type of stable plaque could potentially become symptomatic (ischaemic) at a high percentage of stenosis, e.g. above 70%-75%, which would typically cause angina upon exertion.

The unstable or vulnerable plaque tends to occur in areas of low to moderate stenosis. However, there are some key factors that make the plaque prone to rupture. Firstly, the dynamic composition of the plaque's contents and structure, such as large numbers of foam cells, extracellular cholesterol deposition, and a thinned fibrous cap. The latter being caused by a net loss of collagen synthesis due to SMC depletion, and a net gain in collagen degradation, by mainly macrophage synthesised matrix metalloproteinases (MMPs) that outcompete tissue inhibitors of metalloproteinases (TIMPs), which are released by endothelial cells, SMCs and macrophages. Secondly, the additional influence of extrinsic factors, such as vascular shear forces from high blood pressure, brought on by vasoconstriction stimuli, which involve adrenergic agonists, endothelial dysfunction, and local release of platelet contents.

Other features of an advanced atherosclerotic plaque include calcification, in a process similar to bone formation (Demer 2002). Calcification occurs via medial SMCs that undergo intracellular micro-calcification, in cells adjacent to the outer section of the atheroma, during early stages of fatty streak formation. As the atheroma reaches an advanced stage these particular SMCs die, releasing extracellular calcium deposition (detectable by CT scans), causing a more stiffened artery.

In summary, atherosclerotic plaques are subjected to both pro- and anti-inflammatory mediators that have a direct influence on arterial wall remodelling pathways. There is: 1) a sub-clinical remodelling of the artery wall that enables compensatory enlargement; 2) an erosive or ruptured fibrous cap may lead to either: a) a fibrotic stenosis remodelling process increasing the severity of the stenosis but a so-called stable plaque, b) a partial thrombotic occlusion resulting in unstable angina, or c) an acute MI from a complete and sustained occlusion of the artery.

1.5 Risk factors

CAD is a multi-factorial disease that incorporates genetic, environmental and life style components, with many well recognised risk factors (see **Table 1.1 and Figure 1.4**). The age of onset and severity of CAD are dependent upon these risk factors, and it is important that clinicians and public health physicians have a thorough understanding of these risk factors, because even a small change in life style can decrease an individual's and the population's overall risk of CAD, and help to reduce cardiovascular morbidity and mortality. Risk factors for CAD were not formerly documented until the initial findings from the Framingham Heart Study in the early 1960s (Dawber, Moore & Mann 1957; Kannel et al. 1961; Rose 1964; Dawber & Kannel 1966); which led to the introduction of the 10 year CAD Framingham Risk Score calculator (Wilson et al. 1998). The so-called traditional (conventional or major) risk factors for CAD include modifiable risk factors such as, hypertension, dyslipidaemia, smoking, obesity, diabetes and physical inactivity; and non-modifiable factors such as, age, gender and premature family history. Indeed, the specific 2008 UK prevalence of the CAD modifiable risk factors are listed in **Table 1.2** (World Health Organisation (WHO) 2008), and shows that far more can be done to help improve morbidity and mortality.

The epidemiological evidence for many of the modifiable risk factors, has been supported by numerous, randomised placebo-controlled clinical trials, in CAD risk factors such as hypertension, hypercholesterolaemia and smoking; and has been associated with a 30% to 40% reduced incidence of CAD-associated clinical events (Wald & Law 2003).

Other putative or "emerging" risk factors are also listed in **Table 1.1**, although at the current stage of knowledge it is not known whether all of these are causal risk factors or risk markers. It is important to note that an individual with multiple major CAD risk factors is predisposed to a multiplicative effect. For instance, the Framingham study showed that the risk of an MI was seven times higher for those who had the 3 risk factors hyperlipidaemia, smoking and hypertension versus no risk factors; and four times higher for those who had only two of these risk factors (Kumar et al. 2009).

Table 1.1. Major and emerging life styles and characteristics associated with increased risk of CAD (*****Major/Traditional Risk Factors for CAD)

Modifiable Risk Factors		**Non-modifiable Risk Factors**
Behavioural / Life style	***Medical – biochemical and physiological characteristics***	***Personal characteristics***
Unhealthy Diet***** Smoking***** Physical inactivity***** Excess Alcohol consumption Stress Psycho-social	Hypertension***** Elevated total cholesterol***** Elevated LDL cholesterol***** Low plasma HDL cholesterol***** Diabetes (Hyperglycaemia)***** Hypertriglyceraemia***** Obesity***** Thrombotic factors Metabolic Syndrome Kidney Disease Sleep apnea Pre-eclampsia Biomarkers – C-reactive protein (CRP), Fibrinogen, Homocysteine and Lipoprotein-a [Lp(a)]	Old age Male gender Family history of premature CAD or other atherosclerotic vascular disease (men <55yr & women<65yr)***** Ethnicity

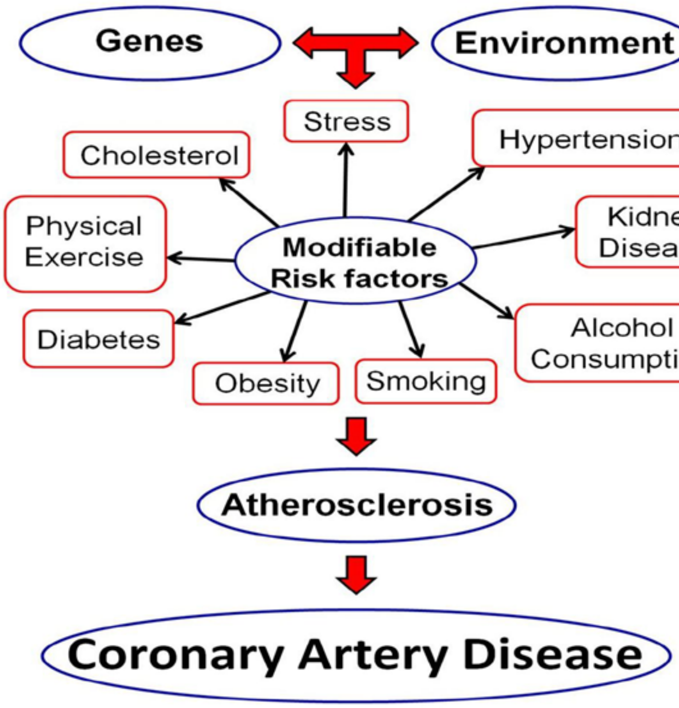


Figure 1.4. Schematic of the interplay between genetic, environmental and modifiable risk factors that lead to CAD.

Table 1.2. Prevalence of modifiable CAD risk factors in the UK (WHO, 2008)

Traditional modifiable	**UK Prevalence (WHO, 2008)**		
CAD risk factors	**Male** % (95% CI)	**Females** % (95% CI)	**Age range**$^{\$}$ (yr+)
Hypertension (BP ≥ 140/90 mmHg, or on medication)	42.2 (36.3-48.7)	32.8 (27.0-39.0)	25+
Dyslipidaemia (Raised total cholesterol (≥5.0mmol/L))	65.2 (48.0-79.7)	61.3 (40.8-77.3)	20+
Current Smoking of any tobacco (2009)	25	23	15+
Overweight/Obesity (Body mass index (BMI) (≥25kg/m^2))	65.6 (62.1-68.9)	57.5 (52.8-61.8)	20+
Physical inactivity (<5 X 30min moderate or < 3 X 20min vigorous activity per week)	58.0$^{\Phi}$ (56.8-59.1)	68.6$^{\Phi}$ (67.6-69.6)	15+
Diabetes (Raised fasting blood glucose (≥7.0mmol/L))	7.8 (3.8-13.4)	5.7 (2.7-10.1)	25+
Alcohol consumption (litres of pure alcohol per person per year)	13.24$^{\sharp}$		15+

Legend. Data taken from world health statistics 2008 (World health organisation (WHO)). $^{\$}$age-standardised estimates, $^{\Phi}$worst figures in Western developed countries, $^{\sharp}$third worst figures in Western developed countries.

1.6 Genetics of CAD

Familial aggregation of CAD has been recognised by physicians for over 100 years (Motulsky & Brunzell 2003). For example, William Osler in 1897 observed that AP presented in several generations; and Charles Hilton Fagge in 1872 described the presence of fatty deposits in both skin and coronary arteries for a multigenerational kindred, and is presumed to have been describing autosomal dominant familial hypercholesterolaemia (FH) (Osler 1897; Fagge 1872).

Twin and family studies have shown us that estimates of heritability for CAD are frequently high, and often in excess of 50% (Marenberg et al. 1994; Lusis, Mar & Pajukanta 2004). A list of heritability estimates for various risk factors of MI are shown in **Table 1.3**. These estimates of heritability must be considered only as approximate, because of assumptions required for their calculation. Heritability will also vary depending upon age (i.e. heritability is higher in those of a younger age for MI), ethnicity, and environment.

Heritability is described as the ‘proportion of inter-individual differences in risk resulting from genetic factors’. The simplest conceptual study design for heritability estimates have been made by comparing monozygotic and dizygotic twins. As monozygotic twins share 100% and dyzygotic twins 50% of their genes, one would expect that if a genetic component for the CAD phenotype exists, it would be more likely seen in monozygotic rather than dizygotic twins. Indeed Zdravkovic and colleagues have compared both monozygotic and dizygotic twins in a large (36 year follow-up) Swedish twin study, providing a 38% to 57% estimate of heritability for CAD (Marenberg et al. 1994; Zdravkovic et al. 2002). These estimates of heritability have led to claims that family history is the most important independent risk factor for CAD (Lusis, Mar & Pajukanta 2004).

The increased relative risk of premature CAD for family members, if a first degree relative is affected with premature CAD, is another key measure of the importance of genetics within families. For example, it has been observed in one study that where there was a first degree relative that had premature CAD, for women under 65 years and men under 55 years - it resulted in a 5- to 7-fold increase in relative

Table 1.3. Genetic and environmental risk factors for CAD with heritability scores

Medical risk factors with a significant genetic component (heritability[2])
Myocardial Infarction (25% to 60%)
Total cholesterol (40% to 60%)
High Density Lipoprotein – cholesterol (HDL-C) (45% to 75%)
Total triglycerides (40% to 80%)
Body mass index (25% to 60%)
Systolic blood pressure (50% to 70%)
Diastolic blood pressure (50% to 65%)
Lp(a) levels (90%)
Homocysteine levels (45%)
Type 2 diabetes (40% to 80%)
Fibrinogen (20% to 50%)
C-reactive protein
Gender
Age
Environmental risk factors with a significant genetic component (heritability)
Smoking (49% to 56%)[3]
Alcohol dependence (35% to 60%)[4]
Physical inactivity (25% to 60%)[5]
Unhealthy diet
Infection
Fetal environment
Air pollution (particulates)

Table adapted from (Lusis 2003). [2]Medical risk factor heritability estimates, in most cases based on multiple studies, are taken from (Lusis et al. 2002; King, Rotter & Motulsky 2002; Jee et al. 2002; Tournier-Lasserve 2002). [3](Broms et al. 2006). [4](Lyons et al. 2006). [5](de Vilhena e Santos et al. 2012).

risk of CAD (Slack & Evans 1966). Additionally, Marenberg and colleagues showed in the same large (26 year follow-up) Swedish twin study (mentioned earlier in relation to Zdravkovic *et al.*) that there was an 8- to 15-fold increase in risk of death from CAD in the second twin, if the first monozygotic twin had already died of MI before the age of 55 years in men, and before 65 years in women (Marenberg et al. 1994).

1.6.1 Mendelian monogenic forms of CAD

Due to its multifactorial and complex nature, the majority of CAD does not segregate as a Mendelian trait, where only one gene with a large effect is wholly responsible for the disorder. However, in some rare instances single gene mutations have been identified, as being sufficient to induce the phenotype of CAD under an autosomal dominant (e.g. FH (Goldstein, Hobbs & Brown 2001)), or autosomal recessive (e.g. sitosterolaemia (Lee et al. 2001; Rios et al. 2010)) Mendelian pattern of segregation. (Autosomal dominant segregation in families means that you will be an affected individual, if you are heterozygous for a genetic risk allele on one of the autosomes (chromosome 1 to 22). Autosomal recessive means you are only affected, if you are homozygous with both copies of the risk allele on one of the autosomes, i.e. one copy of the risk allele would make you a non-affected heterozygous carrier. X-linked inheritance refers to genetic mutations on the X chromosome, females have two X chromosomes, whereas males have one X and one Y chromosome. There are two X-linked inheritance forms – X-linked recessive and X-linked dominant. In the more common X-linked recessive inheritance - females who have only one copy of the mutation, do not usually express the mutation, they tend to be carriers, as it tends to be rare for a woman to get two copies of an X-linked mutation. However, males will always be affected as they only get one copy of the X chromosome, and this means all males will be X-linked recessive affected hemizygotes. In the rarer X-linked dominant disorders - only one copy of the mutation allele is sufficient to cause disease from either parent who has the disorder. Therefore males and females are likely to be equally affected, but males being hemizygous for the X chromosome, i.e. just one copy, tend to show higher gene expression than for females who have two X chromosome copies, but only one that carries the dominant mutation).

Mendelian single gene disorders that either cause CAD or affect risk factors for AS, have provided key insights into CAD biological pathways; and have helped to pave the way for the development of new life-saving treatments (Goldstein, Hobbs & Brown 2001). For instance, there are several examples of monogenic disorders that raise plasma levels of low density lipioprotein cholesterol (LDL-C), by impairing and thus reducing, the number of hepatic LDL receptors (LDLR) that are available to uptake LDL. The most common of these is FH, in which more than 600 mutations that cause the disorder, have been identified within the *LDLR* gene (Goldstein, Hobbs & Brown 2001). This monogenic disorder has a prevalence in the heterozygote form of 1 / 500 in Caucasians (Rader, Cohen & Hobbs 2003).

However, within affected families this disorder is very common, as it segregates with an autosomal dominant pattern. Patients who are heterozygous for FH have a 2-3 fold increase in plasma LDL-C levels, whereas patients who are homozygous have a 6-10 fold increase in plasma LDL-C levels, and usually die in childhood of MI due to severe AS (Bourbon et al. 2009).

Other autosomal dominant Mendelian disorders that raise plasma levels of cholesterol include: 1) familial ligand-defective apolipoprotein B-100 (FDB), in which mutations in the *APOB-100* gene cause a reduction in the binding affinity of apoB-100 to LDLR, and thus a diminished clearance of LDL-C plasma; the prevalence of this disorder in Caucasians is 1 / 1000 (Rader, Cohen & Hobbs 2003); 2) a rare autosomal dominant familial hypercholesterolaemia (HCHOLA3), with a Caucasian heterozygous prevalence of < 1 / 2500 (Rader, Cohen & Hobbs 2003), which is caused by gain-in-function mutations in the *PCSK9* (proprotein convertase subtilisin/kexin type 9) gene. *PCSK9* is a recently discovered subtilase (serine protease) that catabolises LDLRs via lysosomes, during cholesterol homeostasis, and is mainly expressed in the liver and small intestine (Abifadel et al. 2003).

Two rare monogenic lipid disorders, Tangier disease and sitosterolaemia have really helped our comprehension of sterol transport, and both conditions involve mutations within ATP binding cassette (ABC) transporter genes. The role of ABC transporters in CAD first became apparent, when mutations in the *ABCA1* gene were shown to cause Tangier disease. This is a rare autosomal recessive disorder best characterised by a major reduction in HDL-C, deposition of sterol in macrophages, and early AS (Bodzioch et al. 1999; Brooks-Wilson et al. 1999; Rust et al. 1999). The rare autosomal recessive monogenic disorder called sitosterolaemia, is another condition that elevates plasma LDL-C levels. It is caused by loss-of-function homozygous risk allele mutations, within either of two ATP binding cassette (ABC) transporter genes. These two genes encode for ABCG5 and ABCG8 proteins that act in concert, to export cholesterol into the intestinal lumen, so as to reduce cholesterol absorption (Lee et al. 2001; Rios et al. 2010).

Another very rare form of hypercholesterolaemia is ARH (autosomal recessive hypercholesterolaemia), this has a prevalence of < 1 in 10 million. ARH mutations cause a defect in the hepatic adaptor protein, which results in failure of cell surface LDLRs to clear plasma LDL-C (Garcia et al. 2001). Plasma levels of LDL-C then reach levels similar to homozygous FH (Barbagallo et al. 2003). Additionally, several severe HDL-C deficiency familial hypoalphalipoproteinaemia (FHA) syndromes can also lead to early onset CAD. These include autosomal recessive rare mutations in genes *APOA1 (*apolipoprotein A-1), the previously mentioned *ABCA1* gene in Tangier disease, and *LCAT* (lecithin cholesterol acyltransferase) that encode proteins for HDL-C formation, maturation and catabolic breakdown (Esperon et al. 2008).

1.6.2 Genetic architecture of CAD

Epidemiological studies over the past 60 years have identified many risk factors for CAD (**Table 1.3**). These risk factors include those with a significant genetic component in their own right (as shown by percentage estimates of heritability), and those that are mainly explained by environmental exposure (see **Table 1.3**) (Lusis, Mar & Pajukanta 2004). Nonetheless, even for some of the environmental risk factors, such as smoking (Broms et al. 2006), alcohol dependence (Lyons et al. 2006), and physical inactivity (de Vilhena e Santos et al. 2012), recent studies have identified a significant genetic determination. The sheer number of genes likely to be involved becomes ever more apparent, when one considers the number of genes involved with any of the risk factors described in **Table 1.3**; and yet these risk factors do not begin to account for the family history estimates of heritability, which are considered independent of these traditional risk factors.

It is believed that because CAD is a rare disorder before the age of 50 years, the condition is unlikely to have an effect on the success of reproduction, and thus its genetic determination is unlikely to be influenced by pressures of natural selection. Therefore genetic variants that might influence CAD susceptibility, or indeed protect against CAD, will probably have evolved in a neutral manner in the past, and so may appear over a wide spectrum of frequencies (i.e. both as common and rare genetic variants). This reasoning is the basis for the common disease / common variant (CD/CV) hypothesis, which states that a small number of common genetic variants at a frequency above 1-5%, are the main contributors to common complex disorders; and are considered to be relatively stable having undergone little or no selection in earlier ancestral populations, for perhaps 100,000 years (Lander 1996). However, an alternate common disease / rare variant (CD/RV) hypothesis has also been postulated, and states that there are a multiplicity of large effect rare genetic variants (<1%), with perhaps extensive allelic heterogeneity (i.e. a convergence of multiple rare variants at the same locus) or multiple loci (locus heterogeneity), which are the main cause of common disease. These are considered to be either new loci (i.e. just a few generations old, perhaps two centuries) that have not undergone significant negative selection (i.e. influenced by natural selection), or rare due to their deleterious nature (Pritchard 2001). Indeed, it is feasible that exploding population growth has resulted in an accumulation of extremely rare, yet potentially deleterious variation, and that this could explain some of the heritability of many common complex disorders. Moreover, it would be unwise to ignore the CD/RV hypothesis given that the CD/CV hypothesis suffers from the assumption that there will be little allelic heterogeneity; a phenomenon often seen in Mendelian disorders that will likely weaken statistical powers of association (Pritchard 2001).

The true complexity of AS occurs at a cellular and molecular level, and is modulated by genetic variants at every stage of AS progression. Although we know that premature clinical phenotypes of CAD and especially MI, show a strong basis of inheritance - we do not know, other than through rare Mendelian forms, which genetic variants are associated with increased risk. Therefore, given that most risk factors for CAD have a strong genetic component, and that family history estimates of heritability are not fully explained by traditional risk factors; it has been the goal of the past 30 years to identify genetic modulators of CAD, and elucidate how they function - in order that we can develop more relevant and comprehensive prevention and therapy. To this end, a vast number of studies have been performed, to help decipher the molecular mechanisms of CAD explained by heritability estimates, using several different approaches, to prove genetic variants have an influence on AS development and progression.

1.6.3 Polygenic inheritance of CAD

For the most part familial aggregation of CAD has a lack of Mendelian monogenic inheritance (<0.1%). Therefore, in order to investigate polygenic non-Mendelian patterns of inheritance, where each genetic variant has a low phenotypic penetrance (i.e. on its own merit it is unable to cause disease, without other contributory genetic variants, or environmental risk factors), two traditional approaches have been undertaken - the candidate gene approach and genome-wide linkage studies. In addition, recently due to the sequencing of the human genome, the subsequent identification of several million single nucleotide polymorphisms (SNPs), and their related linkage disequilibrium (LD) derived haplotypes (through the SNP Consortium and HapMap Project), a new unbiased approach has come into play, called a genome-wide association (GWA) study. This new approach can identify disease associated variants very quickly, in comparison to traditional genetic methods. In fact, the pace at which new variants are now being discovered has accelerated, at an even faster rate, due to ever improving SNP and haplotype information, statistical methodology (i.e. imputation and meta-analyses), and decreasing costs in high-throughput genotyping and next generation sequencing. In addition, there is a willingness amongst common complex disease study principal investigators worldwide, to form collaborative consortia, e.g. CARDIoGRAMPlusC4D (CARDIoGRAMplusC4D Consortium et al. 2013).

1.6.3.1 Different types of genetic variation

It is known that ~99.5% of the human genome sequence is identical amongst all humans, thus the remaining 0.5% (~16 million bp) is responsible for all human differences, and this includes disease susceptibility. The genetic variants are mainly made up of SNPs; di-, tri-, and tetra-nucleotide repeats called microsatellites; large variants called copy number variants (CNVs) of varying length, mostly >0.5kb caused by deletions, insertions or duplications; and short nucleotide substitutions evenly distributed throughout the genome.

It has been postulated that the genetic component responsible for complex disease is down to multiple low risk genetic variants, of which some are common, and others rare. The most frequent (~80%) non-repetitive sequence variants are SNPs, and these have advantages over previous genetic markers used in genetic analysis (i.e. microsatellite tandem repeats). Firstly, SNPs are biallelic, so their frequencies can be easily estimated in any population. Secondly, SNPs are very stable genetic markers compared to tandem repeat markers, whose high mutation rate can affect genetic analysis in different populations. Thirdly, technologies have been developed that can genotype SNPs simply, accurately and rapidly via automated methods. There are estimated to be on average one SNP for every 1000 base pairs (bp), when any two chromosomes are compared, and that the human genome has approximately 3 million common (>5%) SNPs per individual (International HapMap Consortium 2003); and perhaps as many as 17 million SNPs present in the general population (Roberts et al. 2010). SNPs are distributed evenly throughout the genome, so that every gene will be covered by several SNPs, which allows indirect detection of candidate genetic variants through the LD between a SNP marker of complex disease, and a functional variant of a gene, this is known as the proximity hypothesis (Collins 1999).

Prior to the advent of the GWA study researchers made use of a SNP candidate gene strategy that could be performed in one of two ways: 1) select candidate genes for CAD and focus on just scanning the coding sequence for potentially functional coding SNPs (cSNPs); 2) focus on a CAD candidate gene one by one, and systematically scan all of its genomic sequence for all possible polymorphisms. There are advantages and disadvantages to both these approaches.

The first cSNP approach was seen as a cost effective, and very good way of identifying functional variants, because non-synonymous coding SNPs were seen at the time, as the most likely to be functional. Indeed, non-coding DNA was seen as non-functional, and certainly non-coding DNA outside of the genes was seen as 'junk' DNA. However, in 'hind-sight' the so-called 'junk DNA' is now seen as potentially regulatory and important - and so this form of candidate gene approach fails to identify genetic variants that are located in regulatory non-coding regions. It is hoped that at least one SNP will be detected that is close enough to be in LD with the truly functional variant, and has a presumption that the causal SNP will be amongst the coding region; but lacks power to detect predisposing genetic variants when cSNPs are sparse, with perhaps just 3 or 4 SNPs in a 10kb random sampling of genomic sequence (Nickerson et al. 1998).

The second approach has the advantage of finding all potential functional polymorphisms, if enough individuals are analysed, but was considered costly in the era prior to the advent of high throughput genotyping; and restricted principal investigators, to only looking at regions already known through linkage or disease association (Johnson & Todd 2000). Examples of candidate gene studies that used the entire gene SNP scanning approach include: *ACE* (angiotensinogen converting enzyme), *LPL* (lipoprotein lipase), *APOE* (apolipoprotein E) and *ADRB2* (beta-2-adrenergic receptor), which are involved in lipid metabolism, and the renin angiotensinogen system (RAS) blood pressure pathway (Nickerson et al. 1998; Rieder & Nickerson 2000; Martin et al. 2000; Drysdale et al. 2000). These efforts really were the precursor to GWA studies.

1.6.3.2 Candidate gene approach

The first approach utilised for the discovery of genetic variants that predispose to CAD was the candidate gene approach. Up until recently this was the most widely used method to test whether genes (genotype) selected empirically from known atherogenic pathways, play a role in predisposing to increased CAD (phenotype) risk. This approach is achieved by comparing the allelic frequencies of selected genetic variants of these genes - in cases (those with disease) versus controls (those without disease) - and is called an association study. This approach has been able to take advantage of the milestone DNA technology - polymerase chain reaction (PCR), by genotyping SNPs within the candidate genes selected.

The best way to present statistical results of association studies is as an odds ratio (OR) with 95% confidence intervals (95% CI). Also, the best statistical test with good power under an additive, dominant and recessive genetic model, is the one degree of freedom Cochran-Armitage trend test, because each allelic copy with an OR greater than 1 - will be associated with a higher relative risk of disease.

Some of the key results in the literature for candidate gene studies of this type, are shown in **Table 1.4**. A large number of case-control studies were performed in the 1990s, testing numerous variants in multiple genes, and several of them, showed a significant association with risk for CAD. However, reproducibility through independent replication studies rarely upheld, previously implicated genetic variants of CAD risk (Voetsch & Loscalzo 2004; Morgan et al. 2007). There are several possible reasons for a lack of reproducibility: 1) a lack of statistical power to detect association due to insufficient sample size, i.e. often less than 500 cases and 500 controls; 2) potential for false-positive (type 1 error) due to unrecognised population stratification between cases and controls explaining different allele frequencies; 3) poorly defined homogeneity in CAD phenotypes and truly unaffected controls, due to ascertainment bias (for instance MI survival bias and the fact that all asymptomatic individuals whether case or control will have some level of atherosclerotic risk, but only a small percentage of them will ever develop a MI); and 4) poor *a priori* knowledge as to whether the encoded protein would actually be mechanistically relevant.

Of the few studies that have shown successful replication, the ε4 allele of the *APOE* gene, is one of the most persuasive, e.g. an adequately powered meta-analysis using 48 studies shows an associated 1.4-fold increase in risk of CAD (Song, Stampfer & Liu 2004). Another large meta-analysis of haemostatic genetic variants showed only two variants with a moderately increased risk of CAD, namely variants for factor V G1691A, prothrombin G20210A, and a weak association with the plasminogen activator inhibitor 1 (*PAI-1*) gene variant 4G-668/5G (Ye et al. 2006).

Table 1.4. Key results of studies on candidate genes[†]

Article (year)	Study design	No. of cases/controls	Clinical event	Gene variants associated with CAD	OR (95% CI)
Yamada (2002)	Case–control	2819/2242	MI	Connexin 37 C1019T	1.40 (1.10-1.60)
				PAI-1 4G-668/5G	1.60 (1.20-2.10)
				Stromelysin-1 5A-1171/6A	4.70 (2.00-12.2)
Ozaki (2002, 2004)	Case–control	1133/1006	MI	LTA	1.78 (1.39-2.27)
McCarthy (2004)	Case–control	325/418	MI	THBS2	0.38 (0.14-1.01)
				THBS4	2.22 (0.93-5.28)
				PAI-2	2.35 (0.90-6.12)
Song (2004)	Meta-analysis	15,492/32,965 (48 studies)	CAD	Apolipoprotein E ε4	1.42 (1.26-1.61)
Shiffman (2005)		1345/1843	MI	Palladin	1.40 (NA)
				ROS1	1.75 (NA)
				TAS2R50	1.58 (NA)
				OR13G1	1.40 (NA)
Shiffman (2006)	Case–control	1200/262	MI	VAMP8	1.75 (1.17-2.62)
				HNRPUL1	1.92 (1.28-2.86)
Ye (2006)	Meta-analysis	66,155/91,307 (191 studies)	CAD	Factor V G1691A	1.17 (1.08-1.28)
				Prothrombin G20210A	1.31 (1.12-1.52)
				PAI-1 4G-668/5G	1.06 (1.02-1.10)
Zwicker (2006)	Case–control	1425/1425	MI	THBS1 N700S	1.40 (1.10-1.80) (heterozygotes) 1.90 (1.10-3.30) (homozygotes)
Iakoubova (2006)	Case–control	2903/1080	MI	FCAR	1.68 (1.10-2.57)
Luke (2007)	Case–control	1806/1274	CAD	LPA	3.14 (1.51-6.56)
Iakoubova (2008)	Case–control	3394/1080	CAD	KIF6	1.50 (1.05-2.15)
Shiffman (2008)	Case–control	4522	MI	KIF6	1.29 (1.1-1.52)
				VAMP8	1.2 (1.02-1.41)
				TAS2R50	1.13 (1-1.27)
				LPA	1.62 (1.09-2.42)

CVD, cardiovascular disease; † Table adapted from (Franchini, Peyvandi & Mannucci 2008)

Due to technological advances during the human genome sequencing project, candidate gene association studies were able to take advantage of new microarray technology, by simultaneously genotyping multiple SNPs from multiple candidate genes. The first to use this technology was the GENEQUEST study, where 72 SNPs were genotyped for 62 candidate genes, in 398 sibling pairs with premature CAD. This resulted in three variants in three different thrombospondin genes (*THBS1*, *THBS2* and *THBS4*) being associated with premature MI (Topol et al. 2001). A follow-up study by McCarthy and colleagues, attempted to replicate these findings in another premature CAD association study, using another microarray style study that typed 210 polymorphisms in 111 candidate genes. However, this study was somewhat under-powered, because there were too few subjects and issues with multiple testing. In spite of these limitations, two of the thrombospondin gene variants (*THBS2* and *THBS4*) along with a *PAI-2* gene variant, showed a clear trend towards association with premature CAD, albeit without reaching statistical significance (McCarthy et al. 2004). Moreover, in another well powered study, the *THBS1* N700S gene variant was shown to be associated with an increased risk in young MI survivors (< 45 years) (Zwicker et al. 2006) (see **Table 1.4**).

In 2002, Yamada and colleagues performed another large scale association study that took advantage of the new (at the time) high-throughput microarray technology, but this time they used a two stage approach, to identify genetic variants associated with risk of MI. The first stage involved screening 112 SNPs, in 71 candidate genes, in 909 unrelated patients, and they identified 19 SNPs in men and 18 SNPs in women that showed associated risk with MI. The Yamada group then attempted to replicate their findings in a larger study of 2819 unrelated MI cases and 2242 controls. Two variants, connexin 37 C1019T and *PAI-1* 4G-668/5G, were shown to be significantly associated with MI in men, and one variant stromelysin-1 5A-1171/6A was shown to be significantly associated with MI in women (Yamada et al. 2002) (see **Table 1.4**).

A study by Ozaki and colleagues in 2002 was the first study to undertake a large scale candidate gene approach. They genotyped 65,671 SNPs in a Japanese population, which consisted of 94 MI cases and 658 healthy controls. A significant association was detected for two SNPs within the *LTA* (lymphotoxin alpha) gene, which encodes for a TNF (tumour necrosis factor) ligand family (Ozaki et al. 2002). This discovery was then replicated by the same investigators, along with an association for a SNP within *LGALS2*, a gene that regulates *LTA* (Ozaki et al. 2004).

An additional large scale candidate gene study was performed by Shiffman and colleagues, by firstly genotyping 11,053 SNPs in 6891 genes in 1345 MI cases and 1843 controls, using three sequential MI case-control studies (Shiffman et al. 2005); and secondly genotyping 11,647 SNPs in 7136 genes in 1200 cases and 262 controls, using three sequential early-onset MI case-control studies (Shiffman et al. 2006). These two Shiffman *et al.* studies identified four novel variant MI associations and two novel early-onset MI associations, respectively (Shiffman et al. 2005; Shiffman et al. 2006). Subsequently, five further MI or CAD associated SNPs were replicated by candidate gene case-control studies (Iakoubova et al. 2006; Luke et al. 2007; Iakoubova et al. 2008; Shiffman et al. 2008). Despite these large scale candidate gene studies having limited power, they still managed to detect a small percentage of CAD associated genetic variants (see **Table 1.4**).

Nonetheless, although many candidate gene studies of CAD have been performed, the take home message is that very few have replicated. The reasons for this have in part already been stated, namely, poor study design with respect to study size, population stratification, and multiple testing. However, there is another reason, and that is a poor coverage of markers across any given candidate gene. Only picking one or two variants from a candidate gene, is not truly assessing all the common variants, or indeed rare variants that could potentially be associated with the risk of CAD (this will be addressed later under genome wide association studies).

1.6.3.3 Genome-wide linkage studies

To overcome several of the limitations of candidate gene studies, another approach - genome-wide linkage studies based on affected sibling pairs (ASPs) – gained popularity in the 1990s. This is an unbiased method with no *a priori* assumptions made about genes that might be associated with CAD, but yet has the potential to discover *de novo* atherogenic genetic variants. Initially, common complex diseases such as CAD, were not studied by linkage analysis. This was due to its polygenic nature, its presentation late in life (meaning the proband may have no parents), and it's being potentially fatal, leading to loss of genetic information and survival bias.

However, a statistical approach using a non-parametric (model-free) linkage analysis in ASPs, to look for allele, or chromosomal sharing within known extended families; has since been successfully employed in many human atherosclerotic disease linkage studies (Chen et al. 2007). The basic concepts behind the ASP approach were first set out by Neil Risch, in three back-to-back publications in 1990 (Risch 1990a; Risch 1990b; Risch 1990c). The ASP approach looks at two disease-affected siblings, and evaluates sharing of alleles at multiple sites within the genome, using a linkage mapping set of markers. The principle is one where, the more often the affected siblings share the same allele at a particular site, the more likely the site is close to the disease gene. Alleles shared by ASPs can be either identical-by-state (IBS), or identical-by-descent (IBD), and it is important for analysis purposes to be able to distinguish between the two. IBS alleles appear to be the same, but may be derived from different parents, whereas IBD alleles are derived from the same parent; this is particularly important if only one parent carries the risk allele for disease.

There are statistical methodologies for linkage analysis, which exist to be able to deal with, either IBS or IBD sharing of alleles; and this is important in the case of ASPs, as the parental alleles are unknown (IBD is the more powerful method, but requires more affected relatives). Siblings are expected to share alleles 50% of the time, if there is no linkage. In order to detect linkage we need to calculate a point-wise statistic *p*-value at each marker along the genome, and identify regions that show significant deviation from that expected by random independent assortment, which is a ratio of 1:2:1 (i.e. ASPs are expected to share 0, 1 or 2 parental haplotypes with a frequency of 0.25, 0.5 and 0.25 respectively) when alleles are IBD. However, a multi-point rather than single-point analysis is usually preferred, because it does a better job at extracting IBD sharing information across a chromosomal region (Strachan & Read 2004).

The genome-wide linkage analysis needs to identify the genome-wide significant deviation, which is the probability that deviation will occur somewhere in the whole genome scan. This is represented statistically by a LOD score, the log-likelihood odds ratio (LOD) of excess sharing, as compared with the null hypothesis that there is 50% sharing (i.e. no linkage). The genome-wide statistic LOD score is related to a χ^2 distribution, when 'n' is small, whereas the *p*-value point-wise statistic is related to a normal distribution, when 'n' is large. Statistically, the ASP approach is seen as statistically robust for detecting linkage, because non-parametric statistics using large numbers of ASPs move towards a normal distribution (Strachan & Read 1999).

To achieve a viable linkage analysis, the LOD score significance must have a probability of below 5% ($P<0.05$), to avoid false positive (type 1) errors. Traditionally, a LOD score of 3.0 has been used to indicate significant linkage, and would seem reasonable based on simulated data using 100 ASPs with 10cM microsatellites markers (i.e. short tandem repeat sequences) (Lander & Kruglyak 1995). However, asymptotic expectation, using mathematical theory explained by Lander & Kruglyak 1995, has indicated that to ensure avoidance of a type 1 error, a more stringent LOD score of 3.6, is recommended to achieve genome-wide statistical significance (Lander & Kruglyak 1995). Risch and Merkangas have run simulations to indicate the number of ASPs and Transmission disequilibrium test (TDT) trios (2 parents and proband) required, to achieve a given power and significance level. They showed that ASPs can only realistically be used to detect high susceptibility loci, and that detection of a LOD score of <3.0 for moderate /

weak susceptibility loci requires an unfeasibly high sample size; whereas TDT trios can plausibly detect LOD scores of <2.0 with a realistic sample size (Risch & Merikangas 1996).

The method of choice for performing genome-wide linkage studies in complex diseases such as CAD, is to first recruit hundreds of families with an emphasis for premature CAD case ASPs (I was involved in the recruitment phase of the flagship UK-wide BHF-FHS (British Heart Foundation Family Heart Study) that collected nearly 2000 ASP families) (Samani et al. 2005). Subsequent to recruitment, genome-wide linkage is achieved by scanning 400 evenly spaced microsatellite markers across the genome. The goal was to achieve a genome-wide significant LOD score of >3.5 ($P<1x10^{-6}$), which would indicate a gene that is in linkage disequilibrium (LD) near to, or even within the microsatellite region.

LD is defined as a measure of the non-random association between two alleles at nearby loci; (i.e. if a particular allele at one locus is found together with a specific allele at another locus, on the same chromosome, more often than would be expected if the loci were segregating by independent assortment in a population, the loci are said to be in LD). The LD between any two genetic markers on a chromosome are most commonly measured either by D' (D prime – disequilibrium coefficient) or r^2 (square of the correlation coefficient). A measure of strong LD would reflect a situation where the two adjacent alleles have been rarely separated over time, and are therefore likely to segregate together within the same common haplotype - and hence be linked. This phenomenon can be used to pinpoint genetic risk variants under significant linkage LOD score peaks.

The findings of the main linkage studies for CAD are summarised in **Table 1.5**. The first genome-wide linkage study was performed in Finnish premature CAD patients, and showed significant linkage with two loci, on chr2 and chrX (Pajukanta et al. 2000). A lower threshold genome-wide significance for linkage was shown on chr16, in a large number of Mauritian families (Francke et al. 2001). Wang and colleagues discovered a significant linkage for a locus on chr15, using a single large family, where 13 members were affected with premature CAD, and 9 had suffered with a MI (Wang et al. 2003; Wang et al. 2004). The candidate gene identified for this locus region is *MEF2A* (myocyte enhancer factor 2A), a transcription factor expressed in the endothelium of arteries. A subsequent study has shown in non-angiography tested controls that a missense coding mutation, P279L in *MEF2A* was significantly associated with increased risk of MI, OR=3.06 (Gonzalez et al. 2006).

Table 1.5. Linkage studies in families with CAD

First author (year)	Population origin	Study sample	Clinical event	Chr region	LOD score	Gene identified
Pajukanta (2000)	Finland	156 families (364 individuals)	Premature CAD	2q21-22, Xq23-26	3.7,3.5	No
Francke (2001)	Mauritius	99 families (535 individuals)	CAD	16p13	3.06	No
Broeckel (2002)	Germany	513 families (1406 individuals)	Premature CAD	14q32	3.9	No
Harrap (2002)	Australia	61 families (161 individuals)	CAD	2q36	2.6	No
Wang (2003)	USA	One family (25 individuals)	CAD or MI	15q26	4.19	MEF2
Helgadottir (2004)	Iceland	296 families (2454 individuals)	MI	13q12-13	2.86	ALOX5AP
Hauser (2004)	Euro-American	438 families (1168 individuals)	Premature CAD	3q13	3.3	No
Samani (2005)	UK	1933 families (4175 individuals)	CAD or MI	2q14-21	1.86	No
Helgadottir (2006)	Iceland	296 families (2454 individuals)	MI	12q22	NA	LTA4H
Farrall (2006)	Europe	2036 families (2658 individuals)	CAD or MI	17q21	2.68	No

NA indicates not available. Table taken from (Franchini, Peyvandi & Mannucci 2008).

Another exciting discovery was made by Helgadottir *et al.* in 2004, where an Icelandic family MI linkage study, made use of a more dense 5cM linkage mapping set of 1068 microsatellites, and discovered an initially suggestive linkage signal on chr13. However, with the genotyping of an additional 120 microsatellite markers for this region, via an association study in MI cases and controls, they found a 4-SNP haplotype within the *ALOX5AP* gene (which encodes for the 5-lipoxygenase activating protein FLAP as part of the inflammatory leukotriene B4 production pathway) associated with a two-fold increase in MI risk. In follow-up replication studies, *ALOX5AP* haplotypes were shown to be significantly associated with CAD in an English cohort, and with stroke in Icelandic and Scottish subjects (Helgadottir et al. 2004; Helgadottir et al. 2005). Moreover in 2006, Helgadottir reported on a new gene *LTA4H* (which encodes for leukotriene A4 hydrolase) that was identified using the same genotyping approach (Helgadottir et al. 2006).

The BHF-FHS linkage analysis of premature ASPs from 1933 families across the UK, is one of the largest linkage studies performed, and yet it produced not one significant LOD score, the best was a weak signal on chr2 (LOD=1.86) (Samani et al. 2005). In addition, the remaining linkage studies mentioned in **Table 1.5** showed linkage peaks on chr 2, 3, 14 and 17, but no definite susceptibility gene for CAD risk (Broeckel et al. 2002; Harrap et al. 2002; Hauser et al. 2004; Farrall et al. 2006). In retrospect, although ASP-based genome-wide linkage analysis has many attractive features, it is clear that even with 2000 or more families, the studies lacked sufficient power to definitively identify candidate CAD loci, given its complex polygenic nature.

1.6.3.4 Genome wide association studies

The inherent limitation of candidate gene association studies and the limited power of genome-wide linkage analysis, led to a hiatus in progress in elucidating the genetic basis of common diseases, until the advent of genome-wide association studies. The possibility of whole genome association studies only became a reality, as a consequence of the collective efforts from The Human Genome Project (Lander et al. 2001), The SNP Consortium (Sachidanandam et al. 2001), and The International HapMap Project (International HapMap Consortium et al. 2007) that have identified in a limited set of DNA samples, ~10 million common variants of which most are SNPs. Indeed, it was the key research from earlier studies that reported an unexpected correlation of African, European and Asian haplotype blocks and recombination hot spots (i.e. lengths of chromosomal sequence that showed very little historical recombination, represented by only a few common haplotypes, and flanked by boundaries of high recombination) that led to the initiation of the Human Haplotype Map project (Reich et al. 2001; Daly et al. 2001; Gabriel et al. 2002).

The mapped locations and linkage disequilibrium (LD) patterns of the HapMap project SNPs, as well as, actual (TDT trio) genotype data for 3.8 million common SNPs provides a powerful tool, by which to select tag SNPs that can be used as proxies for the majority of the SNPs that remain. This means that a smaller subset of tag SNPs can be genotyped, and yet because of strong LD, the information on many more SNPs can be obtained. Tagging SNPs tend to be selected on the basis of the minimum number of SNPs that need to be genotyped, in order to capture, all other SNPs at a particular threshold of strong LD (see **Table 1.6**). The highest LD threshold r^2=1, will capture all tag SNPs (i.e. apart from those that are redundant) within HapMap phase II with a minor allele frequency (MAF) ≥0.05, and has the highest power to detect novel loci. However, the number of SNPs that will need to be genotyped is much higher for r^2=1, than if a commonly used tagging SNP threshold of r^2=0.8 was applied (see **Table 1.6**).

Table 1.6. Number of tag SNPs required to capture common (MAF≥0.05) HapMap Phase II SNPs

Threshold	YRI	CEU	CHB/JPT
r2 ≥ 0.5	627,458	290,969	277,831
r2 ≥ 0.8	1,093,422	552,853	520,111
r2 = 1.0	1,616,739	1,024,665	1,078,959

Table taken from (International HapMap Consortium et al. 2007)

At the same time as these scientific advances, technological breakthroughs in microarray technology, enabled the simultaneous high throughput genotyping of up to 1 million SNPs. This means that unbiased GWA studies with no *a priori* hypothesis can now be achieved, and has in fact led to the successful identification of hundreds of novel genomic loci that influence human diseases (Donnelly 2008). In a sense GWA studies capture the best features of the two previous approaches – an association based analysis (allowing greater power) and a genome-wide unbiased approach. The accuracy of these genotyping array platforms is extremely high, but a penalty is the potential for false positive findings (type 1 error). For instance, a 1 million SNP array could produce 50,000 false positives, when performing a case / control association study with a *P*-value threshold of $P<0.05$. This means there is a need to correct for multiple testing. A simple way in which this can be achieved is via the Bonferroni correction. This is calculated by dividing the alpha (α) level of significance for a single hypothesis experiment, i.e. the $P<0.05$ threshold, by the number of SNPs (n) tested; hence, a 1 million SNP array provides a genome-wide significance threshold of $p<5x10^{-8}$ (i.e. α/n). This *P*-value threshold is only relevant in the GWA study discovery phase, subsequent replication studies only require that you divide $P<0.05$ (α), by the number of replication SNPs (n) taken forward (e.g. a 100 SNPs would only require a (α/n) significance level of $P<5x10^{-4}$).

Another issue with GWA studies is deciding the sample size required, because although they have the power to detect small or moderate effects for common SNPs (MAF≥0.05), they do not tend to have the power to detect these subtle effects in SNPs with a MAF<0.05. In general most studies, only have ≥80% power to detect moderate effects in SNPs with a MAF>0.1, at the genome-wide significance level ($P<5x10^{-8}$). The lack of power to detect small effects in common allele frequency SNPs, and moderate effects for low frequency SNPs, is due to a limit in study sample size. It would therefore require a combined discovery stage meta-analysis of several GWA studies, in order to gain the greater power needed to detect small effects in SNPs with a MAF≥0.05, and moderate effects in SNPs with a MAF<0.05, as shown in **Table 1.7**. For this reason, single GWA study statistical strategy has been to use a two stage procedure, whereby the primary discovery stage genome-wide significant findings must be replicated in another study, using different case / control subjects; in order to overt the issue of false positive results.

Limitations of GWA studies include - incomplete capture of all SNPs, resulting in the requirement for imputation or additional genotyping, and that most CNVs are not captured via SNP microarrays.

Table 1.7. Sample size (case–control pair) needed with power ≥80%, to detect genome-wide significance $p<5\times10^{-8}$

MAF	Odds Ratio		
	1.1	1.2	1.3
0.05	88,731	23,537	11,069
0.1	47,015	12,519	5,910
0.15	33,318	8,905	4,219
0.2	26,654	7,151	3,400
0.3	20,466	5,531	2,648
0.4	18,046	4,912	2,367
0.5	17,457	4,785	2,321

Taken from (Roberts et al. 2010)

The main results for GWA studies are listed in **Table 1.8**. Indeed, a relatively strong association at locus 9p21.3 has been shown in multiple GWA studies, in different populations, to increase the risk of CAD and MI (Helgadottir et al. 2007; McPherson et al. 2007; Wellcome Trust Case Control Consortium 2007; Samani et al. 2007).

Table 1.8. Early genome-wide association studies in patients with CAD

Article (year)	Population screened	No. of cases/ controls	Clinical event	Genes/chr loci associated with CAD	OR (95% CI)
Helgadottir (2007)	European	4587/12,767	MI	9p21	2.02 (1.72-2.36)
McPherson (2007)	European	2326/10,427	CAD	9p21	1.20 (1.02-1.42)
WTCCC (2007)	European	1926/2938	CAD	9p21	1.37 (1.26-1.48)
Samani (2007)	European	2801/4582	CAD	9p21	1.36 (1.27-1.46)
				6q25	1.23 (1.15-1.33)
				2q36	1.21 (1.13-1.30)

WTCCC (Wellcome Trust Case Control Consortium). Table adapted from (Franchini, Peyvandi & Mannucci 2008)

1.6.3.5 The British Heart Foundation Family Heart Study and its contribution to genetics of CAD

The primary purpose of the BHF-FHS was to recruit a large collection of ASPs, to map genes relating to CAD risk through a genome-wide linkage analysis approach. We eventually recruited and analysed one of the largest collections for this type of analysis, 1933 families (Samani et al. 2005).

Recruitment for the study was made possible if an individual proband and their sibling(s) had an occurrence of MI or angina (verified by exercise stress test or angiography), or had undergone one of two possible re-vascularisation procedures: 1) coronary artery bypass graft surgery (CABG), or 2) percutaneous transluminal coronary angioplasty (PTCA), before they turned 66 years of age (i.e. premature CAD).

We recruited families with at least one affected sibling with premature CAD, either via a nationwide direct appeal to the public through the media, or by writing to all general practices across the United Kingdom (UK) with study information. Additionally, in the pilot stage of the study, CAD patient databases were utilised to target potential recruits at the two lead centres for the study in Leeds and Leicester. In total, with the recruitment of other family members that included unaffected siblings, we collected 6285 individuals.

1.6.3.5.1 Genome-wide linkage analysis in the BHF-FHS

For the ASP genome-wide linkage analysis we genotyped 416 microsatellite markers in 4216 individuals from 1958 families, and after quality checks the final linkage analysis included 4175 individuals and 1933 families. Linkage analysis identified two linkage peaks within the same region of chromosome 2, one for CAD and the other for MI with LOD scores of 1.86 and 1.15, respectively. The LOD scores for both CAD (2.7) and MI (2.1) improved, but were still not significant at a genome-wide level (3.3), when performing ordered subset analysis in younger individuals, i.e.≤56yrs for CAD and ≤59yrs for MI (Samani et al. 2005). These findings overlapped with two previous linkage discoveries on chromosome 2 (Pajukanta et al. 2000; Wang et al. 2004). Therefore, the findings in this study did not generate any novel linkage peaks of significance, but did replicate previous locus discoveries, and strongly suggested that there is a locus that influences CAD risk within this region of chromosome 2.

1.6.3.5.2 Genome-wide association study in the BHF-FHS: The WTCCC study

The Wellcome Trust Case control Consortium (WTCCC) came into being, as a way of exploiting the potential for undertaking genetic association analysis, at a genome-wide level, using the scientific and technological advances, described earlier. The primary experimental design was to perform a GWA study, on 2000 cases and 3000 shared controls for 7 common complex diseases of public health concern in the UK – bipolar disorder (BD), CAD, Crohn's disease (CD), hypertension (HTN), rheumatoid arthritis (RA), type 1 diabetes (T1D) and type 2 diabetes (T2D) (Wellcome Trust Case Control Consortium 2007).

The genotyping platform used was the Affymetrix GeneChip® Human Mapping 500K Array Set, which enabled the genotyping of 500,568 SNPs. This chip when averaged across SNPs with a MAF of 5%, has 80% power to detect a relative risk of 1.5, and 43% power to detect a relative risk of 1.3, for a threshold *P*-value of 5×10^{-7} (based on estimated simulations with HapMap Caucasian (CEU) LD).

The 1988 CAD cases selected for the WTCCC project were taken primarily from the BHF-FHS, with some coming from the GRACE study (a related study that recruited only one affected sibling with an affected parent, alongside the BHF-FHS recruitment). The selection criteria for cases from these two studies was for unrelated subjects with a history of MI (in the first instance), or a revascularisation procedure before the age of 66 years and a strong family history (Samani et al. 2007). Controls were collected from two sources - a nationwide resource of 1504 subjects from the 1958 Birth Cohort, and a further 1500 samples from blood donors recruited as part of the WTCCC project from Cambridge (UK).

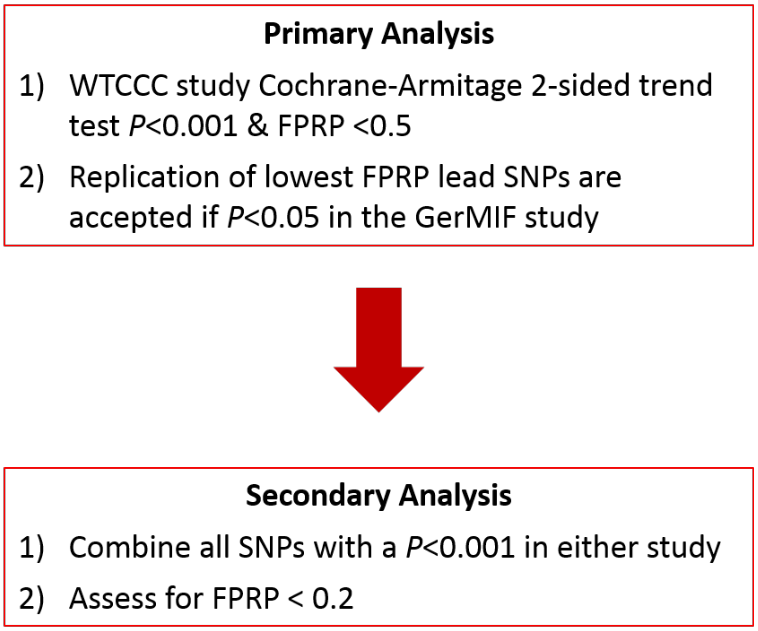
Two analyses were performed on these CAD subjects, the main WTCCC publication analysis involving all 7 complex common diseases that included a separate CAD analysis (Wellcome Trust Case Control Consortium 2007), and the

Samani group publication (Samani et al. 2007) that performed a more robust analysis by specifically analysing the CAD only WTCCC data, in combination with an additional GWA study dataset from Germany (to look for replication of the principle findings). The German MI Family (GerMIF) study consisted of 875 MI cases (<60 years with at least one affected first degree relative with premature CAD), and 1644 healthy controls, who had also been typed with the same Affymetrix array.

Here, I will focus principally on the findings from the Samani group analyses. The ‘primary analysis’ entailed the identification of the strongest SNP association with CAD in the WTCCC sample set, using an additive model two-sided Cochran-Armitage trend test threshold of *P*<0.001, followed by a false positive report probability (FPRP) assessment (it should be noted that the smaller the FPRP value, the more likely the SNP locus is a true positive). A single locus was defined as a SNP showing a FPRP below a threshold of <0.5 (i.e. the likelihood of a true false positive result is greater than 50%), along with other SNPs that reached the *P*<0.001 threshold, 100kb in either direction of this SNP.

Subsequently, replication was sought for each of these defined loci in the German GWA study data set, for the lead SNP (i.e. lowest FPRP) at each locus using a Cochran-Armitage trend test threshold of *P*<0.05 for statistical significance, after Bonferroni adjustment for multiple testing. A secondary analysis was a ‘combined analysis’, where all SNPs that showed moderately significant association with CAD in either study, i.e. *P*<0.001, were combined and assessed for FPRP below 0.2, to ensure a low probability of a false positive result. Pooled OR and their 95% CI were calculated. The primary and secondary analyses are summarised in **Figure 1.5**.

The primary analysis generated 396 moderately significant SNPs with a *P*<0.001, of these 30 SNPs scored a FPRP below 0.5, and clustered into 9 distinct loci (as defined above). Three of these 9 loci passed the replication criteria in the German MI GWA study dataset, even after adjusting for multiple-testing (*P*<0.005): 1) rs1333049 on chromosome 9p21.3; 2) rs6922269 on chromosome 6q25.1; and 3) rs2943436 on chromosome 2q36.3.



Primary Analysis

1) WTCCC study Cochrane-Armitage 2-sided trend test *P*<0.001 & FPRP <0.5
2) Replication of lowest FPRP lead SNPs are accepted if *P*<0.05 in the GerMIF study

Secondary Analysis

1) Combine all SNPs with a *P*<0.001 in either study
2) Assess for FPRP < 0.2

Figure 1.5. Summarised flow chart of the primary and secondary analyses that generated the seven novel CAD associated SNPs.

The strongest result (and the only SNP that reached a Bonferroni corrected genome-wide significance in the WTCCC discovery GWA study of $P<1\text{x}10^{-7}$) in both the WTCCC ($P=1.8\text{x}10^{-14}$) and German ($P=3.4\text{x}10^{-6}$) GWA studies was for rs1333049 (chr9p21.3), with a combined *P*-value of $2.91\text{x}10^{-19}$, and a 36% (95% CI: 27% to 46%) increased risk per copy of the risk allele (C), see **Table 1.9**. This locus was also identified in two other GWA studies published at around the same time, see **Table 1.8** (Helgadottir et al. 2007; McPherson et al. 2007). The lead SNP rs1333049 and 18 other SNPs in strong LD within this region span ~100kb, and include two cyclin dependent kinase inhibitors *CDKN2A* and *CDKN2B*, which play an important role in cell cycle regulation, and may through TGFβ induced cell growth inhibition be involved in the pathogenesis of AS (Lowe & Sherr 2003; Hannon & Beach 1994; Kalinina et al. 2004). However, a long non-coding RNA (lncRNA) called *CDKN2B-AS1* (CDKN2B anti-sense 1, aka ANRIL) of ill-defined function could (with respect to CAD) also have a role to play, perhaps by regulating cell growth by influencing *CDKN2A* and *CDKN2B* expression (Congrains et al. 2012).

The second lead SNP rs6922269 (chr6q25.1) association with CAD showed a combined *P*-value of *P*=2.9x10^{-8}, with a 23% (95% CI: 15% to 33%) increased risk per copy of the risk allele (A), see **Table 1.9**. This SNP is situated within an intron of the gene methylenetetrahydrofolate dehydrogenase (NADP+-dependent) 1-like (*MTHFD1L*), a gene that encodes for the mitochondrial isozyme of C1-tetrahydrofolate (THF) synthase (Prasannan et al. 2003; Walkup & Appling 2005). Proteins of this type synthesise purine and methionine, and so this gene may have an influence on the known risk factor homocysteine (Randak et al. 2000; Fruchart et al. 2004).

Table 1.9 Seven novel loci identified by WTCCC and GerMIF GWA studies

Lead SNP	Chromosome Band	Minor/risk allele in controls	MAF in controls	Combined Analysis		FPRP (<0.2)
				OR (95% CI)	*P*-value	
rs2943634	2q36	A/C	0.34	1.21 (1.13-1.30)	1.61x10^{-07}	0.0190
rs6922269	6q25	A/A	0.25	1.23 (1.15-1.33)	2.90x10^{-08}	0.0002
rs1333049	9p21	C/C	0.47	1.36 (1.27-1.46)	2.91x10^{-19}	0.0006
rs599839	1p13	G/A	0.23	1.29 (1.18-1.40)	4.05x10^{-09}	0.0010
rs17465637	1q41	A/C	0.29	1.20 (1.12-1.30)	1.27x10^{-06}	0.1314
rs501120	10q11	C/T	0.13	1.33 (1.20-1.48)	9.46x10^{-08}	0.0234
rs17228212	15q22	C/C	0.30	1.21 (1.13-1.30)	1.98x10^{-07}	0.0178

Primary analysis GerMIF study replication of WTCCC lead SNPs are listed first, followed by secondary analysis SNPs identified by combined WTCCC and GerMIF studies.

The third SNP showing replication for CAD association, rs2943634 (chr2q36.3) had a combined *P*-value of *P*=1.61x10^{-7}, with a 21% (95% CI: 13% to 30%) increased risk of CAD per copy of the risk allele (C), see **Table 1.9**. This SNP is situated in a poor gene annotated region spanning 233kb. However, a proxy SNP rs7578326 (r^2=0.815, D'=0.920) is situated within the one uncharacterised gene *LOC646736* within this region, recently validated as a lncRNA by Refseq (NCBI Reference Sequence: NR_046102.1); this particular SNP has subsequently been shown to be associated with type 2 diabetes (T2D) (Parikh, Lyssenko & Groop 2009). Of note, the association of this SNP was reduced when adjusting for covariates in the

German MI study, i.e. BMI, hypertension and LDL-C, whereas no cardiovascular risk factors had an effect on rs1333049 or rs6922269.

The combined analysis produced four novel loci (see **Table 1.9**): 1) rs599839 on chromosome 1p13.3 located near to the 3'UTR (3 prime untranslated region) of two genes *CELSR2* (cadherin, EGF LAG seven-pass G-type receptor 2) and *PSRC1* (proline-serine rich coiled-coil 1); 2) rs17465637 on chromosome 1q41 is situated within an intron of the *MIA3* (melanoma inhibitory activity 3) gene; 3) rs501120 on chromosome 10q11.21 is located 100kb downstream of the *CXCL12* (chemokine (C-X-C motif) ligand 12) gene; and 4) rs17228212 on chromosome 15q22.33 is located within an intron of the *SMAD3* gene. All of these are novel loci with respect to CAD association, but have biological plausibility due to their involvement with cell growth or inhibition (Lo et al. 1999; Bosserhoff & Buettner 2002; Qin et al. 2002; Miyazono 2000), processes that could be involved in AS formation, progression and plaque instability (Libby & Theroux 2005).

In summary, **Figure 1.6** shows Manhattan plots for the WTCCC (A) and GerMIF (B) studies, respectively. The –$\log_{10}$ *P*-values are plotted on the Y-axis, and the genomic base pair (bp) location of all Affymetrix 500K array SNPs that passed quality control criteria (i.e. for WTCCC n=377,857 SNPs and for GerMIF n=272,602 SNPs), are plotted on the X-axis (all SNP cluster plots with a *P*-value below 10^{-3} were independently assessed by two individuals at each centre for authenticity).

The WTCCC GWA study approach proved to be an effective way of identifying previously unknown loci associated with CAD. However, none of the seven novel loci discovered have an obvious role in CAD, and require further elucidation.

1.6.3.5.3 Genome-wide association study in the BHF-FHS: CARDIoGRAM and CARDIoGRAMplusC4D.

More recently the CAD GWA studies described for the WTCCC and GerMIF studies have provided yet greater utility, by their being meta-analysed as part of a large consortium of fourteen GWA studies, from either Northern European or Northern American descent called CARDIoGRAM (Coronary ARtery DIsease

Genome-Wide Replication And Meta-Analysis). The purpose of such a large collaborative effort was to generate greater statistical power, so as to enable novel loci identification at a genome-wide level of significance (*P*<5x10^{-8}). CARDIoGRAM's efforts resulted in the validated discovery of thirteen further novel loci associated with CAD (Schunkert et al. 2011).

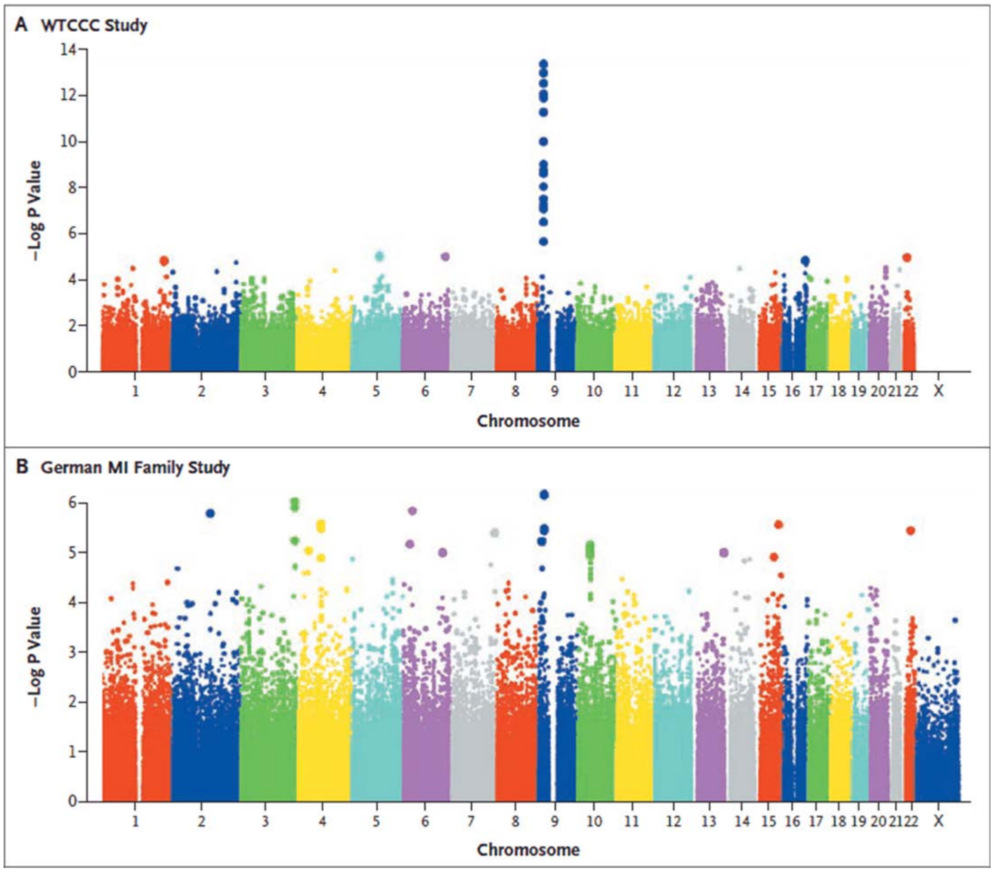


Figure 1.6. Manhattan Plots Showing the association of SNPs with CAD or MI in the GWA study analyses.

The −log *P*-values are plotted on the Y-axis for the association of each SNP with CAD or MI, from two-sided Cochran–Armitage tests for trend. The genomic co-ordinates for each chromosome are displayed along the X-axis, within each chromosome shown, the data are plotted from the p-ter end. The Y-axis scale for associations in the WTCCC study (Panel A) differs from the scale for the GerMIF Study (Panel B) (taken from Samani et al. 2007).

A CARDIoGRAM follow-up study to harness yet greater statistical power to detect novel loci at a validated genome-wide level of significance (*P*<5x10^{-8}), involved the CARDIoGRAMplusC4D collaboration. The goal was to fine-map 30 reported CAD risk loci at genome-wide significance (*P*<5x10^{-8}), and to replicate a LD-pruned (i.e. r^2<0.2) set of 6,222 variants with nominal association (*P*<0.01) from CARDIoGRAM

- with the aim of identifying further susceptibility loci for CAD risk. To this end, a custom Illumina iSelect genotyping ~200,000 SNP array called ‘Metabochip’, was designed prinicipally on the basis of large scale meta-analyses for relevant traits relating to metabolic, atherosclerotic and cardiovascular disease, and included the CARDIoGRAM consortium. Meta-analysis of 6,222 replication SNPs and 20,876 fine-mapping SNPs were carried out in over 190,000 individuals of European and South Asian descent. This grand effort resulted in the identification of an additional fifteen novel loci validated at genome-wide significance ($P<5\text{x}10^{-8}$) (CARDIoGRAMplusC4D Consortium et al. 2013).

1.7 Aims of this thesis

I was part of the Samani group that assembled the BHF-FHS and undertook the initial genome-wide linkage analysis (Samani et al. 2005), as well as, the subsequent GWA study through the WTCCC (Wellcome Trust Case Control Consortium 2007), in collaboration with the GerMIF Study (Samani et al. 2007). This latter work identified 7 novel genetic loci, where variants were associated with increased risk of CAD. None of the loci had an obvious gene, with an immediately known mechanism that could contribute to CAD risk.

Therefore, the **first aim** of my thesis (Chapter 2) was to systematically investigate the association of genetic variants at these loci, with traditional cardiovascular risk factors (BP, cholesterol etc), using a large-population-based resource (the GRAPHIC study) that we had available in Leicester.

These studies showed a novel association of the lead CAD-related SNP variant rs599839 on chromosome 1p13.3 with LDL cholesterol. I therefore next focused (Chapter 2) on this locus to examine its relationship to other phenotypes, including whether the locus influenced pharmacological response to statin therapy, for this I utilised the WOSCOPS (West Of Scotland Coronary Prevention Study). I also (Chapter 3) carried out fine-mapping of the locus using data from the 50K-IBC array, and I undertook *in silico* analysis to identify the putative functional variant. However, during the course of this analysis, the functional variant at this locus was reported by Musunuru et al. (Musunuru et al. 2010).

Therefore, for the final part of my thesis (Chapter 4), I switched to another novel and interesting locus for CAD identified by the CARDIoGRAM consortium that performed a meta-analysis of our original WTCCC CAD GWA study, along with thirteen other CAD GWA studies of Northern European ancestry (Schunkert et al. 2011). This locus on Chr13q34, marked by lead SNP rs4773144 is located distally within the shared bidirectional promoter of major basement membrane genes collagen(IV) alpha1 (*COL4A1*) and collagen(IV) alpha2 (*COL4A2*). I undertook bioinformatics and functional genomics experimentation of this locus.

2. Chapter 2

Assessment of seven novel genetic loci associated with CAD for intermediate phenotypes

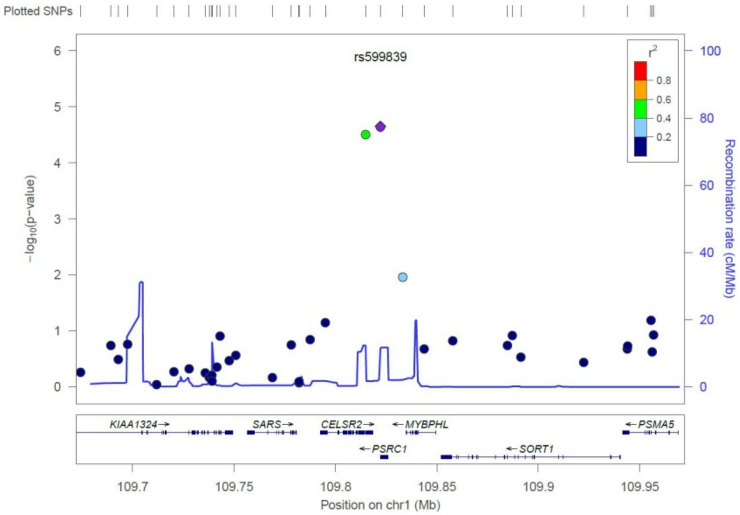
2.1 Introduction

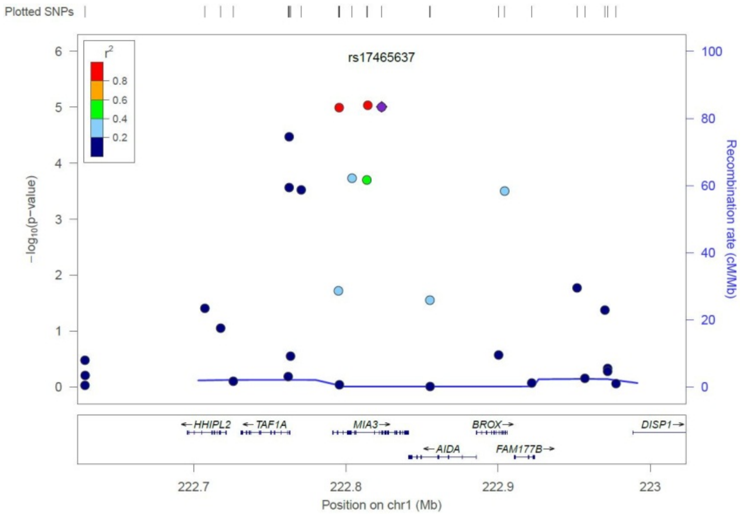
The combined analyses of two North European Caucasian CAD GWA studies, from the WTCCC study (with 1926 CAD cases and 2938 non-CAD controls) and the GerMIF study (with 875 MI cases and 1644 non-MI controls), have identified seven novel chromosomal loci with strong statistical association for CAD risk. The individual WTCCC and GerMIF studies, as well as, the combined association results are shown in **Table 2.1** (Samani et al. 2007). These include lead SNP variants: rs599839 on chr1p13.3 (*CELSR2* / *PSRC1*); rs17465637 on chr1q41 (*MIA3*); rs2943634 on chr2q36.3 (*LOC646736* – validated uncharacterised lncRNA); rs6922269 on chr6q25.1 (*MTHFD1L*); rs1333049 on chr9p21.3 (*CDKN2B-AS1*, *CDKN2A* and *CDKN2B*); rs501120 on chr10q11.21 (non-genic, *CXCL12* is the nearest gene, 100kb upstream); and rs17228212 on chr15q22.33 (*SMAD3*). In addition, **Figure 2.1** shows the WTCCC chromosome regional *Manhattan* plots at each CAD-associated locus, to provide the genetic architectural context of each locus, for any SNP associations at a distance of at least 100 kb upstream or downstream of the lead SNP (Samani et al. 2007).

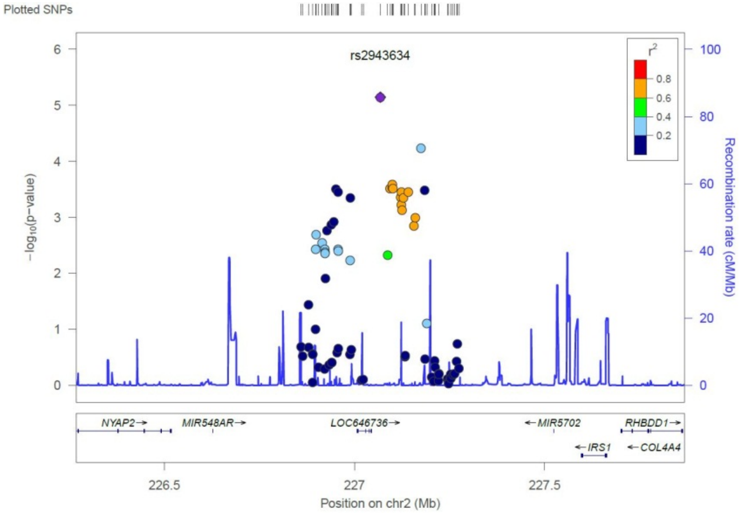
The functional mechanisms behind these CAD associated novel chromosomal loci are unknown. As discussed in the last Chapter, for some of the loci there appear to be potential mechanisms based on current understanding of the pathogenesis of CAD. For example, the chromosome 9p21.3 locus contains two cyclin dependent kinase inhibitors which could affect smooth muscle cell proliferation (a key step in plaque development), while the chromosome 6q25.1 lead variant sits within a gene (*MTHFD1L*), which may be involved in homocysteine metabolism, long postulated to play a role in atherogenesis (McCully 1993). However, the first step is to assess whether any of the loci also associate with traditional quantitative risk factors for CAD, such as blood pressure (BP) and lipid traits, because these may prove to be intermediate phenotypes in relevant causal pathways.

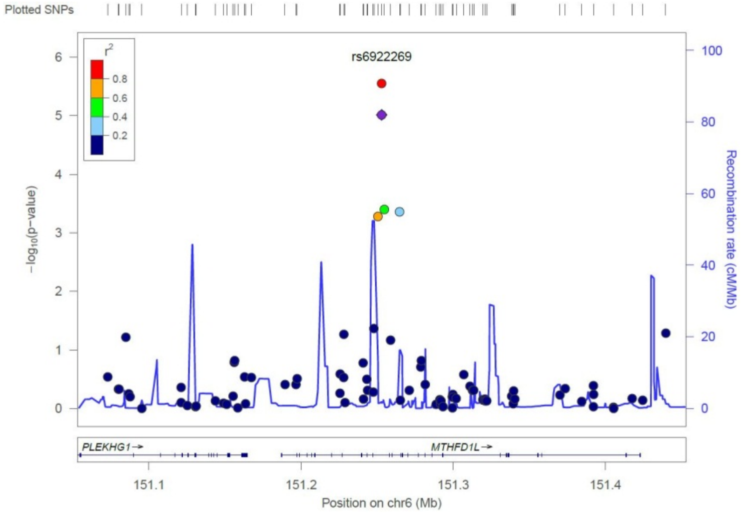
Table 2.1 The WTCCC and GerMIF study individual and combined CAD GWA study association results for seven novel loci

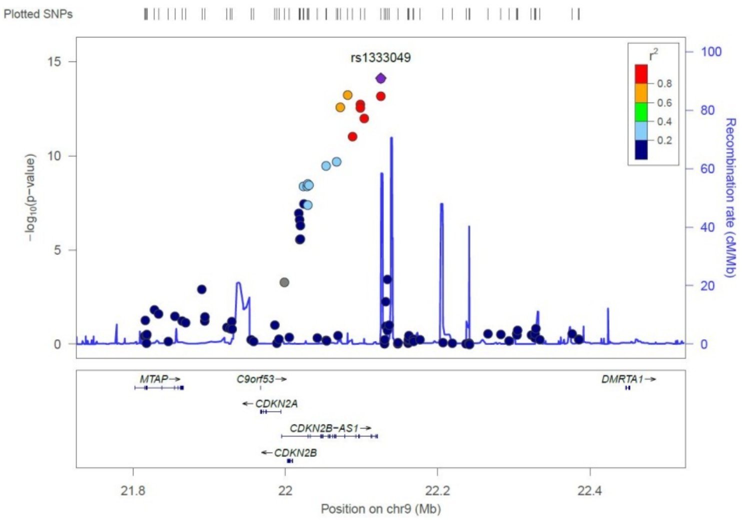
Chr	**SNP**	**Position**	**Minor/ Risk Allele**	**WTCCC**			**GerMIF**			**Combined**		
				MAF	**OR (95% CI)**	***P*-value**	**MAF**	**OR (95% CI)**	***P*-value**	**OR (95% CI)**	***P*-value**	**FPRP**
1	rs599839	109534208	G/A	0.23	1.24 (1.12,1.38)	2.19x10^{-05}	0.22	1.39 (1.19,1.63)	3.17x10^{-05}	1.29 (1.18,1.40)	4.05x10^{-09}	0.001
1	rs17465637	219211924	A/C	0.29	1.23 (1.12,1.34)	1.00x10^{-05}	0.26	1.15 (1.01,1.32)	3.82x10^{-02}	1.20 (1.12,1.30)	1.27x10^{-06}	0.131
2	rs2943634	226893585	A/C	0.34	1.22 (1.11,1.33)	1.19x10^{-05}	0.37	1.20 (1.06,1.35)	4.33x10^{-03}	1.21 (1.13,1.30)	1.61x10^{-07}	0.019
6	rs6922269	151345099	A/A	0.25	1.23 (1.13,1.35)	6.33x10^{-06}	0.26	1.24 (1.09,1.41)	1.14x10^{-03}	1.23 (1.15,1.33)	2.90x10^{-08}	0.002
9	rs1333049	22115503	C/C	0.47	1.37 (1.26,1.48)	1.80x10^{-14}	0.48	1.33 (1.18,1.51)	6.80x10^{-06}	1.36 (1.27,1.46)	2.91x10^{-19}	0.000
10	rs501120	44073873	C/T	0.13	1.24 (1.09,1.41)	1.31x10^{-03}	0.16	1.54 (1.28,1.86)	5.28x10^{-06}	1.33 (1.20,1.48)	9.46x10^{-08}	0.023
15	rs17228212	65245693	C/C	0.30	1.19 (1.09,1.30)	1.18x10^{-04}	0.26	1.26 (1.11,1.44)	3.09x10^{-04}	1.21 (1.13,1.30)	1.98x10^{-07}	0.018

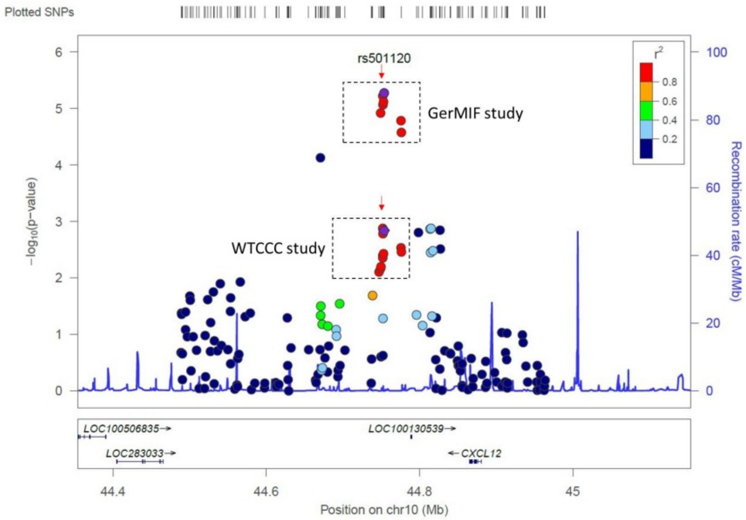

Plotted SNPs
rs599839
r²
0.8
0.6
0.4
0.2
−log10(p−value)
Recombination rate (cM/Mb)
0
1
2
3
4
5
6
0
20
40
60
80
100
KIAA1324→
SARS→
CELSR2→
←MYBPHL
←PSMA5
←PSRC1
←SORT1
109.7
109.75
109.8
109.85
109.9
109.95
Position on chr1 (Mb)


Plotted SNPs
rs17465637
r²
0.8
0.6
0.4
0.2
−log10(p−value)
Recombination rate (cM/Mb)
0
1
2
3
4
5
6
0
20
40
60
80
100
←HHIPL2
←TAF1A
MIA3→
BROX→
DISP1→
←AIDA
FAM177B→
222.7
222.8
222.9
223
Position on chr1 (Mb)


Plotted SNPs
rs2943634
r²
0.8
0.6
0.4
0.2
−log10(p−value)
Recombination rate (cM/Mb)
0
1
2
3
4
5
6
0
20
40
60
80
100
NYAP2→
MIR548AR→
LOC646736→
←MIR5702
RHBDD1→
←IRS1
←COL4A4
226.5
227
227.5
Position on chr2 (Mb)


Plotted SNPs
rs6922269
r²
0.8
0.6
0.4
0.2
−log10(p−value)
Recombination rate (cM/Mb)
0
1
2
3
4
5
6
0
20
40
60
80
100
PLEKHG1→
MTHFD1L→
151.1
151.2
151.3
151.4
Position on chr6 (Mb)





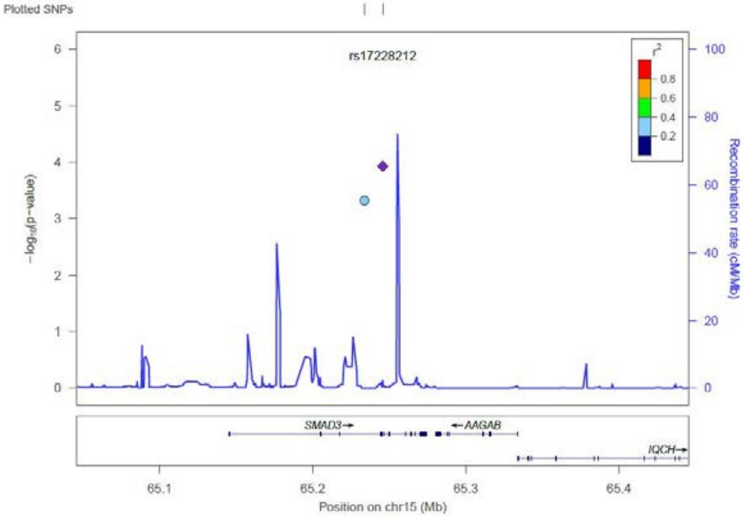


Figure 2.1. Seven novel loci regional Manhattan plots of CAD WTCCC Affymetrix 500K GWA study - Cochran-Armitage trend test –log10 *P*-values ($P<10^{-3}$) for CAD association (Y-axis) *vs.* chromosome base pair position (X-axis). The purple diamond represents the lead SNP at each locus. (For rs501120 the GerMIFstudy data is shown too, because for this locus the replication cohort is the lead SNP; the rs501120 for WTCCC is given as a purple circle, and red arrows mark the lead SNP for each study). The pairwise LD r^2 with the lead SNP is represented by colour-coding. Recombination rates are plotted on the right hand side of the Y-axis. Plots for **Figure 2.1** are generated by the webtool LocusZoom, hosted via the Center for Statistical Genetics, University of Michigan via the website https://statgen.sph.umich.edu/locuszoom /genform .php?type=yourdata (Version 1.1: June 2011) (Pruim et al. 2010).

The ideal population to test for association of genetic variants with quantitative traits is in one that is representative of the general population to avoid confounding factors, such as medication and other co-morbidities that exist in CAD populations. I was able to investigate the association of lead variants at seven novel loci with quantitative traits with cardiovascular relevance in a general population of families of North European Caucasian ancestry from the UK called the GRAPHIC (Genetic Regulation of Arterial Pressure of Humans In the Community) study. In addition, I was able to replicate our key findings in another independent collection of healthy unrelated control subjects, as well as MI cases from the GerMIF study from Germany.

2.2 Materials and methods

2.2.1 Study populations

2.2.1.1 GRAPHIC Study:

The GRAPHIC study comprises 2037 subjects from 520 nuclear families that were recruited from the general population in Leicestershire (UK), between 2002 and 2005. The details of the GRAPHIC recruitment and phenotyping have been reported previously (Tobin et al. 2008). In brief, the nuclear families comprise of two parents (aged 40-60 years) and two adult offspring (aged ≥18 years), identified through general practice (GP) surgeries in Leicestershire. The participants were assessed for several phenotypes of potential relevance to CAD, including body mass index (BMI), waist-hip ratio (WHR), clinic and 24-hour (hr) ambulatory systolic and diastolic BP, non-fasting blood glucose, total cholesterol (TC) and high density lipoprotein cholesterol (HDL-C), blood urate, creatinine clearance; and self-reported smoking status and alcohol intake.

Participants underwent an interview and examination by trained research nurses that included taking a detailed medical history, collecting blood and urine samples, and their measurement using standard operating procedures. Clinic BP was recorded three times (using an Omron HEM-705CP digital BP monitor with an appropriate size cuff), using the non-dominant arm, with an interval of at least three minutes between measures. Clinic BP was defined as the mean of the second and third readings. The 24-hr ambulatory BP was measured using a Spacelabs 90207 monitor (Spacelabs, Wokingham, UK) for 26-hr. The first two hours of each record was discarded to avoid any alerting response. The monitor recorded BP at 30-minute intervals between 08:00 and 21:59, and at 1-hr intervals between 22:00 and 07:59. In addition, BP readings were adjusted for those on anti-hypertensive medication using a semi-parametric algorithm (Levy et al. 2000; Tobin et al. 2005). All blood and urine assays were carried out in the University Hospitals of Leicester biochemistry laboratory, using conventional enzymatic assays on the Abbott Aeroset 2.0 Analyser (Abbott Laboratories, IL, USA). Creatinine clearance was calculated using the formula: creatinine clearance = (urine creatinine concentration x urine volume per unit of time) / plasma creatinine concentration.

Additional GRAPHIC lipid phenotypes were obtained on stored samples in the period between June – December 2008, by shipping –80^{o}C frozen (at the time of collection) 1mL plasma samples on dry-ice to LipoFIT® GmbH (Regensburg, Germany), to measure lipoprotein subclasses by nuclear magnetic resonance (NMR) spectroscopy, as previously described (Kaess et al. 2008). Briefly, gradient weighted NMR spectra of blood plasma were recorded, which lead to characteristic overall profiles of the lipoprotein signals. Using the gradient weighted spectra, spectral regions ranging from 1.5 to 0.7 ppm were modeled into a set of 15 lipoprotein subclasses. Mean particle size and particle number were then computed from the distribution of the different lipid classes: high density lipoprotein (HDL), low density lipoprotein (LDL), IDL (Intermediate density lipoprotein), very low density lipoprotein (VLDL) and chylomicrons. For sample validation we also reconstructed TC, LDL-cholesterol (LDL-C), HDL-C and triglycerides (TG) from NMR data, and measured TC, LDL-C, HDL-C and TG by conventional enzymatic assays. The 12 cholesterol measured lipoprotein sub-fractions (chylomicrons were excluded as they contain very little cholesterol, and were only measured for TG) used in the analysis were classified according to **Table 2.2** (Kaess et al. 2008).

Table 2.2. Classification and properties of 12 NMR measured lipoprotein subclasses

Lipoprotein Subfraction	Particle Size (nm)	Mean Density (g/mL)	Mass (g/nmol)
HDL A (small)	7–8.5	1.2	0.01
HDL B (medium)	8.5–10	1.12	0.02
HDL C (large)	10–13	1.09	0.03
HDL D (very large)	13–16	1.063	0.05
LDL A (very small)	16–19	1.06	0.1
LDL B (small)	19–21	1.045	0.15
LDL C (medium)	21–22	1.035	0.2
LDL D (large)	22–25	1.027	0.26
LDL E (very large)	25–30	1.019	0.39
IDL	30–40	1.015	0.58
VLDL A (small)	40–60	1.01	1.62
VLDL B (large)	60–80	1.006	4.87

Legend. NMR measurements (LipoFIT® Analytic GmbH; Regensburg, Germany) for lipoprotein subclasses: HDL (high density lipoproteins); LDL (low density lipoproteins); IDL (Intermediate density lipoproteins); VLDL (very low density lipoproteins). (**Table 2.2.** adapted from (Kaess et al. 2008))

Before we analysed the GRAPHIC dataset, we tested all 2037 recruited samples with complete phenotypic data for Mendelian errors (that might indicate unrelatedness within families, or lab-based sampling errors). This was achieved by genotyping 3 highly polymorphic tandem repeat microsatellite markers and multiple SNPs, and resulted in the removal of 15 individuals prior to analysis (Tobin et al. 2008), leaving 2022 individuals for the final analysis.

2.2.1.2 GerMIF Study

The GerMIF study provided the GRAPHIC cardiovascular quantitative trait analysis with 847 healthy unrelated German controls and 1090 MI cases, for the purposes of replication of any intermediate phenotype associations found in the GRAPHIC discovery study. The 1090 MI subjects were unrelated (i.e. one subject was selected from each German MI family recruited) and selected if they had suffered a MI < 65 years. GerMIF study proband patients were recruited if they had suffered a MI before the age of 60 years (validated through hospital records), and also had a first degree relative (preferably a sibling) with premature CAD, such as - a MI, CABG or PTCA before the age of 70 years. All proband subjects were of German Caucasian descent and 70% of cases were recruited from near Augsburg and Southern Germany, between 1997 and 2002. The healthy German controls were aged between 29 and 91 years, recruited from the same centre as the unrelated MI patients in the GerMIF study, i.e. subjects that took part in the MONICA/KORA Augsburg study in the years 1994/5 and a follow-up study in the years 2004/5, and were without CAD or other major comorbidities. The healthy subject recruits underwent a physical examination, blood tests, and a standardised interview (Samani et al. 2007; Sedlacek et al. 2007). Serum TC and HDL-C were measured using a standard enzymatic method, CHOD-PAP (cholesterol oxidase phenol 4-aminoantipyrine peroxidise), HDL-C was determined after precipitation with phosphotungstate/Mg^{2+}, and LDL-C was calculated after precipitation with dextrane sulphate in the supernatant, as previously described (Holmer et al. 2000).

2.2.1.3 WOSCOP Study

The WOSCOP (West Of Scotland COronary Prevention) study was a randomised, double-blinded and placebo controlled study, for the purpose of determining whether pravastatin treatment in men with moderate hypercholesterolaemia and without a history of MI, would reduce the combined incidence of non-fatal MI and death from CAD. The study included 6595 males between 45 and 64 years, who had a measured fasting base-line LDL-C of ≥4.0 and ≤6.0 mmol/L, with no history of MI that self-administered each evening, either 40mg of pravastatin, or placebo for 5 years. The primary study endpoint was the occurrence of a non-fatal MI, or death from CAD as a first event. The findings from this study were that pravastatin rather than placebo lowered plasma levels of TC and LDL-C, as well as TG, but raised levels of HDL-C. This resulted in the reduced incidence of a range of coronary endpoints (i.e. nonfatal MI, death from CAD, death from other CVD causes, and the frequency of coronary revascularisation procedures), and reduced the coronary risk associated with adverse lipid levels. This was a landmark study that showed for the first time that statin treatment of patients with hypercholesterolaemia without symptomatic CAD, not only lowers LDL-C, but also reduced fatal and non-fatal cardiovascular events (Shepherd et al. 1995). The end of study number of cases was n=580, each case was age and smoking status matched with two controls (selected from the original 6595 males) with a total of n=1160 controls, only n=1624 had DNA samples collected (Freeman et al. 2003).

The purpose of mentioning this study is that I genotyped all seven GWA study CAD-associated SNPs, in 1606 WOSCOP study subjects (n=768 pravastatin and n=838 placebo treated subjects), available within the Samani group for collaborative study purposes. Due to the key finding that rs599839 was associated with TC for my primary analysis in the GRAPHIC study, and LDL-C in my replication of GerMIF study subjects. I was particularly interested in seeing whether the amount of LDL-C lowering achieved by subjects taking a statin, would also be affected by rs599839 SNP genotypes (i.e. through a pharmaco-genetic response), as this would provide a mechanistic pathway for the 1p13.3 locus.

2.2.1.4 Phenotypic study differences

The GRAPHIC study is the primary cardiovascular intermediate phenotype analysis study selected to assess CAD risk associated rs599839, as it is representative of the general population for two familial generations, and is free from co-morbidities and bias. The unrelated German healthy controls are an appropriate replication control for the GRAPHIC parents in particular, but have a range of ages that could be seen as replication of the GRAPHIC offspring generation too. An age-matched unrelated German MI cases replication group were also used to see if the same intermediate phenotype effects were present in MI cases as well as healthy subjects. This was important, as the CAD risk genotype effect in the first place was identified in the GerMIF study, but more especially because our own ‘in house’ WTCCC CAD study was poorly phenotyped for cholesterol, and no direct measure of LDL-C was taken for the primary GRAPHIC study. It should also be acknowledged that the randomised double-blinded statin-placebo WOSCOP study cases differ from the German MI cases by phenotype as WOSCOP study cases are inclusive of CAD cases (i.e. those that have undergone CABG and PTCA procedures) with and without MI survival, whereas the German MI cases have a survival bias and no CAD only cases.

2.2.2 Genotyping

The WTCCC CAD GWA study lead SNPs for each of the seven novel loci with strong association for CAD risk (Samani et al. 2007): rs599839 (1p13.3), rs17465637 (1q41), rs2943634 (2q36.3), rs6922269 (6q25.1), rs1333049 (9p21.3), rs501120 (10q11.21) and rs17228212 (15q22.33) (as described in the Introduction) were genotyped using a TaqMan® SNP genotyping assay. The GRAPHIC and WOSCOP study subjects underwent allelic discrimination for each of the seven SNPs using 15ng of DNA, 36mM of each primer pair (forward and reverse), 8mM of specific VIC®dye-allele-1 and FAM™dye-allele-2 reporter/quencher 3' MGB (3 prime Minor groove binder) fluorescent probes, and TaqMan® Universal PCR Master Mix, No AmpErase® UNG containing: AmpliTaq Gold® DNA Polymerase, dNTPs (deoxynucleotide triphosphates) and ROX™dye as a passive reference (Applied Biosystems (ABI), CA, USA). Polymerase chain reaction (PCR) was performed on a GeneAmp® PCR system 9700 (ABI, CA, USA) using 384 well plates, and a cycling protocol of 95°C for 10min, followed by 45 cycles of 92°C for 15 s and 60°C for 1min. Fluorescence was detected post-PCR using the ABI Prism® 7900HT Sequence Detector System (SDS), and genotypes called using ABI Prism® SDS software version 2.1 (ABI, CA, USA). In the German subjects, the seven SNPs were typed using a similar protocol.

2.2.3 Statistical analysis

Association tests for quantitative traits were performed using generalised estimating equations (GEE), with an exchangeable correlation structure, and robust standard errors to account for correlation induced by family structure (Burton, Gurrin & Sly 1998). Hardy-Weinberg equilibrium (HWE) and minor allele frequencies (MAF) were assessed for the unrelated parental generation. All analyses included age, sex and the SNP of interest as covariates, and an additive genetic model was assumed. Prior to analyses, all quantitative traits were assessed for whether they were distributed normally (none required transformation). OR estimates were generated for binary traits, by adapting the statistical models used to incorporate a logit link function. All analyses were carried out using STATA software, release 9.1 (STATACorp LP, College Station TX, USA).

No further covariates were adjusted for in the primary analyses, so that any potential intermediate phenotype on a causal pathway between SNP and CAD remained detectable. However, where the primary analyses showed an association with a particular SNP, then the impact of additional covariates were adjusted for, and included BMI, WHR, excess alcohol consumption, smoking status, blood glucose and diabetes.

You will see from the results section that rs599839 was the only SNP to show a significant association in the primary analysis. For this reason a comparison was made in the GerMIF study (for the purposes of replication) with TC, LDL-C and HDL-C for rs599839, using a multiple linear regression with covariates including age, sex, genotype and rs599839. The results show adjusted means and standard deviations (SD) for TC, LDL-C and HDL-C, under an additive genetic model. Due to a high proportion of German MI subjects taking lipid-lowering drugs, an additional analysis was performed that separated those who were taking lipid-lowering treatment, from those who were not. OR for binary traits were calculated in the same manner as for the GRAPHIC study.

The subsequent GRAPHIC clinical chemistry and NMR lipoprotein subclass analysis for rs599839 (as a consequence of the primary GRAPHIC analysis and German analysis replication) was performed in STATA in the same way as for the original aforementioned seven SNPs and cardiovascular traits.

2.2.4 Power calculations

A priori power calculations for the GRAPHIC study were undertaken before recruitment, to estimate the power to detect genetic variant impact on cardiovascular quantitative traits (in particular BP). For example it was shown that there was >99% power to detect at least 5 of 6 polymorphisms with a prevalence of 0.2 and an effect size of 0.25 SD, based on an aimed for recruitment of 450 families (**Table 2.3**).

Post-hoc power calculations for genotype effects on cardiovascular traits in the GRAPHIC study were generated retrospectively, using only the parental data. The reason for calculating the power *post-hoc*, is because the sample size is now fixed, therefore, I can now look at the power to detect specific effect sizes in the GRAPHIC cohort. It should be noted that the sample size is larger than that used in the original *a priori* power calculations, so I am hoping to determine the minimum effect-size that can be detected in the current study.

Table 2.3. *A priori* power calculations for the GRAPHIC study trait blood pressure

Effect size (mm Hg)	**Prevalence = 0.2** Allele frequency§ = 10.6%	**Prevalence = 0.1** Allele frequency§ = 5.1%	**Prevalence = 0.05** Allele frequency§ = 2.5%
5.50 (1/2 of a SD)	>99%* **>99%†** *>99%‡*	>99% **>99%** *>99%*	>99% **>99%** *>99%*
3.67 (1/3 of a SD)	>99% **>99%** *>99%*	>99% **>99%** *96%*	78% **98%** *61%*
2.75 (1/4 of a SD)	>99% **>99%** *>99%*	85% **>99%** *77%*	41% **47%** *4%*

*Power to detect (at *P*<0.001) a single polymorphism at given prevalence and effect size
†Power to detect at least three of six polymorphisms all at given prevalence and effect size
‡*Power to detect at least five of six polymorphisms all at given prevalence and effect size*
§Allele frequency corresponding to stated 'prevalence' assuming a bi-allelic locus with a dominant allele in Hardy Weinberg equilibrium. SD – standard deviation.

2.3 Results

2.3.1 Demographics

The main cardiovascular quantitative trait characteristics for the GRAPHIC study are summarised in **Table 2.4**. The mean age of the parental generation was 52.8 years, whilst that of the offspring generation was 25.5 years. Hypertension was reported in 15.9% of subjects, with 6.7% taking anti-hypertensive medication. Forty-one subjects (2%) were taking lipid-lowering treatment, and 75 subjects (3.7%) reported a history of previous cardiovascular disease. The mean TC levels were 5.64mmol/L and 4.51mmol/L in the parental and offspring generations, respectively.

The main cardiovascular quantitative trait characteristics for the GerMIF study are summarised in **Table 2.5**. Subjects from the GerMIF study (both MI cases and heathy controls) were older than for the GRAPHIC study, and with a higher mean cholesterol level. In general, the German healthy controls had similar demographics to the GRAPHIC parents, but hypertension was much higher in the German healthy control samples (~51%) compared to GRAPHIC parents of a similar age (~26%). In addition, hypertension was higher still for German MI cases (~91%), and this was much higher than for the WTCCC CAD cases (~43%)

(Samani et al. 2007). As a consequence hypertension treatment was also much higher in the GerMIF study subjects (~38% in healthy controls and ~87% in MI cases) than for GRAPHIC parental subjects (~15%). High levels of lipid-lowering medication in the MI subjects (63.1%), resulted in lower mean TC levels in MI cases (5.88mmol/L) than in the mean TC levels for control subjects (6.19mmol/L), as only 9.1% were on lipid-lowering medication.

The base-line characteristics for the WOSCOP study are summarisd in **Table 2.6**. WOSCOP study subjects were analysed in particular to see whether there would be a rs599839 genotype specific effect on the amount of lipid lowering brought about by statin treatment. In comparison with the German MI cases, the WOSCOP study CAD cases baseline mean levels for TC (5.88 vs. 7.1mmol/L) and LDL-C (3.91 vs. 5mmol/L) are higher. This is probably for the most part due to the fact that none of WOSCOP study CAD subjects had been taking lipid lowering medication at the time of recruitment. Indeed, the German MI cases were taking lipid lowering medication in 63.1% of subjects, therefore it is feasible that the WOSCOP study cholesterol is more likely to match German MI cases prior to treatment. Smoking status was higher for German MI cases than for WOSCOP study cases (71% vs 53%). All other characteristics reported for WOSCOP study subjects are similar to the German MI cases.

Table 2.4. Characteristics of the GRAPHIC study population

Variable	**Fathers** (n = 516)	**Mothers** (n = 516)	**Sons** (n = 513)	**Daughters** (n = 492)	**All subjects** (n = 2,037)
Age (years): mean (SD)	53.72 (4.28)	51.90 (4.36)	25.04 (5.05)	25.91 (5.43)	39.32 (14.51)
Ever smoker %; current smoker %	56.6; 14.8	43.8; 12.4	39.6; 28.7	38.2; 24.4	44.6; 20.0
Excess alcohol intake# n (%)	81 (17.7%)	42 (9%)	142 (30.9%)	36 (8.1%)	301 (16.4%)
Body mass index (kg/m^2): mean (SD)	27.83 (3.95)	27.08 (4.55)	24.93 (4.13)	24.53 (4.94)	26.11 (4.61)
Waist–hip ratio: mean (SD)	0.93 (0.068)	0.81 (0.064)	0.86 (0.064)	0.78 (0.067)	0.85 (0.086)
History of hypertension: n (%)	139 (26.9)	133 (25.8)	19 (3.7)	33 (6.7)	324 (15.9)
Current antihypertensive treatment: n (%)	79 (15.3)	53 (10.3)	3 (0.6)	1 (0.2)	136 (6.7)
24-hr SBP (mmHg): mean (SD)	124.29 (11.44)	117.04 (11.48)	120.76 (8.04)	112.77 (7.21)	118.79 (10.65)
24-hr DBP (mmHg): mean (SD)	77.66 (7.23)	71.72 (7.56)	69.2 (6.46)	67.93 (5.18)	71.68 (7.66)
Total cholesterol (mmol/L): mean (SD)	5.59 (0.99)	5.69 (0.97)	4.52 (0.90)	4.50 (0.83)	5.08 (1.08)
HDL cholesterol (mmol/L): mean (SD)	1.32 (0.30)	1.64 (0.40)	1.30 (0.28)	1.47 (0.36)	1.43 (0.36)
Glucose (mmol/L): mean (SD)	5.31 (1.56)	5.09 (1.27)	4.87 (0.85)	4.71 (0.82)	5.00 (1.19)
Urate (mmol/L): mean (SD)	319.6 (66.6)	228.2 (60.1)	317.3 (66.0)	217.1 (49.8)	271.2 (77.7)
Creatinine clearance (mL/min): mean (SD)	128.2 (33.0)	105.2 (32.1)	138.5 (42.1)	114.5 (34.0)	121.6 (37.7)

All variable data are provided as means and standard deviations or counts and percentages; lipid phenotypes: TC and HDL-C were measured by enzymatic clinical chemistry; all BP traits have been adjusted for anti-hypertensive medication (Levy et al. 2000; Tobin et al. 2005). # Reported weekly alcohol intake ≥21 units in women and ≥28 units in men.

Table 2.5. Characteristics of the German MI Family Study controls and cases

Variable	Controls (n = 847)	MI cases (n = 1090)
Men (%)	41.5	68.4
Age (years): mean (SD)	57.09 (9.95)	58.59 (8.53)
Ever smoker (%); current smoker (%)	50.6; 18.6	70.5; 11.8
Body mass index (kg/m^2): mean (SD)	26.73 (4.07)	27.42 (3.65)
Current antihypertensive treatment (%)	38.4	86.6
SBP (mm Hg): mean (SD)	134.36 (17.66)	138.11 (20.11)
DBP (mm Hg): mean (SD)	82.21 (9.89)	82.31 (10.32)
Hypertension[#] (%)	51.2	91
Current lipid-lowering medication (%)	9.1	63.1
Total cholesterol (mmol/L): mean (SD)	6.19 (1.10)	5.88 (1.16)
LDL cholesterol (mmol/L): mean (SD)	3.83 (0.88)	3.91 (1.09)
HDL cholesterol (mmol/L): mean (SD)	1.56 (0.40)	1.31 (0.36)
Diabetes %	5.7	16

[#] Hypertension was defined as SBP > 140, DBP > 90, or the intake of anti-hypertensive medication

Table 2.6 Baseline characteristics for the WOSCOP study

Characteristic	Case (n=498)	Control (n=1108)	*P*-val
Age (years)	56.9 (5.1)	56.7 (5.2)	a
Body mass index (kg/m^2)	26.0 (3.2)	25.6 (3.2)	0.04
Smokers (%)[a]	266 (53%)	606 (55%)	a
Alcohol consumption (units/week)	11 (13)	11 (13)	0.55
Total cholesterol (mmol/L)	7.08 (0.61)	7.02 (0.57)	0.09
Triaglycerides (mmol/L)	1.96 (0.83)	1.84 (0.77)	0.05
LDL cholesterol (mmol/L)	5.02 (0.46)	4.95 (0.44)	0.04
HDL cholesterol (mmol/L)	1.07 (0.22)	1.14 (0.25)	<0.001
LDL diameter (nm)	26.33 (0.85)	26.40 (0.89)	0.17

Data are provided as mean (standard deviation) for continuous variables and as number of subjects (%) for categorical variables. [a] Cases and controls were matched for age and smoking (Freeman et al. 2003).

2.3.2 Genotyping

The distribution of genotypes and allele frequencies in the GRAPHIC Study for the seven variants analysed are shown in **Table 2.7**. All SNPs except rs17228212 (*P*=0.02) were in HWE at the 5% level of significance, but none were out of HWE if we consider a less stringent threshold *of P*<0.01, considered due to multiple testing of seven SNPs. The allele frequencies were similar to our previous findings in Northern European populations in the WTCCC GWA Study (Samani et al. 2007). The lowest risk allele frequency was 27%, and the highest was 87%.

Table 2.7. Distribution of genotypes for the seven analysed SNPs in the GRAPHIC study

Chr.	Position	SNP	Major allele	Minor allele	Risk allele	Genotypes in parents			Minor frequency	Hardy –Weinberg *P*-value
						Hom Maj	Het	Hom Min		
1	109534208	rs599839	A	G	A	632	334	61	0.22	0.06
1	219211924	rs17465637	C	A	C	542	404	80	0.27	0.7
2	226893585	rs2943634	C	A	C	422	476	114	0.35	0.24
6	151345099	rs6922269	G	A	A	541	411	70	0.27	0.48
9	22115503	rs1333049	G	C	C	256	507	259	0.5	0.8
10	44073873	rs501120	A	G	A	787	219	19	0.13	0.41
15	65245693	rs17228212	T	C	C	502	412	117	0.31	0.02

The data are from the parental generation in GRAPHIC.

Hom Maj = Homozygote for the major allele, *Het* = heterozygotes, *Hom Min* = homozygote for the minor allele,

Risk allele = the allele found to be associated with increased risk of CAD in the GWA study analysis (Samani et al. 2007)

2.3.3 Primary analysis of genetic variant associations with quantitative cardiovascular traits

I analysed the association of the seven SNP variants with ten quantitative traits for cardiovascular risk factors in GRAPHIC: BMI, WHR, systolic and diastolic 24-hr ambulatory BP, 24-hr pulse pressure (PP), blood glucose, TC and HDL-C, blood urate and creatinine clearance, adjusting for age, sex and family structure (see ***Appendix 6.1***, **Table S2.1**: Association findings for all ten quantitative traits in the GRAPHIC study for the seven analysed SNPs). Associations showing at least a nominal level of significance of $P<0.05$ are shown in **Table 2.8**.

The most striking result was the strong statistically significant association of rs599839 on chromosome 1p13.3 with TC, and statistical significance remained even after adjusting for multiple testing via a stringent Bonferroni correction (i.e. 7 SNPs x 10 cardiovascular traits, n=70, which equates to a Bonferroni corrected significance threshold of $P<7.14\text{x}10^{-4}$). The major allele (A), which conferred an increased risk of CAD in our combined GWA studies (Samani et al. 2007), was associated with a 0.17 mmol/L (95% CI: 0.10, 0.24) higher serum cholesterol level per copy of the risk allele (A) with a *P*-value of $P=3.84\text{x}10^{-6}$; see **Table 2.8** and **Figure 2.2A**. An effect size of this magnitude is equivalent to an approximate 6.9% higher cholesterol level in AA major homozygote subjects as compared to GG minor homozygote subjects. Of note, the effect of rs599839 on TC was observed in both the parental and offspring generations. The effect size was somewhat greater in the parental generation [parents: 0.20 mmol/L (95% CI: 0.11, 0.30), $P=4.1\times10^{-5}$ per copy of the risk allele (A); offspring: 0.14 mmol/L (95% CI: 0.06, 0.23), $P=0.001$ per copy of the risk allele (A)], but the difference between generations was not significant ($P>0.05$). There was no interaction with gender ($P=0.60$); the rs599839 risk allele (A) effect size was 0.16 mmol/L (95% CI: 0.07, 0.26) in females and 0.19 mmol/L (95% CI: 0.10, 0.29) in males. Also, the association between rs599839 and TC was not attenuated (and remained significant, $P=1.71\times10^{-6}$) even after adjustment for BMI, WHR, excess alcohol consumption, smoking status (ever smoked, current smoking), blood glucose and diabetes status. No association was observed between rs599839 and HDL-C (**Figure 2.2C** and **Table 2.9**).

Table 2.8. Cardiovascular risk factors showing association with CAD-associated genetic variants in the GRAPHIC study

Chr.	Position	SNP	Risk allele	Minor allele	Phenotype	β-Coefficient	SE	Lower 95% CI	Upper 95% CI	*P*-value
1	109534208	rs599839	A	G	Total Cholesterol (mmol/L)	0.168	0.036	0.096	0.239	3.8×10^{-6}
6	151345099	rs6922269	A	A	Glucose (mmol/L)	0.092	0.045	0.003	0.181	0.042
6	151345099	rs6922269	A	A	Mean 24h PP (mmHg)	-0.504	0.252	-0.997	-0.011	0.045
9	22115503	rs1333049	C	C	Waist-hip ratio	0.005	0.002	0.001	0.01	0.023
9	22115503	rs1333049	C	C	Body Mass Index (kg/m^2)	0.304	0.15	0.009	0.598	0.043

The coefficients are presented in the direction of the effect of the risk allele (i.e., a positive value indicating a higher level in carriers of the risk allele; (PP = pulse pressure).

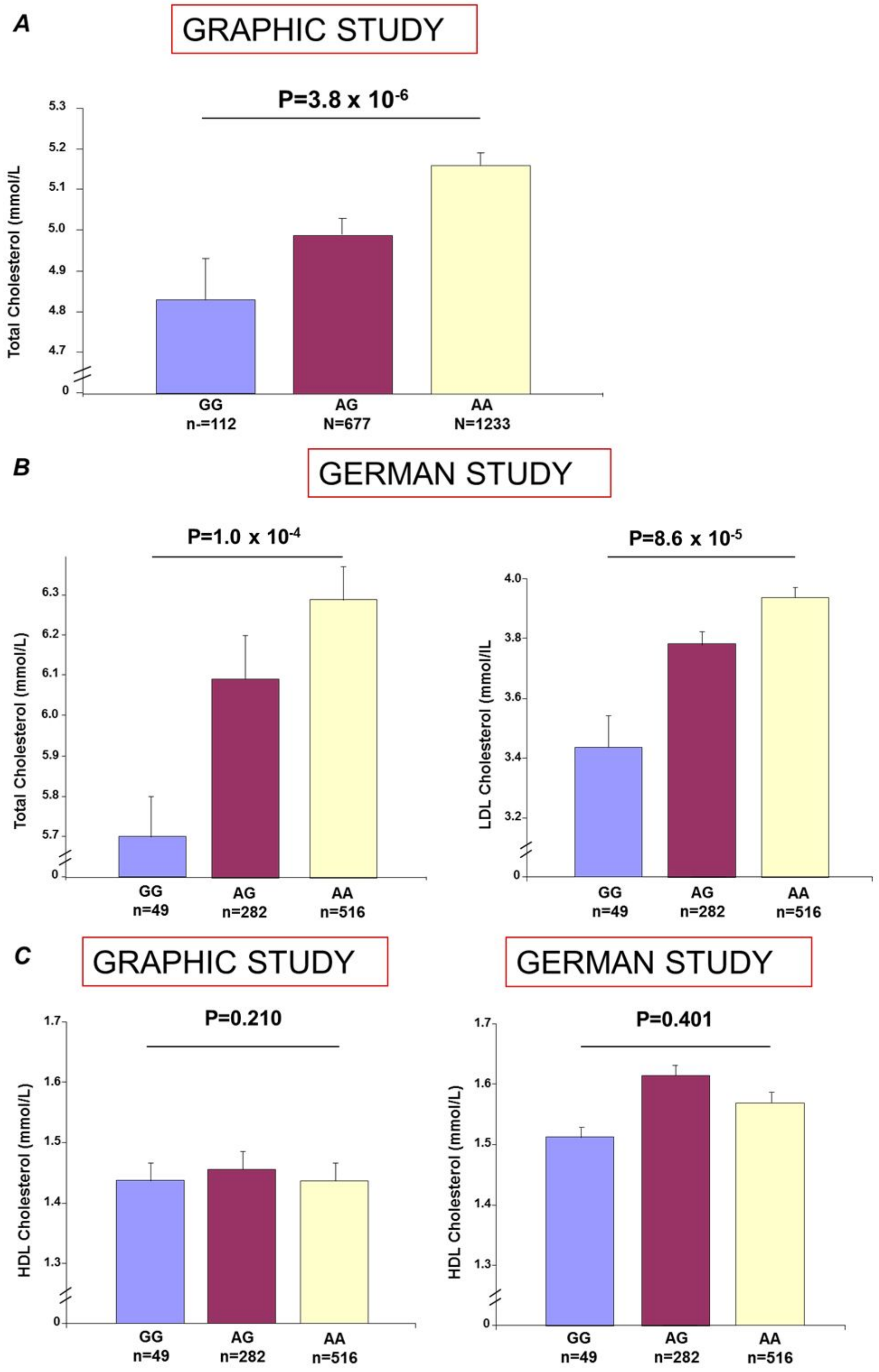


Figure 2.2. Cholesterol association with rs599839 Taqman genotypes**. *A.*** GRAPHIC study TC, ***B.*** GerMIF study controls TC and LDL-C, ***C.*** GRAPHIC and GerMIF controls studies HDL-C.

Table 2.9. Distribution of serum lipid levels in the GRAPHIC Study and the German cohort according to genotypes for rs599839

Parameter	GRAPHIC Study			*P*-value	German controls			*P*-value
	GG (112)	AG (677)	AA (1233)		GG (49)	AG (282)	AA (516)	
Total cholesterol (mmol/L): mean (SD)	4.83 (1.08)	4.99 (1.03)	5.16 (1.10)	3.8×10^{-6}	5.70 (1.14)	6.10 (1.14)	6.28 (1.14)	1.0×10^{-4}
LDL cholesterol (mmol/L): mean (SD)	–	–	–	–	3.42 (0.75)	3.76 (0.75)	3.91 (0.74)	8.56×10^{-5}
HDL cholesterol (mmol/L): mean (SD)	1.43 (0.35)	1.45 (0.37)	1.43 (0.36)	0.21	1.51 (0.14)	1.60 (0.14)	1.55 (0.14)	0.401

Complete genotype, TC and HDL-C data were available for 2,022 / 2,037 (99.3%) of GRAPHIC study subjects. LDL-C data are not available for the GRAPHIC cohort. The *P*-values are after adjustment for age, gender and familial correlations in the GRAPHIC Study and after age and gender in the German controls.

2.3.4 Replication in German data

The rs599839 association with TC was replicated in the GerMIF study control subjects, see **Table 2.9 and Figure 2.2B**. The major allele (A) was associated with an age and sex adjusted 0.24mmol/L (95% CI: 0.09, 0.36), *P*=1.0x10^{-4} rise in TC per copy of the risk allele, which equates to an 8.4% higher TC in AA major homozygotes compared with GG minor homozygotes. Unlike the GRAPHIC study (primary analysis), the German controls also had LDL-C measurements, and analysis showed a highly significant association with LDL-C (**Table 2.9** and **Figure 2.2B**). After adjusting for age and sex, each copy of the risk allele (A) was associated with a 0.19mmol/L higher LDL-C (95% CI: 0.09, 0.29); *P*=8.56x10^{-5}. Again, as in the GRAPHIC primary analysis, no association was found between rs599839 and HDL-C in the German controls (**Table 2.9 and Figure 2.2C**).

An adjustment for the small percentage of subjects on lipid-lowering medications made little difference to the GRAPHIC primary analysis (removal of n=41 subjects slightly reduced the per risk allele (A) TC association to 0.16mmol/L, *P*=7.81x10^{-6}; whereas an all sample inclusive multiplicative adjustment for those on lipid lowering medication (Asselbergs et al. 2012), had a negligible effect on the per risk allele (A) TC association 0.17mmol/L, *P*=3.14x10^{-6}) and German (German data not available) control findings for TC. When pooled the estimated effect size of the per copy risk allele on TC for both the GRAPHIC and German control samples was 0.19mmol/L, *P*=1.7x10^{-9}.

In order to further investigate the association of rs599839 with TC and LDL-C, the SNP rs599839 was genotyped in 1090 MI subjects from the GerMIF study. The risk allele (A) was significantly associated with raised LDL-C, and a per allele copy effect size after adjustment for age and sex of 0.19mmol/L (95% CI: 0.07, 0.31), *P*=0.0026, but no association was found for TC (*P*=0.21) or HDL-C (*P*=0.16), see **Table 2.10**. Due to a high proportion of MI cases taking lipid-lowering medication (61.4%), the data was further analysed by partitioning the data by treatment (i.e. with or without lipid-lowering medication). The risk allele (A) showed a significant association in subjects with raised TC and LDL-C that were not on medication, but an attenuation occurred in those subjects who were on medication, see **Table 2.10**. However, the interaction of genotype with statin was shown not to be significant for TC (*P*=0.158), or LDL-C (*P*=0.252).

Table 2.10. Distribution of serum lipid levels by rs599839 genotype in the German MI Cases, also stratified by lipid-lowering medication

Parameter	Total German MI Cases			*P*-value	German MI Cases without medication			*P*-value	German MI Cases with medication			*P*-value
	GG (36)	AG (342)	AA (712)		GG (20)	AG (136)	AA (265)		GG (16)	AG (206)	AA (447)	
Total cholesterol (mmol/L): mean (SD)	5.89 (1.32)	5.81 (1.32)	5.92 (1.32)	0.2116	5.92 (1.13)	6.00 (1.12)	6.25 (1.06)	0.025	5.83 (1.37)	5.68 (1.37)	5.74 (1.36)	0.7289
LDL cholesterol (mmol/L): mean (SD)	3.81 (1.18)	3.77 (1.18)	3.99 (1.18)	0.0026	3.85 (1.13)	4.02 (1.13)	4.34 (1.13)	0.0029	3.71 (1.14)	3.61 (1.14)	3.80 (1.14)	0.0685
HDL cholesterol (mmol/L): mean (SD)	1.40 (0.12)	1.31 (0.12)	1.30 (0.12)	0.1609	1.30 (0.12)	1.29 (0.12)	1.31 (0.12)	0.7622	1.50 (0.11)	1.33 (0.11)	1.30 (0.11)	0.0414

Complete genotypes, TC, LDL-C and HDL-C were available for 1090 German MI cases. *P*-values are adjusted for age and sex.

2.3.5 LDL particle size and sub-fraction analysis

In light of the German replication cohort identifying that LDL-C rather than TC shows a stronger association with rs599839 risk allele (A), a secondary analysis of GRAPHIC study plasma samples was undertaken to measure LDL-C by enzymatic clinical chemistry (as described for the GerMIF study) and cholesterol measured lipoprotein subclasses by NMR (LipoFIT® Analytic GmbH; Regensburg, Germany). This was to ensure that LDL-C was indeed associated with rs599839 in the GRAPHIC study, and to see if the effect of the 1p13.3 locus could be better defined by other lipoprotein cholesterol subclasses. The association results of GRAPHIC rs599839 genotypes with LDL-C and lipoprotein subclass phenotypes are shown in **Table 2.11**, and show that LDL-C level is the most strongly associated phenotype with rs599839 risk allele (A) resulting in a rise of 0.15 mmol/L (95% CI: 0.09, 0.20); P=9.9x10^{-8}. Interestingly, LDL-C subclasses D and B also show an association with rs599839, with a per risk allele (A) increase of 0.021 mmol/L (95% CI: 0.007, 0.036); P=0.005 for LDL-C subclass D and 0.011 (95% CI: 0.003, 0.018); P=0.008 for LDL-C subclass B. The LDL-C association with rs599839 in GRAPHIC remained significant even after adjustment for multiple testing using a conservative Bonferroni corrected *P*-value of P<6.5x10^{-4} (i.e. 0.05/(7 SNP x 11 cardiovascular phenotypes)). However, when applying an adjustment for multiple testing using a conservative Bonferroni corrected *P*-value of P<2.8x10^{-3} (i.e. 0.05 / (1 SNP x 18 lipoprotein subclass phenotypes)), on the LDL-C subclasses D and B for the 12 cholesterol measured subclasses, along with mean HDL, LDL and VLDL particle size, and particle number, the association becomes suggestive (i.e. borderline significant). As for the other lipoprotein subclass phenotypes: LDL-C subclasses E and C, HDL-C subclasses D and B, and LDL particle number all showed nominal significance P<0.05. The remaining HDL and LDL subclasses, IDL and VLDL subclasses, as well as LDL size, plus HDL and VLDL number and size, showed no association with rs599839 genotypes.

Table 2.11. Additional GRAPHIC clinical chemistry and NMR lipid measurements

Chr.1	Phenotypes	β-coefficient (SE)	Lower 95% CI	Upper 95% CI	*P*-value
rs599839	LDL-C (mmol/L)	0.145 (0.027)	0.092	0.198	$9.97x10^{-8}$
	Total-C (mmol/L)	0.169 (0.036)	0.098	0.240	$3.14x10^{-6}$#
	LDL-C subclass D (mmol/L)	0.021 (0.008)	0.007	0.036	0.005
	LDL-C subclass B (mmol/L)	0.011 (0.004)	0.003	0.018	0.008
	LDL particle number	21.07 (8.945)	3.536	38.60	0.02
	LDL-C subclass E (mmol/L)	0.020 (0.009)	0.002	0.038	0.03
	HDL-C subclass D (mmol/L)	0.004 (0.002)	0.0004	0.007	0.03
	LDL-C subclass C (mmol/L)	0.011 (0.005)	0.001	0.021	0.03
	HDL-C subclass B (mmol/L)	-0.010 (0.005)	-0.020	-0.0002	0.04
	IDL-C subclass (mmol/L)	0.449 (0.294)	-0.127	1.024	0.13
	HDL-C subclass C (mmol/L)	-0.006 (0.004)	-0.014	0.002	0.13
	LDL particle size (nm)	0.023 (0.016)	-0.008	0.054	0.14
	HDL-C (mmol/L)	-0.020 (0.014)	-0.048	0.008	0.16
	HDL particle size (nm)	-0.014 (0.010)	-0.033	0.006	0.17
	VLDL particle number	3.480 (2.610)	-1.636	8.595	0.18
	VLDL-C subclass A (mmol/L)	0.265 (0.219)	-0.164	0.694	0.23
	HDL-C subclass A (mmol/L)	0.005 (0.005)	-0.004	0.014	0.31
	VLDL particle size (nm)	-0.029 (0.032)	-0.092	0.034	0.37
	LDL-C subclass A (mmol/L)	0.004 (0.004)	-0.005	0.012	0.38
	VLDL-C subclass B (mmol/L)	0.096 (0.161)	-0.220	0.411	0.55
	HDL particle number	-23.85 (134.5)	-287.6	239.9	0.86

# Measurement of serum TC from Leicester chemical pathology laboratory at the time of recruitment adjusted for lipid-lowering medication (Asselbergs et al. 2012). All other data were measured from -80°C frozen once plasma stocks (4 years after recruitment completion of GRAPHIC samples) either by enzymatic chemistry or NMR (LipoFIT® Analytic GmbH; Regensburg, Germany). The β-coefficients are presented in the direction of the effect of the rs599839 risk allele (A) (i.e., a positive value indicating a higher level in carriers of the risk allele.

2.3.6 Analysis of the association of rs599839 on lipid-lowering effects of statin therapy: the WOSCOP Study

The German MI data suggested that those on lipid-lowering medication attenuated the effects of the 1p13.3 locus on LDL-C (**Table 2.10.)**, but (as previously mentioned) the German data showed no significant statin effect on TC (*P*=0.158) and LDL-C (*P*=0.252), by rs599839 genotype. Therefore, to be certain that rs599839 genotypes do not affect the amount of lipid-lowering from statin treatment, I genotyped the seven WTCCC GWA study variants in a large (n=1607) unbiased randomised double-blind pravastatin / placebo prospective study - the WOSCOP study.

The distribution of genotypes and allele frequencies in the WOSCOP study for the seven novel variants are shown in **Table 2.12**. All SNPs except rs17465637 (*P*=0.035) and rs6922269 (*P*=0.042) were in HWE at a conservative threshold of *P*<0.05; but none were out of HWE at a less stringent HWE of *P*<0.01, considered because of multiple testing of seven SNPs. Allele frequencies in the control samples were similar to those found in the GRAPHIC study. A Cochrane-Armitage trend test CAD case-control analysis was performed on all SNPs and only rs1333049 (*P*=0.028) was found to be significant at a nominal 5% level of significance. However, if we take into account multiple testing under a conservative Bonferroni corrected threshold of P<7.14x10^{-3}, none of the SNPs remained significant.

Two further WOSCOP study analyses were performed using linear regression. Firstly, to assess the effect of baseline TC, LDL-C, HDL-C and TG measurements by rs599839 genotype; and secondly, (more especially) in follow-up of the finding that lipid lowering drug medicated only German MI subjects saw an attenuation in the association with the 1p13.3 locus, to assess the effect of statin on levels of change in TC, LDL-C, HDLC and TG from baseline by rs599839 genotype. As shown in **Figure 2.3.**, the risk allele (A) shows a nominally significant association with an increase in baseline TC (*P*=0.037) and LDL-C (*P*=0.019) and a decrease in baseline HDL-C (*P*=0.028), plus a slight baseline increase but no association with

TG, under an additive model of genetic inheritance. **Figure 2.4** shows that pravastatin was responsible for a decrease in TC (*P*=0.649) and LDL-C (*P*=0.423), when compared to baseline levels, but there was no association with rs599839 genotypes. However, risk allele (A) was associated with an increase in HDL-C (*P*=0.044) by pravastatin, when compared to baseline levels, albeit at a nominal level of significance. Finally, pravastatin caused a slight reduction in TG when compared to baseline levels, but there was no association with rs599839 genotype.

Table 2.12. Distribution of WOSCOP study (case-control) genotypes and analysis for the 7 novel WTCCC CAD risk loci

SNP	Minor/ Risk Allele	WOSCOP study control genotypes			WOSCOP study CAD genotypes			MAF Controls	OR allelic	Lower 95% CI	Upper 95% CI	CA Trend *P*-value	HWE *P*-value Controls
		Maj Hom	Het	Min Hom	Maj Hom	Het	Min Hom						
rs599839	G/A	729	351	29	329	148	20	0.184	0.969	0.800	1.174	0.747	0.082
rs17465637	A/C	563	433	112	264	188	42	0.296	1.109	0.939	1.311	0.236	0.035
rs2943634	A/C	499	471	137	214	227	56	0.336	0.980	0.837	1.148	0.804	0.117
rs6922269	A/A	558	475	74	230	225	40	0.281	1.137	0.965	1.339	0.112	0.042
rs1333049	C/C	299	548	262	107	258	132	0.483	1.182	1.018	1.373	0.028	0.724
rs501120	C/T	854	238	18	381	106	9	0.123	0.986	0.786	1.236	0.901	0.762
rs17228212	C/C	503	487	115	216	222	58	0.324	1.076	0.918	1.261	0.364	0.857

Legend. The WOSCOP case-control study consisted of n=1607 genotyped subjects (n=1110 controls and n=497 CAD cases). *Risk allele* = the allele found to be associated with increased risk of CAD in the GWA study analysis (Samani et al. 2007); *Hom Maj* = Homozygote for the major allele, *Het* = heterozygotes, *Hom Min* = homozygote for the minor allele; MAF = minor allele frequency; OR = odds ratio; 95% CI = 95% confidence intervals; CA Trend = Cochrane-Armitage trend test; HWE = Hardy-Weinberg equilibrium.

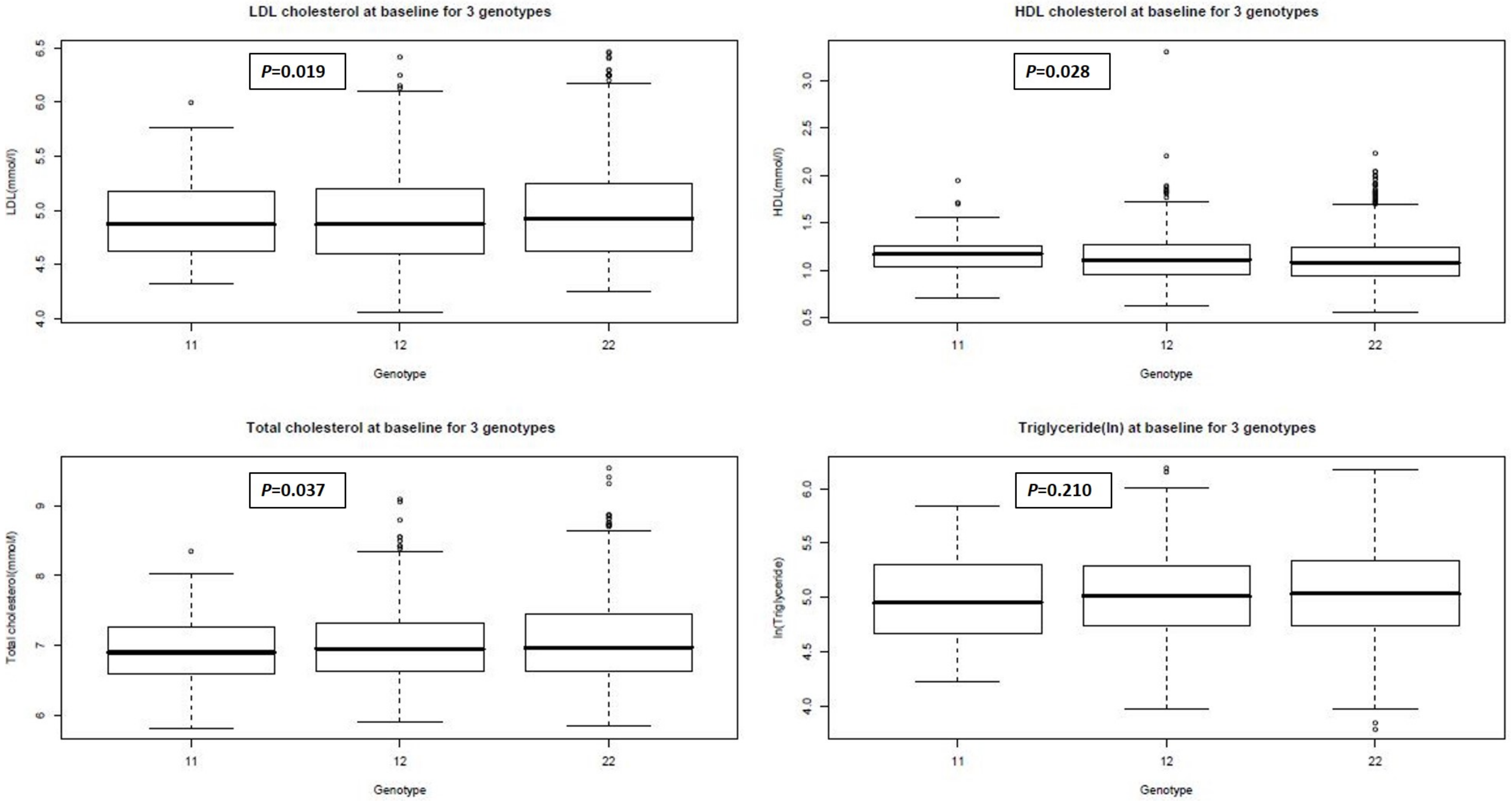


Figure 2.3. WOSCOP study box plots for LDL-C, HDL-C, TC and TG at baseline measurement levels by rs599839 genotypes (11=GG, 12=AG, and 22=AA; A is the risk allele). A box plot represents five summaries: 1) sample minimum (the smallest value observed); 2) lower quartile Q1 (the 25th percentile); 3) the median Q2 (the 50th percentile); 4) upper quartile Q3 (the 75th percentile); 5) sample maximum (the largest value observed); beyond the five summaries are outliers represented by open circles.

N.B. I would have liked to have created a figure with bar graphs something akin to those shown in Fig. 2.2, but unfortunately I was not privy to the raw data used to generate these box-plots.

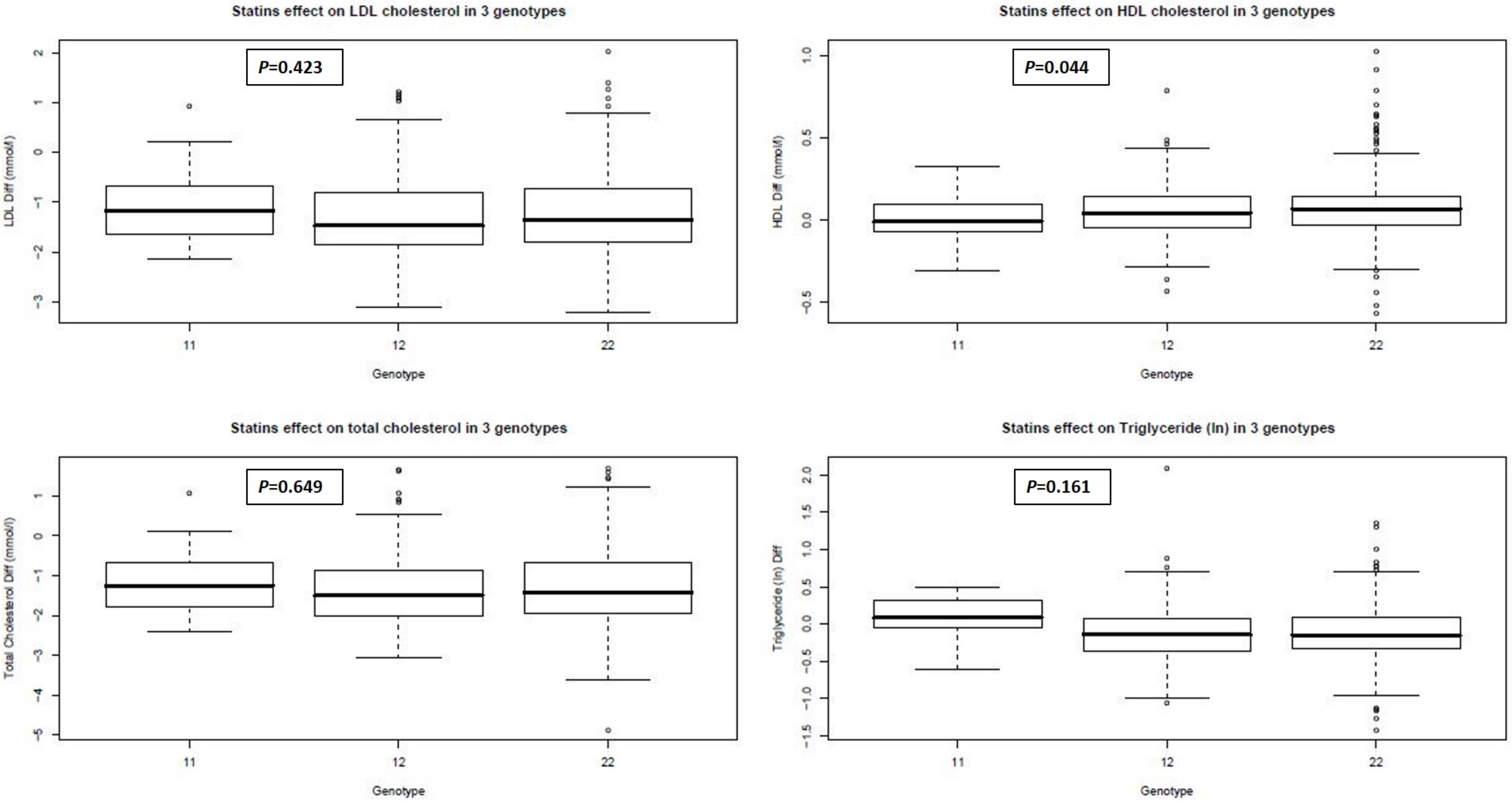


Figure 2.4. WOSCOP study box plots to show the effects of Pravastatin on LDL-C, HDL-C, TC and TG change compared to baseline measurements by rs599839 genotypes (11=GG, 12=AG, and 22=AA; A is the risk allele). A box plot represents five summaries: 1) sample minimum (the smallest value observed); 2) lower quartile Q1 (the 25th percentile); 3) the median Q2 (the 50th percentile); 4) upper quartile Q3 (the 75th percentile); 5) sample maximum (the largest value observed); beyond the five summaries are outliers represented by open circles.

N.B. I would have liked to have created a figure with bar graphs something akin to those shown in Fig. 2.2, but unfortunately I was not privy to the raw data used to generate these box-plots.

2.3.7 Post-hoc power calculations

Post-hoc power calculations were generated from the mean and standard deviation of the completed study. The 80% and 90% power estimate thresholds to detect a statistically significant effect size (i.e. β-coefficient estimates), for each of the cardiovascular traits tested for genetic association with a biallelic SNP (of any allele frequency), using a Bonferroni corrected significance level of α=0.002 (0.05/23) are shown in **Table 2.13**.

Table 2.13 Post-hoc power calculations for detectable effect sizes

Cardiovascular phenotype	obs	mean	SD	min	max	diff80	diff90
Total-C (mmol/L)	1024	5.685	0.999	3.200	10.100	0.125	0.139
HDL-C (mmol/L)	1024	1.480	0.386	0.500	3.200	0.048	0.054
LDL-C (mmol/L)	995	3.267	0.730	1.297	6.032	0.091	0.102
mean 24hr SBP (mmHg)	1023	121.98	13.118	94.663	190.80	1.640	1.824
mean 24hr DBP (mmHg)	1023	75.592	8.484	52.458	121.01	1.060	1.180
mean 24hr PP (mmHg)	1023	67.946	10.263	40.500	103.00	1.283	1.427
BMI (kg/m^2)	1024	27.446	4.274	17.135	48.000	0.534	0.594
WHR	1016	0.872	0.088	0.622	1.143	0.011	0.012
blood glucose (mmol/L)	1024	5.190	1.416	2.000	21.700	0.177	0.197
blood urate (mmol/L)	1024	273.55	78.154	75.000	553.00	9.768	10.867
CCR (mL/min)	969	116.73	34.620	30.637	395.20	4.327	4.814
HDL-C subclass A (mmol/L)	982	0.710	0.118	0.362	1.138	0.015	0.016
HDL-C subclass B (mmol/L)	982	0.359	0.135	0.028	0.861	0.017	0.019
HDL-C subclass C (mmol/L)	982	0.245	0.103	0.064	0.680	0.013	0.014
HDL-C subclass D (mmol/L)	982	0.261	0.043	0.140	0.408	0.005	0.006
LDL-C subclass A (mmol/L)	982	0.418	0.124	0.132	0.963	0.016	0.017
LDL-C subclass B (mmol/L)	982	0.595	0.116	0.177	1.273	0.015	0.016
LDL-C subclass C (mmol/L)	982	0.643	0.150	0.313	1.515	0.019	0.021
LDL-C subclass D (mmol/L)	981	0.672	0.215	0.008	2.006	0.027	0.030
LDL-C subclass E (mmol/L)	982	0.753	0.271	0.272	2.217	0.034	0.038
IDL-C subclass (mmol/L)	982	19.784	7.839	3.000	60.000	0.980	1.090
VLDL-C subclass A (mmol/L)	982	11.335	6.306	1.860	44.150	0.788	0.877
VLDL-C subclass B (mmol/L)	982	7.442	4.284	0.850	34.730	0.535	0.596

Legend. obs=observations, SD=standard deviation, min=lowest individual obs, max=highest individual obs, diff80=effect size estimate for 80% power, diff90=effect size estimate for 90% power.

2.3.8 Other cardiovascular trait associations

Although, there were several other CAD lead SNPs that showed nominal associations with quantitative traits (see **Table 2.8**), none reached statistical significance based on the multiple-testing of seven SNPs and ten cardiovascular traits (Bonferroni corrected *P*-value threshold *P*=7.14x10^{-4}), and are likely to be chance results. This is supported by the fact that in the case of rs6922269, the 24-hr ambulatory PP trend was in the opposite direction to that for CAD risk. In addition, the associations for rs6922269 with glucose and rs1333049 with BMI were tested in the German cohort and neither showed a similar association (rs6922269 and glucose, coefficient was −0.10 (95% CI: −0.33, 0.14), *P*=0.421; rs1333049 and BMI, coefficient was −0.25 (95% CI: −0.63, 0.14), *P*=0.207).

2.4 Discussion

The advent of the GWA study has begun a new era in genomics, culminating in the recent identification of several novel loci that strongly associate with CAD risk (Helgadottir et al. 2007; McPherson et al. 2007; Wellcome Trust Case Control Consortium 2007; Samani et al. 2007), and sparking a fresh perspective on the genetic basis of this common complex condition. Although replication in further studies is of importance, it is also relevant to identify potential mechanisms by which, the novel loci discovered might mediate their effect on CAD. Therefore, an immediate question to ask is whether these novel loci can affect, and therefore work through, well known traditional cardiovascular risk factors. To this end, I have investigated whether any of the seven novel loci associated with CAD, as identified by our combined UK and German GWA study analyses (Samani et al. 2007), are also associated with quantitative traits that are known to affect cardiovascular risk.

The key result from this study is that the CAD risk allele (A) for SNP rs599839 on chromosome 1p13.3, in a strong LD block region that includes the two genes *CELSR2* and *PSRC1*, is strongly associated with higher levels of serum TC (0.17mmol/L*, P*=$3.8\text{x}10^{-6}$) and LDL-C (0.15mmol/L, *P*=$9.97\text{x}10^{-8}$), which remained significant even after a Bonferroni correction for multiple testing (*P*<$7.14\text{x}10^{-4}$). It should be noted here that an attempt to find subtle association effects in twelve cholesterol measured lipoprotein subclasses for HDL, LDL, IDL and VLDL (along with six additional variables for mean particle size and total particle numbers for HDL, LDL and VLDL) measured by NMR (see **Table 2.2**) at this locus, identified two LDL-C subclasses - LDL-C subclass D (*P*=0.005) and B (*P*=0.008) as being significantly associated with rs599839. Furthermore, these two LDL subclasses D and B almost remained significant, even after a conservative adjustment for multiple testing using a Bonferroni corrected threshold of *P*<0.003. However, given that the cholesterol measured lipoprotein subclasses could be considered as not being wholly independent covariates, and that the use of the Bonferroni correction in this instance, could be seen as an overly conservative *P*-value correction, I would suggest that the two LDL-C subclasses D and B, should be considered as truly significant associations.

The SNP rs599839 is located just 10bp beyond the 3’UTR of the *PSRC1* gene and ~3.7kb downstream of the 3’UTR of the *CELSR2* gene (see **Figure 2.1**), and maps to the telomeric section (i.e. the section of the locus region located toward the telomere of chromosome 1p short-arm) of a 150kb chromosome 1p13.3 locus region (encompassing 4 genes *CELSR2*, *PSRC1*, *MYBPHL* and *SORT1*) that is consistently associated with LDL-C in recent GWA studies (Kathiresan et al. 2008; Willer et al. 2008; Sandhu et al. 2008; Prospective Studies Collaboration et al. 2007). *In silico* bioinformatics annotation and any scientific literature regarding the genes situated in the 1p13.3 locus, will be discussed in more detail in the next chapter.

I observed a consistent association between the risk allele (A) of SNP rs599839 with an increase in TC and LDL-C, not only in a family-based study representative of the general public, but also in cohorts of unrelated healthy subjects, as well as, MI subjects that were recruited for the purposes of case-control studies. In addition, this same strong association with TC and LDL-C has been recently identified in four other GWA studies (Kathiresan et al. 2008; Willer et al. 2008; Sandhu et al. 2008; Wallace et al. 2008). The overall findings from these investigations are that the locus marked by SNP rs599839 on chromosome 1p13.3 represents a novel locus involved in cholesterol and more specifically in pathway mechanisms that involve LDL-C metabolism.

Due to the multifactorial nature of the complex trait CAD it was to be expected that the effect of the rs599839 chr1p13.3 locus on TC and LDL-C is modest: 0.17-0.24 mmol/L (3.6-4.2%) and 0.15-0.19 mmol/L (4.6-6.0%) respectively, per copy of the rs599839 SNP risk allele (A). However, as the risk allele for this locus is the major allele, approximately 60% of individuals are homozygous for the risk allele (A). Therefore from a public health perspective, the potential relevance of the rs599839 chr1p13.3 effect on LDL-C is substantial. Of note, a meta-analysis of 61 prospective observational studies by the Prospective Studies Collaboration have shown that a 1 mmol/L higher TC level in those of age 40 to 49 years is associated with a 55% (95% CI: 53-58%) higher risk of CAD, and in those of age 50 to 59 years is associated with a 43% (95% CI: 42-45%) increased risk in CAD in both sexes (Prospective Studies Collaboration et al. 2007). We have observed in our combined GWA studies a 29% (95% CI: 18-40%) increase in risk of CAD per copy of the risk allele (A) in subjects with an average age of 50 years (Samani et al. 2007). The impact of the rs599839 chr1p13.3 locus on CAD risk is in fact double the effect observed on cholesterol for this locus. However, it should be noted that the CAD risk estimate has large 95% CI, and could indicate an inaccurate measure of CAD risk. Indeed, a replication by the CAD consortium reported an OR and 95% CI of 1.13 and 1.08-1.19 respectively, which is more a kin to the CAD risk generated by TC and LDL-C, and could support the notion that our original combined GWA studies (WTCCC and GerMIF) is an over estimation of CAD risk (Coronary Artery Disease Consortium et al. 2009). But more importantly, the influence of this locus on cholesterol could be present since birth - and therefore its effect on affecting CAD risk could have started from a very young age - such that these observational studies may well be under estimating, the true effect of this level of average cholesterol difference that has taken place over decades on CAD risk. Thus, in order to truly measure the CAD risk impact of the rs599839 locus, very large longitudinal prospective studies are needed that generate large numbers of CAD subjects and are able to observe the long-term effects of the locus on cholesterol, to ascertain the extent to which the risk of CAD is attributable to LDL-C.

To explore in more detail how the rs599839 locus affects LDL-C, I examined the association of genotype with LDL particle number and with LDL mean particle sizes measured in the GRAPHIC subjects. The findings from this analysis clearly indicate that the locus does not primarily influence plasma LDL-C through an effect on LDL particle number. Interestingly, there appeared to be a greater association with LDL-B and LDL-D particles rather than LDL-A, C or E.

The reason for measuring the lipoprotein sub-fractions (A to E), is due to previous reports that small particle size dense LDL subclass particles are more atherogenic, than large buoyant LDL subclass particles (Berneis & Krauss 2002). Indeed, epidemiological studies have shown that small dense LDL is associated with increased risk of CAD and clinical events. However, it should be noted that small dense LDL are only one component of a complex physiological syndrome (i.e. increased TG, decreased HDL, insulin resistance, and increased central obesity), and so this has prevented a definitive independent risk of CAD from small dense LDL particles. Nevertheless, lipid lowering intervention studies have shown that small dense LDL predicts angiographic changes in response to statins, and the conversion of small dense LDL to large buoyant LDL has been associated with CAD regression (Hokanson 2000).

In this analysis, LDL-C subclasses D and B were significantly associated to about the same level *(P*=0.005 and *P*=0.008 respectively), whereas LDL subclasses E and C were only nominally significant, and the smallest LDL subclass A was insignificant. So how do these findings correlate with the scientific literature? In general, the scientific literature shows that there is a lack of standardisation in terms of LDL-C subclass measurement and nomenclature surrounding the description of particle size and density, which makes it hard to make direct comparisons between studies (see **Table 2.14** and **2.15**). Nevertheless, when looking back to earlier comparisons of LDL-C particle size and number it was considered that those with CAD were associated with a Pattern B (<25.5nm) profile type with large numbers of small dense LDL-C particles; whereas healthy individuals were considered to have a Pattern A (>25.5nm) profile type with a predominance of large buoyant LDL-C particles (see **Table 2.14** Subclass* and Subclass‡). Under these parameters both LDL subclasses D and B would be considered small and dense, and therefore a significant increase in cholesterol for the risk allele rs599839 fits well with the atherogenic nature of such LDL subclasses. Yet again, although the GRAPHIC LipoFIT® NMR subclassifications differ somewhat from previous classifications (see Table **2.15**).The CAD associations found in the EPIC-Norfolk study by Harchaoui *et al.*, do seem to overlap with our findings. The GRAPHIC LDL subclass D (22-25nm) particle size overlaps with the Harchaoui *et al.* IDL (23-27nm), and the LDL subclass B

(19-21nm) particle size overlaps with the Harchaoui *et al.* small LDL (18-21.2nm).

Table 2.14 Comparison of different LDL subclass definitions

Subclass*	Density range (g/mL) §	Peak Size range (nm)*	Subclass‡	Density range (g/mL)	Subclass‡	Size range (nm)	Subclass§	Weighted score
I	1.025-1.033	26.5-28.5	1	1.025-1.034	A1	>26.0	1 Large	1.00-2.60
IIA	1.033-1.038	26.0-26.5	2	1.034-1.044	A2	>25.5	2 intermediate	2.60-3.80
IIB		25.5-26.0					3 intermediate	
IIIA	1.038-1.050	24.7-25.5			B3	<25.5	4 small	3.80-5.60
IIIB		24.2-24.7					5 small	
IVA	1.050-1.063	23.2-24.2	3	1.044-1.060	B4	<24.7	6 very small	>5.60
IVB		22.0-23.2					7 very small	

Legend. Table adapted from Krauss & Blanche, 1992. Data provided from three sources – Griffin *et al.* ‡, Campos *et al.* §, Williams *et al.** (Griffin et al. 1990; Campos et al. 1992; Williams, Vranizan & Krauss 1992)

Table 2.15 Comparison of NMR based LDL subclass definitions

GRAPHIC NMR Subclass Results				Harchaoui *et al.* 2007 (CAD cohort)			Kuller *et al.* 2002	
Subclass	Density	Size	*P*-value	Subclass	Size	*P*-value	Subclass	Size
IDL	1.015	30-40	0.13	IDL	23-27	0.003	Small VLDL	27-35
LDL E	1.019	25-30	0.03				IDL	23-27
LDL D	1.027	22-25	0.005				Large LDL	21.3-23
LDL C	1.035	21-22	0.03	Large LDL	21.2-23	0.6	medium LDL	19.8-21.2
LDL B	1.045	19-21	0.008	small LDL	18-21.2	<0.0001	Small LDL	18.3-19.7
LDL A	1.060	16-19	0.38					

Legend. Comparison of GRAPHIC LipoFIT® NMR subclassifications vs. Liposcience Inc. (El Harchaoui et al. 2007) and LipoMed Inc. (Kuller et al. 2002).

In the German MI cases, I observed an attenuated LDL-C difference between genotypes in those taking lipid-lowering medications (see **Table 2.10.**) This raised the possibility that the genotype at the rs599839 locus may influence the amount of lipid lowering achieved by statin therapy with a greater lowering in those carrying the risk allele resulting in a lesser difference between the genotype groups.

To investigate this, I took advantage of our previous collaboration of the Samani group with the WOSCOP study investigators (Brouilette et al. 2003) , and genotyped rs599839 in subjects in the statin-arm of the this randomised clinical trial and compared the difference in baseline and on-treatment LDL-C values between the genotype groups (**Figure 2.4**). It is clear from this analysis that the genetic variant rs599839 shows no association with statin treatment (*P*=0.423), despite a suggestion from the non-randomised German MI cases, the rs599839 SNP within the 1p13.3 locus does not influence the amount of LDL-C lowering achieved by pravastatin. A potentially important implication of this observation is that the rs599839 locus likely affects LDL-C through a pathway distinct from that which is targeted by statins and hence understanding the mechanism may provide a new target for lowering LDL-C and consequently CAD risk. However, it should be noted that by design the WOSCOP study only included subjects with moderate LDL-C measurements at baseline, and are atypical of the general population, and indeed of the German MI case subjects. Indeed, baseline LDL-C was found to not be a good predictor of outcome in the WOSCOP study (West of Scotland Coronary Prevention Study Group 1998). Therefore, although this was a randomised trial it may not be the best cohort to compare with GRAPHIC and the GerMIF study. However, putting these issues aside, the WOSCOP study findings do confer with a reported lack of statin effect association in the GerMIF study on TC (*P*=0.158) and LDL-C (*P*=0.252), by rs599839 genotype. Therefore, it is most likely that the LDL-C associated 1p13.3 locus works through a pathway mechanism other than that affected by statins.

Although I observed a number of nominal associations between the other six CAD-associated SNPs and cardiovascular risk traits in the GRAPHIC Study, the level of significance of these associations (when taking multiple-testing into account), the lack of replication in the German cohort for some of the associations, and the fact that one of the associations (rs6922269 and 24-hr ambulatory PP) is in the opposite direction to the association with CAD, leads one to the conclusion that these findings are unlikely to be true associations. However, due to limitations of my study design, one cannot totally rule out the possibility that one or more of these loci, could act via some of these cardiovascular traits. Hence, although it is unlikely I have any false positive (Type I errors) results, it is plausible that I may have false negative (Type II errors) within my results, for some intermediate phenotypes tested by genotype. Indeed, this was born out by the section discussed in parenthesis at the end of this chapter, for rs2943634, see ***Appendix* 6.1, Table S2.2**.

The study limitations included firstly that although my study was adequately powered for the detection of modest (hypothesis generating) effect sizes, it had little power to detect small effect sizes, or those that involved unforeseen interactions with other traits or the environment. The *post-hoc* effect sizes that the GRAPHIC Study had >80% and >90% power to detect at a Bonferroni corrected alpha of 0.002, for the quantitative traits analysed are shown in **Table 2.13**. Secondly, because my aim was to analyse the SNPs identified as CAD-associated in the GWA studies, I have not analysed the full genetic architecture of these regions, i.e. performed fine-mapping (a methodology that is described in Chapter 3). Thirdly, GRAPHIC blood samples were collected under non-fasting conditions, which may have impacted on the association with food-intake related traits such as glucose. Therefore it is important that additional studies on the effects of these loci on both the currently measured cardiovascular risk factors and others are performed - although it would seem from my findings that at least some of these novel loci, act through as yet unknown mechanisms.

In summary, in the work presented in this Chapter, I report that a novel locus on chromosome 1p13.3 that was recently identified to be associated with CAD risk - also associates with the plasma levels of TC and LDL-C. These findings indicate a need for further focus on this locus to identify the gene and understand the mechanism by which it affects plasma LDL-C. My efforts to do this are described in the next Chapter.

(Since the completion of my analysis of the seven loci, several large scale GWA study meta-analyses have been reported for major cardiovascular risk factor traits. For completeness, I examined the association of the lead SNPs for the seven CAD loci in these datasets. The results are summarised in the ***Appendix* 6.1, Table S2.2**. Despite the much larger sizes of these analyses and greater power, only the association of the chromosome 1 rs599839 locus with LDL-C, and a MAGIC (Meta-Analysis of Glucose and Insulin related traits Consortium) consortium association of the chromosome 2 rs2943634 locus, which reached genome-wide significance for increased fasting insulin (*P*=1.49x10^{-13}), increased TG (*P*=5.17x10^{-8}) and reduced HDL-C (*P*=2.33x10^{-9}), none of the other loci shows a significant association with these cardiovascular traits, suggesting that they indeed act in an independent fashion from that of conventional risk factors).

3. Chapter 3

Further characterisation of CAD and LDL-C associated locus on chromosome 1p13.3 by refinement and in silico bioinformatics analyses.

3.1 Introduction:

The key findings of the previous chapter were that the lead SNP rs599839, within the CAD risk associated chr1p13.3 locus, is also associated with known traditional CAD risk intermediary phenotypes - total cholesterol (TC) and low density lipoprotein cholesterol (LDL-C). Indeed, each copy of rs599839 risk allele (A) was associated with an approximate 0.17-0.24 mmol/L (3.6-4.2%) and 0.15-0.19 mmol/L (4.6-6.0%) increase in TC and LDL-C levels respectively, in the family-based GRAPHIC study and healthy controls GerMIF study, both representative of the general population (Samani et al. 2008). Interestingly, an additional discovery was that two LDL-C subclasses D and B were found to associate with rs599839, whereas LDL-C subclasses C and E only showed nominal significance before adjustment for multiple testing, and subclass A was non-significant.

Below **Figure 3.1** shows the regional *Manhattan* plot of CAD associations at 1p13.3 in the WTCCC GWA study for all the SNPs available on the Affymetrix 500K array that was utilised. The plot is centred upon the lead SNP rs599839 (identified by a purple diamond) and includes all other SNP *P*-values within a 300kb window (i.e., 150kb upstream or downstream of the lead SNP). Additionally, the plot is annotated along the X-axis chromosome location with the local vicinity candidate genes. Only one other SNP rs4970834 is shown to be in moderate pairwise LD (r^2=0.59) with rs599839, this SNP is intronic and located within the *CELSR2* gene.

As described in the previous chapter, rs599839 is located within the 3’ flanking region of two genes *CELSR2* (~3.7kb downstream of the 3’UTR) and *PSRC1* (just

10bp downstream of the 3'UTR). Until now, neither gene has been connected or implicated with having a role in CAD or cholesterol metabolism.

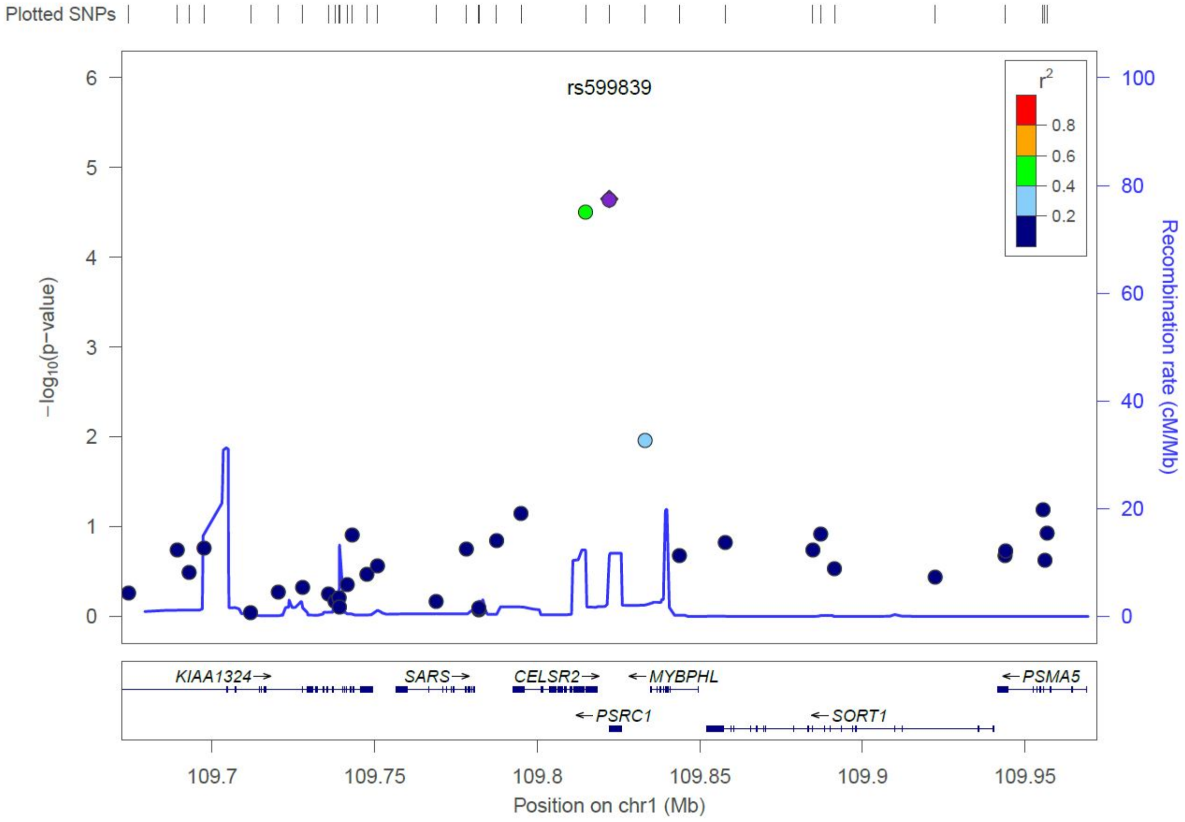


Figure 3.1. Regional Manhattan plot of Affymetrix 500K WTCCC GWA study lead CAD associated SNP rs599839 within the 1p13.3 locus. Plotting Cochran-Armitage trend test –log10 *P*-values for CAD association (Y-axis) *vs.* chromosome base pair position (X-axis). The purple diamond represents the lead SNP at this locus. The pairwise LD r^2 with the lead SNP is represented by colour-coding. Recombination rates are plotted on the right hand side of the Y-axis. The plot is also annotated with (The plot is generated by the web tool LocusZoom, hosted via the Center for Statistical Genetics, University of Michigan via the website https://statgen.sph.umich.edu/locuszoom/genform.php?type=yourdata; Version 1.1: June 2011 (Pruim et al. 2010)).

CELSR2 (cadherin, EGF LAG seven-pass G-type receptor 2) encodes a flamingo/CELSR subfamily cadherin (part of the cadherin superfamily) that is primarily expressed in the brain and testis. This flamingo subfamily contains atypical non-classic cadherins that do not interact with catenins. This particular flamingo cadherin along with others in the subfamily is located within the plasma membrane and has nine cadherin domains, seven epidermal growth factor-like repeats and two laminin A G-type repeats in the ectodomain. In addition, this protein also contains a unique subfamily characteristic – seven transmembrane domains. The suspected role of this protein is to act as a cell surface receptor and perform contact-mediated communication, via homophilic cadherin domain binding, and EGF-like cell adhesion - enabling G-protein coupled (Ca^{2+}-dependent) receptor-ligand interactions to occur (RefSeq, July 2008). The function of the CELSR2 flamingo cadherin protein is still unknown, but it is likely to play an important role in cell-cell signalling during the development of the nervous system (Takeichi 2007; Shima et al. 2007), and may also play a role in cell cycle control and cell proliferation (Takeichi 2007; Wu & Maniatis 1999; Vincent, Skaug & Scherer 2000).

PSRC1 (a proline-serine rich protein that contains a coiled-coil region, and six SH3-domain binding motifs PXXP, i.e. P=proline, X=any amino acid) encodes a cytoplasmic proline-rich protein, which according to human and mouse homologue studies is likely to be regulated by the transcription factor p53 (a well-known tumour suppressor), such that over expression of *PSRC1* will result in p53-mediated inhibition of cell growth (Hsieh, Lo & Wang 2002). In addition, the protein may function as a microtubule destabiliser by controlling spindle dynamics and mitotic progression by recruiting and regulating microtubule depolymerases, e.g. Kif2a (Jang et al. 2008). In contrast, another functional role of PSRC1 is to interact with MAPRE3 (microtubule-associated protein, RP/EB family, member 3) and activate the β-catenin pathway and increase expression of cyclin D1 (CCND1) ((Hsieh et al. 2007).

In the case of two other nearby genes *MYBPHL* (myosin binding protein H-like) and *SARS* (serine-tRNA ligase, cytoplasmic) there is little known about their biological function. Structurally, the MYBPHL gene encodes a protein with two immunoglobulin superfamily domains and a fibronectin III domain (RefSeq, July 2008); whereas SARS consists of two distinct domains, a catalytic core and an N-terminal involved in t-RNA binding. The role of encoded protein MYBPHL is unknown, whereas SARS catalyses the attachment of serine to tRNA to form tRNA (serine) in the cytoplasm, and belongs to the class II amino-acyl tRNA family (Rigler et al. 1970).

Finally, another nearby gene further downstream of *CELSR2*, is the gene *SORT1* that encodes for the multi-ligand type-1 receptor protein sortilin 1. The sortilin 1 receptor is a member of the vsp10p domain gene family. The vsp10p domain contains a propeptide that must be cleaved by furin (an endopeptidase) in the trans-Golgi network (TGN) to enable activation of the mature sortilin 1 protein (Munck Petersen et al. 1999). Sortilin 1 functions as a sorting receptor in the Golgi apparatus and as a clearance receptor at the cell surface. Sortilin 1 is required for protein transport from the Golgi apparatus to the lysosomes or endosomes. Examples include the sortilin 1 cell surface binding of lipoprotein lipase and its subsequently mediated endocytosis and degradation (Nielsen et al. 1999); and the action of sortilin 1 as a receptor for neurotensin, and its subsequent internalisation and trafficking to the TGN (Morinville et al. 2004). Sortilin 1 is also involved in nervous system development and maintenance due to its distinct role in pro-neurotrophin induced apoptosis. Pro-neurotrophin forms a death-signalling receptor complex with sortilin 1 and p75 neurotrophin receptor (Jansen et al. 2007). Yet another role for sortilin 1 is its requirement for the formation of GLUT4 (Glucose transporter 4) storage vesicles (the major insulin-response compartments) in adipocytes during adipogenesis (Hou & Pessin 2007).

In order to be able to identify which gene(s) are responsible for the CAD association via the identified intermediary phenotypes TC and LDL-C, it is important to first find if there are already any known expression quantitative trait loci (eQTL). Interestingly, a recent strong eQTL association was discovered at SNP rs599839 for three of the candidate genes described above – *CELSR2*, *PSRC1* and *SORT1* in a hepatic tissue specific gene expression microarray study involving a relatively large 427 subject human liver cohort (HLC) (Schadt et al. 2008). Specifically, this study showed in the HLC that *CELSR2* (P=4.31x10^{-23}), *PSRC1* (P=2.17x10^{-53}), and *SORT1* (P=1.52x10^{-56}) are all expressed at a significantly lower level in those subject carrying the rs599839 risk allele (A), as compared to those carriers of the minor allele (G) (Schadt et al. 2008). Intriguingly, on the basis of further liver expression studies using a mouse cross model (BXH/wt) specifically designed to study metabolic traits the authors show that *psrc1* gene expression has a strong positive correlation with plasma LDL-C, such that low *psrc1* expression is associated with low LDL-C; whereas *celsr2* and *sort1* gene expression have a strong negative correlation with plasma LDL-C levels, such that low *celsr2* and *sort1* expression is associated with high LDL-C levels. This gene expression data suggests that *CELSR2* and *SORT1* are the main candidates for increased LDL-C and CAD risk, and that *PSRC1* is perhaps protective (Schadt et al. 2008). Despite the use of a humanised mouse model for CAD via induced atherosclerosis, it may not necessarily be a good model for this locus, and as such this finding may offer no beneficial insight or simply indicate an added layer of complexity to the regulation of LDL-C at this locus. Further in depth investigation is required to better understand the regulation of these genes in relation to their influence on LDL-C regulation.

As a member of the Cardiogenics consortium, I was privy to, two further sets of pre-publication eQTL data relating to locus chr1p13.3. The first was lymphocyte gene expression data from the Caucasian German population-based control study MONICA (Monitoring of Trends and Determinations of Cardiovascular Disease) / KORA (Cooperative Research in the Region of Augsburg) F3 (n=190) that showed a significant association between rs599839 and *SORT1* (*P*=0.01), an additive trend towards significance for *PSRC1* (*P*=0.10), and no association with *CELSR2* (*P*=0.72), (Linsel-Nitschke et al. 2010). The second was monocyte and macrophage gene expression data from a specific Cardiogenics study that recruited Caucasian patients with CAD (n=459) and healthy controls (n=459) from five centres across Northern and Western Europe (Schunkert et al. 2011). Gene expression data for the rs599839 proxy SNP rs646776 showed in monocytes, a strong association with *PSRC1* (*P*=$7.7\text{x}10^{-21}$; ILMN_1671843), expression but no significance with *SORT1* (*P*=0.74; ILMN_1707077) and no expression with *CELSR2*; and in macrophages a strong association with *PSRC1* (*P*=$4.8\text{x}10^{-20}$; ILMN_1671843), moderate association with *SORT1* (*P*=$2.8\text{x}10^{-4}$; ILMN_1707077) and again no expression with *CELSR2* (Schunkert et al. 2011). Taken together the eQTL data from these different cell types suggest a key role for *PSRC1* and *SORT1,* and perhaps to a lesser degree *CELSR2* in CAD association through the intermediary risk factor LDL-C. Moreover, there is evidentiary support that the 1p13.3 locus mechanistically is influencing all three genes, and as such they all warrant further investigation.

The main limitation of the WTCCC CAD GWA study is the limited number of SNPs typed on the platform. Indeed, the Affymetrix 500K array platform only contains genotyping data on 19 SNPs (of which six are monomorphic in our North European population) within a 168kb region that includes the three candidate genes with eQTL data. It is clear that there is at least one other SNP in strong pairwise LD (i.e. r^2>0.8) with rs599839, because there are several other early GWA studies for lipids that have simultaneously discovered not only rs599839, but also another SNP in very strong association with LDL-C at this 1p13.3 locus, i.e. rs646776, a SNP located just beyond the 3'UTR of *CELSR2*. The two SNPs rs599839 and rs646776 have a pairwise LD score of r^2=0.89 (D'=1.0) in the HapMap Phase II (release 21a) CEU population (Utah residents with Northern and Western European ancestry from the CEPH collection) (Kathiresan et al. 2008; Sandhu et al. 2008).

3.1.1 Aims:

Hence the purpose of the experiments carried out in this chapter, were:

1. To refine (fine-map) the 1p13.3 locus identified by the rs599839 signal for LDL-C and CAD by genotyping additional SNPs to provide greater coverage.
2. To use *in silico* bioinformatics approaches to identify the most likely causal SNP(s)
3. To undertake functional experiments on the putative causal SNP(s)

3.2 Methods and Materials

3.2.1 Study populations

3.2.1.1 The GRAPHIC Study for locus 1p13.3 LDL-C fine-mapping

The GRAPHIC study includes 2037 subjects from 520 nuclear families that were recruited from the general population in Leicestershire, United Kingdom (UK). The details of the GRAPHIC recruitment and phenotyping were described in the previous chapter under the section *3.2 Methods and Materials*. For the purposes of this analysis the only cardiovascular phenotype to be considered will be LDL-C.

3.2.1.2 Metabochip Study for locus 1p13.3 CAD fine-mapping

The UK CAD samples to be genotyped in the Metabochip fine-mapping study include n=1926 (WTCCC-CAD1) case subjects (first genotyped in the original WTCCC CAD GWA study on the Affymetrix 500K array), and an additional sub-set of n=1303 CAD case subjects (WTCCC-CAD2). All UK CAD samples were sourced from the BHF-FHS (or lead BHF-FHS centre subsidiary CAD studies from Leeds and Leicester), and totalled n=3219 CAD subjects. The non-CAD controls subjects were all sourced from the 1958BC, with n=1500 coming from the original WTCCC CAD GWA study, and n=4246 additional control subjects (CARDIoGRAMplusC4D Consortium et al. 2013).

3.2.2 Genotyping and quality control

3.2.2.1 50K IBCv2 HumanCVD Beadchip

The advent of a custom-designed 50K SNP genotyping assay aimed at genotyping both common (MAF>0.02) and functional (MAF>0.01) SNPs for >2100 candidate genes and loci related to cardiovascular, inflammatory and metabolic pathways and phenotypes, known as the 50K IBC (ITMAT-Broad-CARe) HumanCVD BeadChip array (Illumina, Inc.), has enabled the possibility of cardiovascular disease candidate gene refinement (Keating et al. 2008). The *a priori* candidate gene selection was led by investigators from the Institute of Translational Medicine and Therapeutics (ITMAT) of the University of Pennsylvania, the Broad Institute and by the National Heart Lung Institute (NHBLI) supported candidate-gene Association resource (CARe). The candidate genes were prioritised in to three tiers for SNP selection: (i) genes and loci with a high likelihood of functional significance or shown to be associated with phenotypes of importance were tagged with a $r^2 \geq 0.8$ for HapMap populations with MAF>0.02 (Tier 1); (ii) loci that are potentially involved in cardiovascular phenotypes or well-known loci that require large numbers of tagging SNPs were selected with $r^2 \geq 0.5$ for MAF>0.05 (Tier 2); (iii) includes for the most part larger (>100kb) genes with lower interest; only non-synonymous and known functional variants with MAF>0.01 were selected (Tier 3). Additionally, the IBC array contains SNPs informative of ancestry (to adjust for population stratification), copy number variations (CNVs), and duplicates for assay quality control (QC).

The second version of the 50K IBC array (50K IBCv2) was released commercially with the name HumanCVD beadchip (Illumina, Inc. CA). The GRAPHIC Study was genotyped with this array to fine-map the LDL-C associated 1p13.3 locus, identified by SNP rs599839. The HumanCVD beadchip contains 94 SNPs that span 155kb of genomic DNA, within the 1p13.3 locus that includes four candidate genes: *CELSR2*, *PSRC1*, *MYBPHL* and *SORT1,* as well as, 5' flanking sequence upstream of *CELSR2* and *SORT1*, selected through tier 1 (i.e. $r^2 \geq 0.8$), tagging coverage (see above). This is a gain of 75 SNPs when compared to the 19 SNPs available on the Affymetrix 500K array, used for the original GWA study.

The GRAPHIC study genomic DNA samples (n=2037) were diluted to a normalised concentration of 50ng/µL using a NanoDrop ND-8000 UV-Vis Spectrophotometer (NanoDrop Technologies Inc.) and genotyped for ~50,000 SNPs using the 50K IBCv2 HumanCVD Beadchip at the Genomics Core facility (NUCLEUS) in the Department of Genetics, University of Leicester. The manufacturer's (Infinium® II assay technology) three-day protocol for amplification and hybridisation to the 50K IBCv2 HumanCVD Beadchip (Illumina, Inc. CA) was adhered to. In brief, approximately 200ng (4µL at 50ng/µL) of genomic DNA was added to a whole genome amplification reaction producing fragments of approximately 1.5–2kb in length. Enzymatic fragmentation, followed by purification, produces 200–600bp fragments for hybridization to the beadchips. Each bead contains many oligonucleotides to measure the presence of a single SNP allele, with approximately 30 technical replicates of each bead randomly distributed on the beadchip. During the chip's manufacturer QC performed by Illumina (Illumina Inc., CA), the location of the bead replicates are identified for each chip and distributed on a DVD (as *.dmap files) with the beadchip. Each beadchip contains wells allowing 12 samples to run simultaneously. The 50K IBCv2 HumanCVD Beadchips were scanned on an Illumina Beadstation 500 benchtop system (Illumina, Inc. CA).

Investigator analysis of genotyping and QC measures were performed using BeadStudio software 'Genotyping Module v3.2' that utilises GenCall software algorithms for clustering, calling and scoring genotypes (information on how to analyse the 50K IBCv2 HumanCVD Beadchip is available at this URL: ftp://ftp.illumina.com/BeadStudioUserGuides/ (Username = Guest, Password = illumina), and important supportive technotes, such as 'infinium genotyping data analysis' and 'TOPBOT technote' are available from within the BeadStudio software Illumina Portal). The full study dataset of 2037 GRAPHIC samples were used to generate a customised cluster file (*.egt filename) in BeadStudio to ensure good quality cluster plots. The quality of the cluster plots were visually inspected by two independent investigators (myself and a colleague).

3.2.2.2 Illumina custom-designed iSelect Metabochip

In August 2009, 217,697 SNPs were identified in follow-up to genome-wide meta-analyses for metabolic and cardiovascular disease traits and were submitted to the Illumina infinium custom-design iSelect BeadChip pipeline. A total of 196,725 SNPs passed QC and were manufactured for the Metabochip array. The collaborative partnership (including CARDIoGRAM) have named this custom array “Metabochip”. The utility of this Beadchip array is that it has incorporated 6,222 CAD replication SNPs that have been LD-pruned (based on HapMap phase2 data) at r^2<0.2 for SNPs at a nominal significance of *P*<0.01 from CARDIoGRAM, and a further 20,876 refinement SNPs for 30 genome-wide significant (P<5x10^{-8}) associated SNPs (Schunkert et al. 2011; Voight et al. 2012). For the purposes of fine-mapping the CAD associated chr1p13.3 locus identified by rs599839, there are n=438 SNPs on the Metabochip, tagged at a pairwise LD $r^2 \geq 0.8$ threshold across a genomic region that includes the genes *CELSR2*, *PSRC1*, *MYBPHL*, *SORT1,* plus a 10kb 5' flanking region upstream of *CELSR2* and *SORT1*, spanning a total genomic region of ~168kb. The number of SNPs fine-mapping the four candidate genes within the 1p13.3 locus on the Metabochip array, is a gain of 419 SNPs when compared to the 19 SNPs that were typed on the original WTCCC Affymetrix 500K array

The Metabochip array was genotyped at the Sanger Institute, Genome Campus (Hinxton, Cambs, UK). The Sanger Genome Campus pipe-line involves customer shipment of ~2-3µg of Nanodrop spectrophotometer normalised genomic DNA at a final concentration of between 70-100ng/µL in sealed ABgene (AB-0765) 0.8mL microplates on dry-ice to the Sanger Institute Genome Campus Logistics Facility. Sanger DNA QC procedures are adhered to before samples are made available for Illumina HD Infinium Beadchip genotyping, these include normalising the genomic DNA to 50ng/µL concentration via pico-green readings, running a Sequenom quality scoring array to validate gender and ethnicity, and finally an agarose gel run to check for DNA degradation. Genomic DNA samples for the BHF-FHS and subsidiary CAD cohorts were shipped from Leicester or Leeds (UK, CARDIoGRAM centres). The 1958BC samples were supplied to the Sanger Centre after an application approval for use by the Metabochip consortium.

3.2.3 Statistical analysis

Exclusion filters were applied to the 50K IBCv2 HumanCVD Beadchip GRAPHIC study genotyping data to remove unreliable samples or SNPs. SNPs were excluded if they were admixture or ancestral controls called Ancestry Informative Markers (AIMs), the full list 'HumanCVD_AIMs_SNP_list.xls', is available at URL: ftp://ftp.illumina.com/Whole%20Genome%20Genotyping%20Files/ArchivedHumanProducts/HumanCVDSNP55/ (username = guest, password = illumina). SNPs were also excluded if the genotype call rate was <90%, the Hardy-Weinberg equilibrium threshold was <0.0001 (using a χ^2 test), and the MAF was <0.01 as defined by the BeadStudio software. Further samples were excluded if samples had ambiguous genotypes according to gender or Mendelian errors. SNPs and samples passing these thresholds were analysed using the statistical software Stata11/IC (StataCorp LP) for associations with LDL-C with adjustment for age, age^2 (to account for the two generation family structure of GRAPHIC) and gender, as well as lipid lowering, (Asselbergs et al. 2012) under an additive model, using generalised estimating equations (GEE) with exchangeable correlation structure to account for familial correlations (Burton, Gurrin & Sly 1998). Genetic estimated effect sizes are shown as: β-coefficients per minor allele copy of each SNP with their standard error (SE) and 95% confidence intervals (CI).

The same exclusion criteria were used for Metabochip, except 3 different clustering and calling software algorithms: ILLUMINUS (Teo et al. 2007), GenoSNP (Giannoulatou et al. 2008) and GenCall (http://www.illumina.com/Documents/products/technotes/technote_gencall_data_analysis_software.pdf), were used, to help call rare variants more accurately. It should be mentioned here that all clustering algorithms are considered comparable for common alleles MAF ≥0.05. The results reported here are using ILLUMINUS (Teo et al. 2007) and this data analysis is independent of and pre-dates the CARDIoGRAMplusC4D analyses (CARDIoGRAMplusC4D Consortium et al. 2013). Association analysis for CAD Metabochip genotypes were undertaken in PLINK (http://pngu.mgh.harvard.edu/~purcell/plink/) (Purcell et al. 2007) by logistic regression adjusting for age and sex, under an additive genetic model. Effect sizes are given as odds ratios (OR) and 95% confidence intervals (CI). To account for multiple testing, a conservative Bonferroni corrected *P*-value (0.05 / no. of SNPs being tested) was considered to be significant and *P*-values of between $0.01<P<0.05$ were considered to be suggestive of nominal significance level associations.

3.2.4 In silico bioinformatics methods

3.2.4.1 50K (IBCv2), HumanCVD Beadchip annotation

All SNPs on the 50K IBCv2 HumanCVD Beadchip were annotated using publicly available databases including the IBC array design information at http://bmic.upenn.edu/cvdsnp/ [N.B. this website no longer exists], the dbSNP database build 129 (NCBI build 36.3) at: http://www.ncbi.nlm.nih.gov/projects/SNP/ dbSNP.cgi? list=rslist and the SNP Function Portal at: http://brainarray.mbni.med.umich.edu/Brainarray/Database/SearchSNP/snpfunc.aspx (Wang et al. 2006).

3.2.4.2 In silico bioinformatics assessment of locus 1p13.3 candidate SNPs

Appropriate *in silico* bioinformatics were performed on the LDL-C and CAD fine-mapping analyses candidate SNP findings, using *in silico* web-based tools such as the ENCyclopaedia Of DNA Elements (ENCODE) project, as well as putative transcription factor binding site (TFBS) and MicroRNA target site (miRNA-TS) prediction tools.

3.3 Results

3.3.1 Subjects

The characteristics of the GRAPHIC study subjects included in the analysis are shown in **Table 3.1**. Participants were aged between 18 and 61 years. Approximately 6.7% and 2% of participants were taking antihypertensive and lipid lowering medication, respectively. The characteristics of the CAD case-control samples analysed using the Metabochip array are shown in **Table 3.2**. Only gender and age were available for the 1958BC control populations.

Table 3.1. Characteristics of the GRAPHIC population

Variable	Fathers n=512	Mothers n=512	Sons n=509	Daughters n=491	All Subjects n=2024
Age (years): mean (SD)	53.76 (4.27)	51.84 (4.36)	25.04 (5.05)	25.93 (5.42)	39.30 (14.50)
Body Mass Index (kg/m^2): mean (SD)	27.82 (3.95)	27.07 (4.55)	24.92 (4.14)	24.56 (4.95)	26.11 (4.62)
Waist : hip ratio: mean (SD)	0.93 (0.068)	0.81 (0.064)	0.86 (0.065)	0.78 (0.067)	0.85 (0.087)
Clinic SBP (mm Hg): mean (SD)	140.18 (20.16)	130.26 (19.92)	127.83 (13.14)	113.88 (11.96)	128.18 (19.18)
Mean 24hr SBP (mm Hg): mean (SD)	125.87 (12.59)	118.10 (12.48)	120.83 (8.21)	112.81 (7.23)	119.47(11.45)
Clinic DBP (mm Hg): mean (SD)	87.27 (11.51)	81.81 (10.95)	75.95 (9.68)	73.61 (8.66)	79.72 (11.56)
Mean 24hr DBP (mm Hg): mean (SD)	78.66 (7.83)	72.53 (8.00)	69.25 (6.56)	67.94 (5.20)	72.15(8.14)
Total (mmol/L): mean (SD)	5.76 (1.05)	5.91 (1.02)	4.64 (0.92)	4.67 (0.87)	5.26 (1.14)
LDL cholesterol (mmol/L): mean (SD)	3.31 (0.72)	3.28 (0.75)	2.52 (0.68)	2.5 (0.59)	2.91 (0.79)
HDL cholesterol (mmol/L): mean (SD)	1.29 (0.3)	1.62 (0.38)	1.31 (0.27)	1.47 (0.34)	1.42 (0.35)
Estimated GFR (mL/min/1.73m^2): mean (SD)	78.71 (11.71)	73.61 (10.22)	96.26 (13.08)	87.16 (12.97)	83.87 (14.81)
C-reactive protein (mg/L): mean (SD)	0.28 (0.95)	0.26 (0.39)	0.16 (0.37)	0.26 (0.43)	0.24 (0.59)
Current Antihypertensive Treatment: n (%)	78 (15.2)	53 (10.3)	3 (0.6)	1 (0.2)	135 (6.7)
Current Lipid Lowering Treatment: n (%)	28 (5.47)	12 (2.34)	0 (0.0)	1 (0.2)	41 (2.03)
Ever Smoker %; current smoker %	56.45; 14.8	43.95; 12.4	39.69; 28.7	38.29; 24.4	44.66; 20.0

Legend. All variable data are provided as means and standard deviations or counts and percentages; covariates: total cholesterol (TC), high density lipoprotein cholesterol (HDL-C) and low density lipoprotein cholesterol (LDL-C) were measured by enzymatic clinical chemistry; all lipid and blood pressure traits have been adjusted for lipid lowering or anti-hypertensive medication (Asselbergs et al. 2012); (Levy et al. 2000; Tobin et al. 2005). This table for the GRAPHIC study characteristics **Table 3.1** differs from the previous chapter **Table 2.3** in subject numbers, because it includes LDL-C measurements that were not originally measured, and it only includes subjects with complete characteristic measurements, and complete HumanCVD 50K Beadchip genotype data.

Table 3.2. Characteristics of WTCCC-CAD2 cases and 1958BC controls

Variable	CAD Cases	Controls
Number	3188	6000
Age at first event (years): mean ± sd	51.3±8.7	n/a
Age at recruitment (years): mean ± sd	58.5±8.7	44±0.0
Ethnicity	Caucasian	n/a
Gender (% Male)	2573 (79.8)	3000 (50)
Myocardial infarction (%)	2547 (79.9)	n/a
Diabetes mellitus (%)	344 (10.8)	n/a
Hypertension (%)	1320 (41.4)	n/a
Hyperlipidaemia (%)	2459 (77.1)	n/a
Body Mass Index (kg/m^2): mean ± sd	28.1±4.6	n/a
Ever Smoked (%)	2429 (76.2)	n/a

Legend. All variable data are provided as means and standard deviations or counts and percentages. Metabochip includes CAD samples from UK based cohorts - Family Heart Study (BHF-FHS) (Samani et al. 2005), GRACE study (Alfakih et al. 2007), Space Rocket Trial (Hall et al. 2009), Opera (Pearson et al. 2011) and PRAMIS study (Brouilette et al. 2003); Controls are from the 1958 Birth Cohort (http://www.b58cgene.sgul.ac.uk/followup.php); Hypertension (systolic blood pressure ≥140 or diastolic blood pressure ≥90 or receiving anti-hypertensive treatment); Hyperlipidemia (TC ≥200 mg/dL, or LDL-C ≥130 mg/dL, or receiving lipid lowering treatment); Body mass index (>30 kg/m^2).

3.3.2 Genotyping quality control for 50K IBCv2 HumanCVD Beadchip

The SNP QC exclusion process of the GRAPHIC genotyped 50K IBCv2 HumanCVD Beadchip, resulted in the following: Of 49,094 genotyped SNPs, 1775 were excluded because they belonged to admixture and ancestry informative control markers. Duplicate SNPs and those identified as copy number variants (n=106) were also removed. Additional filters resulted in exclusion of 13,636 further SNPs because of a low MAF (<0.01) (n=12,443), location on sex chromosomes (n=638), low (<90%) genotyping call rate (n=424), >10 Mendelian errors suggestive of poor genotyping quality (which was possible to check because of the family-based structure of GRAPHIC) (n=20), Hardy-Weinberg equilibrium violation (p<0.0001) (n=107), and the lack of an unambiguous reference SNP (rs) identification number (n=4). This left 33,577 SNPs that passed all exclusion filters available for analysis with n=2024 GRAPHIC study subjects. We annotated 32,939 SNPs to 3036 candidate genes and the remaining 638 SNPs were mapped to hypothetical genes.

After QC filtering and annotation, the number of SNPs available for the chr1p13.3 locus fine-mapping that span a 155kb genomic region were reduced from n=94 to n=75.

3.3.3 Chr1p13.3 fine-mapping SNP association with lipid traits in GRAPHIC

The GRAPHIC study 50K HumanCVD array distribution of *P*-value LDL-C associations adjusted for age and gender for all chr1p13.3 locus refinement SNPs (n=75) spanning four candidate genes (*CELSR2, PSRC1, MYBPHL, SORT1*) are shown in **Table 3.3** and **Figure 3.2**. The key finding of the GRAPHIC fine-mapping study was that SNP rs629301 and three other almost perfect proxy (i.e. $r^2 \approx 1$) SNPs: rs7528419, rs646776 and rs12740374 were more strongly associated with LDL-C than the CAD and LDL-C associated discovery SNP rs599839. Indeed, the lead fine-mapping SNP rs629301 risk allele (A) (allele frequency 0.78) was associated with a 0.156mmol/L (95% CI: 0.103-0.208mmol/L) increase in plasma LDL-C level per allele copy (P=5.36x10^{-9}); whereas SNP rs599839 although still strongly associated, had a slightly reduced level of significance, with a 0.145mmol/L (95% CI: 0.092-0.198mmol/L) increase in plasma LDL-C level per risk allele (A) copy (P=9.97x10^{-8}).

Interestingly, when adjusting for multiple testing via a Bonferroni correction of P<6.66x10^{-4} (0.05/75), there are 15 fine-mapped SNPs spanning 23,445bp (from within an intronic location of *CELSR2* to a location just beyond the 3'UTR of *PSRC1*), which remain significant for LDL-C association (N.B. 30 SNPs over a range of 122,256bp showed at least nominal significance at 0.01<P<0.05).

Table 3.3 SNPs within the 1p13.3 locus showing association with LDL-Cholesterol in the GRAPHIC study.

Chr	Location (bp)	SNP	Risk/ Minor Allele	MAF	HWE *P*-val	Gene	Function Class	β-coefficient (SE)	Lower 95% CI	Upper 95% CI	*P*-val	LD (r^2)
1	109619829	rs629301	A/C	0.22	0.14	CELSR2	3'UTR	-0.156 (0.027)	-0.208	-0.103	5.36x10^{-09}	1.000
1	109618715	rs7528419	A/G	0.22	0.14	CELSR2	3'UTR	-0.154 (0.027)	-0.207	-0.102	6.84x10^{-09}	0.997
1	109620053	rs646776	A/G	0.22	0.20	CELSR2	3'	-0.154 (0.027)	-0.206	-0.102	7.44x10^{-09}	0.995
1	109619113	rs12740374	C/A	0.22	0.14	CELSR2	3'UTR	-0.153 (0.027)	-0.205	-0.101	9.34x10^{-09}	0.995
1	109616775	rs611917	A/G	0.32	0.07	CELSR2	intron	-0.139 (0.025)	-0.188	-0.089	3.47x10^{-08}	0.563
1	109623034	rs602633	C/A	0.21	0.11	PSRC1	3'	-0.149 (0.028)	-0.204	-0.095	8.20x10^{-08}	0.941
1	109622830	rs583104	A/C	0.22	0.07	PSRC1	3'	-0.145 (0.027)	-0.199	-0.092	9.18x10^{-08}	0.958
1	109623689	rs599839	A/G	0.22	0.06	PSRC1	3'	-0.145 (0.027)	-0.198	-0.092	9.97x10^{-08}	0.955
1	109608806	rs6657811	A/T	0.13	0.17	CELSR2	intron	-0.158 (0.033)	-0.222	-0.094	1.25x10^{-06}	0.480
1	109616403	rs4970834	G/A	0.18	0.83	CELSR2	intron	-0.126 (0.028)	-0.182	-0.071	8.37x10^{-06}	0.688
1	109600244	rs10858082	A/G	0.44	0.01	CELSR2	intron	-0.107 (0.024)	-0.154	-0.059	1.10x10^{-05}	0.297
1	109608357	rs6698843	A/A	0.47	0.06	CELSR2	exon	0.094 (0.024)	0.046	0.142	1.19x10^{-04}	0.207
1	109606169	rs4970833	A/A	0.47	0.05	CELSR2	intron	0.093 (0.024)	0.045	0.141	1.37x10^{-04}	0.208
1	109608622	rs6689614	A/A	0.47	0.04	CELSR2	exon	0.093 (0.024)	0.045	0.141	1.38x10^{-04}	0.209
1	109622442	rs17035949	T/G	0.04	0.44	PSRC1	3'	-0.180 (0.051)	-0.280	-0.080	4.08x10^{-04}	0.192
1	109633806	rs655246	G/A	0.46	0.90	MYBPHL	3'	-0.081 (0.024)	-0.128	-0.033	9.39x10^{-04}	0.151
1	109591318	rs585362	A/G	0.16	0.65	CELSR2	5'	-0.101 (0.031)	-0.162	-0.039	1.33x10^{-03}	0.193
1	109597131	rs437444	G/A	0.04	0.39	CELSR2	exon	-0.183 (0.057)	-0.294	-0.071	1.37x10^{-03}	0.121
1	109596549	rs413380	G/A	0.04	0.39	CELSR2	exon	-0.182 (0.057)	-0.294	-0.071	1.39x10^{-03}	0.121
1	109627659	rs657420	A/G	0.47	0.90	PSRC1	5'	-0.077 (0.024)	-0.125	-0.030	1.45x10^{-03}	0.179
1	109713574	rs17646665	A/G	0.06	0.79	SORT1	intron	-0.132 (0.043)	-0.216	-0.049	1.97x10^{-03}	0.074
1	109688714	rs4970843	G/G	0.47	0.80	SORT1	intron	0.057 (0.024)	0.011	0.103	0.02	0.108
1	109652649	rs413582	A/A	0.48	0.95	MYBPHL	5'	0.056 (0.024)	0.010	0.103	0.02	0.109
1	109600326	rs626387	G/A	0.14	1.00	CELSR2	intron	-0.077 (0.032)	-0.141	-0.013	0.02	0.003
1	109640441	rs629001	A/G	0.07	0.81	MYBPHL	exon	-0.097 (0.044)	-0.182	-0.011	0.03	0.101

1	109615242	rs17035665	G/A	0.19	0.16	CELSR2	intron	-0.063 (0.030)	-0.122	-0.004	0.04	0.146
1	109637280	rs17645143	A/G	0.29	0.88	MYBPHL	intron	-0.052 (0.025)	-0.101	-0.003	0.04	0.003
1	109630036	rs688498	T/A	0.29	0.94	PSRC1	5'	-0.051 (0.025)	-0.101	-0.002	0.04	0.002
1	109641419	rs7515901	G/A	0.16	1.00	MYBPHL	intron	-0.062 (0.031)	-0.124	-0.001	0.05	0.067
1	109636461	rs17645031	G/A	0.08	0.23	MYBPHL	3'	-0.083 (0.042)	-0.166	-0.001	0.05	0.210
1	109636686	rs41306199	G/A	0.08	0.30	MYBPHL	3'UTR	-0.077 (0.043)	-0.161	0.007	0.07	0.203
1	109639787	rs12127701	A/G	0.07	0.62	MYBPHL	intron	-0.078 (0.045)	-0.167	0.011	0.08	0.094
1	109624032	rs14000	A/G	0.11	0.53	PSRC1	3'UTR	-0.047 (0.038)	-0.122	0.028	0.22	0.026
1	109619793	rs658435	G/A	0.11	0.32	CELSR2	3'UTR	-0.044 (0.039)	-0.120	0.032	0.26	0.023
1	109657829	rs464218	A/G	0.46	0.71	SORT1	3'	-0.025 (0.024)	-0.072	0.022	0.29	0.040
1	109607836	rs653635	A/G	0.10	0.87	CELSR2	exon	-0.042 (0.040)	-0.119	0.036	0.29	0.025
1	109623927	rs10410	G/A	0.11	0.26	PSRC1	3'UTR	-0.039 (0.039)	-0.114	0.037	0.32	0.024
1	109640027	rs630822	C/G	0.001	1.00	MYBPHL	intron	-0.223 (0.232)	-0.679	0.233	0.34	0.002
1	109640956	rs1278285	G/A	0.001	1.00	MYBPHL	exon	-0.223 (0.233)	-0.679	0.233	0.34	0.002
1	109640117	rs630395	A/G	0.001	1.00	MYBPHL	intron	-0.223 (0.233)	-0.679	0.233	0.34	0.002
1	109725200	rs12037569	C/A	0.16	0.64	SORT1	intron	-0.030 (0.033)	-0.094	0.034	0.36	0.001
1	109683773	rs10858084	T/A	0.28	0.64	SORT1	intron	-0.023 (0.026)	-0.073	0.027	0.37	0.036
1	109590236	rs17035443	G/A	0.18	0.92	CELSR2	5'	-0.025 (0.028)	-0.080	0.030	0.38	0.033
1	109673310	rs10745352	G/A	0.28	0.64	SORT1	intron	-0.022 (0.025)	-0.072	0.028	0.39	0.035
1	109612504	rs17035630	A/A	0.12	0.47	CELSR2	intron	0.027 (0.032)	-0.036	0.090	0.40	0.028
1	109602150	rs655334	A/T	0.10	0.30	CELSR2	intron	-0.033 (0.040)	-0.111	0.044	0.40	0.026
1	109672382	rs6695482	G/C	0.28	0.64	SORT1	intron	-0.020 (0.026)	-0.071	0.030	0.42	0.036
1	109698123	rs4970751	C/A	0.28	0.64	SORT1	intron	-0.020 (0.026)	-0.070	0.030	0.43	0.035
1	109737101	rs10858091	G/A	0.28	0.64	SORT1	intron	-0.020 (0.025)	-0.070	0.030	0.43	0.035
1	109745416	rs10858092	A/G	0.28	0.64	SORT1	5'	-0.020 (0.025)	-0.070	0.030	0.43	0.035
1	109698626	rs11142	G/A	0.28	0.64	SORT1	exon	-0.020 (0.025)	-0.070	0.030	0.43	0.035
1	109659642	rs370088	G/A	0.28	0.64	SORT1	intron	-0.020 (0.025)	-0.070	0.030	0.43	0.035
1	109736950	rs3768494	G/A	0.28	0.64	SORT1	intron	-0.020 (0.025)	-0.070	0.030	0.43	0.035
1	109667753	rs3768497	G/A	0.28	0.64	SORT1	intron	-0.020 (0.025)	-0.070	0.030	0.43	0.035

1	109732375	rs4970752	T/C	0.28	0.64	SORT1	intron	-0.020 (0.025)	-0.070	0.030	0.43	0.035
1	109745601	rs7518013	G/A	0.28	0.64	SORT1	5'	-0.020 (0.025)	-0.070	0.030	0.43	0.035
1	109661667	rs3853498	G/A	0.27	0.64	SORT1	intron	-0.020 (0.026)	-0.070	0.030	0.43	0.036
1	109665792	rs3853500	G/A	0.27	0.64	SORT1	intron	-0.020 (0.026)	-0.070	0.030	0.43	0.036
1	109686298	rs2228604	G/T	0.27	0.58	SORT1	exon	-0.020 (0.026)	-0.070	0.030	0.44	0.037
1	109742656	rs1880670	A/G	0.28	0.64	SORT1	5'	-0.020 (0.026)	-0.070	0.030	0.44	0.035
1	109652650	rs444387	G/A	0.27	0.64	SORT1	3'	-0.020 (0.026)	-0.070	0.030	0.44	0.036
1	109612067	rs2281894	C/A	0.19	0.31	CELSR2	exon	-0.020 (0.027)	-0.072	0.032	0.44	0.022
1	109679029	rs10745353	G/A	0.27	0.69	SORT1	intron	-0.020 (0.026)	-0.070	0.030	0.44	0.036
1	109682244	rs11102972	A/G	0.22	1.00	SORT1	intron	-0.020 (0.028)	-0.074	0.035	0.47	0.001
1	109694733	rs1278664	C/A	0.22	1.00	SORT1	intron	-0.020 (0.028)	-0.074	0.035	0.47	0.001
1	109646055	rs652651	G/A	0.27	0.64	MYBPHL	intron	-0.018 (0.026)	-0.068	0.032	0.48	0.037
1	109650249	rs592107	A/T	0.27	0.69	MYBPHL	intron	-0.017 (0.026)	-0.068	0.033	0.50	0.037
1	109711536	rs11581665	G/A	0.14	0.61	SORT1	intron	-0.020 (0.033)	-0.085	0.045	0.54	0.025
1	109645298	rs11583969	A/A	0.07	0.62	MYBPHL	intron	0.029 (0.049)	-0.068	0.126	0.56	0.004
1	109696216	rs7536292	A/G	0.19	0.76	SORT1	intron	-0.016 (0.031)	-0.077	0.045	0.61	0.002
1	109692656	rs10127790	T/C	0.28	0.70	SORT1	intron	-0.011 (0.025)	-0.060	0.039	0.68	0.032
1	109659338	rs17585355	A/C	0.06	0.38	SORT1	intron	-0.017 (0.048)	-0.111	0.076	0.71	0.016
1	109641261	rs3850615	C/A	0.07	0.43	MYBPHL	exon	-0.016 (0.045)	-0.104	0.071	0.72	0.016
1	109692946	rs3879448	G/C	0.28	0.64	SORT1	intron	-0.008 (0.025)	-0.057	0.042	0.76	0.032
1	109721048	rs17646731	G/A	0.05	0.76	SORT1	intron	-0.008 (0.050)	-0.106	0.089	0.86	0.095

Legend. Chr – chromosome; SNP – single nucleotide polymorphism; Location – Build 36.3; MAF (Parents) – minor allele frequency; HWE *P*-val (Parents) – statistical significance of Hardy-Weinberg equilibrium; β-coefficient – estimated quantitative effect of each SNP minor allele copy on LDL-C adjusted for age, age^2, sex and lipid lowering medication (Asselbergs et al. 2012); a negative β-estimate indicates lower LDL-C (mmol/L) in carriers of the minor allele, but indicates increased LDL-C (mmol/L) in carriers of the risk allele; SE – standard error; *P*-val – statistical significance of association. The lead LDL-C associated SNP rs629301 is highlighted in bold red text and the CAD GWA study discovery SNP rs599839 is highlighted in normal red text.

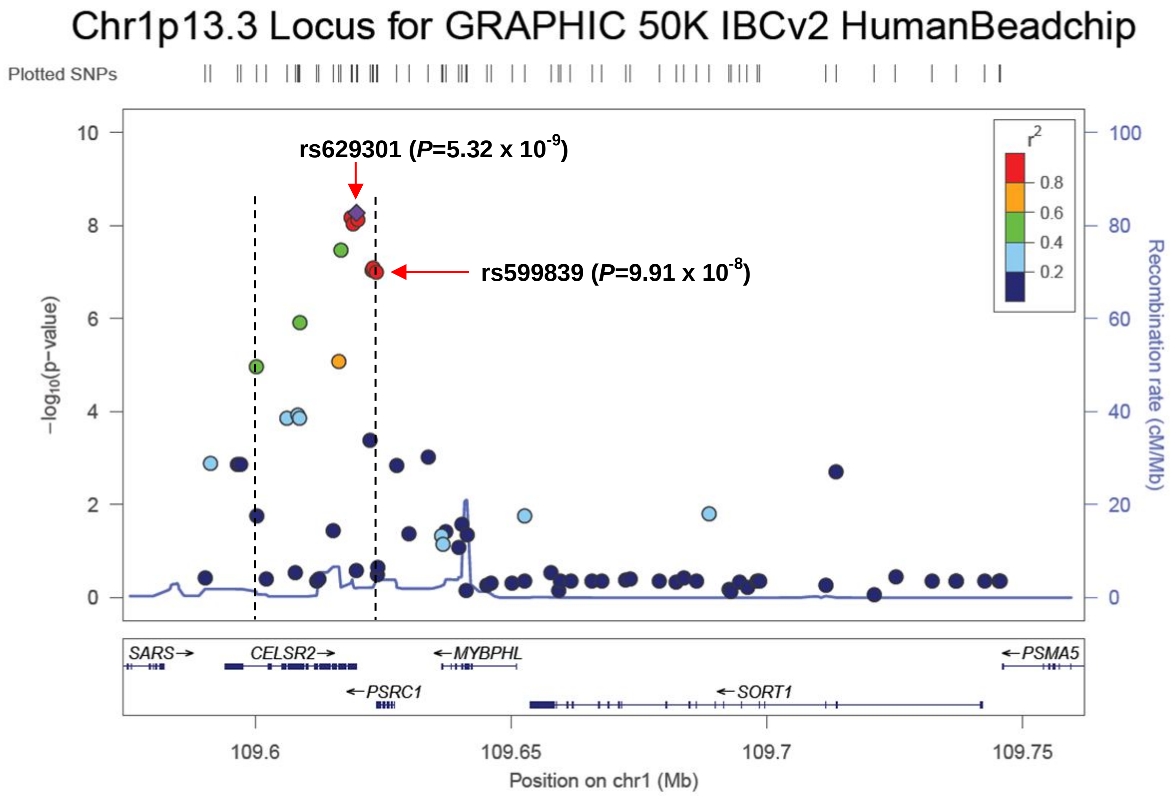


Figure 3.2. Regional association plot at locus chr1p13.3 for LDL-cholesterol in the GRAPHIC study on the 50K IBCv2 HumanCVD Beadchip**.**

Legend. The associations for the individual SNPs shown in **Table 3.3** are plotted as –$\log_{10}$ *P*-values on the Y-axis (left-hand-side) versus chromosome base pair position on the X-axis (Recombination rate (hotspots) is also plotted on the Y-axis (right-hand-side), shown in light-blue). The purple diamond plots the most significant SNP rs629301 for association with LDL-C. A colour-coded scale of pairwise LD r^2 scores between the lead SNP and 74 other plotted SNPs is shown for the locus region. The vertical dotted-lines indicate the distance over which the lead SNP rs629301 has a pairwise LD of $r^2 \geq 0.5$. The X-axis spans the SNP locations over 155,365 bp. The gene location and strand direction are shown with arrows. The four candidate genes are: *CELSR2* (cadherin, EGF LAG seven-pass G-type receptor 2), *PSRC1* (proline/serine-rich coiled-coil 1), *MYBPHL* (myosin binding protein H-like), *SORT1* (sortilin 1). (**Figure 3.2.** is generated by the web tool LocusZoom, hosted via the Center for Statistical Genetics, University of Michigan at https://statgen.sph.umich.edu/locuszoom/genform.php?type=yourdata; Version 1.1: June 2011 (Pruim et al. 2010)).

In particular, **Table 3.3** and **Figure 3.2** show that there are now eight SNPs with a *P*-value for LDL-C association of $P<1x10^{-7}$, and that these SNPs are highly correlated. There are seven SNPs in very strong LD ($r^2>0.94$), and one SNP in moderately strong LD ($r^2=0.56$). In situations such as this, it is useful to perform a *conditional* analysis, where you adjust for the lead SNP (i.e. rs629301) in the model. This is to test whether any of the additional SNPs also associated with LDL-C are independent of the lead SNP, or whether it is as a result of simply being in correlation with the lead SNP (i.e. in LD). Thus a conditional analysis was performed to see if the observed association is an independent signal. To this end, **Table 3.4**, provides us with the evidence that there is only one independent signal for LDL-C, because the *P*-value association disappears for all fifteen of the Bonferroni corrected significant SNPs, when conditioned on rs629301 and the original CAD/LDL-C lead SNP rs599839. The only potential for additional independent signals are seen at a nominal $0.01<P<0.05$ threshold, where the unconditioned analysis showed no association with LDL-C; and two potentially co-dependent SNPs (i.e. where the association has not dropped as much as one might expect when conditioning on the lead SNP) - rs611917 (*CELSR2* intronic) and rs17646665 (*SORT1* intronic). However, the take home message from the conditional analysis is that the most likely SNPs to be causally driving the LDL-C association in the GRAPHIC study are one of rs629301, rs7528419, rs12740374 and rs646776, within or just beyond the 3'UTR of *CELSR2*.

Table 3.4. Conditional analysis of lead SNPs rs629301 and rs599839 within the 1p13.3 locus showing association with LDL-Cholesterol in the GRAPHIC study

SNP	Gene	Func. Class	*P*-val GEE	LD (r^2) with rs629301	*P*-val Conditioned on rs629301	Comment for Conditioning on rs629301	LD (r^2) with rs599839	*P*-val Conditioned on rs599839	Comment for Conditioning on rs599839
rs629301	CELSR2	3'UTR	5.36x10^{-09}	1.000	9.68x10^{-09}	Conditioned SNP	0.955	0.027	Strong LD Sig.↓
rs7528419	CELSR2	3'UTR	6.84x10^{-09}	0.997	NA	Multi-collinearity	0.952	0.065	Strong LD Sig.↓
rs646776	CELSR2	3'	7.44x10^{-09}	0.995	0.37	Strong LD Sig.↓	0.960	0.059	Strong LD Sig.↓
rs12740374	CELSR2	3'UTR	9.34x10^{-09}	0.995	NA	Multi-collinearity	0.950	0.115	Strong LD Sig.↓
rs611917	CELSR2	intron	3.47x10^{-08}	0.563	0.03	Moderate LD Sig.↓	0.542	5.49x10^{-03}	Moderate LD Sig.↓
rs602633	PSRC1	3'	8.20x10^{-08}	0.941	0.74	Strong LD Sig.↓	0.947	0.253	Strong LD Sig.↓
rs583104	PSRC1	3'	9.18x10^{-08}	0.958	0.23	Strong LD Sig.↓	0.997	N	Multi-collinearity
rs599839	PSRC1	3'	9.97x10^{-08}	0.955	0.33	Strong LD Sig.↓	1.000	4.32x10^{-07}	Conditioned SNP
rs6657811	CELSR2	intron	1.25x10^{-06}	0.480	0.14	Moderate LD Sig.↓	0.489	0.052	Moderate LD Sig.↓
rs4970834	CELSR2	intron	8.37x10^{-06}	0.688	0.60	Moderate LD Sig.↓	0.679	0.823	Moderate LD Sig.↓
rs10858082	CELSR2	intron	1.10x10^{-05}	0.297	0.04	Weak LD Sig.↓	0.295	0.031	Weak LD Sig.↓
rs6698843	CELSR2	exon	1.19x10^{-04}	0.207	0.09	Weak LD Sig.↓	0.209	0.048	Weak LD Sig.↓
rs4970833	CELSR2	intron	1.37x10^{-04}	0.208	0.09	Weak LD Sig.↓	0.212	0.055	Weak LD Sig.↓
rs6689614	CELSR2	exon	1.38x10^{-04}	0.209	0.10	Weak LD Sig.↓	0.212	0.055	Weak LD Sig.↓
rs17035949	PSRC1	3'	4.08x10^{-04}	0.192	0.41		0.194	0.258	
rs655246	MYBPHL	3'	9.39x10^{-04}	0.151	0.17		0.155	0.110	
rs585362	CELSR2	5'	1.33x10^{-03}	0.193	0.31		0.188	0.186	
rs437444	CELSR2	exon	1.37x10^{-03}	0.121	0.24		0.138	0.214	
rs413380	CELSR2	exon	1.39x10^{-03}	0.121	0.24		0.138	0.215	
rs657420	PSRC1	5'	1.45x10^{-03}	0.179	0.24		0.188	0.161	
rs17646665	SORT1	intron	1.97x10^{-03}	0.074	0.02	Co-dependent Sig.↓	0.078	9.93x10^{-03}	Co-dependent Sig.↓
rs4970843	SORT1	intron	0.02	0.108	0.27		0.110	0.276	
rs413582	MYBPHL	5'	0.02	0.109	0.29		0.111	0.302	
rs626387	CELSR2	intron	0.02	0.003	0.05	Co-dependent Sig.↓	0.005	0.037	Co-dependent Sig.↓

rs629001	MYBPHL	exon	0.03	0.101	0.40		0.108	0.322	
rs17035665	CELSR2	intron	0.04	0.146	0.91		0.143	0.653	
rs17645143	MYBPHL	intron	0.04	0.003	0.02	Independent, P<0.05	0.002	0.011	Independent P<0.05
rs688498	PSRC1	5'	0.04	0.002	0.02	Independent, P<0.05	0.002	0.014	Independent P<0.05
rs7515901	MYBPHL	intron	0.05	0.067	0.84		0.070	0.831	
rs17645031	MYBPHL	3'	0.05	0.210	0.56	Weak LD Sig.↓	0.202	0.604	Weak LD Sig.↓
rs41306199	MYBPHL	3'UTR	0.07	0.203	0.47	Weak LD Sig.↓	0.196	0.490	Weak LD Sig.↓
rs12127701	MYBPHL	intron	0.08	0.094	0.67		0.101	0.596	
rs14000	PSRC1	3'UTR	0.22	0.026	0.03	Independent, P<0.05	0.026	0.022	Independent P<0.05
rs658435	CELSR2	3'UTR	0.26	0.023	0.04	Independent, P<0.05	0.022	0.031	Independent P<0.05
rs464218	SORT1	3'	0.29	0.040	0.86		0.040	0.948	
rs653635	CELSR2	exon	0.29	0.025	0.05	Independent, P=0.05	0.024	0.043	Independent P<0.05
rs10410	PSRC1	3'UTR	0.32	0.024	0.05	Independent, P=0.05	0.024	0.040	Independent P<0.05

Legend. Conditioned SNP text high-lighted in red, SNPs in strong LD (r^2>0.8) with diminished *P*-val association (or Multi-collinearity) are highlighted in yellow, SNPs with moderate LD (0.5<r^2<0.8) with diminished *P*-val association are highlighted in amber; SNPs in weak LD (0.2<r^2<0.5) with diminished *P*-val association are highlighted in green; SNPs showing co-dependent *P*-val association are highlighted in purple (i.e. only part of the association seen with this SNP is driven by the SNP being conditioned on); SNPs showing an independent association (*P*<0.05) are highlighted in blue (i.e. *P*-value lower than that seen in the unconditioned analysis) with conditioned SNP).

3.3.4 Genotyping quality control for Metabochip

The Metabochip CAD samples submitted to the Sanger Genome Campus logistics facility pipe-line for DNA QC resulted in the removal of 53 samples, reducing the submitted CAD sample number from 3219 to 3166 samples that underwent genotyping. The 3166 genotyped samples (including 4 lab fails) were submitted for ILLUMINUS (Teo et al. 2007) software clustering, resulting in the removal of a further 41 samples, 36 of these failed the required >90% call rate threshold, leaving 3125 samples for analysis. All 1958BC control samples, 6000 (including 28 lab fails) were used for ILLUMINUS (Teo et al. 2007) software clustering, resulting in the removal of 137 samples, 131 of these failed the required >90% call rate threshold, leaving 5841 samples for analysis.

A logistic regression analysis was undertaken for 3125 CAD cases and 5841 controls using PLINK (Purcell et al. 2007); again exclusion filters were applied for missingness of individuals ≥2% (PLINK command: - - MIND 0.02) (n=113), missingness of SNP genotypes < 90% (PLINK command: - - GENO 0.1) (n=145), Hardy-Weinberg equilibrium violation (p<0.0001) (n=514), and SNPs with low MAF <0.01 (n=6,582). After frequency and genotype pruning, there remained 11,656 chromosome 1 SNPs, 3000 CAD cases and 5628 controls in the analysis. Of note, the number of SNPs that span four candidate genes (*CELSR2, PSRC1, MYBPHL, SORT1*) within the chr1p13.3 locus region that also incorporates a 10kb 5' flanking region upstream of *CELSR2* and *SORT1* included n=438 SNPs prior to quality control (QC) exclusions, and n=195 SNPs after QC exclusions, mainly due to removal of SNPs with very low MAFs (<0.01).

3.3.5 Chr1p13.3 fine-mapping SNP association with CAD in Metabochip

The Metabochip distribution of *P*-value CAD associations adjusted for age and gender for all chr1p13.3 locus refinement SNPs (n=195) are shown in **Table 3.5** and **Figure 3.3**. The main finding in the Metabochip fine-mapping CAD association study was that SNP rs7528419 and three other almost perfect proxy (i.e. $r^2 \approx 1$) SNPs: rs646776, rs12740374 and rs629301 were more strongly associated with CAD risk than the CAD GWA study discovery SNP rs599839. The most significant

SNP was rs7528419 where the risk allele (A) (allele frequency 0.79) was associated with a 24.2% (95% CI: 16.2–31.3) increase in CAD risk per allele copy (*P*=4.88x10^{-8}); whereas SNP rs599839 although still strongly associated with CAD, had a slightly lower level of association, with a 23.2% (95% CI: 15.3–30.4) increase in CAD risk per copy of the risk allele (A), (*P*=1.27x10^{-7}).

Interestingly, rs7528419 remained significant even after a Bonferroni correction for multiple genotype testing, which lowered the *P*-value threshold for statistical significance to *P*<2.56x10^{-4} (0.05/195). Indeed, under this adjusted level of significance, 23 fine-mapped SNPs spanning 38,518bp (again as with the LDL-C GRAPHIC study association this locates to an intronic location of *CELSR2* to a location just beyond the 3'UTR of *PSRC1*) reached a Bonferroni corrected significance for LDL-C association (Of note, 62 SNPs over a range of 159,719bp showed at least nominal significance at 0.01<*P*<0.05).

Table 3.5. SNPs within the 1p13.3 locus showing association with CAD in the Metabochip study.

Chr.	Location (bp)	SNP	Risk allele	Minor allele	MAF (Unaffected)	HWE (Unaffected)	Gene	Function class	Odds Ratio (95% CI)	*P*-value	LD (r$^{2)}$
1	109618715	rs7528419	A	G	0.211	0.307	CELSR2	3'UTR	0.758 (0.687-0.838)	4.88x10^{-08}	1.000
1	109620053	rs646776	A	G	0.211	0.344	CELSR2	3'	0.759 (0.687-0.838)	5.04x10^{-08}	0.997
1	109619113	rs12740374	C	A	0.210	0.362	CELSR2	3'UTR	0.759 (0.687-0.838)	5.25x10^{-08}	0.997
1	109622830	rs583104	A	C	0.216	0.456	PSRC1	3'	0.767 (0.695-0.846)	1.19x10^{-07}	0.954
1	109619361	rs660240	G	A	0.202	0.558	CELSR2	3'UTR	0.762 (0.689-0.843)	1.24x10^{-07}	0.946
1	109623689	rs599839	A	G	0.217	0.480	PSRC1	3'	0.768 (0.696-0.847)	1.27x10^{-07}	0.946
1	109619681	rs57677983	A	G	0.202	0.482	CELSR2	3'UTR	0.763 (0.690-0.844)	1.40x10^{-07}	0.945
1	109619829	rs629301	A	C	0.211	0.849	CELSR2	3'UTR	0.767 (0.695-0.847)	1.54x10^{-07}	0.993
1	109623034	rs602633	C	A	0.205	0.908	PSRC1	3'	0.767 (0.694-0.848)	2.03x10^{-07}	0.915
1	109623666	rs1277930	A	G	0.218	0.737	PSRC1	3'	0.773 (0.701-0.852)	2.35x10^{-07}	0.943
1	109616403	rs4970834	G	A	0.177	0.966	CELSR2	intron	0.775 (0.697-0.862)	2.62x10^{-06}	0.654
1	109618768	rs11102967	A	G	0.319	0.594	CELSR2	3'UTR	0.816 (0.749-0.889)	3.06x10^{-06}	0.569
1	109616775	rs611917	A	G	0.307	0.809	CELSR2	intron	0.817 (0.749-0.890)	4.21x10^{-06}	0.533
1	109606169	rs4970833	A	A	0.463	0.303	CELSR2	intron	1.183 (1.095-1.280)	2.33x10^{-05}	0.176
1	109585171	rs579947	A	A	0.449	0.410	CELSR2	5'	1.181 (1.093-1.278)	2.98x10^{-05}	0.103
1	109585549	rs34293021	A	A	0.451	0.490	CELSR2	5'	1.180 (1.091-1.276)	3.38x10^{-05}	0.105
1	109585986	rs688386	A	A	0.450	0.595	CELSR2	5'	1.180 (1.091-1.276)	3.39x10^{-05}	0.106
1	109614543	rs3895559	A	A	0.465	0.397	CELSR2	intron	1.179 (1.090-1.275)	3.63x10^{-05}	0.178
1	109586461	rs11102965	G	G	0.450	0.473	CELSR2	5'	1.179 (1.090-1.275)	3.65x10^{-05}	0.106
1	109608357	rs6698843	A	A	0.463	0.290	CELSR2	exon	1.179 (1.090-1.274)	3.69x10^{-05}	0.176
1	109608622	rs6689614	A	A	0.463	0.303	CELSR2	exon	1.178 (1.090-1.274)	3.75x10^{-05}	0.176
1	109607965	rs1337248	A	C	0.418	0.669	CELSR2	intron	0.854 (0.789-0.925)	1.01x10^{-04}	0.234
1	109600244	rs10858082	A	G	0.427	0.873	CELSR2	intron	0.861 (0.795-0.932)	2.21x10^{-04}	0.260
1	109601844	rs12746961	G	A	0.091	0.941	CELSR2	intron	0.770 (0.669-0.887)	2.97x10^{-04}	0.312
1	109615242	rs17035665	G	A	0.204	0.258	CELSR2	intron	0.833 (0.754-0.921)	3.47x10^{-04}	0.112
1	109589976	rs10776804	T	T	0.500	0.295	CELSR2	5'	0.867 (0.802-0.938)	3.50x10^{-04}	0.085

1	109622407	rs11577931	A	G	0.086	0.531	PSRC1	3'	0.769 (0.665-0.890)	4.40x10^{-04}	0.345
1	109621504	rs55882046	G	A	0.086	0.583	CELSR2	3'	0.778 (0.672-0.899)	7.01x10^{-04}	0.349
1	109585078	rs4246519	G	A	0.472	0.271	CELSR2	5'	0.873 (0.807-0.945)	7.56x10^{-04}	0.014
1	109606406	rs66795935	G	C	0.092	0.711	CELSR2	intron	0.787 (0.684-0.906)	8.28x10^{-04}	0.316
1	109611723	rs72703203	G	A	0.042	0.878	CELSR2	exon	0.709 (0.576-0.873)	1.19x10^{-03}	0.148
1	109611724	rs72703204	G	A	0.041	0.878	CELSR2	exon	0.709 (0.576-0.873)	1.20x10^{-03}	0.148
1	109632748	rs115967373	G	A	0.083	1.000	MYBPHL	3'	0.787 (0.678-0.913)	1.59x10^{-03}	0.193
1	109625773	rs35358959	G	A	0.076	0.286	PSRC1	exon	0.779 (0.667-0.910)	1.66x10^{-03}	0.194
1	109636686	rs41306199	G	A	0.081	0.737	MYBPHL	exon	0.787 (0.678-0.915)	1.76x10^{-03}	0.187
1	109627659	rs657420	A	G	0.468	0.282	PSRC1	5'	0.884 (0.817-0.956)	2.00x10^{-03}	0.173
1	109591318	rs585362	A	G	0.150	0.921	CELSR2	5'	0.840 (0.750-0.940)	2.39x10^{-03}	0.172
1	109615033	rs41279716	T	A	0.206	0.186	CELSR2	intron	0.858 (0.777-0.947)	2.43x10^{-03}	0.101
1	109636461	rs17645031	G	A	0.083	1.000	MYBPHL	3'	0.796 (0.687-0.924)	2.61x10^{-03}	0.193
1	109634710	rs17584208	G	A	0.083	1.000	MYBPHL	3'	0.797 (0.687-0.925)	2.79x10^{-03}	0.193
1	109633806	rs655246	G	A	0.468	0.401	MYBPHL	3'	0.894 (0.826-0.967)	5.01x10^{-03}	0.135
1	109652649	rs413582	A	A	0.487	0.218	SORT1	3'	1.114 (1.030-1.205)	6.72x10^{-03}	0.093
1	109658366	rs661278	C	A	0.210	0.173	SORT1	intron	0.875 (0.794-0.964)	6.88x10^{-03}	0.077
1	109641692	rs76186504	G	A	0.023	0.786	MYBPHL	exon	0.683 (0.518-0.901)	6.93x10^{-03}	0.074
1	109688714	rs4970843	G	G	0.484	0.294	SORT1	intron	1.112 (1.028-1.203)	7.91x10^{-03}	0.092
1	109585093	rs4246520	T	A	0.425	0.936	CELSR2	5'	0.905 (0.836-0.980)	0.01	0.005
1	109636910	rs590502	C	A	0.455	0.445	MYBPHL	intron	0.907 (0.838-0.982)	0.02	0.102
1	109696112	rs75430477	A	C	0.160	0.444	SORT1	intron	0.875 (0.785-0.975)	0.02	0.072
1	109615008	rs41279714	G	A	0.040	1.000	CELSR2	intron	0.775 (0.627-0.957)	0.02	0.152
1	109662603	rs116149872	G	C	0.054	0.139	SORT1	intron	0.808 (0.677-0.965)	0.02	0.104
1	109696304	rs116611120	A	C	0.055	0.141	SORT1	intron	0.809 (0.678-0.966)	0.02	0.104
1	109722744	rs115796838	G	A	0.055	0.140	SORT1	intron	0.810 (0.679-0.967)	0.02	0.105
1	109698623	rs72646560	A	G	0.054	0.139	SORT1	exon	0.811 (0.679-0.967)	0.02	0.105
1	109685336	rs114191770	A	G	0.054	0.140	SORT1	intron	0.811 (0.679-0.968)	0.02	0.104
1	109584784	rs11102964	A	G	0.168	0.231	CELSR2	5'	0.882 (0.792-0.983)	0.02	0.084

1	109664544	rs72703223	A	A	0.015	0.283	SORT1	intron	1.423 (1.045-1.937)	0.02	0.002
1	109647801	rs407102	A	G	0.287	0.949	MYBPHL	intron	0.911 (0.834-0.994)	0.04	0.035
1	109744447	rs72703257	A	A	0.031	0.643	SORT1	5'	1.258 (1.014-1.560)	0.04	0.006
1	109646623	rs1278286	A	A	0.140	0.656	MYBPHL	intron	1.126 (1.007-1.259)	0.04	0.019
1	109657829	rs464218	A	G	0.450	0.510	SORT1	3'UTR	0.921 (0.851-0.997)	0.04	0.036
1	109697491	rs6683212	C	A	0.277	0.476	SORT1	intron	0.913 (0.836-0.997)	0.04	0.036
1	109584728	rs4623734	C	A	0.263	0.016	CELSR2	5'	0.911 (0.832-0.998)	0.05	0.015
1	109634677	rs78127234	G	A	0.013	0.288	MYBPHL	3'	0.697 (0.480-1.010)	0.06	0.041
1	109612067	rs2281894	C	A	0.186	0.162	CELSR2	exon	0.905 (0.817-1.003)	0.06	0.022
1	109747618	rs4970846	C	A	0.279	0.419	SORT1	5'	0.920 (0.843-1.005)	0.06	0.037
1	109634413	rs17036080	G	A	0.013	0.277	MYBPHL	3'	0.705 (0.486-1.022)	0.07	0.042
1	109643794	rs680434	A	G	0.282	0.585	MYBPHL	intron	0.921 (0.844-1.006)	0.07	0.034
1	109585036	rs61799429	C	G	0.206	0.478	CELSR2	5'	0.912 (0.826-1.006)	0.07	0.051
1	109643185	rs585545	T	A	0.265	0.973	MYBPHL	intron	0.920 (0.841-1.006)	0.07	0.040
1	109632768	rs6699122	G	A	0.013	0.273	MYBPHL	3'	0.710 (0.490-1.029)	0.07	0.043
1	109667753	rs3768497	G	A	0.278	0.538	SORT1	intron	0.922 (0.844-1.007)	0.07	0.036
1	109673310	rs10745352	G	A	0.278	0.538	SORT1	intron	0.922 (0.844-1.007)	0.07	0.035
1	109679029	rs10745353	G	A	0.275	0.280	SORT1	intron	0.922 (0.844-1.007)	0.07	0.036
1	109661667	rs3853498	G	A	0.277	0.517	SORT1	intron	0.922 (0.845-1.007)	0.07	0.036
1	109710672	rs3879450	A	G	0.278	0.538	SORT1	intron	0.923 (0.845-1.007)	0.07	0.036
1	109672382	rs6695482	G	C	0.278	0.538	SORT1	intron	0.923 (0.845-1.008)	0.07	0.035
1	109665897	rs3853501	A	G	0.278	0.538	SORT1	intron	0.923 (0.845-1.008)	0.07	0.036
1	109742656	rs1880670	A	G	0.278	0.475	SORT1	5'	0.923 (0.845-1.008)	0.07	0.036
1	109656927	rs461200	A	G	0.277	0.417	SORT1	3'UTR	0.923 (0.845-1.008)	0.07	0.036
1	109716598	rs10858086	A	C	0.277	0.538	SORT1	intron	0.923 (0.845-1.008)	0.07	0.036
1	109683773	rs10858084	T	A	0.277	0.399	SORT1	intron	0.923 (0.845-1.008)	0.07	0.036
1	109642152	rs600806	A	G	0.273	0.471	MYBPHL	intron	0.922 (0.844-1.008)	0.07	0.041
1	109738949	rs1343465	G	A	0.278	0.476	SORT1	intron	0.923 (0.845-1.008)	0.07	0.036
1	109737101	rs10858091	G	A	0.277	0.517	SORT1	intron	0.923 (0.845-1.008)	0.07	0.035

1	109665792	rs3853500	G	A	0.278	0.582	SORT1	intron	0.923 (0.845-1.008)	0.07	0.036
1	109668092	rs3768496	A	T	0.278	0.581	SORT1	intron	0.923 (0.845-1.008)	0.08	0.036
1	109745601	rs7518013	G	A	0.277	0.436	SORT1	5'	0.923 (0.845-1.008)	0.08	0.035
1	109625485	rs72703210	A	G	0.043	0.089	PSRC1	intron	0.835 (0.685-1.019)	0.08	0.064
1	109725109	rs10858088	A	G	0.278	0.476	SORT1	intron	0.923 (0.845-1.008)	0.08	0.035
1	109716506	rs10858085	G	A	0.277	0.559	SORT1	intron	0.923 (0.845-1.008)	0.08	0.036
1	109652650	rs444387	G	A	0.278	0.517	SORT1	3'	0.924 (0.846-1.009)	0.08	0.036
1	109745416	rs10858092	A	G	0.278	0.475	SORT1	5'	0.924 (0.846-1.009)	0.08	0.035
1	109681072	rs4603158	A	G	0.277	0.538	SORT1	intron	0.924 (0.846-1.009)	0.08	0.035
1	109686298	rs2228604	C	A	0.277	0.517	SORT1	exon	0.924 (0.846-1.009)	0.08	0.036
1	109733431	rs10745354	A	G	0.277	0.346	SORT1	intron	0.924 (0.846-1.009)	0.08	0.036
1	109650777	rs443345	G	A	0.277	0.516	MYBPHL	intron	0.924 (0.846-1.009)	0.08	0.036
1	109736848	rs3768495	A	G	0.277	0.537	SORT1	intron	0.924 (0.846-1.009)	0.08	0.036
1	109736950	rs3768494	G	A	0.277	0.516	SORT1	intron	0.924 (0.846-1.010)	0.08	0.035
1	109692946	rs3879448	G	C	0.286	0.610	SORT1	intron	0.925 (0.848-1.009)	0.08	0.031
1	109728351	rs10858089	G	A	0.277	0.399	SORT1	intron	0.925 (0.847-1.010)	0.08	0.035
1	109706494	rs3879449	A	T	0.277	0.517	SORT1	intron	0.925 (0.847-1.010)	0.08	0.036
1	109646055	rs652651	G	A	0.277	0.363	MYBPHL	intron	0.925 (0.847-1.010)	0.08	0.035
1	109698626	rs11142	G	A	0.276	0.455	SORT1	exon	0.925 (0.847-1.010)	0.08	0.035
1	109659642	rs370088	G	A	0.278	0.559	SORT1	intron	0.925 (0.847-1.010)	0.08	0.036
1	109685099	rs74584797	A	G	0.278	0.456	SORT1	intron	0.925 (0.847-1.010)	0.08	0.036
1	109698123	rs4970751	C	A	0.277	0.436	SORT1	intron	0.925 (0.847-1.010)	0.08	0.035
1	109643182	rs681242	G	A	0.262	0.840	MYBPHL	intron	0.924 (0.844-1.011)	0.08	0.041
1	109643283	rs680767	G	A	0.262	0.893	MYBPHL	intron	0.924 (0.845-1.011)	0.09	0.041
1	109750531	rs4120621	G	A	0.277	0.417	SORT1	5'	0.926 (0.848-1.011)	0.09	0.036
1	109597576	rs2359414	G	A	0.249	0.427	CELSR2	intron	0.925 (0.844-1.014)	0.10	0.003
1	109589016	rs17035415	C	A	0.199	0.241	CELSR2	5'	0.919 (0.832-1.016)	0.10	0.040
1	109699790	rs12027996	C	G	0.276	0.515	SORT1	intron	0.932 (0.853-1.018)	0.12	0.035
1	109641261	rs3850615	A	A	0.077	0.781	MYBPHL	exon	1.122 (0.971-1.297)	0.12	0.013

1	109590870	rs4970748	T	A	0.201	0.335	CELSR2	5'	0.924 (0.837-1.021)	0.12	0.039
1	109639787	rs12127701	A	G	0.068	1.000	MYBPHL	intron	0.884 (0.752-1.038)	0.13	0.079
1	109616891	rs6670347	A	G	0.039	0.308	CELSR2	intron	0.855 (0.692-1.055)	0.14	0.148
1	109713574	rs17646665	A	G	0.063	0.440	SORT1	intron	0.885 (0.750-1.044)	0.15	0.059
1	109617269	rs6667814	A	G	0.039	0.306	CELSR2	intron	0.857 (0.694-1.058)	0.15	0.148
1	109616597	rs79868705	G	A	0.039	0.304	CELSR2	intron	0.858 (0.695-1.059)	0.15	0.148
1	109616656	rs79482788	G	A	0.038	0.304	CELSR2	intron	0.860 (0.696-1.061)	0.16	0.147
1	109595646	rs41279704	C	C	0.017	0.653	CELSR2	exon	1.237 (0.917-1.670)	0.16	0.005
1	109625060	rs76057315	A	A	0.014	0.629	PSRC1	exon	1.249 (0.907-1.720)	0.17	0.003
1	109619651	rs77579579	G	A	0.038	0.295	CELSR2	3'UTR	0.865 (0.700-1.068)	0.18	0.147
1	109618822	rs41279722	C	C	0.012	0.505	PSRC1	3’UTR	1.271 (0.891-1.814)	0.19	0.003
1	109641419	rs7515901	G	A	0.164	0.119	MYBPHL	intron	0.931 (0.837-1.036)	0.19	0.052
1	109723900	rs1149175	G	A	0.132	1.000	SORT1	intron	0.925 (0.822-1.040)	0.19	0.017
1	109640441	rs629001	A	G	0.069	0.763	MYBPHL	exon	0.902 (0.770-1.057)	0.20	0.079
1	109711536	rs11581665	G	A	0.132	1.000	SORT1	intron	0.926 (0.824-1.042)	0.20	0.017
1	109752130	rs1149174	G	A	0.132	1.000	SORT1	5'	0.927 (0.825-1.043)	0.21	0.017
1	109653316	rs72647819	G	A	0.023	1.000	SORT1	3'	0.840 (0.640-1.103)	0.21	0.005
1	109621633	rs6677122	G	A	0.040	0.178	CELSR2	3'	0.879 (0.715-1.081)	0.22	0.140
1	109642846	rs76569103	C	G	0.012	1.000	MYBPHL	intron	0.790 (0.541-1.155)	0.22	$1x10^{-04}$
1	109734973	rs56072034	G	A	0.023	1.000	SORT1	intron	0.846 (0.645-1.110)	0.23	0.004
1	109656543	rs72647804	A	G	0.044	0.882	SORT1	3'UTR	0.888 (0.732-1.079)	0.23	0.009
1	109697527	rs17646453	A	A	0.023	0.764	SORT1	intron	1.164 (0.900-1.506)	0.25	0.007
1	109601609	rs78852738	C	C	0.030	0.354	CELSR2	intron	1.135 (0.907-1.421)	0.27	0.007
1	109599286	rs6693893	A	G	0.033	1.000	CELSR2	intron	0.897 (0.715-1.125)	0.35	0.090
1	109604166	rs72975220	G	A	0.033	1.000	CELSR2	intron	0.898 (0.715-1.127)	0.35	0.089
1	109638946	rs57607242	A	C	0.076	0.461	MYBPHL	intron	0.933 (0.802-1.086)	0.37	0.067
1	109659338	rs17585355	C	C	0.064	1.000	SORT1	intron	1.074 (0.916-1.260)	0.38	0.009
1	109645298	rs11583969	G	A	0.065	0.600	MYBPHL	intron	0.931 (0.794-1.092)	0.38	0.006
1	109642441	rs115030251	G	A	0.131	0.954	MYBPHL	intron	0.950 (0.845-1.068)	0.39	0.018

1	109637904	rs10858083	A	T	0.076	0.576	MYBPHL	intron	0.937 (0.805-1.090)	0.40	0.067
1	109619002	rs1277917	C	G	0.141	0.237	PSRC1	3’UTR	0.955 (0.853-1.070)	0.43	0.004
1	109593079	rs434524	A	G	0.033	1.000	CELSR2	5'	0.917 (0.732-1.150)	0.45	0.084
1	109741188	rs116362610	C	G	0.020	0.309	SORT1	intron	0.899 (0.674-1.198)	0.47	0.003
1	109686459	rs114259186	A	C	0.020	0.309	SORT1	intron	0.899 (0.675-1.198)	0.47	0.003
1	109682244	rs11102972	A	G	0.222	0.821	SORT1	intron	0.967 (0.879-1.063)	0.49	0.001
1	109742026	rs72646553	G	C	0.222	0.733	SORT1	exon	0.970 (0.882-1.067)	0.53	0.001
1	109638365	rs646335	C	G	0.374	0.597	MYBPHL	3' UTR	0.974 (0.898-1.057)	0.53	0.006
1	109678408	rs7550401	G	A	0.222	0.850	SORT1	intron	0.970 (0.882-1.067)	0.54	0.001
1	109597131	rs437444	G	A	0.033	1.000	CELSR2	exon	0.934 (0.745-1.170)	0.55	0.085
1	109696216	rs7536292	A	G	0.175	0.501	SORT1	intron	0.970 (0.875-1.074)	0.55	4x10^{-04}
1	109694733	rs1278664	C	A	0.221	0.704	SORT1	intron	0.972 (0.884-1.069)	0.56	0.001
1	109723293	rs191809754	A	A	0.012	0.001	SORT1	intron	1.106 (0.787-1.556)	0.56	3x10^{-04}
1	109636906	rs2282292	G	C	0.134	0.617	MYBPHL	intron	0.967 (0.862-1.085)	0.57	0.003
1	109591809	rs74113799	G	A	0.032	1.000	CELSR2	5'	0.936 (0.747-1.173)	0.57	0.085
1	109599150	rs145496681	A	A	0.104	0.671	CELSR2	intron	1.037 (0.912-1.178)	0.58	0.031
1	109599151	rs182402368	A	A	0.104	0.671	CELSR2	intron	1.036 (0.912-1.178)	0.59	0.031
1	109595775	rs454107	G	A	0.032	1.000	CELSR2	exon	0.939 (0.749-1.177)	0.59	0.085
1	109596549	rs413380	G	A	0.032	1.000	CELSR2	exon	0.942 (0.752-1.180)	0.60	0.085
1	109728122	rs116363925	A	A	0.066	0.751	SORT1	intron	1.039 (0.887-1.217)	0.64	0.010
1	109602037	rs623371	A	A	0.101	0.663	CELSR2	intron	1.029 (0.903-1.171)	0.67	0.029
1	109600767	rs649281	G	G	0.101	0.771	CELSR2	intron	1.028 (0.903-1.171)	0.67	0.029
1	109603451	rs606434	G	G	0.101	0.662	CELSR2	intron	1.026 (0.901-1.169)	0.70	0.029
1	109641477	rs603624	G	C	0.136	0.469	MYBPHL	intron	0.978 (0.873-1.095)	0.70	0.003
1	109601654	rs653138	G	G	0.101	0.716	CELSR2	intron	1.026 (0.901-1.168)	0.70	0.029
1	109603554	rs606007	A	A	0.101	0.662	CELSR2	intron	1.025 (0.900-1.168)	0.70	0.029
1	109601741	rs634495	G	G	0.101	0.717	CELSR2	intron	1.025 (0.900-1.168)	0.71	0.029
1	109733513	rs116649254	G	G	0.056	0.382	SORT1	intron	1.032 (0.872-1.222)	0.72	0.004
1	109603923	rs592360	A	A	0.101	0.560	CELSR2	intron	1.022 (0.898-1.164)	0.74	0.029

1	109635562	rs17036085	A	G	0.012	0.239	MYBPHL	3'	0.942 (0.652-1.361)	0.75	0.016
1	109602150	rs655334	T	T	0.100	0.662	CELSR2	intron	1.020 (0.895-1.162)	0.77	0.029
1	109603756	rs591789	A	A	0.100	0.609	CELSR2	intron	1.018 (0.894-1.160)	0.78	0.029
1	109609640	rs608196	A	A	0.099	0.506	CELSR2	intron	1.018 (0.893-1.161)	0.79	0.029
1	109637280	rs17645143	A	G	0.297	0.685	MYBPHL	intron	0.989 (0.907-1.078)	0.80	0.005
1	109601098	rs640137	A	A	0.101	0.827	CELSR2	intron	1.017 (0.893-1.159)	0.80	0.029
1	109638358	rs3738777	A	G	0.298	0.732	MYBPHL	3' UTR	0.989 (0.907-1.078)	0.80	0.005
1	109634661	rs191675995	A	C	0.296	0.852	MYBPHL	3'	0.990 (0.908-1.079)	0.82	0.005
1	109607836	rs653635	G	G	0.098	0.155	CELSR2	exon	1.013 (0.887-1.158)	0.85	0.028
1	109727352	rs17037403	G	A	0.019	0.466	SORT1	intron	0.973 (0.731-1.295)	0.85	0.003
1	109612504	rs17035630	G	A	0.121	0.143	CELSR2	intron	0.989 (0.875-1.116)	0.85	0.035
1	109725200	rs12037569	C	A	0.154	0.486	SORT1	intron	0.990 (0.890-1.102)	0.86	0.003
1	109600326	rs626387	G	A	0.134	0.173	CELSR2	intron	0.991 (0.883-1.111)	0.87	$4x10^{-05}$
1	109710272	rs115730757	A	A	0.023	1.000	SORT1	intron	1.021 (0.790-1.320)	0.87	0.008
1	109603918	rs592355	G	A	0.133	0.124	CELSR2	intron	0.992 (0.884-1.113)	0.89	$2x10^{-05}$
1	109601909	rs654106	T	A	0.134	0.156	CELSR2	intron	0.993 (0.885-1.114)	0.91	$2x10^{-05}$
1	109584152	rs116645638	A	A	0.055	0.210	CELSR2	5'	1.010 (0.852-1.197)	0.91	0.001
1	109623927	rs10410	A	A	0.106	0.835	PSRC1	3' UTR	1.007 (0.886-1.144)	0.92	0.033
1	109619793	rs658435	A	A	0.103	0.943	PSRC1	3’UTR	1.006 (0.884-1.145)	0.93	0.031
1	109624032	rs14000	A	G	0.107	0.837	PSRC1	3' UTR	0.997 (0.878-1.132)	0.97	0.033
1	109639738	rs35267391	A	A	0.014	0.635	MYBPHL	intron	1.007 (0.720-1.407)	0.97	0.023
1	109723294	rs183660926	A	G	0.153	0.484	SORT1	intron	0.999 (0.897-1.112)	0.98	0.003
1	109696279	rs112987420	C	G	0.015	1.000	SORT1	intron	0.998 (0.721-1.382)	0.99	0.009
1	109638652	rs56237434	A	A	0.032	0.299	MYBPHL	intron	1.000 (0.803-1.246)	1.00	0.008

Legend. Chr – chromosome; SNP – single nucleotide polymorphism; Location – Build 36.3; MAF – minor allele frequency; HWE – statistical significance of Hardy-Weinberg equilibrium; Odds Ratio – estimated effect size of each SNP minor allele copy on CAD (adjusted for age and sex); an odds ratio below 1.0 indicates a decreased risk of CAD in carriers of the minor allele, whereas an odds ratio greater than 1.00 indicates an increased risk of CAD in carriers of the minor allele; 95% confidence intervals (95% CI); *P*-value – statistical significance of association. The lead Metabochip CAD associated SNP rs7528419 is highlighted in bold red text and the CAD GWA study discovery SNP rs599839 is highlighted in normal red text.

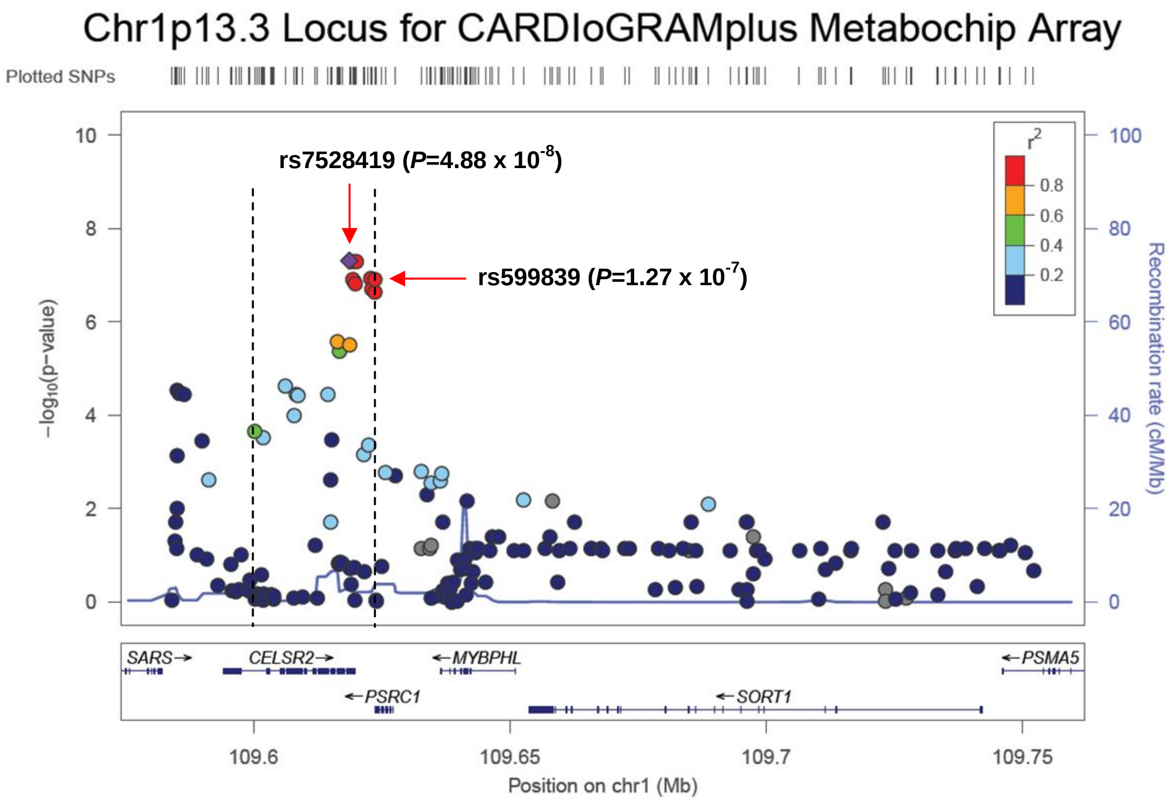


Figure 3.3. Regional Association plot at locus chr1p13.3 for CAD in the Metabochip study**.**

Legend. The associations for the individual SNPs shown in **Table 3.5** are plotted as –log$_{10}$ *P*-values on the Y-axis (left-hand-side) versus chromosome base pair position on the X-axis (Recombination rate (hotspots) are also plotted on the Y-axis (right-hand-side), shown in light-blue). The purple diamond plots the most significant SNP rs7528419 for association with CAD. A colour-coded scale of pairwise LD r^2 scores between the lead SNP and 194 other plotted SNPs are shown for the locus region. The vertical dotted-lines indicate the distance over which the lead SNP rs629301 has a pairwise LD of r^2≥0.5. The X-axis spans the SNP locations over 167,400bp. The gene location and strand direction are shown with arrows. The four candidate genes are: *CELSR2* (cadherin, EGF LAG seven-pass G-type receptor 2), *PSRC1* (proline/serine-rich coiled-coil 1), *MYBPHL* (myosin binding protein H-like), *SORT1* (sortilin 1). (**Figure 3.3.** is generated by the webtool LocusZoom, hosted via the Center for Statistical Genetics, University of Michigan at https://statgen.sph.umich.edu/locuszoom/genform.php?type=yourdata (Version 1.1: June 2011) (Pruim et al. 2010).

In particular, **Table 3.5** and **Figure 3.3** show that there are now ten SNPs with a *P*-value for CAD association of $P<1\times10^{-7}$, and that these SNPs are highly correlated, i.e. all ten SNPs are in very strong LD ($r^2>0.915$). As with the LDL-C GRAPHIC study a *conditional* analysis was performed, adjusting for the new lead SNP rs7528419 and the original CAD association SNP rs599839 in the model. Thus a conditional analysis was performed to see if the observed association is an independent signal. To this end, **Table 3.6**, provides us with the evidence that there is only one independent signal for CAD (the same as was seen for the GRAPHIC LDL-C association study), because the *P*-value association disappears for all 23 of the Bonferroni corrected significant SNPs. The only potential for additional independent signals are seen at a $0.001<P<0.01$ threshold for a potentially co-dependent SNP rs4246519 in the distal promoter of *CELSR2*, and at a nominal $0.01<P<0.05$ threshold, for two potentially co-dependent SNPs rs4246520 in the distal promoter region of *CELSR2* and rs72703223 an intronic SNP within *SORT1* (i.e. where the association has not dropped as much as one might expect when conditioning on the lead SNP). However, the take home message from the conditional analysis is that the most likely SNPs to be causally driving the CAD association in the Metabochip study are one of rs629301, rs7528419, rs12740374 and rs646776, within or just beyond the 3'UTR of *CELSR2* (the same as was seen for the GRAPHIC study LDL-C association).

Table 3.6. Conditional analysis of lead SNPs rs7528419 and rs599839 within the 1p13.3 locus showing association with CAD in the Metabochip study.

SNP	Gene	Func. class	*P*-val GEE	LD (r^2) with rs7528419	*P*-val conditioned on rs7528419	Comment for conditioning on rs7528419	LD (r^2) with rs599839	*P*-val conditioned on rs599839	Comment for conditioning on rs599839
rs7528419	CELSR2	3'UTR	4.88x10^{-08}	1.000	NA	Multi-collinearity	0.944	0.160	Strong LD Sig.↓
rs646776	CELSR2	3'	5.04x10^{-08}	0.997	NA	Multi-collinearity	0.945	0.141	Strong LD Sig.↓
rs12740374	CELSR2	3'UTR	5.25x10^{-08}	0.997	NA	Multi-collinearity	0.942	0.139	Strong LD Sig.↓
rs583104	PSRC1	3'	1.19x10^{-07}	0.954	0.704	Strong LD Sig.↓	0.989	NA	Multi-collinearity
rs660240	CELSR2	3'UTR	1.24x10^{-07}	0.946	0.997	Strong LD Sig.↓	0.900	0.352	Strong LD Sig.↓
rs599839	PSRC1	3'	1.27x10^{-07}	0.946	0.907	Strong LD Sig.↓	1.000	NA	Multi-collinearity
rs57677983	CELSR2	3'UTR	1.40x10^{-07}	0.945	0.917	Strong LD Sig.↓	0.900	0.394	Strong LD Sig.↓
rs629301	CELSR2	3'UTR	1.54x10^{-07}	0.993	NA	Multi-collinearity	0.944	0.292	Strong LD Sig.↓
rs602633	PSRC1	3'	2.03x10^{-07}	0.915	0.864	Strong LD Sig.↓	0.934	0.725	Strong LD Sig.↓
rs1277930	PSRC1	3'	2.35x10^{-07}	0.943	0.645	Strong LD Sig.↓	0.999	NA	Multi-collinearity
rs4970834	CELSR2	intron	2.62x10^{-06}	0.654	0.605	Moderate LD Sig.↓	0.669	0.469	Moderate LD Sig.↓
rs11102967	CELSR2	3'UTR	3.06x10^{-06}	0.569	0.422	Moderate LD Sig.↓	0.560	0.288	Moderate LD Sig.↓
rs611917	CELSR2	intron	4.21x10^{-06}	0.533	0.367	Moderate LD Sig.↓	0.505	0.235	Moderate LD Sig.↓
rs4970833	CELSR2	intron	2.33x10^{-05}	0.176	0.035	Very weak LD Sig.↓	0.179	0.027	Very weak LD Sig.↓
rs579947	CELSR2	5'	2.98x10^{-05}	0.103	0.012	Very weak LD Sig.↓	0.105	9.14x10^{-03}	Very weak LD Sig.↓
rs34293021	CELSR2	5'	3.38x10^{-05}	0.105	0.013	Very weak LD Sig.↓	0.107	0.010	Very weak LD Sig.↓
rs688386	CELSR2	5'	3.39x10^{-05}	0.106	0.014	Very weak LD Sig.↓	0.109	0.011	Very weak LD Sig.↓
rs3895559	CELSR2	intron	3.63x10^{-05}	0.178	0.042	Very weak LD Sig.↓	0.149	0.032	Very weak LD Sig.↓
rs11102965	CELSR2	5'	3.65x10^{-05}	0.106	0.015	Very weak LD Sig.↓	0.109	0.011	Very weak LD Sig.↓
rs6698843	CELSR2	exon	3.69x10^{-05}	0.176	0.048	Very weak LD Sig.↓	0.178	0.037	Very weak LD Sig.↓
rs6689614	CELSR2	exon	3.75x10^{-05}	0.176	0.047	Very weak LD Sig.↓	0.179	0.037	Very weak LD Sig.↓
rs1337248	CELSR2	intron	1.01x10^{-04}	0.234	0.163	Weak LD Sig.↓	0.237	0.144	Weak LD Sig.↓
rs10858082	CELSR2	intron	2.21x10^{-04}	0.260	0.306	Weak LD Sig.↓	0.262	0.266	Weak LD Sig.↓

rs12746961	CELSR2	intron	2.97×10^{-04}	0.312	0.455	Weak LD Sig.↓	0.304	0.401	Weak LD Sig.↓
rs17035665	CELSR2	intron	3.47×10^{-04}	0.112	0.059	Very weak LD Sig.↓	0.030	0.047	Very weak LD Sig.↓
rs10776804	CELSR2	5'	3.50×10^{-04}	0.085	0.040	Very weak LD Sig.↓	0.089	0.036	Very weak LD Sig.↓
rs11577931	PSRC1	3'	4.40×10^{-04}	0.345	0.661	Weak LD Sig.↓	0.337	0.575	Weak LD Sig.↓
rs55882046	CELSR2	3'	7.01×10^{-04}	0.349	0.801	Weak LD Sig.↓	0.333	0.675	Weak LD Sig.↓
rs4246519	CELSR2	5'	7.56×10^{-04}	0.014	6.91×10^{-03}	Co-dependent; $P<0.01$	0.008	6.42×10^{-03}	Co-dependent; $P<0.01$
rs66795935	CELSR2	intron	8.28×10^{-04}	0.316	0.710	Weak LD Sig.↓	0.308	0.647	Weak LD Sig.↓
rs72703203	CELSR2	exon	1.19×10^{-03}	0.148	0.163		0.148	0.165	
rs72703204	CELSR2	exon	1.20×10^{-03}	0.148	0.164		0.149	0.166	
rs115967373	MYBPHL	3'	1.59×10^{-03}	0.193	0.402		0.190	0.328	
rs35358959	PSRC1	exon	1.66×10^{-03}	0.194	0.407		0.192	0.331	
rs41306199	MYBPHL	exon	1.76×10^{-03}	0.187	0.400		0.183	0.325	
rs657420	PSRC1	5'	2.00×10^{-03}	0.173	0.386		0.183	0.359	
rs585362	CELSR2	5'	2.39×10^{-03}	0.172	0.392		0.175	0.384	
rs41279716	CELSR2	intron	2.43×10^{-03}	0.101	0.162		0.097	0.135	
rs17645031	MYBPHL	3'	2.61×10^{-03}	0.193	0.500		0.190	0.413	
rs17584208	MYBPHL	3'	2.79×10^{-03}	0.193	0.523		0.190	0.434	
rs655246	MYBPHL	3'	5.01×10^{-03}	0.135	0.402		0.139	0.374	
rs413582	SORT1	3'	6.72×10^{-03}	0.093	0.282		0.096	0.239	
rs661278	SORT1	intron	6.88×10^{-03}	0.077	0.202		0.081	0.172	
rs76186504	MYBPHL	exon	6.93×10^{-03}	0.074	0.202		0.073	0.164	
rs4970843	SORT1	intron	7.91×10^{-03}	0.092	0.304		0.095	0.259	
rs4246520	CELSR2	5'	0.01	0.005	0.041	Co-dependent; $P<0.05$	0.002	0.043	Co-dependent; $P<0.05$
rs590502	MYBPHL	intron	0.02	0.102	0.496		0.106	0.476	
rs75430477	SORT1	intron	0.02	0.072	0.312		0.078	0.269	
rs41279714	CELSR2	intron	0.02	0.152	0.761		0.152	0.703	
rs116149872	SORT1	intron	0.02	0.104	0.549		0.103	0.471	
rs116611120	SORT1	intron	0.02	0.104	0.559		0.103	0.480	
rs115796838	SORT1	intron	0.02	0.105	0.566		0.104	0.487	

rs72646560	SORT1	exon	0.02	0.105	0.573		0.104	0.493	
rs114191770	SORT1	intron	0.02	0.104	0.566		0.103	0.487	
rs11102964	CELSR2	5'	0.02	0.084	0.460		0.084	0.431	
rs72703223	SORT1	intron	0.02	0.002	0.045	Co-dependent; P<0.05	0.002	0.035	Co-dependent; P<0.05

Legend. Conditioned SNP text high-lighted in red, SNPs in strong LD (r^2>0.8) with diminished *P*-val association (or Multi-collinearity) are highlighted in yellow, SNPs with moderate LD (0.5<r^2<0.8) with diminished *P*-val association are highlighted in amber; SNPs in weak LD (0.2<r^2<0.5) with diminished *P*-val association are highlighted in green; SNPs in very weak LD (0.05<r^2<0.2) with diminished *P*-val association are highlighted in dark green; SNPs showing co-dependent *P*-val association are highlighted in purple (i.e. only part of the association seen with this SNP is driven by the SNP being conditioned on); SNPs showing an independent association (P<0.05) are highlighted in blue (i.e. *P*-value lower than that seen in the unconditioned analysis) with conditioned SNP).

3.3.6 Overall conclusions from the LDL-C and CAD refinement

As a consequence of the IBCv2 50K HumanCVD beadchip GRAPHIC LDL-C and Metabochip CAD fine-mapping association studies, four lead SNPs ($r^2 \approx 1$) have emerged as the most likely causal candidate SNP(s), but there are a further six correlated SNPs in strong LD, $r^2 > 0.91$, which could still play a causal role. Below in **Figure 3.4** and **Figure 3.5** are two regional association plots from each study analysed, i.e. GRAPHIC and Metabochip respectively. These figures indicate more precisely the genomic location of the lead candidate four SNPs, and some additional local vicinity SNPs (inclusive of the six proxy SNPs in strong pairwise LD $r^2 > 0.91$) within ~5kb upstream or downstream of the lead SNP.

It can be seen clearly that there are four SNP with the strongest *P*-value LDL-C association, clustered together within and just beyond the 3'UTR of *CELSR2,* for the GRAPHIC study (see **Figure 3.4**). All but one of the four candidate SNPs identified during the GRAPHIC LDL-C association analysis (i.e. not rs629301) are the most strongly associated SNPs for CAD (see **Figure 3.5**). However, given the almost perfect pairwise LD ($r^2 \approx 1$) that rs629301 shares with the other three lead candidate SNPs (i.e. rs7528419, rs12740374 and rs646776), it is most likely that a slight genotype calling error has occurred for the SNP rs629301 in the CAD association Metabochip analysis. The additional more weakly associated SNPs are located within *CELSR2*, or just beyond the 3'UTR of *PSRC1*. (N.B. one additional SNP rs611917 in moderate LD, i.e. $r^2 = 0.56$, indicated by a green dot in **Figure 3.4** and **3.5**, is as strongly associated as those SNPs in strong LD $r^2 > 0.94$ in GRAPHIC, but seems to drop off in Metabochip, again this could be due to a slight genotype calling error).

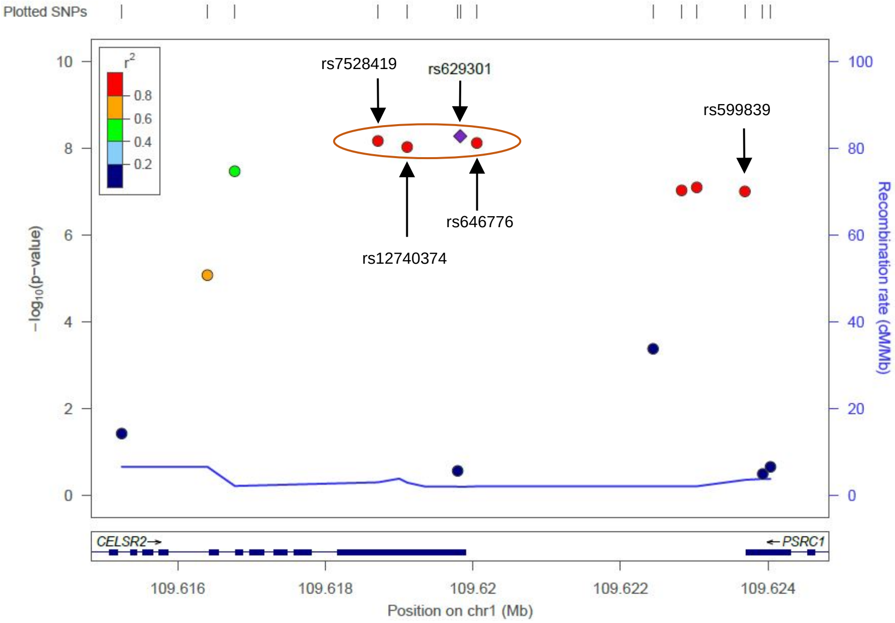


Figure 3.4. Focussed regional association plot for LDL-C in the GRAPHIC study

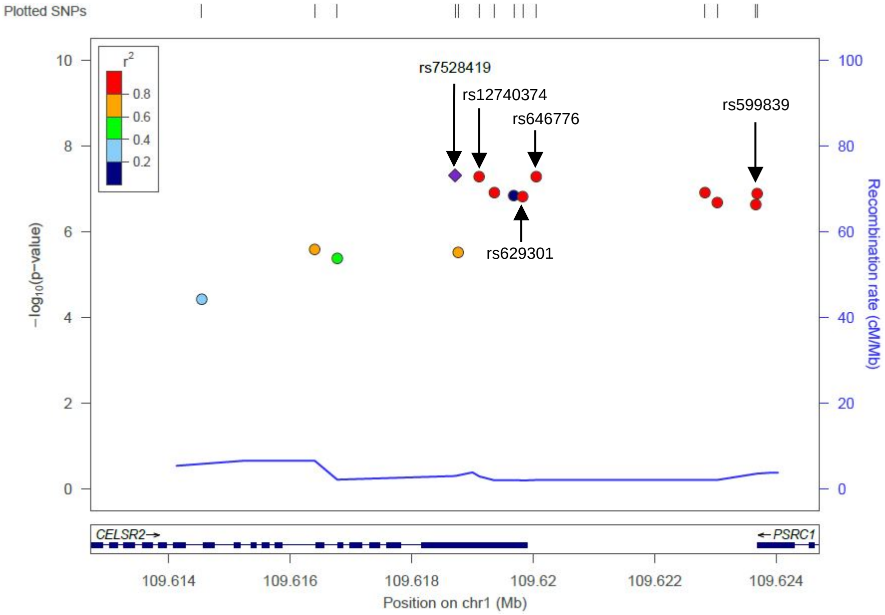


Figure 3.5. Focussed regional association plot for CAD in the Metabochip study.

3.3.7 In silico bioinformatics results

In follow-up to the fine-mapping localisation of the four lead SNPs (rs7528419, rs12740374, rs629301 and rs646776), I then performed a focussed *in silico* analysis to capture any further evidence that would help identify the most likely causal SNP. The types of *in silico* bioinformatics analyses, methods and resources used, and main findings are shown in a large summary **Table 3.7**, and refer to the following **Figures 3.6A** and **3.6B** and **Tables 3.8, 3.9,** and **3.10**.

Table 3.7. Summarised *in silico* assessment of ten correlated candidate 1p13.3 locus SNPs using ENCODE, TFBS and miRNA-TS.

Type of Analysis	Methods and Resources	Main Findings
ENCODE Analysis: *In silico* assessment of the encyclopaedia of DNA elements (ENCODE) - a project to identify all functional elements in the human genome sequence. **Rationale:** The purpose for looking at the ENCODE data is to identify transcriptional regulatory regions that align with or close to the four candidate SNPs and any SNPs nearby that are also in strong pairwise LD (r^2>0.9).	**Resource:** ENCODE data was accessed through the UCSC (University of California Santa Cruz) Genome Browser **Methods:** **DNase-HS**: DNaseI hypersensitivity sites identified by DNaseI nuclease activity and next generation genome wide sequencing, signifies general chromatin accessibility, following binding of trans-acting factors in place of a canonical nucleosome, and indicates locations of active *cis*-regulatory sequences (Crawford et al. 2006). **FAIRE-seq**: Formaldehyde assisted isolation of regulatory elements (FAIRE), followed by next generation sequencing is a method to isolate and identify nucleosome-depleted or -destabilised regions of the genome (Giresi et al. 2007). A low variability technique for any cell type as an alternative to DNase-seq. FAIRE-seq is more sensitive to distal regulatory regions and DNase-seq is more sensitive to promoter regions, but generally they cross-validate each other. Along with DNaseI HS, FAIRE has led to the discovery of functional regulatory elements that include enhancers, silencers, insulators, promotors, locus control regions and novel elements. Sensitivity can be greater than DNase-seq with deeper sequencing.	**Visualisation of ENCODE analysis results:** See **Figures 3.6A and 3.6B** for details **Key findings:** 1) A strong H3K4Me1 signal indicative of an enhancer region spans all four lead candidate SNPs. 2) A strong H3K27Ac signal indicative of transcriptional regulation is located in the vicinity of candidate SNPs - rs7528419 and rs12740374. 3) There are no H3 histone marks beyond the 3'UTR region of *PSRC1*, suggesting that rs599839 and two other SNPs in strong LD with this SNP are not likely to be causal via transcriptional regulation. 4) A CTCF insulator region is situated midway between the 3' flanking region of both *CELSR2* and *PSRC1*, this is supported by tight signal peaks from multiple ENCODE tracks in multiple cell types via DNaseI HS, FAIRE, TFBS ChIP-seq and DNaseI footprinting. This region does not contain any candidate SNPs.

Type of Analysis	Methods and Resources	Main Findings
ENCODE Analysis: *…Continued*	**Methods** (*Continued):* **ChIP-seq**: Chromatin immunoprecipitation (ChIP) with antibodies specific to a protein, followed by sequencing of the precipitated DNA is a method to identify the specific location of proteins that are directly or indirectly bound to genomic DNA. By identifying the binding location of sequence-specific transcription factors (Euskirchen et al. 2007), general transcription machinery components, and chromatin factors such as H3 histone modifications (methylation or acetylation), which influence gene expression by regulating how accessible chromatin is to transcription (Bernstein et al. 2005). Specific histone marks are identified throughout the genome using ChIP-seq and include H3K4Me1 (enhancer associated), H3K4Me3 (promoters – active or poised to be activated) and H3K27Ac (indicates regulatory regions of transcription). Resolution of ChIP-seq is about 200bp. **DNase-DGF:** base pair resolution DNaseI digital footprinting makes use of *'double hit'* DNaseI HS sites DNase-seq data, by identifying narrow depleted regions of high depth enriched deep sequencing reads mapped in the forward and reverse direction. This method enables the detection of reliable protein-binding footprints (Sabo et al. 2006). **Vertebrate Multiz Alignment:** Comparative genomics cross-species conservation across 46 species.	**Key findings** (C*ontinued):* 5) Low resolution DNaseI HS, FAIRE and TFBS ChIP-seq show their highest peak signals align with or near to SNP rs12740374 in multiple cell types. However, there are still signals from all these tracks that align with rs646776, and another highly correlated proxy rs660240 (This SNP is situated between rs12740374 and rs629301). As for rs7528419 and rs629301 there is little evidence of signal alignment with these tracks. Again rs599839 and 3 other nearby SNPs show no ENCODE signal evidence. 6) High resolution DNAseI footprinting shows very strong alignment with rs12740374 across multiple cell types, a little evidence of alignment with rs646776 and rs660240, but no evidence of alignments with rs7528419, rs629301 and rs599839 (and SNPs nearby). 7) Of the TFBS ChIP-seq signals that align strongly with rs12740374, there are two particularly important transcription regulatory factors. These include Pol2 (RNA Polymerase II) and TBP (TATA box protein) in HepG2 and HeLa cells. Given this locus is a LDL-C associated locus the liver cell association is of particular interest due to its role in cholesterol metabolism and degradation.

Type of Analysis	Methods and Resources	Main Findings
TFBS analysis: Transcription factor binding site (TFBS) predictions using web based tools. **Rationale:** Transcription factor binding sites (TFBS) are short, sequence specific, genomic DNA target regulatory regions that are recognised by transcription factors. The role of the transcription factor protein is to either enhance or repress transcription, and this may occur alone, or as part of a multicomponent complex. The TFBS maybe canonical or non-canonical, and will recruit transcription factors with varying degrees of affinity, depending on the sequence of nucleotides. As regards putative function for the CAD/LDL-C candidate 1p13.3 locus SNPs, it is important to see whether any of	**Resource:** JASPAR website: http://jaspar.genereg.net/ (Portales-Casamar et al. 2010). MatInspector (version 8.02, Feb 2010 released) accessed via the Genomatrix software suite, GmbH website: http://www.genomatix.de/index.html **Methods**: ***JASPAR*** is an open access database resource for eukaryote DNA TFBS profiles, typically modelled as position specific score matrices (PSSM), also known as position weight matrices (PWM) analogous to binding energy affinity (Stormo 2000; Maerkl & Quake 2007). The JASPAR CORE *Vertebrata* (2010 release) database consists of a high quality, non-redundant set of 129 PSSM profiles, derived from literature published experimentally proven nucleotide sequence binding to transcription factors in eukaryotes. The JASPAR database resource is complemented by a web interface for browsing, searching, subset selection and *fasta* sequence analysis utility that is ideal for assessing regulatory regions of genomic sequence. The output score results are based on differences in similarity between the PSSM model binding affinity scores for each SNP allele (i.e. major *vs* minor) for the same TFBS on the same orientation DNA strand.	**Visualisation of TFBS analysis results:** See **Tables 3.7** and **3.8** for details **Key findings:** 1) Putative TFBS predicted by JASPAR for the GRAPHIC study lead SNP rs629301 risk allele (T) include several TF, but the three main ones with the largest binding affinity score differences (>4.0; 18.2-28.4%) were for a disruption of MEF2, and a creation of YY1 and GATA2 binding sites. Interestingly, the MatInspector TFBS predictions for rs629301 concurred with JASPAR for MEF2 and YY1, but MatInspector also showed the risk allele (T) created a MARCA3 binding site too. 2) The Metabochip study lead SNP rs7528419 risk allele (A) creates a putative Elk1 binding site, using JASPAR TFBS prediction software, but a less impressive score difference was observed between the risk and minor alleles (2.0; 6.6%). The MatInspector TFBS software observed no binding affinity matrix score >80% for rs7528419. 3) The first of two additional proxy (LD: $r^2 \approx 1$) SNPs rs12740374 disrupted four TFBS: AML1, HLF, SPIB and CEBP in the presence of the risk allele (G) using JASPAR; each had a large

Type of Analysis	Methods and Resources	Main Findings
TFBS analysis *(Continued):* the SNPs are located with a TFBS, and whether the major or minor allele creates or disrupts the binding of a particular transcription factor, which may in turn modify transcriptional regulation. To this end, there are many predictive databases of canonical TFBS that can be assessed for homology with the genomic DNA sequence that includes the candidate SNP variant of interest, to look for differential binding affinity.	**Methods** *(Continued):* ***MatInspector*** is a commercial well-established, web--based tool (first released in 1995). It utilises a matrix database called MatBase (Matrix family library version 8.2, released Jan 2010) that consists of a large library of 1166 PWMs within 386 family descriptions for TFBS. A sequence analysis utility enables assessment of specific genomic sequence for TFBS at the four candidate SNPs of interest. Specificity is improved by incorporating a conservation profile and four base core region, so that a matrix similarity score can be calculated as a measure of conservation at any given nucleotide (Cartharius et al. 2005; Quandt et al. 1995).	**Key findings** *(Continued):* matrix score difference when compared with the minor allele (>3.8; 10.7-16.4%). For CEBP the binding site disruption occurred on both the –ve and +ve strands, but more strongly for the minus strand. MatInspector was found to concur with the CEBPA and HLF observations, but again there were additional putative TFBS created by the rs12740374 risk allele (G), EKLF and GCM1. 4) The final proxy SNP (LD: $r^2 \approx 1$) rs646776 risk allele (A) creates a putative FOXL1, and disrupts MXF_1-4 and YY1 binding sites; and large binding affinity scores were observed (>4.1; 19.1-21.1%). However, none of these TFBS predictions were observed with MatInspector. Indeed, for MatInspector the rs646776 risk allele (A) creates a putative PAX3, and disrupts a P53 binding site.

Type of Analysis	Methods and Resources	Main Findings
MiRNA-TS analysis: MicroRNA target site (MiRNA-TS) predictions using web based tools. **Rationale:** The four lead candidate SNPs associated with CAD/LDL-C identified by the Metabochip and GRAPHIC studies respectively, are located within or just beyond the 3'UTR of *CELSR2*. It is well-known that the 3'UTR is the main target of microRNA, for the purpose of mRNA decay or degradation as a means of controlling gene expression levels. Therefore, it is important to assess each of these four SNPs to see if they disrupt or create putative miRNA-TS.	**Resource:** 1) Microcosm website: http://www.ebi.ac.uk/enright srv/microcosm/htdocs/targets/ v5/, using miRanda 3.0, Oct 2007, with 851 miRNAs supplied by miRBase (http://www.mirbase.org/, release 15, April 2010) (Griffiths-Jones et al. 2008; Griffiths-Jones et al. 2006) 2) TargetScan 5.1 website: http://www.targetscan.org/vert_50/, using release data June, 2010 (Friedman et al. 2009) 3) DIANA-microT3 website: http://diana.cslab.ece.ntua.gr/microT/, using the miRBase (N.B. release date version is unclear) (Maragkakis et al. 2009); 4) MicroRNA.org website: http://www.microrna.org/microrna/home.do, using miRanda-mirSVR, last updated Nov 2010, with 1100 human miRNAs (Betel et al. 2008; Betel et al. 2010); 5) MicroSNiPer website: (http://cbdb.nimh.nih.gov/microsniper/, using 939 human miRNAs from the miRBase, release 15, April 2010. [N.B. a SNP allele specific miRNA-TS prediction program].	**Visualisation of miRNA-TS analysis results:** See **Table 3.9** for details **Key findings:** 1) **O**nly MicroSNiPer and MicroRNA.org web-based tools were capable of detecting 5' seed region miRNA-TS. The miRNA-TS identified through 5' seed regions were unique for the risk and minor allele across all four candidate SNPs, for creating or disrupting miRNA-TS. 2) All miRNA-TS prediction programs detected 3' region miRNA-TS, except TargetScan 5.1. Moreover, in many cases the 3' region detected miRNA-TS was detected by both the risk and the minor allele of a given candidate SNP. 3) There was no overlap between the 5' seed region and 3' region identified miRNA-TS 4) At least one independent miRNA-TS was predicted to be created or disrupted in the presence of the risk allele at each of the four candidate SNPs tested.

Type of Analysis	Methods and Resources	Main Findings
MiRNA-TS analysis: *…Continued*	**Methods** *(Continued):* The above miRNA-TS *in silico* web-based tools were utilised to identify putative miRNA-TS that could be created or disrupted by one or more of the four candidate 1p13.3 locus SNPs. Key genomic DNA:microRNA sequence alignment features are taken into account when scoring target prediction sites, and these include: 1) 5' end Watson-Crick complementarity at the so-called 'seed region' of microRNAs (miRNAs), allowing only 1 or 2 mismatches (such as a G:U 'wobble') usually between nucleotide positions 2-7 (but also between 2-8 or 2-9); 2) assessment of 3'UTR potential target sites for conservation in orthologous transcripts from other species, requiring detection at the same position in a cross-species orthologous 3'UTR alignment by a miRNA from the same family, in at least two species; 3) 3' complementarity within nucleotides 12-17, (particularly 13-16); 4) relative position within the 3'UTR and mRNA local context AU rich motifs near to the miRNA-TS, which affect secondary mRNA structure (Friedman et al. 2009; Grimson et al. 2007).	**Key findings** *(Continued):* 5) The rs7528419 risk allele (A) creates two potential miRNA-TS hsa-miR-1254/ 4293, but disrupts hsa-miR-3940. 6) The rs12740374 risk allele (G) creates a potential miRNA-TS hsa-miR-29a, but disrupts hsa-miR-3908. 7) The rs629301 risk allele (T) creates four potential miRNA-TS hsa-miR-10a/ 10b/495/ 1273c, but disrupts ten potential miRNA-TS hsa-miR-1279/3145/138-2*/ 454*/518a-3p/518b/518c/518d-3p/424*/ 7-2*. 8) The rs646776 risk allele (A) creates five potential miRNA-TS hsa-miR-486-3p/ 541/1285/3689a-3p/4255, but disrupts four potential miRNA-TS hsa-miR-376c/ 885-3p/940/1827. 9) The take home message is that none of these miRNA-TS results point to an obviously identifiable causal SNP.

Legend. Summary breakdown of three types of *in silico* analyses: 1) ENCODE, 2) TFBS and 3) miRNA-TS. Inclusive of methods and resources, and the main findings for each analysis.

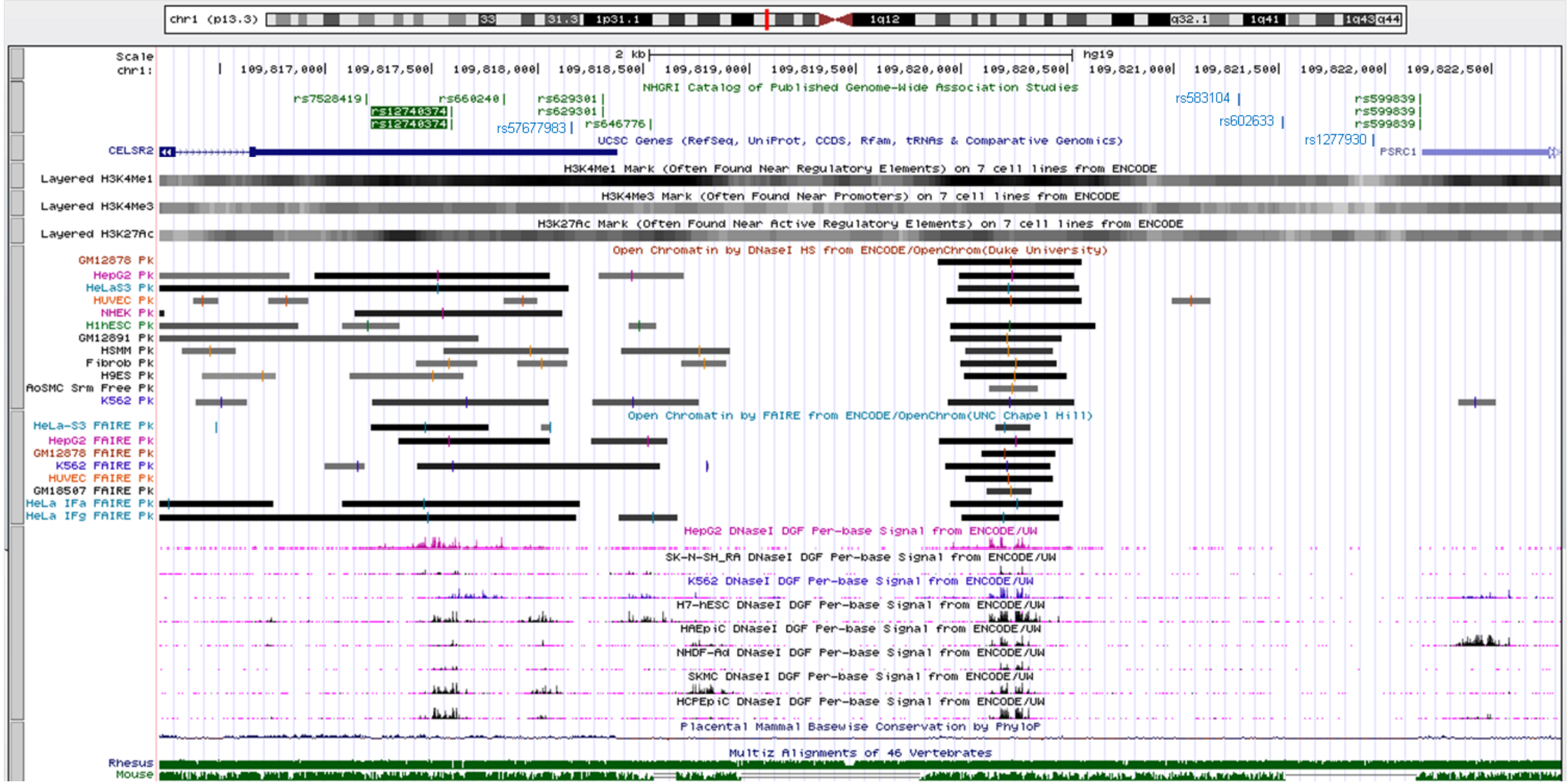

chr1 (p13.3)
33
31.3
1p31.1
1q12
q32.1
1q41
1q43
q44
Scale
chr1:
2 kb
hg19
109,817,000
109,817,500
109,818,000
109,818,500
109,819,000
109,819,500
109,820,000
109,820,500
109,821,000
109,821,500
109,822,000
109,822,500
NHGRI Catalog of Published Genome-Wide Association Studies
rs7528419
rs660240
rs629301
rs12740374
rs12740374
rs629301
rs57677983
rs646776
rs583104
rs602633
rs1277930
rs599839
rs599839
rs599839
UCSC Genes (RefSeq, UniProt, CCDS, Rfam, tRNAs & Comparative Genomics)
CELSR2
PSRC1
Layered H3K4Me1
H3K4Me1 Mark (Often Found Near Regulatory Elements) on 7 cell lines from ENCODE
Layered H3K4Me3
H3K4Me3 Mark (Often Found Near Promoters) on 7 cell lines from ENCODE
Layered H3K27Ac
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
Open Chromatin by DNaseI HS from ENCODE/OpenChrom(Duke University)
GM12878 Pk
HepG2 Pk
HeLaS3 Pk
HUVEC Pk
NHEK Pk
H1hESC Pk
GM12891 Pk
HSMM Pk
Fibrob Pk
H9ES Pk
AoSMC Srm Free Pk
K562 Pk
Open Chromatin by FAIRE from ENCODE/OpenChrom(UNC Chapel Hill)
HeLa-S3 FAIRE Pk
HepG2 FAIRE Pk
GM12878 FAIRE Pk
K562 FAIRE Pk
HUVEC FAIRE Pk
GM18507 FAIRE Pk
HeLa IFa FAIRE Pk
HeLa IFg FAIRE Pk
HepG2 DNaseI DGF Per-base Signal from ENCODE/UW
SK-N-SH_RA DNaseI DGF Per-base Signal from ENCODE/UW
K562 DNaseI DGF Per-base Signal from ENCODE/UW
H7-hESC DNaseI DGF Per-base Signal from ENCODE/UW
HAEpiC DNaseI DGF Per-base Signal from ENCODE/UW
NHDF-Ad DNaseI DGF Per-base Signal from ENCODE/UW
SKMC DNaseI DGF Per-base Signal from ENCODE/UW
HCPEpiC DNaseI DGF Per-base Signal from ENCODE/UW
Placental Mammal Basewise Conservation by PhyloP
Multiz Alignments of 46 Vertebrates
Rhesus
Mouse

Figure 3.6A. Cis-regulatory ENCODE track alignments from for Chr1p13.3 LDL-C and CAD-associated locus candidate SNPs. ENCODE alignment tracks using the UCSC Genome Browser HG19 assembly are shown for Histone marks, DNaseI HS, FAIRE, DNaseI DGF and Multiz alignment for conservation. Track 1) shows the '10' most significantly associated SNPs for CAD and LDL-C, the published GWA study SNPs are coloured green and non-published SNPs are blue. Track 2) shows UCSC gene annotation for the 3'UTR and 3' flanking regions of genes *CELSR2* and *PSRC1*; Track 3) shows peaks for H3K4me1, H3K4me3 and H3K27Ac histone marks, the strongest signal is indicated by dark shading; Track 4) shows open chromatin by FAIRE-seq, dense peaks are highest when dark and vertical lines indicate the peak summits; Track 5) shows open chromatin DNaseI HS, dense peaks are highest when dark and vertical lines indicate peak summits; Track 6) shows single nucleotide resolution DNaseI footprint signals indicative of DNA-bound regulatory protein; Track 7) shows Rhesus Monkey and Mouse conservation alignments.

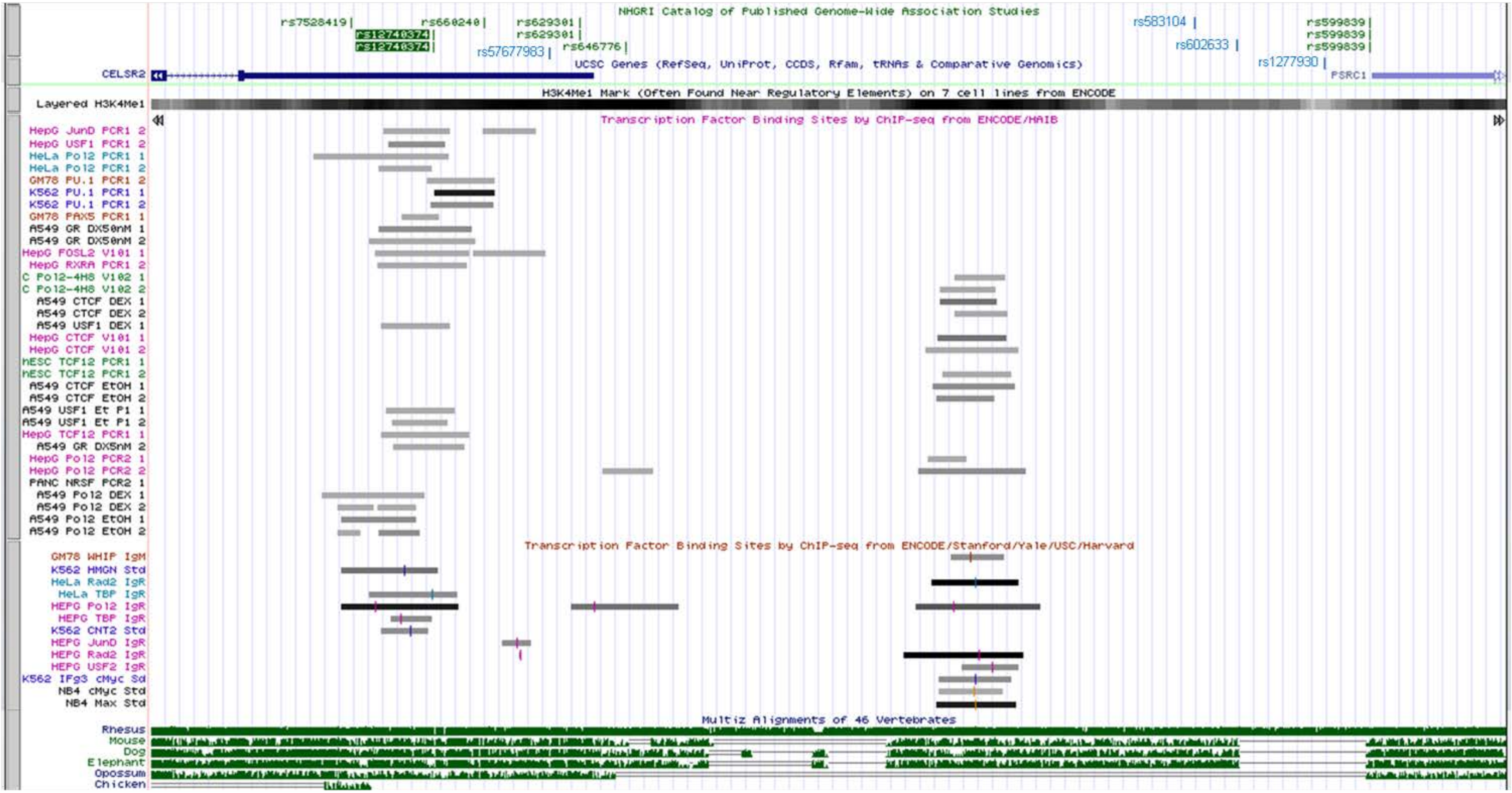


Figure 3.6B. Cis-regulatory ENCODE track alignments from for Chr1p13.3 LDL-C and CAD-associated locus candidate SNPs. ENCODE alignment tracks within the UCSC Genome Browser HG19 assembly focussed on TFBS ChIP-seq. Track 1) shows the '10' most significantly associated SNPs for CAD and LDL-C, the published GWA study SNPs are coloured green and non-published SNPs are blue. Track 2) shows UCSC gene annotation for the 3'UTR and 3' flanking regions of genes *CELSR2* and *PSRC1*; Track 3) shows peaks for enhancer H3K4me1 histone marks, the strongest signal is indicated by dark shading; Track 4) shows Transcription factor binding sites identified by ChIP-seq from the HAIB, dense peaks are highest when dark; Track 5) shows TFBS ChIP-seq from SYUH, dense peaks are highest when dark and vertical lines indicate the peak summits; Track 6) shows Multiz alignments of 6 species that show at least some small level of conservation, as far as opposum.

Table 3.8. Putative transcription factor binding sites from the JASPAR *in silico* prediction software

SNP ID	Transcription factor	Strand	Risk allele	Risk allele score	Minor allele	Minor allele score	Score diff	Rel diff	Con
rs629301	FOXD1	+	GTAAATAT	9.335 (89.3%)	GTAAATAG	8.188 (85.5%)	1.147	3.9	yes
	FOXC1	+	ATATGGTA	2.916 (74.5%)	ATAGGGTA	4.501 (83.4%)	-1.59	8.9	yes
	GATA-2	-	CCATA	3.363 (80.9%)	CCCTA	-1.523 (52.5%)	4.886	28.4	yes
	GATA-3	+	TGGTAG	3.693 (80.3%)	GGGTAG	1.816 (72.4%)	1.877	7.9	yes
	MEF2	-	ATATTTACAG	7.980 (81.1%)	CTATTTACAG	12.140 (89.8%)	-4.16	8.6	yes
	NKX3-1	-	ATATTTA	8.333 (88.3%)	CTATTTA	5.596 (78.5%)	2.737	9.8	yes
	TBP	+	CTGTAAATATGGTAG	5.391 (79.5%)	CTGTAAATAGGGTAG	6.941 (82.5%)	-1.55	3.1	yes
	Yin-Yang	-	ACCATA	7.027 (93.1%)	ACCCTA	2.899 (72.0%)	4.128	21.1	yes
rs7528419	Elk-1	-	CAGCCTGAAG	8.974 (91.0%)	CAGCCCGAAG	6.966 (84.4%)	2.008	6.6	no
rs12740374	AML-1~	+	CCTGAGGGT	3.278 (73.5%)	CCTGAGGTT	7.124 (84.4%)	-3.846	10.9	no
	HLF	+	GGGTGCTCAATC	4.435 (74.1%)	GGTTGCTCAATC	8.525 (84.8%)	-4.09	10.7	no
	SPI-B	-	TGAGCAC	0.140 (64.7%)	TGAGCAA	4.940 (81.1%)	-4.8	16.4	no
	cEBP	-	TTGAGCACCCTC	5.893 (79.2%)	TTGAGCAACCTC	10.060 (92.1%)	-4.167	12.9	no
	cEBP	+	GTGCTCAATCAA	3.308 (71.2%)	TTGCTCAATCAA	7.228 (83.3%)	-3.92	12.1	no
rs646776	FOXL1	+	TGGACATA	6.554 (90.2%)	TGGACATG	2.279 (70.8%)	4.275	19.4	no
	GATA-3	+	ACATAG	3.966 (81.4%)	ACATGG	1.190 (69.8%)	2.776	11.6	no
	MZF_1-4	+	TAGGCA	0.676 (61.6%)	TGGGCA	4.846 (80.7%)	-4.17	19.1	no
	Yin-Yang	-	CCTATG	1.107 (62.8%)	CCCATG	5.236 (83.9%)	-4.129	21.1	no

Legend: Transcription factors high-lighted in yellow show a strong 'binding affinity' score difference; the risk allele and minor allele provide the base nucleotide sequence for the matrices and the major and minor SNP alleles are high-lighted in red; risk allele score and minor allele score show the binding affinity score and in brackets the matrix similarity % score for each allele; Score diff – this is the difference between the risk allele score and minor allele score; Rel diff – this is the % similarity difference between the risk and minor allele; Con=conserved; Source data – *in silico* TFBS prediction web-based tool JASPAR (http://jaspar.genereg.net/).

Table 3.9. MatInspector Version 8.02 (release date February 2010) TFBS predictions for locus chr1p13.3

1p13.3 locus SNP	Matrix Family	Detailed Family Information	Matrix (Library v8.2)	from	to	anchor^	Strand	Matrix Sim$	Sequence Red=Ci-Vector >60%# Capitals=Core Sequence*
Rs629301_A	V$YY1F	Activator/repressor binding to transcription initiation site	V$YY2.01	14	34	24	(+)	0.961	cgctaCCATatttacagcaac
Rs629301_A	V$RUSH	SWI/SNF related nucleophosphoproteins with a RING finger DNA binding motif	V$SMARCA3.01	17	27	22	(+)	0.963	taCCATattta
Rs629301_C	V$MEF2	MEF2, myocyte-specific enhancer binding factor	V$SL1.01	14	36	25	(+)	0.895	cgctaccCTATttacagcaacaa
Rs629301_C	V$MEF2	MEF2, myocyte-specific enhancer binding factor	V$MEF2.03	15	37	26	(-)	0.901	gttgttgctgtaaATAGggtagc
Rs12740374_G	V$KLFS	Krueppel like transcription factors	V$EKLF.01	32	48	40	(+)	0.956	gccctgaGGGTgctcaa
Rs12740374_G	V$GCMF	Chorion-specific transcription factors with a GCM DNA binding domain	V$GCM1.01	33	43	38	(-)	0.915	caCCCTcaggg
Rs12740374_T	V$PARF	PAR/bZIP family	V$HLF.01	37	53	45	(-)	0.864	cttgattgaGCAAcctc
Rs12740374_T	V$CEBP	Ccaat/Enhancer Binding Protein	V$CEBPA.01	38	52	45	(-)	0.967	ttgattgaGCAAcct
Rs646776_A	V$PAX3	PAX-3 binding sites	V$PAX3.02	37	55	46	(-)	0.860	cctgTCCCtctgcctatgt
Rs646776_G	V$P53F	p53 tumor suppressor	V$P53.02	34	56	45	(-)	0.928	gcctgtccctctgccCATGtcca

Legend: $The TFBS **matrix similarity** is calculated as follows: A perfect match to the matrix gets a score of 1.00 (each sequence position corresponds to the highest conserved nucleotide at that position in the matrix), a "good" match to the matrix usually has a similarity of **>0.80**. Mismatches in highly conserved positions of the matrix decrease the matrix similarity more than mismatches in less conserved regions. In MatInspector, a green background in the matrix similarity column marks a similarity above optimized (i.e. a "good" match), a red background marks a similarity below optimized; ***The core sequence** of a matrix is defined as the (usually 4) highest conserved, consecutive positions of the matrix, shown as CAPITAL LETTERS; #The **Ci-vector** (consensus index vector) for the matrix represents the degree of conservation of each position within the matrix. The maximum **Ci-value** of 100 is reached by a position with total conservation of one nucleotide, whereas the minimum value of 0 only occurs at a position with equal distribution of all four nucleotides and gaps. Base pairs marked red are important, i.e. they appear in a position where the matrix exhibits a high conservation profile (ci-value > 60); **^Anchor**: All matrices in a family are of the same (uneven) length and have an anchor position assigned which is the centre position of the matrix. This assures that matrices of a family match exactly at the same position (Cartharius et al. 2005; Quandt et al. 1995).

Table 3.10. Putative MicroRNA target sites for the lead chr1p13.3 SNPs within the 3’UTR of *CELSR2*

SNP	3'UTR	Position	Risk allele	Minor allele	Targets the 3' region	Targets the 5' seed region	Prediction Program
rs7528419	CELSR2	521	A			hsa-miR-1254; 4293	MicroSNiPer
				G		hsa-miR-3940	MicroSNiPer
rs12740374	CELSR2	919	G		hsa-miR-218; 636; 649	hsa-miR-29a	MicroSNiPer
			G		hsa-miR-218; 636		DIANA microT
			G		hsa-miR-218; 636		MicroRNA.org
			G		hsa-miR-218		Microcosm
				T	hsa-miR-218; 636; 649	hsa-miR-3908	MicroSNiPer
rs629301	CELSR2	1635	T		hsa-miR-548o; 548a-3p; 495; 1273c	hsa-miR-10a; 10b	MicroSNiPer
				G	hsa-miR-548o; 548a-3p	hsa-miR-1279; 3145	MicroSNiPer
				G	hsa-miR-518a-3p; 518b; 518c; 518d-3p		Microcosm
				G	hsa-miR-424*;7-2*	hsa-miR-138-2*; 454*	MicroRNA.org
rs646776	Just beyond 3’UTR CELSR2	1859	A		hsa-miR-490-5p; 592; 939; 3689a-3p; 4255	hsa-miR-486-3p; 541; 1285	MicroSNiPer
				G	hsa-miR-490-5p; 592; 939	hsa-miR-376c;885-3p; 940; 1827	MicroSNiPer

Legend. SNP – four lead chr1p13.3 SNPs; 3’UTR – main target for miRNAs; Position – location within the 3’UTR; Targets the 3’ region – miRNA nucleotides 12-17 bind to SNPs located within genomic DNA; Targets the 5’ seed region – main anchor for miRNA binding to SNPs located within genomic DNA. The SNP alleles high-lighted in green are the UCSC ref seq alleles that the five microRNA target site (miRNA-TS) prediction programs used. The miRNAs high-lighted in red have a mirSVR (mRNA down regulation) score below -0.1 (the MicroRNA.org threshold for true positive target sites).

3.3.8 Overall conclusions from the in silico bioinformatics refinement

The ENCODE *in silico* bioinformatics analyses indicate through DNase-seq, FAIRE-seq and ChIP-seq that rs12740374 is the SNP with the strongest overall evidence for being the causal variant (see **Table 3.7, Figures 3.6A** and **3.6B**), which best explains LDL-C and CAD association at the 1p13.3 locus. However, two further adjacent SNPs rs660240 and rs646776 (i.e. within or just beyond the *CELSR2* 3'UTR respectively), cannot be entirely ruled out, because they too showed signals for potential regulatory protein binding (see **Figures 3.6A** and **3.6B**). Another important observation was that the ENCODE data alignments to rs599839 (the initially identified CAD associated SNP), and three other SNPs adjacent to the *PSRC1* 3'UTR, showed no evidence for regulatory protein binding (see **Figures 3.6A** and **3.6B**).

Further non-ENCODE, *in silico* TFBS prediction analyses of the four post fine-mapping lead candidate SNPs, identified supporting evidence that three of the four were potentially causal (see **Tables 3.7, 3.8 and 3.9**). This is based on weak evidence for rs7528419 predictions, but much stronger evidence for the other three SNPs (rs12740374, rs629301 and rs646776) in terms of PSSM binding energy affinity differences of >3.8 (8.6-28.4%), observed between the risk and minor allele for multiple TFBS. Given that the intermediate phenotype for CAD association is (most likely to be) LDL-C, another interesting observation is that two liver specific TFBS predictions, i.e. HLF (hepatic leukaemia factor) and CEBPA (CCAAT/enhancer binding protein alpha), show that the rs12740374 risk allele disrupted binding energy affinity, when compared to the minor allele. This finding may have some additional significance given the rs12740374 ENCODE ChIP-seq alignment, potentially near to two key transcription regulatory proteins Pol2 (RNA Pol II) and TBP in hepatocellular carcinoma cells (HepG2) (see **Figure 3.6B**).

As for the miRNA-TS analysis (see **Tables 3.7** and **3.10**), it was not possible to distinguish between the four candidate SNPs in terms of which was the more likely to be causal. This is because all four SNPs showed binding feasibility to 5' seed or 3' regions of many different miRNAs depending on whether the risk or minor allele were present.

A further assessment of the 1000 genomes project, did not identify any additional SNPs in almost perfect pairwise LD ($r^2 \approx 1$) with the four candidate SNPs already identified by fine-mapping. Only rs660240 and rs57677983 (now known as rs3832016) were identified by Metabochip study fine-mapping to be in very strong pairwise LD (r^2=0.94), located within the *CELSR2* 3'UTR between rs12740374 and rs629301.

Therefore, based on the strongest evidence when considering both the fine-mapping and *in silico* analyses together, the most likely 1p13.3 locus causal SNP is rs12740374. However, one cannot totally rule out the other three lead fine-mapping candidate SNPs (rs646776, rs629301 and rs7528419 - in order of strength of evidence), or two other strongly correlated adjacent SNPs rs660240 and rs57677983 based on their location, and *in silico* evidence for rs660240. Indeed, it is even feasible that more than one of these SNPs are causal.

3.3.9 Planned functional experiments

Based on the fine-mapping and *in silico* bioinformatics analyses, it is most likely that the *CELSR2* 3'UTR, or a region just beyond, harbours the casual SNP - by creating or disrupting a regulatory protein sequence specific genomic DNA binding site (i.e. a TFBS). In fact, there is even strong evidence to suggest that one particular SNP rs12740374, could be the causal SNP. Nonetheless, there are a further five nearby proxy SNPs with weaker evidence (rs646776, rs629301, rs7528419, rs660240 and rs57677983) that could also be causal. In order to prove categorically, which SNP(s) are truly causal, it is important to carry out functional wet-lab experiments, to test whether one of the SNPs has influence over a TFBS. Therefore, the first experiment I planned to carry out was an electrophoretic mobility shift assay (EMSA). This would crudely assess whether two short DNA oligonucleotides (i.e. 20-30bp) that contain either the risk allele or the minor allele for each candidate SNP, would show differential binding to a protein (or protein complex) from an extracted cell lysate, such as HepG2. A liver cell line would be selected, because liver was previously shown to harbour a very strong *cis*-acting eQTL for the proxy ($r^2 \approx 1$) SNP rs646776 within the 1p13.3 locus (Schadt et al. 2008; Kathiresan et al. 2009). Super shift EMSA experiments can then be performed, if any of the SNP containing oligonucleotides shows clear differential binding affinity, using putative transcription factor antibodies, selected on the basis of the *in silico* analysis findings, e.g. well-known liver enriched transcription factors could be targeted first. A useful (more subtle) complementary experiment that I planned to perform was a luciferase expression assay. Firstly, I would amplify via polymerase chain reaction (PCR) a *CELSR2* 3'UTR genomic DNA fragment that contains either a major or a minor six candidate SNP haplotype (i.e. an approximate 2kb fragment from the stop codon of *CELSR2* to a location just beyond rs646776). Secondly, I would sub-clone these two haplotypes into a luciferase promoter vector (e.g. pGL3-promoter, Promega) in the cloning region just beyond the *luc* reporter gene stop codon, in both the *plus*-strand and *minus*-strand orientation. Thirdly, I would then transfect the four luciferase vector constructs into a liver cell line, such as HepG2 cells. Finally, I would measure luciferase activity to see if there is a difference in luciferase gene expression. If a difference in expression is identified, then site-directed mutagenesis can be performed on each candidate SNP, to help pinpoint the causal SNP.

Unfortunately, before I was able to make a start on these functional experiments, an article was published by Musunuru and colleagues that identified the causal SNP as being rs12740374, the same SNP identified through my fine mapping and *in silico* analyses. They showed that rs12740374 risk allele (G) causes a disruption to the binding of transcription factor CEBPA, in liver cells, which decreases gene expression of *SORT1*, which in turn affects intracellular apolipoprotein B (apoB) processing, resulting in increased levels of serum LDL-C, and the subsequent increase in CAD risk (Musunuru et al. 2010).

3.4 Discussion:

This chapter has high-lighted the challenges of moving from a common disease GWA study locus signal, to a definitive causal SNP and mechanism. The earliest GWA studies have made this challenge all the more difficult, because the coverage of the genome was far from comprehensive. Indeed, the signal identified at a given locus may be the true causal SNP, or it may merely be a marker for the true causal SNP. Moreover, the initial signal may be a marker for more than one phenotypic association within the same locus (i.e. pleiotrophic).

In order to overcome these early GWA study shortcomings, it is important to refine the signal, to tease out the true causal SNP. This can be achieved by genotyping all other SNPs within the near genomic location, for instance within 50kb upstream or downstream of the lead signal. This will hopefully capture all haplotype block signals within the same locus, and will also capture all SNPs, in at least weak pairwise LD, i.e. $r^2>0.2$, with the initially identified lead signal SNP. Additionally, the narrowing of the signal is helped all the more, if the initially identified lead SNP is also associated with a putative quantitative, intermediate phenotype that is likely to be causal.

The rs599839 signal refinement undertaken for the 1p13.3 locus, was able to take advantage of a gene-centric custom designed, HumanCVD beadchip (Illumina, Inc.), also known as the IBCv2 50K beadchip (Keating et al. 2008). This beadchip was specifically designed to fine-map the most important candidate CAD-associated genes using a pairwise LD ($r^2\geq0.8$) tagging threshold. The HumanCVD beadchip was ideally suited for fine-mapping the rs599839 signal, because it included comprehensive coverage of all independent (n=94) SNPs that span four nearby genes *CELSR2*, *PSRC1*, *MYBPHL* and *SORT1*. The previous chapter identified LDL-C as an intermediate quantitative phenotype in the GRAPHIC study, a family-based study representative of the general population. Therefore, the GRAPHIC study was again used for fine-mapping purposes using the HumanCVD beadchip. Four SNPs (rs629301, rs7528419, rs646776 and rs12740374) in almost perfect pairwise LD ($r^2\approx1$), led by rs629301, were most strongly associated with LDL-C, and surpassed the initial signal SNP rs599839. The pairwise LD between rs629301 and rs599839 was still very strong at $r^2=0.95$. These four candidate SNPs span 1338bp, and are located within, or just beyond the 3'UTR of *CELSR2*. The nearest SNP rs646776 is approximately 3.6kb from rs599839, which resides just beyond the 3'UTR of *PSRC1*.

Yet further fine-mapping was made possible for the CAD-associated signal rs599839, through the Metabochip study. Here a 200K custom-designed iSELECT array, called Metabochip (Illumina, Inc.) was in part designed to replicate 6,222 CARDIoGRAM (*P*<0.01) discovery SNPs, and also to fine-map thirty CAD-associated loci (Schunkert et al. 2011; Voight et al. 2012). Moreover, the fine-mapping was able to include some additional 1000 genomes pilot study SNPs, before utilising a pairwise LD ($r^2 \geq 0.8$) tagging threshold. The Metabochip study described here consisted of an extended WTCCC case-control CAD cohort, which included WTCCC CAD (n=3219) and 1958BC (n=6000) control subjects to enhance statistical power (Samani et al. 2007). As a consequence, the Metabochip was fine-mapped with n=438 SNPs for the same four genes as were genotyped for GRAPHIC, but with some additional flanking sequence. Four SNPs (rs7528419, rs646776, rs12740374 and rs629301) in almost perfect pairwise LD ($r^2 \approx 1$), led by rs7528419, were the most strongly associated with CAD. These were the same SNPs that were most strongly associated with LDL-C in the GRAPHIC study fine-mapping analysis. Again the four lead SNPs have a very strong pairwise LD ($r^2=0.95$) with the original lead SNP signal rs599839.

The next step was to ensure that the four lead SNPs were indicative of an independent signal. This was achieved by performing a conditional analysis that included the lead SNP in the analysis model. This was performed on both the fine-mapping studies, to see if the associated SNPs discovered were the only independent signals identifiable within the 1p13.3 locus. Conditional analyses were performed on the lead SNP rs629301 in the GRAPHIC study and rs7528419 in Metabochip study, but also for the original lead signal SNP rs599839, in both studies. There were no independent signals (beyond the four lead SNPs) amongst any SNPs in strong LD between the stop codons of genes *CELSR2* and *PSRC1*. The only independent signals identified beyond the stop codons of *CELSR2* and *PSRC1* were small and would not reach Bonferroni corrected significance thresholds for multiple testing, the strongest of these was in the 5’ flanking region of *CELSR2 for* rs4246519 *(*$0.001<P<0.01$).

The two complementary LDL-C and CAD association fine-mapping studies have reduced the number of likely causal SNP possibilities to four. However, the role of fine-mapping is now limited at this stage, because it cannot distinguish between these four candidate SNPs, which are in almost perfect pairwise LD ($r^2 \approx 1$). Therefore, although the location has been narrowed down to a location within, or just beyond the 3’UTR of *CELSR2*, it is still unclear as to which SNP is truly causal. One possible technique that could help to narrow the signal further is trans-ethnic fine-mapping. This requires genotyping and pairwise LD data for non-Caucasian populations. Here African populations can be useful as these are much older populations, which could help breakdown the pairwise LD and pinpoint the causal SNP. To this end, Yoruba in Ibadan, Nigeria (YRI) 1000 genomes pilot study, pairwise LD data (available via the web-based tool SNAP, Broad Institute), for proxy SNPs shows the separation of the four lead candidate SNPs, and another highly correlated Metabochip study SNP rs660240, into two haplotype blocks separated by LD of r^2=0.6. The first haplotype block includes: rs7528419 and rs12740374, and second includes: rs646776, rs629301 and rs660240 (of note, a third haplotype block also exists that includes rs599839). In the appendix are two Haploview web-based tool generated images showing the haplotype blocks for HapMap (phase 3) Caucasians of Northern & Western European ancestry, Utah (CEU) and African ancestry in Southwest USA (ASW), showing the breakdown in LD for the two populations (see ***Appendix*** **6.2, Figures S3.1** and **S3.2**, respectively). However, in order to use this trans-ethnic fine-mapping technique more effectively, one would ideally need to genotype these SNPs in similar studies to those described here, but in appropriately selected non-Caucasian populations, i.e. a general population with LDL-C measurements, or a CAD case-control study.

The WTCCC recently published a paper by Maller and colleagues, in parallel with this work that tried to fine-map the 1p13.3 locus, along with 15 other loci using a custom-designed iSELECT Illumina (Inc.) beadchip (Wellcome Trust Case Control Consortium et al. 2012). Prior to chip design, they first re-sequenced 32 unrelated CEU individuals across 16 loci (regions) from type 2 diabetes (T2D), CAD and autoimmune Graves' disease, to identify any additional fine-mapping SNPs. Their lead discovery SNP at the 1p13.3 locus was rs3832016 (aka rs57677983 identified in the Metabochip study analysis of this chapter), and this was the only additional SNP they genotyped as a consequence of their re-sequencing experiment. One other SNP rs660240, matched the significant effect size of rs3832016 ($r^2 \approx 1$ in the Metabochip study data) in the article, and only three other SNPs were mentioned, two SNPs rs599839 and rs611917 showed a lower but significant effect size, and one SNP rs12740374 failed to genotype. This suggests that Maller *et al.* discovered no further SNPs beyond those discussed in this chapter (Wellcome Trust Case Control Consortium et al. 2012).

The next steps I took to identify the causal SNP(s) at the 1p13.3 locus, were to try and predict function, through the utility of *in silico* bioinformatics, web-based tools. On the basis of a strong eQTL at the 1p13.3 locus, identified through the association of rs646776 (one of the four lead LDL-C and CAD candidate SNPs identified through fine-mapping) with the regulation of liver tissue gene expression in three local vicinity genes: *CELSR2*, *PSRC1* and *SORT1* - the function of the causal SNP is likely to affect transcriptional regulation of gene expression. Therefore, the *in silico* web-based tools employed in this chapter include ENCODE, and predictions for TFBS and miRNA-TS; as these will help identify *cis*-regulatory DNA binding elements.

ENCODE accessible through the UCSC genome browser is a particularly useful tool, as it aligns to the genome, large libraries of *in vivo* DNase-seq, ChIP-seq and FAIRE-seq - experimentally identified (captured) protein-DNA binding across multiple primary tissues and cell types. These techniques are utilised to identify *cis*-regulatory elements for protein binding to regions of the genome, where the chromatin is open and active. DNaseI is a nuclease that can only cut genomic DNA that is exposed, i.e. where the protein is not binding. This technique can be used with next generation sequencing to identify hypersensitivity sites, and also through deep coverage and dual directional sequencing, identify base-pair level resolution footprints of where the regulatory protein is binding. A newer technique - FAIRE, is capable of identifying DNA-protein binding of nucleosome depleted, or nucleosome disrupted regulatory genomic regions - by making use of the preferential formaldehyde cross-linking to nucleosome bound DNA over nucleosome unbound DNA. Of note, DNase-seq and FAIRE-seq tend to complement each other in most cases, but both are essential because the techniques have their own particular strengths in terms of DNA-protein binding recognition, i.e. FAIRE-seq is more sensitive to distal regulatory regions, and DNase-seq is more sensitive to promoter regions. Another method of wide ranging utility is ChIP-seq, as this can identify specific chromatin-associated protein types, for example particular epigenetic histone modifications, transcription machinery proteins and TFBS, across the genome. The limitation here is that you need to know the protein that is bound to the DNA in order to identify it, as the method of capture is immuno-precipitation, after a specific antibody has bound the protein. Histone marks are very useful as they identify a range of chromatin active and inactive states (and differentiate between promoter and enhancer regions), and TFBS can be compared to consensus PSSM predictions identified, by both *in vivo* and *in vitro* experiments. Each technique in itself has its own limitations, for instance DNase-seq, FAIRE-seq and ChIP-seq have a resolution of about 200bp. However, when considered together with DNaseI footprinting, which has single base-pair resolution, ENCODE alignment tracks become very powerful as a prediction of function; and will continue to become more powerful with the addition of extra ChIP-seq TFBS in multiple cell types under various stimulated conditions (e.g. using INFγ stimulation for STAT1). Therefore, ENCODE is a very powerful tool that can give important clues for specific functional genomics designed experiments, such as EMSA and luciferase reporter assays.

It is clear from the results section that ENCODE can identify a strong likelihood for regulatory DNA-protein binding sites, at any of the four lead candidate SNPs, and those in their close vicinity, i.e. rs660240 and rs57677983. For example, the first key finding is that all six SNPs are aligned with an enhancer region as identified through a wide dense H3K4Me1 histone mark signature across seven cell types. This was further supported through intermittent dense H3K27Ac, and weak H3K4Me3 histone mark signatures. The reason for all six SNPs being potentially causal is because individually, DNase-seq, ChIP-seq and FAIRE-seq only have a resolution accuracy of approximately 200bp. However, by considering all three low resolution methodologies together, in multiple cell types, in conjunction with the high (base-pair level) resolution of DNaseI footprinting - the ENCODE evidence strongly points towards one of the lead fine-mapping SNPs, namely rs12740374, as being the most likely causal SNP for the putative enhancer. Yet another, potentially significant finding through the TFBS ChIP-seq evidence, is that two key transcription machinery proteins RNA Pol II and TBP, align most strongly with rs12740374 in HepG2 and HeLa cells. Indeed, given the previous eQTL evidence provided by Schadt *et al.* (Schadt et al. 2008), and the known relevance of liver with regards to cholesterol metabolism and degradation – the evidence points all the more strongly towards a role for rs12740374 and hepatocytes, for transcriptional regulation at the 1p13.3 locus. Nevertheless, one cannot fully rule out any of the other five proxy (r^2>0.95) SNPs, because they are in such close proximity to rs12740374, within or just beyond the 3'UTR of *CELSR2*. Indeed, two of the nearby SNPs rs646776 ($r^2 \approx 1$) and rs660240 (r^2=0.95), show clear evidence of DNA-protein binding even at a base-pair level through DNaseI footprinting, and they have ChIP-seq TFBS alignments in HepG2 cells; in fact RNA Pol II aligns with both rs646776 and rs629301 too. Another key discovery through the assessment of ENCODE, is that there is no evidence of DNA-protein regulatory binding whatsoever, for the initially observed lead SNP rs599839, or indeed for the three other SNPs (rs602633, rs583108 and rs127903) in almost perfect LD ($r^2 \approx 1$) with rs599839, which reside close to the 3'UTR of *PSRC1*.

The evidentiary support of ENCODE, shows the power of the fine-mapping approach for the 1p13.3 locus. However, when one looks to the ChIP-seq TFBS and transcriptional machinery binding predictions, one is left with a rather unclear picture as to which transcription factor might be functional. This is because the *in vivo* experiments carried out here are merely of a snap-shot in time, and not truly representative of the dynamic conditions that will affect the cell's cycle under a multitude of stimuli. This is made particularly clear, when one makes use of TFBS and miRNA-TS prediction tools, where the rates of false-positive discovery run very high. For example, any one of the four lead candidate SNPs identified by fine-mapping, can be a canonical or non-canonical TFBS or miRNA-TS, for a multitude of transcription factors or miRNAs. This is because evolution has provided us with the redundancy, i.e. adaptability, to cope with diverse environmental changes, which are in turn influenced by each individual's genetic predisposition (both genomic and epigenetic). For three of the four lead candidate fine-mapping SNPs (i.e. rs12740374, rs629301 and rs646776, but not rs7528419) there are several TFBS that show large binding energy affinity differences, between the risk allele and minor allele. However, none of the putative TFBS predictions agree with the ENCODE ChIP-seq identified TFBS for these SNPs. This does not necessarily mean the putative TFBS are incorrect, it simply shows there has not been a ChIP-seq experiment performed that captured one of these TFBS. At this time, and even to date the TFBS ChIP-seq data resource, is rather in its infancy with regards to capturing all possible *in vivo* TFBS, across different cell types. It should be noted that there is yet another interesting discovery to be followed up, as a consequence of the putative TFBS predictions. This is with regards to the fact that the risk allele of rs12740374 disrupts HLF and CEBPA putative TFBS, and these are well-known liver enriched transcription factors. This again provides added evidentiary support to the notion that rs12740374 could be the lead SNP, as suggested by the ENCODE data.

Finally, as regards the *in silico* bioinformatics analyses, miRNA-TS predictions were assessed, because the 3'UTR is a known preferential target for miRNAs. However, the analysis as regards identifying a causal SNP was nigh on impossible to decipher, because for all SNPs the risk allele and minor allele created miRNA-TS via the 5' seed region for different miRNAs. In addition, for those SNPs that were situated within the miRNA-TS 3' region, the change of SNP allele made no difference to the binding affinity of several miRNA-TS.

Unfortunately, even with a stack of evidence pointing towards one or more SNPs causing the LDL-C and CAD, the *in silico* bioinformatics alone is not going to pinpoint the true mechanism. Therefore, it is imperative that functional genomics experiments are performed as were mentioned earlier, i.e. EMSA that shows differential DNA-protein binding, and luciferase reporter assays that show differential gene expression, to be able to identify a definitive causal SNP(s).

Given the empirical liver eQTL data, the location of the lead candidate SNPs identified by fine-mapping, and the *in silico* bioinformatics findings (e.g. histone mark signatures for an enhancer and base-pair level DNaseI footprinting) at the 1p13.3 locus - it is most likely that the *CELSR2* 3'UTR, or a region just beyond, harbours the casual SNP. This would be most likely to occur via a SNP change that creates or disrupts, a regulatory protein sequence specific genomic DNA binding site (such as a TFBS). The specific functional experiments that I was going to perform with regards to six putative causal SNPs (i.e. the four lead candidate LDL-C/CAD associated SNPs, and the two proxy SNPs located between rs12740374 and rs629301) included: 1) an EMSA, this method enables you to test the binding affinity of extracted protein cell lysate from a cultured liver cell line (e.g. HepG2), with two short oligonucleotides (~20-30bp), both containing a putative functional SNP of interest, but differing for the major or minor allele; and 2) a luciferase reporter assay, this method enables the assessment of a major or minor haplotype genomic sequence (i.e. ~2kb, inclusive of the same six putative candidate LDL-C/CAD associated SNPs, assessed using EMSA), which can be sub-cloned into a luciferase vector (e.g. pGL3-promoter (Promega), just beyond the stop codon of the *luc* gene, which in turn can be transfected into cultured HepG2 cells, lysed after 48 hours and measured for luciferase activity, i.e. expression of luciferase. Any binding or expression changes would be suggestive of a transcriptional regulatory protein factor complex. Unfortunately, my planned functional experiments were made somewhat redundant, because another research group published an article in *Nature*, describing how rs12740374 is the causal SNP (Musunuru et al. 2010).

The first step Musunuru *et al.* took towards identifying the causal SNP was to assess the LDL-C subclasses using two alternatively measured lipoprotein subclass cohorts: the Malmo¨ Diet and Cancer Study – Cardiovascular Cohort (MDC-CC) by ion-mobility (Musunuru et al. 2009) and the Pharmacogenomics and Risk of Cardiovascular Disease (PARC) study by gradient gel electrophoresis (Siri-Tarino et al. 2009). They found that rs646776, was most highly associated with changes in the very small LDL (LDL-VS) lipoprotein subclass (20% increase in major vs. versus minor allele homozygotes with P=1.1x10^{-11} in MDC-CC; 37% increase with P=8.0x10^{-11} in PARC); and progressively smaller changes were observed with larger LDL subclasses. The MDC-CC study is similar to the GRAPHIC study, as they are both recruited from community based healthy subjects, whereas the PARC study was recruited for moderately overweight males. In GRAPHIC, the strongest effect *P*-values were for NMR measured LDL-C subclasses D (P=0.005) and B (P=0.008), but both fell short of Bonferroni *P*-value corrected significance, as described in chapter 2. The highly significant LDL-VS subclass in MDC-CC is matched by the GRAPHIC LDL-C subclass B, but the major vs. minor allele homozygotes ratio in GRAPHIC, although in the same direction, is much smaller at 6%. The moderately significant LDL-L / IDL-S subclasses for MDC-CC seem to match the GRAPHIC LDL-C subclass D, both in terms of significance, and major vs. minor allele homozygotes ratio, i.e. ~6% (see **Table 3.11**).

Table 3.11. Comparison of MDC-CC/PARC vs. GRAPHIC study LDL-C subclasses

MDC-CC - Ion mobility

rs646776	N	LDL-VS 18.0-20.8nm	LDL-S 20.8-21.4nm	LDL-M 21.4-22.0nm	LDL-L 22.0-23.3nm	IDL-S 23.3-25.0nm	IDL-L 25.0-29.6nm
Maj Hom	2689	114	82.3	126	441	121	216
Min Hom	279	94.8	74.7	116	416	115	213
Ratio		1.2	1.1	1.09	1.06	1.05	1.01
***P*-value**		**$1.1\text{x}10^{-11}$**	0.03	0.02	0.0004	0.0002	0.24

PARC study - Gradient gel electrophoresis

rs646776	N	LDL-VS 18.0-20.8nm	LDL-S 20.8-21.4nm	LDL-M 21.4-22.0nm	LDL-L 22.0-23.3nm
Maj Hom	1196	14.4	17.1	26.4	56.3
Min Hom	75	10.5	14.2	24.5	58.4
Ratio		1.37	1.2	1.08	0.96
***P*-value**		**$8.0\text{x}10^{-11}$**	0.16	0.48	0.29

GRAPHIC study - Nuclear magnetic resonance

rs646776	N	LDL-A 16-19nm	LDL-B 19-21nm	LDL-C 21-22nm	LDL-D 22-25nm	LDL-E 25-30nm
Maj Hom	1220	16.14	22.94	24.93	25.88	29.29
Min Hom	102	15.49	21.64	23.45	24.21	26.77
Ratio		1.042	1.06	1.063	1.069	1.094
***P*- value**		0.38	0.008	0.03	0.005	0.03

Legend. Comparison of three studies that have used different methods for measuring and categorising LDL-C subclasses and their association with rs646776 (Musunuru et al. 2009; Siri-Tarino et al. 2009; Petersen et al. 2012). Ratio is for the major vs. minor allele homozygotes of rs646776, *P*-values in bold are still significant after multiple testing Bonferroni correction.

A more recent study by Petersen *et al.* using the same NMR measuring company LipoFIT® (as for GRAPHIC) in a healthy subject German study called KORA, also identified LDL-C subclass B as being significant (*P*=$1.46\text{x}10^{-5}$) for the proxy SNP rs629301, even at a Bonferroni corrected *P*-value (Petersen et al. 2012). The KORA study used GRAPHIC as a replication study (*P*=0.0095), which agreed with my findings for LDL-C subclass B (*P*=0.008). The KORA study did not show any other LDL-C subclasses to be significant after Bonferroni *P*-value correction, but did show LDL-C subclass E to be suggestive of significance (*P*=$5.97\text{x}10^{-4}$), there was no GRAPHIC replication performed for this finding (Petersen et al. 2012). Overall, the GRAPHIC and KORA NMR LDL-C subclass data agree with the MDC-CC and PARC studies, in terms of small dense LDL being the most significant subclass at the 1p13.3 locus, albeit at a much lower level in GRAPHIC. However, the data for the other LDL-C subclasses at the 1p13.3 locus are in disagreement, and are therefore inconclusive.

The second step, Musunuru *et al.* took towards identifying the causal SNP was to assess the eQTL association at rs646776 in an extended liver cohort and in two forms of adipose tissue (subcutaneous and omental), based on the association of the candidate SNP with LDL-C. The result of this analysis was a confirmation of the liver eQTL, but no eQTL was observed for adipose tissues (Musunuru et al. 2010). My approach as regards looking for an eQTL in other cell types was to make use of *in* silico data, made available through the eQTL Chicago Browser (http://eqtl.uchicago.edu/cgi-bin/gbrowse/eqtl/), this resource does not show adipose cell eQTL data, but *in silico* data is now available publicly via an eQTL Genevar software resource at the Sanger Institute (http://www.sanger.ac.uk/resources/software/genevar/).

The third step taken by Musunuru *et al.* was to perform a strategy initiated in this chapter, i.e. to fine-map the region using GWA study array data, so they could identify the shortest stretch of genomic DNA that would include all proxy SNPs, with the lead discovery SNPs rs599839/rs646776 (Samani et al. 2007; Kathiresan et al. 2008). The reason for doing this was so they could use two human bacterial artificial chromosomes (BACs), for the major and minor haplotypes (~6.1kb) that span the genome from the stop codon of *CELSR2* to the stop codon of *PSRC1*, for the purpose of sub-cloning into a Firefly luciferase vector pGL3-promoter (Promega), just beyond the *luc* gene in both orientations (Musunuru et al. 2010). This luciferase vector construct was then transfected into a cultured human liver carcinoma cell line, Hep3B. The result was a much higher expression of luciferase level in the minor haplotype than the major haplotype, and was in accord with the liver eQTL. Musunuru *et al.* then truncated the 6.1kb insert to a smaller 2.1kb insert inclusive of the *CELSR2* 3'UTR, and used site-directed mutagenesis to switch each potentially causal SNP in the minor haplotype from the minor allele to the major allele. This brought about the observation that rs12740374 alone, was causing the increased expression in the minor haplotype (Musunuru et al. 2010).

Musunuru *et al.* then used a trans-ethnic fine-mapping approach to show that rs12740374 was the lead SNP for LDL-C association in a large African American association study (n=9000) (Musunuru et al. 2010).

Having pinpointed rs12740374 as the putative causal SNP, Musunuru *et al.* identified the transcription factor that was disrupted by the major risk allele for rs12740374, as being CEBPA (Musunuru et al. 2010). They recognised, as I had that CEBPA was a potential TFBS, and that this transcription factor was liver specific. They then went on to compare binding of CEBPA to short oligonucleotide sequences containing either the major or the minor allele of rs12740374 using an EMSA with cultured human hepatoma HepG2 cell extracted, nuclear cell lysate proteins (Musunuru et al. 2010). Musunuru *et al.* showed that the rs12740374 minor allele sequence when in a DNA-protein complex could shift as far as a known consensus CEBPA oligonucleotide sequence, whereas the rs12740374 major allele shifted very little. The further addition of two different CEBPA specific antibodies resulted in impaired protein binding (Musunuru et al. 2010). Some additional confirmatory experiments were performed. For example, luciferase assays, where the sub-cloned minor haplotype was subjected to site-directed mutagenesis of other core consensus CEBPA transcription factor nucleotides; which confirmed the requirement of these mutated nucleotides for transcription of the minor haplotype. Another confirmatory experiment was one where they showed CEBPA binds to rs12740374 using *in vivo* ChIP (Musunuru et al. 2010).

The use of the luciferase reporter assay and an EMSA to identify the causal SNP rs12740374, were two tools that I had been planning to use myself. However, Musunuru *et al.* used these tools in a way that differed from what I had planned to do. I had planned to first use the EMSA assay to see if one of the lead candidate SNPs (or two nearby proxy SNPs) within the *CELSR2* 3'UTR, would indicate clearly that one of the SNPs showed a differential binding to a protein, or protein complex extracted from a liver cell lysate. My approach would have been a more crude approach, but potentially quicker if rs12740374 had been the only SNP to show differential binding, and it could have shown the *CELSR2* 3’UTR was the relevant genomic sequence insert to use in the luciferase assay sub-cloning step too. Musunuru *et al.* first used a larger 6.1kb insert that included the intergenic region between *CELSR2* and *PSRC1* and the 3'UTR of *PSRC1*, rather than focus purely on the 3'UTR of *CELSR2*. However, this approach was probably performed by Musunuru *et al.* to be able to prove that a portion of genomic sequence, which included only rs599839 and three further proxy SNPs ($r^2 \approx 1$) were not capable of altering luciferase activity. Once the rs12740374 was selected as the most probable causal SNP, the first logical transcription factor to try would have been a liver specific transcription factor, so I would have tried CEBPA and HLF first, before attempting others.

Musunuru *et al.* then went on to show through a series of human and mouse liver cell experiments that a non-coding genetic variant rs12740374 at locus 1p13.3 strongly affects LDL-C and CAD risk by liver specific transcriptional regulation of the *SORT1* gene via the creation or disruption of a TFBS for CEBPA, or other family members of the C/EBP family (Musunuru et al. 2010). This pathway discovery is of great clinical relevance as a therapeutic target, because it shows a 40% greater risk in the major allele homozygotes over the minor allele homozygotes. An effect size of this magnitude is on a par with common variants in *LDLR* and *PCSK9*, and is larger than the effects of common variants on *HMGCR* (the gene targeted by the statin drugs) (Samani et al. 2007; Myocardial Infarction Genetics Consortium et al. 2009). In addition, as this is a common variant with a minor allele frequency of ~22% in Caucasians, and also common in other ethnicities, i.e. Africans, Chinese, South Asians and Hispanics, it is a genetic risk variant with important global utility.

SORT1 (as described earlier in the introduction of this chapter) encodes for the human protein sortilin 1. In brief, sortilin 1 is a type 1 transmembrane protein, which acts as a multi-ligand receptor, capable of binding to a number of unrelated ligands that engage in many cellular pathways, but it cannot act as a signalling receptor. Sortilin 1 is mainly found in the trans-Golgi network (TGN) and early endosomes, but it can also be found on the plasma membrane. The main role of sortilin 1 is to transport ligands between the TGN and early endosomes or lysosomes. However, it can also bind and internalise ligands via receptor-mediated endocytosis, across the cell membrane. The mature protein is achieved in the Golgi apparatus following furin cleavage of a propeptide; and consists of a large luminal domain, a transmembrane domain, and a short

C-terminal cytoplasmic tail. The luminal domain shares homology at two cysteine-rich regions in the yeast Vps10p sorting receptor protein, and is the region that binds to ligands (Petersen et al. 1997).

From GWA studies, the 1p13.3 *SORT1* locus shows the strongest association of any genetic loci to serum lipoproteins (Teslovich et al. 2010). One of the key roles of sortilin 1 in the liver, in relation to its effect on LDL-C, is its involvement in degrading nascent VLDL particles before they are secreted, and henceforth reducing levels of serum LDL-C (Musunuru et al. 2010). However, the effect of sortilin 1 on VLDL synthesis, determined by Musunuru *et al.* overexpression and knockdown studies in mouse models, was contradicted by Kjolby *et al.*; who reported that sortilin 1 increased the levels of VLDL, and therefore increased the levels of serum LDL-C (Kjolby et al. 2010). The discrepancy is probably due to genetic differences between the mouse models. In particular, Kjolby *et al.* used a whole body *Sort1*$^{-/-}$ knockout mouse rather than a liver specific siRNA knockdown of humanised knockout mouse models. This could have led to compensatory lipid pathway effects on lipid circulation, perhaps through adipose tissues or the small intestine. However, the findings by Musunuru *et al.* are far more convincing as the true novel mechanistic lipoprotein pathway.

The transcriptional effect on *SORT1* clearly has an effect on the expression of two other genes at the 1p13.3 locus in hepatocytes, namely *CELSR2* and *PSRC1*. However, this effect does not seem to be related to serum levels of LDL-C or CAD, e.g. overexpression of *Psrc1* in a mouse model had no effect on cholesterol levels (Musunuru et al. 2010). These two genes were described in the introduction of this chapter. Briefly, the specific function of *CELSR2* is unknown, but it seems to be another transmembrane multiple ligand receptor involved in many pathways, including cell adhesion at laminin sites of the basement membrane - an important site for arteries and hence the AS plaque. However, *PSRC1* is a p53 mediated cell growth suppressor that destabilises the microtubule during mitosis. In the case of two other nearby genes *MYBPHL* and *SARS,* there is little known about their biological function.

As a consequence of the Musunuru *et al.* studies identifying the functional causal SNP at chromosomal locus 1p13.3, my supervisor and I felt that there was little point in continuing to focus on this locus. Therefore, as a means of enabling me to undertake functional experiments that I was planning to perform, I switched my attentions to another interesting locus that had emerged from the GWA study meta-analysis of CAD (inclusive of the early WTCCC and GerMIF GWA studies). The locus selected and the studies undertaken are described in the next chapter.

4. Chapter 4

Functional studies of the chromosome 13q34 CAD locus that includes the COL4A1 and COL4A2 collagen genes

4.1 Introduction:

As described in Chapter 1, combined analysis of the WTCCC CAD study and the GerMIF study identified seven loci associated with CAD (Samani et al. 2007). Following this a number of further individual GWA studies, with replication of principle findings, increased the number of loci associated with CAD to twelve (Myocardial Infarction Genetics Consortium et al. 2009; Erdmann et al. 2009; Tregouet et al. 2009). These initial studies, as well as those of other complex traits (Wellcome Trust Case Control Consortium 2007), showed that the risk associated with common variants studied using these arrays, was modest at around 1.1-1.4 per copy of a risk allele (Wellcome Trust Case Control Consortium 2007). Coupled with the penalty for multiple testing requiring very stringent *P*-values to be reached (P<5x10^{-8}), it was recognised that individual studies only had power to detect the strongest effects. For example, the WTCCC CAD study for a *P*-value threshold of P=5×10^{-7}, had 43% and 80% power to detect an odds ratio of 1.3 and 1.5, respectively, for SNP minor allele frequencies >5% (Wellcome Trust Case Control Consortium 2007). Therefore for many diseases and traits, groups with GWA study data rapidly coalesced into consortia, to leverage greater power (Zeggini et al. 2008; International Consortium for Blood Pressure Genome-Wide Association Studies et al. 2011; Lango Allen et al. 2010; Global Lipids Genetics Consortium et al. 2013).

For coronary disease the main consortium formed was the CARDIoGRAM consortium. CARDIoGRAM brought together 14 CAD GWA studies of Caucasian ancestry, comprising 22,233 cases and 64,762 controls (including the WTCCC CAD study and the GerMIF study). The CARDIoGRAM analysis was comprised of two stages: a meta-analysis of the GWA studies, and a replication in new cohorts of the most interesting SNP associations found in the meta-analysis. Importantly, prior to the meta-analysis each study was imputed, using one of three algorithms:

MACH (Li et al. 2010), IMPUTE (Marchini et al. 2007), or BIMBAM (Servin & Stephens 2007). The two purposes for imputation were: 1) to standardise any SNP differences between the genotyping arrays used in each study, and 2) to enable HapMap phase II imputed fine-mapping of approximately 2.2 million SNP genotypes. The meta-analysis stage generated 23 novel loci at a chosen pre-replication threshold of P<5x10^{-6}. The subsequent replication stage involved 58,623 subjects (29,894 cases and 28,729 controls) again of Caucasian descent. The CARDIoGRAM investigators used two pre-defined criteria, to define a validated locus - (i) a Bonferroni corrected P-value threshold of P<2.17x10^{-3} (i.e. 0.05/23) in the replication phase; and (ii) a combined meta-analysis discovery and replication of P<5x10^{-8} for genome-wide statistical significance (Preuss et al 2010). On this basis, the CARDIoGRAM meta-analysis identified 13 novel validated loci for CAD (see **Table 4.1**), as well as confirming 10 of the 12 previously reported loci (Schunkert et al. 2011). The 13 new loci showed risk allele frequencies ranging from 0.13 to 0.91, and were associated with a 6% to 17% increase in the risk of CAD per allele (see **Table 4.1**).

Three of the novel CARDIoGRAM loci were also associated with traditional risk factors. The rs964184 risk allele on chromosome 11q23.3 (gene region: *ZNF259*, *APOA5-APOA4-APOC3-APOA1*) was associated with increased LDL-C and decreased HDL-C (and previously with TG) (Kathiresan et al. 2009); the rs579459 risk allele on chromosome 9q34.2 (*ABO*) was associated with increased LDL-C and TC; and the rs12413409 risk allele on chromosome 10q34.32 (gene region: *CYP17A1-CNNM2-NT5C2*) was associated with hypertension (Schunkert et al. 2011).

One of the interesting novel CARDIoGRAM loci was that on chromosome 13q34, where the lead SNP (rs4773144) lay within intron 3 of the *plus*-strand orientated collagen type IV alpha 2 (*COL4A2*) gene, approximately 1kb downstream of the main *COL4A2* transcription start site (TSS(+1)); and within the 5' flanking, distal promoter region of the *minus*-strand orientated collagen type IV alpha 1 (*COL4A1*)

Table 4.1. Novel 13 loci discovered by CARDIoGRAM

Chr. Band	SNP	Gene(s) in region	Risk allele frequency (risk allele)	Meta-analysis		Replication		Combined analysis	
				***P*-val**	***N***	***P*-val**	***N***	**OR (95% CI)**	***P*-val**
1p32.2	rs17114036	*PPAP2B*	0.91 (A)	1.43×10^{-8}	80,870	3.18×10^{-12}	52,356	1.17 (1.13–1.22)	3.81×10^{-19}
6p21.31	rs17609940	*ANKS1A*	0.75 (G)	2.21×10^{-6}	83,997	1.18×10^{-3}	53,415	1.07 (1.05–1.10)	1.36×10^{-8}
6q23.2	rs12190287	*TCF21*	0.62 (C)	4.64×10^{-11}	78,290	3.25×10^{-4}	52,598	1.08 (1.06–1.10)	1.07×10^{-12}
7q32.2	rs11556924	*ZC3HC1*	0.62 (C)	2.22×10^{-9}	80,011	7.37×10^{-10}	54,189	1.09 (1.07–1.12)	9.18×10^{-18}
9q34.2	rs579459	*ABO*	0.21 (C)	1.16×10^{-7}	77,138	7.02×10^{-8}	46,840	1.10 (1.07–1.13)	4.08×10^{-14}
10q24.32	rs12413409	*CYP17A1, CNNM2, NT5C2*	0.89 (G)	1.47×10^{-6}	80,940	1.38×10^{-4}	48,801	1.12 (1.08–1.16)	1.03×10^{-9}
11q23.3	rs964184	*ZNF259, APOA5-A4-C3-A1*	0.13 (G)	8.02×10^{-10}	82,562	2.20×10^{-9}	52,930	1.13 (1.10–1.16)	1.02×10^{-17}
13q34	rs4773144	*COL4A1, COL4A2*	0.44 (G)	4.15×10^{-7}	77,113	1.31×10^{-3}	37,618	1.07 (1.05–1.09)	3.84×10^{-9}
14q32.2	rs2895811	*HHIPL1*	0.43 (C)	2.67×10^{-7}	63,184	4.59×10^{-5}	51,054	1.07 (1.05–1.10)	1.14×10^{-10}
15q25.1	rs3825807	*ADAMTS7*	0.57 (A)	9.63×10^{-6}	80,849	1.39×10^{-8}	48,803	1.08 (1.06–1.10)	1.07×10^{-12}
17p13.3	rs216172	*SMG6, SRR*	0.37 (C)	6.22×10^{-7}	57,235	2.11×10^{-4}	54,303	1.07 (1.05–1.09)	1.15×10^{-9}
17p11.2	rs12936587	*RASD1, SMCR3, PEMT*	0.56 (G)	4.89×10^{-7}	76,952	1.35×10^{-4}	52,648	1.07 (1.05–1.09)	4.45×10^{-10}
17q21.32	rs46522	*UBE2Z, GIP, ATP5G1, SNF8*	0.53 (T)	3.57×10^{-6}	83,867	8.88×10^{-4}	53,766	1.06 (1.04–1.08)	1.81×10^{-8}

Legend. Chr. Band – chromosome G-band; SNP – single nucleotide polymorphism; *P*-value – statistical significance of association; *N* – total number of cases and controls for CARDIoGRAM analysis first or second stage; OR - Odds Ratio, i.e. estimated effect size of each SNP risk allele copy on CAD (adjusted for age and sex); 95% CI – 95% confidence intervals; Meta-analysis – CARDIoGRAM first stage meta-analysis requiring a *P*<5x10^{-6}; Replication – CARDIoGRAM second stage replication analysis of first stage meta-analysis SNPs requiring a *P*<2.17x10^{-3} (i.e. a Bonferroni corrected *P*-value of 0.05/23); Combined analysis – CARDIoGRAM combined analysis of first stage meta-analysis and second stage replication requiring a *P*<5x10^{-8} (i.e. genome wide statistical significance).

gene, approximately 1kb upstream of the *COL4A1* TSS(+1). A regional ‘Manhattan’ plot for this locus is shown in **Figure 4.1**.

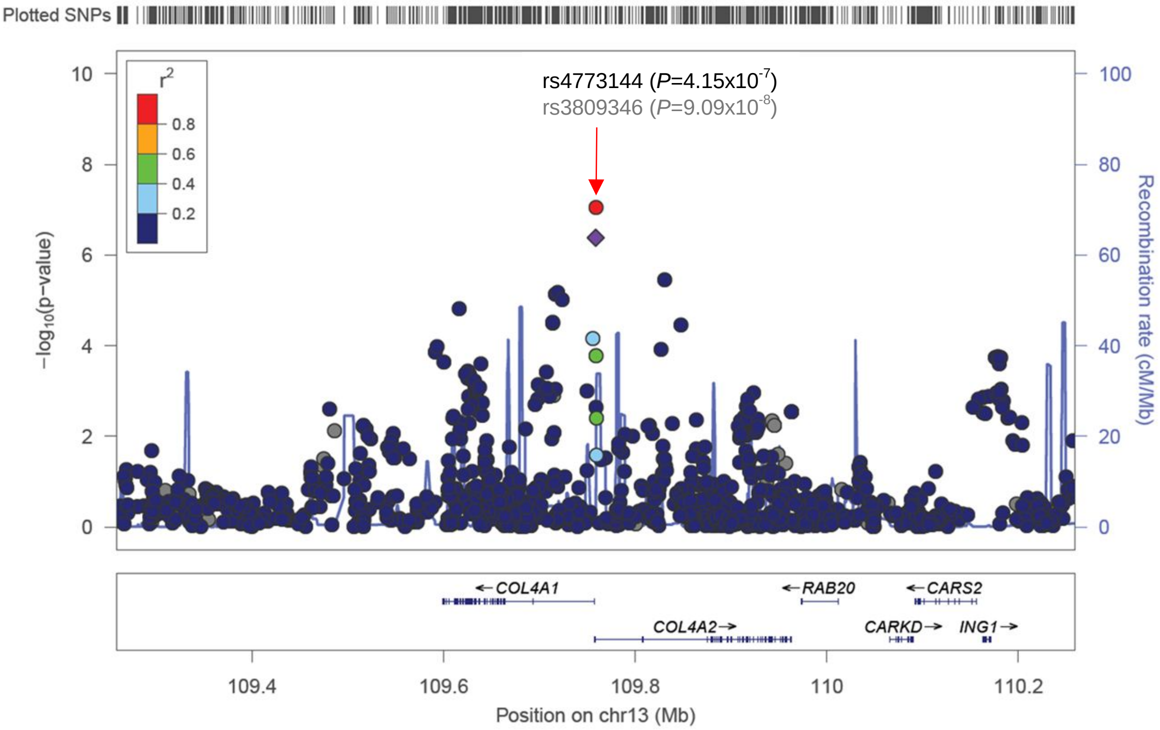


Figure 4.1. Regional CAD association plot for rs4773144 at chr13q34 for the CARDIoGRAM consortium meta-analysis.

Legend. The associations for the individual 2.2 million imputed SNPs used in the CARDIoGRAM analysis are plotted as –log$_{10}$ *P*-values on the Y-axis, (left-hand-side) versus chromosome base pair position on the X-axis, (Recombination rate (hotspots) are also plotted on the Y-axis (right-hand-side), shown in light-blue). The purple diamond plots the lead CARDIoGRAM meta-analysis SNP rs4773144 for association with CAD. A colour-coded scale of r^2 pairwise (LD) between the lead SNP and all other plotted imputed SNPs are shown for the locus region. Another SNP rs3809346 coded as a red circle for strong pairwise LD r^2>0.8, is shown at a higher *P*-value than the lead SNP, but was not considered to be the lead SNP, because it was present in only 7 of the discovery genotyped GWA studies, and so did not pass the pre-determined 8 genotyped studies threshold needed for replication. The X-axis spans the SNP locations over a 1Mb region. The gene location and strand direction are shown with arrows. The two candidate genes are: *COL4A1* (collagen type IV alpha 1) and *COL4A2* (collagen type IV alpha 2). **Figure 4.1.** is generated is generated by the webtool LocusZoom, hosted via the Center for Statistical Genetics, University of Michigan at https://statgen.sph.umich.edu /locuszoom/genform.php?type=yourdata (Version 1.1: June 2011) (Pruim et al. 2010).

At the same time as the CARDIoGRAM Study, a separate consortium (the IBC 50K CAD Consortium) was undertaking meta-analyses of CAD case-control studies that had typed the IBC 50K array (see Chapter 3 for description of the array). A primary meta-analysis based on separately analysed ethnicities included: (a) ten European Caucasian cohorts (11,202 cases and 30,733 controls); and (b) two South Asian cohorts PROMIS and LOLIPOP (4,394 cases and 4,259 controls). Subsequently, a secondary (combined) meta-analysis was performed on all twelve studies, and adjusted for ethnicity. The SNP rs4773144 had a European Caucasian discovery meta-analysis odds ratio of 1.08 (95% CI: 1.04-1.12), P=5.5x10^{-5}, a South Asian odds ratio of 1.13 (95% CI: 1.06-1.20), P=7.2x10^{-5}, and a combined discovery meta-analysis odds ratio of 1.09 (95% CI: 1.06-1.13), P=3.5x10^{-8} (IBC 50K CAD Consortium 2011). Although three of the European Caucasian studies (WTCCC, GerMIF and Penncath) overlapped between the CARDIoGRAM and the IBC 50K CAD analyses, the concordant findings between the two meta-analyses, provided further support of the involvement of this locus in CAD.

The two genes *COL4A1* and *COL4A2* are known to be transcribed from a head-to-head, shared common bidirectional promoter, with only 127bp of the common promoter separating the first exon of each gene. *In vitro* studies, performed in the late 1980s and early 1990s, have shown that there are at least three core promoter binding motifs, within the 127bp common promoter: a CCAAT motif, a GC box (Poschl, Pollner & Kuhn 1988), and a later identified CCCTCCC motif of specific sequence C_5TC_7, termed a 'CTC box' (Fischer et al. 1993). As with many bidirectional promoters there is no canonical TATA box, a known AT-rich binding element that is typically located 25-30bp upstream of the TSS(+1). The sequence motif 'TATAA' is a binding site for transcription factor TFIID, an important protein that is known to form part of the pre-initiation complex (PIC), and guide the precise location of transcriptional initiation by RNA polymerase II (Sawadogo & Sentenac 1990). However, within the predominantly GC-rich common promoter, there are two alternate AT-rich short sequences located 25-30bp upstream of exon 1 in both genes. In particular, *in vitro* studies, such as DNaseI footprinting and site-directed mutagenesis have shown that CCAAT binding factor, SP1 and CTC box binding factor (CTCBF) can bind to these motifs, as well as, the two short AT-rich sequences and downstream GC boxes, particularly those within exon 1 of *COL4A2*; proving that these *cis*-regulatory elements are essential for efficient transcription of both genes, albeit with unequal gene promoter activity depending on the *cis*-acting element mutated (Fischer et al. 1993; Schmidt et al. 1993; Heikkila, Soininen & Tryggvason 1993). Yet another interesting observation occurs within intron 3 (just beyond exon 3) of *COL4A2*, which is the presence of an experimentally proven silencer *cis*-regulatory element that inhibits transcription of both genes (Poschl, Pollner & Kuhn 1988). This silencing element termed the COL4 silencer, has been narrowed down to a 25bp region (5'-CGCGCTTGGACTTGCGCGCCCGAGA-3') that is recognised by a nuclear protein called SILBF (Haniel et al. 1995). In addition, to this silencer element, there is also a putative canonical SP1, GC box *cis*-regulatory element within intron 3 of *COL4A2*, just before the SNP rs3809346 (Poschl, Pollner & Kuhn 1988). Furthermore, the CARDIoGRAM CAD-associated SNPs rs4773144 and rs3809346 are only 202bp and 433bp downstream of the COL4 silencer, respectively. Therefore, it is feasible that the rs4773144 CAD-associated locus is either situated within (or is a marker for) another relevant *cis*-regulatory element, capable of influencing transcription of either one, or both *COL4A1* and *COL4A2* genes. **Figure 4.2** shows an adapted schematic image of the shared common bidirectional promoter with experimentally identified *cis*-regulatory elements (Kuhn 1995; Pollner et al. 1997); inclusive of the CAD-associated candidate SNPs rs4773144 and rs3809346, and the putative canonical SP1 TFBS.

The co-expressed *COL4A1* and *COL4A2* genes are post-transcriptionally modified and translated in a highly regulated manner, into the most abundant and ubiquitously expressed isoforms of non-fibrillar collagen type IV (COL(IV)) alpha peptide chains, alpha 1 (α1) and alpha 2 (α2). The α1 and α2 chains are translated within the rough endoplasmic reticulum (RER) (Myllyharju & Kivirikko 2001), where

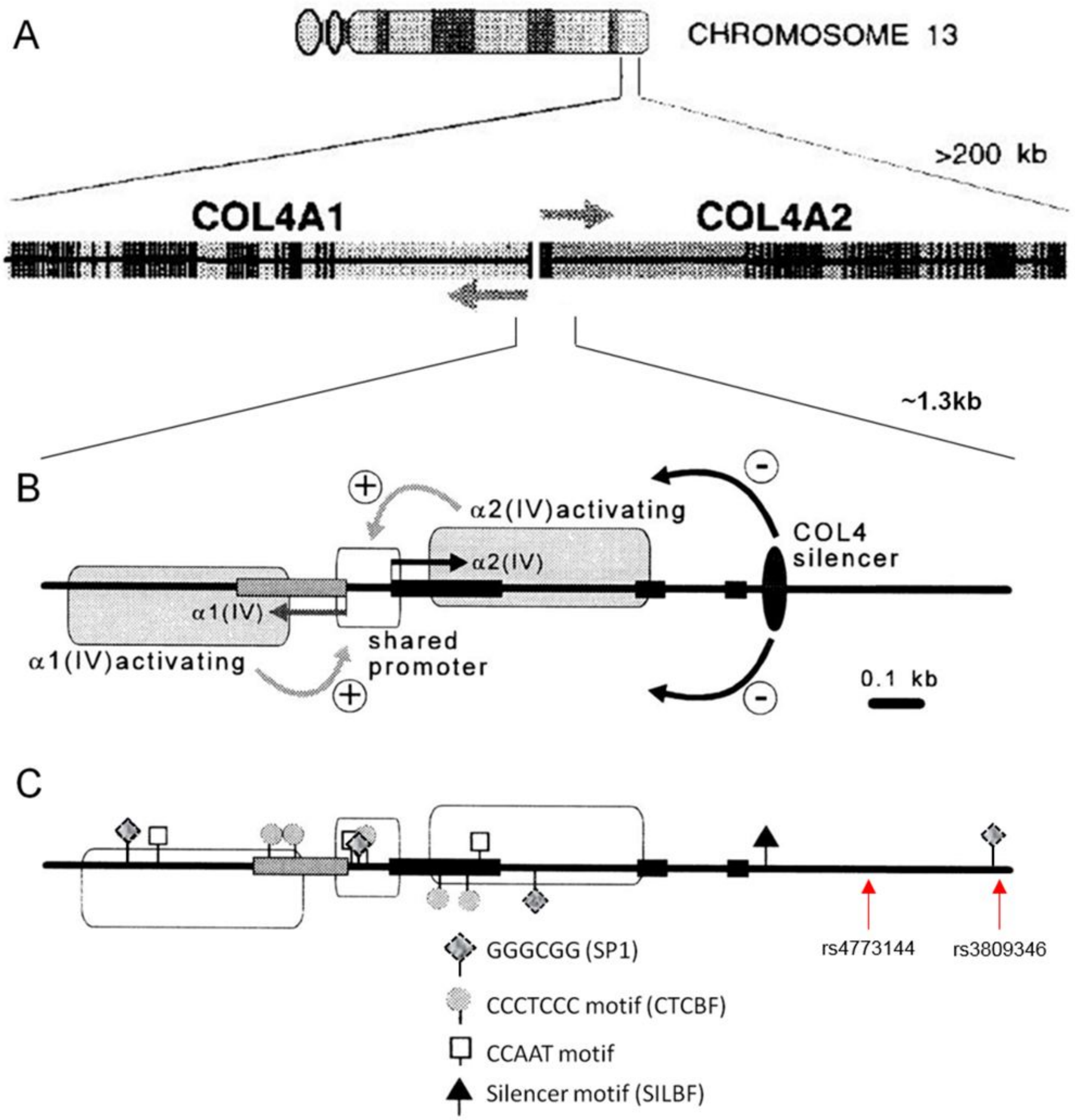


Figure 4.2. Schematic image of the *COL4A1* and *COL4A2* head-to-head shared common bidirectional promoter; adapted from figures taken from (Kuhn 1995; Pollner et al. 1997).

Legend. A) Shows the position of the two genes within the telomeric region of 13q34, and their head-to-head arrangement; **B**) shows the shared common bidirectional promoter and downstream regions that activate or repress transcription in both genes; **C**) shows the specific *cis*-acting elements responsible for the activating and silencer regions high-lighted in section **B**, in addition there is a putative canonical SP1 site just before rs3809346. Also, the location of CARDIoGRAM CAD-associated SNPs rs4773144 and rs3809346 are indicated by red arrows. The exons in section **B** and **C** are represented by a single grey box for *COL4A1* and three black boxes for *COL4A2*.

they form 2:1 ratio COL(IV) (α1α1α2) heterotrimeric protomers, the predominant building blocks that make up 50% of the basement membranes (BM). Subsequently, during development these heterotrimeric protomers are transported from the endoplasmic reticulum (ER) through the Golgi apparatus, where they begin to aggregate laterally, granulate and form secretory vesicles, ready for secretion into the extracellular matrix (ECM). Once in the ECM the heterotrimeric protomers form the major suprastructural component of all BM, in the shape of a network lattice, responsible for membrane strength and integrity (LeBleu, Macdonald & Kalluri 2007; Myllyharju & Kivirikko 2004).

In blood vessels, fibrillar interstitial collagen type I (COL(I)) and collagen type III (COL(III)) are predominant, and provide tensile strength to the outer vessel wall. However, COL(IV) is non-fibrillar, and is unique amongst the 28 known collagen types (Prockop & Kivirikko 1995), because it is only secreted to become an important component of the vascular BM. The vascular BM is located in two relevant areas as regards AS progression: (a) beneath the endothelium, and (b) around medial vascular SMCs. BM provides an important anchoring substrate and permeability barrier to the ingress of cells, and other components from the vascular lumen into the vascular wall, but it also helps maintain the medial layer vascular SMCs in a quiescent contractile state. Disruption of the endothelium anchored BM barrier, and the subsequent breakdown of the BM (COL(IV)) and ECM (COL(I)) that surround medial SMCs, by intimal activated macrophages that induct the adaptive immune response - are early features of atherosclerosis. However, it is thought that the main causes of BM (COL(IV)) degradation surrounding medial SMCs are MMP-2 and MMP-9 endogenously secreted by SMCs themselves (Aguilera et al. 2003). Moreover, as a consequence of SMC activation to a migratory and proliferative state, both fibrillar interstitial and non-fibrillar collagen types, are synthesised by vascular SMCs during the development of the atherosclerotic plaque fibrous cap, and therefore collagen is an integral component of the plaque, which may define its stability (see Chapter 1). In particular, the synthesis of collagen within the plaque can be viewed as a reaction to vessel wall injury.

Further detail regarding the location and role of the BM, within the progression of the atherosclerotic plaque, comes from the interesting observations made in the 1980s and 1990s by immunohistochemical analyses (of atherosclerotic plaque progression), using COL(IV) specific antibodies. For instance, early stage lesions, such as fatty streaks, detected BM around spindle-shaped cells, presumed to be proliferated media SMCs within the intima that had secreted COL(IV); whereas advanced stage lesions detected a multi-layer thickening of BM around extremely elongated SMCs, deep within the fibrous cap (Shekhonin et al. 1985; Shekhonin et al. 1987; Voss & Rauterberg 1986; Katsuda et al. 1992). This advanced stage BM thickening discovery was also shown, by electron microscopy dissections of occlusive fibrous atherosclerotic plaques, taken from post bypass surgery patients (Ross et al. 1984). These fibrous cap observations are most likely indicative of a 'senescence associated secretory phenotype', which serves to revert the SMCs to a more medial quiescent contractile state (Ross et al. 1984; Hirose et al. 1999). Thus preventing apoptotic release of SMC contents, which would release calcified minerals, lipids and tissue factor that would in turn feed the necrotic core via failed efferocytosis.

Furthermore, mutations in the *COL4A1* and *COL4A2* genes lead to diseases affecting the vessel wall, including rare Mendelian conditions, such as encephaloclastic porencephaly, small arterial vessel brain disease (such as single or recurrent haemorrhagic vascular stroke), and an autosomal dominant syndrome, involving hereditary, angiopathy, nephropathy, aneurysms and muscle cramps (HANAC) (Plaisier & Ronco 1993; Plaisier et al. 2007). In addition, more recently, a common non-synonymous coding SNP rs3742207 A$\rightarrow$C (Gln1334His) in the *COL4A1* gene was found to have a strong replicated association (i.e. it reached a GWA study and replication combined analysis genome wide significance; P=5.16x10^{-8}) with increased arterial stiffness (a known risk predictor of first-onset cardiovascular disease events) in a Caucasian founder population (Tarasov et al. 2009). This same SNP was previously shown to associate modestly with MI (P=1.8x10^{-3}, just short of the Bonferroni corrected P-value significance threshold, P<1.25x10^{-3}) in a non-replicated, Japanese candidate gene case-control study

(Yamada et al. 2008). The CARDIoGRAM meta-analysis for rs3742207 was supportive of the Tarasov *et al.* findings, with a suggestive CAD association at P=5x10^{-3}, but it did not reach the pre-defined *P*-value threshold for follow-up CARDIoGRAM (stage two) replication. However, for the CARDIoGRAM MI meta-analysis, rs3742207 did not even reach nominal significance, i.e. P<0.05. These latter reports, set a precedence for the plausibility that rs4773144, rs3809346, or another SNP in strong pairwise LD, could be functionally responsible for CAD risk.

4.1.1 Aims

Hence, I thought that it would be interesting to undertake further functional genomic analysis of this locus. The main goals of this work were:

1. To localise the likely causal SNP via *in silico* analysis and validate it functionally.
2. To investigate if the CAD-associated variant affects expression of *COL4A1* and/or *COL4A2*.
3. To see if the CAD associated variant also affects the vessel wall's response to injury as occurs after coronary angioplasty.

4.2.1 In silico bioinformatics methods

4.2.1.1 Refinement of the rs4773144 identified COL4A2 locus

The Broad Institute's web-based software interface SNAP – Proxy Search tool (http://www.broadinstitute.org/mpg/snap/ldsearch.php) was used to acquire all rs4773144 proxy SNPs with a pairwise LD threshold of $r^2 \geq 0.5$, within a 1Mb region (centred upon rs4773144), for the 1000 genomes pilot 1 data set. Any SNPs that had a pairwise LD ($r^2 \geq 0.8$) were considered of particular importance as regards their being potentially functional. The results of this analysis are shown in **Table 4.2** (Results section 4.3.1), and shape the methodological direction of both subsequent *in silico* bioinformatics and functional genomics analyses.

Additionally, I assessed the Metabochip array (see Chapter 3) SNP list to ensure there were no further SNPs. I also checked for Metabochip study association of these SNPs with CAD in the WTCCC CAD expanded subjects. However, it should be noted that for CARDIoGRAM, the 13q34 locus was not associated with CAD, as an individual study using the WTCCC CAD cohort, so no association was expected. A subsequent collaborative Metabochip array CAD multi-consortia meta-analysis was published by CARDIoGRAMplusC4D, at a later date, the results of which are reported also (CARDIoGRAMplusC4D Consortium et al. 2013).

Table 4.2. Pilot study 1000 genomes proxy SNPs for rs4773144.

CARDIoGRAM Lead SNP	Proxy SNP	Distance (bp)	R^2 (LD)	D' (LD)	Chr	Genomic Position (HG19)
rs4773144	rs4773141	6359	0.68	0.92	chr13	110954353
rs4773144	rs4773143	27	0.97	1	chr13	110960685
rs4773144	rs4773144	0	1	1	chr13	110960712
rs4773144	rs7986871	77	0.97	1	chr13	110960789
rs4773144	rs3809346	231	0.97	1	chr13	110960943
rs4773144	rs2391824	570	0.6	0.91	chr13	110961282
rs4773144	rs1360154	756	0.53	1	chr13	110961468

Legend. This table is adapted from the Broad Institute SNAP web tool output and shows the r^2 and D' pairwise LD between rs4773144 and its proxy SNPs, Chr – chromosome, genomic position uses the up-to-date human genome build 19 (HG19). The four SNPs that represent perfect proxies to each other are highlighted in red and span 258bp.

The specific genomic location of the lead CAD-associated SNP rs4773144 in relation to the *COL4A1* and *COL4A2* bidirectional promoter was achieved by using the UCSC genome browser human genome, build 19 (GRCh37/hg19) (URL: http://genome-euro.ucsc.edu/cgi-bin/hgGateway). In particular, the distance from the TSS(+1) of *COL4A1* and *COL4A2* from rs4773144, and any further 1000 genomes discovered proxy SNPs were reported (1000 Genomes Project Consortium et al. 2012).

4.2.1.2 Assessment of known gene expression data for a 13q34 locus eQTL

The Chicago eQTL browser database web interface (http://eqtl.uchicago.edu/cgi-bin/gbrowse/eqtl/) was used to look for any already known 13q34 locus eQTLs - related to any of the rs4773144 proxy SNPs ($r^2 \geq 0.5$). The available cell/tissue types registered in the Chicage eQTL database include: LCLs, monocytes, liver, fibroblasts, cortex (brain) and T-cells.

Moreover, an assessment was made of the monocyte and macrophage gene expression data available from the Cardiogenics study (previously mentioned in Chapter 3) (Schunkert et al. 2011).

4.2.1.3. In silico bioinformatics assessment of the 13q34 candidate SNPs

Appropriate *in silico* bioinformatics were performed on the CAD associated SNPs rs4773144 and rs3809346 identified through CARDIoGRAM and any further SNPs identified through *in silico* refinement. The *in silico* web-based tools used were ENCODE, RegulomeDB, and putative TFBS prediction software e.g. JASPAR (Portales-Casamar et al. 2010), MatInspector (Cartharius et al. 2005; Quandt et al. 1995) and TRANSFAC® (Matys et al. 2006).

4.3.1 In silico bioinformatics results

4.3.1.1 Refinement of the rs4773144 identified COL4A2 locus

The genetic architectural refinement of CAD-associated locus rs4773144 within intron 3 of *COL4A2* was achieved by using the BROAD institute web-based tool SNAP. A search for rs4773144 proxy SNPs within the 1000 genomes pilot data, identified the already known CARDIoGRAM SNP rs3809346, and a further two candidate SNPs - rs4773143 and rs7986871, as all being in almost perfect pairwise LD, i.e. $r^2 \approx 1$. Additionally, there were also three SNP (rs4773141, rs2391824 and rs1360154) in moderate pairwise LD ($0.5 < r^2 < 0.8$) (see **Table 4.2**). All proxy SNPs were located within intron 3 of *COL4A2* over a distance of 783bp, except rs4773141, which is located within intron 1 of *COL4A1*. The four SNPs (rs4773143, rs4773144, rs7986871 and rs3809346) in almost perfect pairwise LD within intron 3, span a region of 258bp, and are between 215bp and 473bp, downstream of the *COL4A2* exon 3. This analysis therefore suggests that the causal SNP lies within a very narrow 250bp region.

Use of the UCSC genome browser allowed the distance of the 4-SNP CAD-associated haplotype (rs4773143, rs4773144, rs7986871 and rs3809346) to be calculated from the TSS(+1) of *COL4A1* as -1189bp to -1447bp, and from the TSS(+1) of *COL4A2* as +1054bp to +1312bp.

4.3.1.2 Assessment of known gene expression data for a 13q34 locus eQTL

The Chicago eQTL browser database showed no eQTL for any of the rs4773144 proxy SNPs ($r^2 \geq 0.5$), or indeed any SNP in close proximity to the rs4773144 CAD-associated locus for the cell types registered in their database (i.e. LCLs, monocytes, liver, fibroblasts, cortex (brain) and T-cells).

An assessment of the monocyte and macrophage Cardiogenics gene expression and genotyping data discovered that only the rs4773144 proxy SNP rs2391824 (r^2=0.6) was present on the Illumina 660W-Quad beadchip array used, but there was no gene expression measured for the *COL4A1* and *COL4A2* probes (Ilmn_1653028 COL4A1_13 and ilmn_1724994 COL4A2_13 respectively) in monocytes and macrophages. Therefore, no Cardiogenics eQTL was detected in these cell types.

This lead to a wet-lab investigation into whether a 13q34 locus candidate SNP eQTL is present in RNA extracted from healthy adult kidney tissue (see **section 4.2.2**).

4.3.1.3 In silico bioinformatics assessment of the 13q34 candidate SNPs

In follow-up to the fine-mapping localisation of four candidate SNPs (rs4773143, rs4773144, rs7986871 and rs3809346) - I then performed a focussed *in silico* analysis, to capture any further *cis*-regulatory evidence that would help identify the most likely causal SNP. The types of *in silico* bioinformatics analyses, methods and resources used, and main findings are shown in a large summary **Table 4.3**, and refer to the following **Figures 4.3A-C** and **4.4** for ENCODE data, **Tables 4.4A-B** for TFBS data, **Figures 4.5A-D** for RegulomeDB output.

Table 4.3. Summarised *in silico* assessment of 4 candidate 13q34 locus SNPs using ENCODE, TFBS and RegulomeDB.

Type of Analysis	Methods and Resources	Main Findings
ENCODE Analysis: *In silico* assessment of the encyclopaedia of DNA elements (ENCODE) - a project to identify all functional elements in the human genome sequence. **Rationale:** The purpose for looking at the ENCODE data, is to identify 13q34 locus *cis*-regulatory regions that align with or close to, the four candidate SNPs with a near perfect pairwise LD, i.e. $r^2 \approx 1$ (rs4773143, rs4773144, rs7986871, rs3809346). Additionally, the two SNPs in moderately strong pairwise LD ($0.5<r^2<0.8$) with the	**Resource:** ENCODE data was accessed through the UCSC Genome Browser. The genome assemblies used were human (HG19) and mouse (mmp-9). **Methods:** **DNase-HS**: DNaseI hypersensitivity sites identified by DNaseI nuclease activity and next generation genome wide sequencing, signifies general chromatin accessibility, following binding of trans-acting factors in place of a canonical nucleosome, and indicates locations of active *cis*-regulatory sequences (Crawford et al. 2006). **FAIRE-seq**: Formaldehyde assisted isolation of regulatory elements (FAIRE), followed by next generation sequencing is a method to isolate and identify nucleosome-depleted / -destabilised regions of the genome (Giresi et al. 2007). A low variability technique for any cell type as an alternative to DNase-seq. FAIRE-seq is more sensitive to distal regulatory regions and DNase-seq is more sensitive to promoter regions, but generally they cross-validate each other. Along with DNaseI HS, FAIRE has led to the discovery of functional regulatory elements that include enhancers, silencers, insulators, promotors, locus control regions and novel elements. Sensitivity can be greater than DNase-seq with deeper sequencing.	**Visualisation of ENCODE analysis results:** See **Figures 4.3A, 4.3B, 4.3C** and **4.4** for details **Key findings:** 1) A strong H3K4Me3 signal indicative of an active or poised promoter, spans the 4 candidate SNPs across 7 human cell lines (**Fig.4.3.A/C**), and is conserved in mouse heart/kidney tissue (**Fig.4.4**). 2) A moderately strong H3K27Ac signal indicative of transcriptional regulatory element binding, and some weak enhancer binding is observed for H3K4Me1 all 4 candidate SNPs across 7 human cell lines, and is conserved in mouse heart/kidney/MEF tissue (**Fig.4.4**). 3) These histone mark signatures are indicative of a conserved open active chromatin promoter region that spans the 4 candidate SNPs on interest. 4) There is weaker H3K4Me3 and H3K27Ac for the 2 SNPs in moderate pairwise LD with rs4773144, and a stronger H3K4Me1 enhancer signal for rs1360154 only in human 7 cell lines (Fig 4.3C). Mouse shows weaker H3K4Me3 signal in heart/kidney/MEF (**Fig.4.4**). 5) There is a 'dip' in histone mark signatures at a region inclusive of and just beyond rs3809346 (**Fig.4.3**). 6) Strong signal density is observed for DNaseI HS sites in multiple cell types and these align with the active/poised promoter histone mark H3K4Me3. Summits density peaks for DNaseI HS sites align particularly with the127bp COL4A1/A2 common core promoter, and also with a region just beyond SNP rs3809346.

Type of Analysis	Methods and Resources	Main Findings
ENCODE Analysis: *…Continued* lead CARDIoGRAM SNP, were also looked at, i.e. rs23918241 and rs1360154.	**Methods** (*Continued):* **ChIP-seq**: Chromatin immunoprecipitation (ChIP) with antibodies specific to a protein, followed by sequencing of the precipitated DNA is a method to identify the specific location of proteins that are directly or indirectly bound to genomic DNA. By identifying the binding location of sequence-specific transcription factors (Euskirchen et al. 2007), general transcription machinery components, and chromatin factors such as H3 histone modifications (methylation or acetylation), which influence gene expression by regulating how accessible chromatin is to transcription (Bernstein et al. 2005). Specific histone marks are identified throughout the genome using ChIP-seq and include H3K4Me1 (enhancer associated), H3K4Me3 (promoters – active or poised to be activated) and H3K27Ac (indicates regulatory regions of transcription). Transcription factor ChIP-seq data tracks are either shown as individual cells from single centres, or as one or more cells from combined ENCODE centres. Resolution of ChIP-seq is about 200bp. **DNase-DGF:** base pair resolution DNaseI digital footprinting makes use of *'double hit'* DNaseI HS sites DNase-seq data, by identifying narrow depleted regions of high depth enriched deep sequencing reads mapped in the forward and reverse direction. This method enables the detection of reliable protein-binding footprints (Sabo et al. 2006).	**Key findings** (C*ontinued):* 7) Two FAIRE-seq strong density signals span 2 SNPs rs7986871 and rs3809346, and one spans all four candidate SNPs.The summits of the peaks all align near to or just beyond rs3809346 **(Fig.4.3A)**. 8) Strong DNaseI footprint signals align most strongly with the main 127bp COL4A1/A2 common core promoter, and a region just downstream of *COL4A2* rs3809346 in multiple cell types (fibroblasts, myocytes, endothelial and epithelial cells) There are also some medium strength DNaseI footprints near to rs3809346, rs4773143, rs4773144, and just beyond *COL4A2* exon 3, where the known COL4 silencer is located (**Fig.4.3A/B**). 9) Taken together the DNaseI HS, FAIRE and DGF alignments indicate open chromatin, with a strong likelihood of DNA-protein binding at the 127bp common core promoter and just beyond SNP rs3809346, with likely further DNA-protein binding with the known silencer, and three of four candidate SNPs; i.e. rs4773143, rs4773144 and rs3809346, but not rs7986871 (**Fig.4.3A/B**).

Type of Analysis	Methods and Resources	Main Findings
ENCODE Analysis: *…Continued*	**Methods** (*Continued):* **RNA-seq:** RNA sequencing is a genomic mapping and quantifying method capable of capturing at any given time, all RNA transcripts within a cell. The RNA transcripts are reverse transcribed into cDNA, and then directly sequenced using next generation sequencing to a high depth. The method is particularly useful for the identification of any rare *de novo* transcripts, i.e. alternatively spliced isoforms. The reads may be single reads (i.e. one-end reads), or paired-end reads from randomly primed cDNA. (Morozova, Hirst & Marra 2009). **CpG islands:** CpG islands are regions of the genome where an unusually high ratio of CG dinucleotides exist, when compared to the genome as a whole. A CpG is rare in vertebrate DNA, because the Cs in such an arrangement tend to be methylated. CpGs tend to be rare unless there is selective pressure to keep them. It just so happens that CpG islands are typically common near TSS(+1), and so they may be associated with gene promoter regions, and influence gene expression. A CpG island in ENCODE is depicted if it fits three criteria: a) >50% GC content, b) a length >200bp, and c) a ratio >0.6 of observed CG dinucleotides to the expected, based on the GC content in the segment (Gardiner-Garden & Frommer 1987). **Vertebrate Multiz Alignment:** Comparative genomics cross-species conservation across 46 species.	**Key findings** (C*ontinued):* 10) The main ENCODE TFBS ChIP-seq track for multiple centres identifies only HUVEC Pol2 (RNA Pol II) in alignment with the 4 candidate SNPs (**Fig.4.3B**). This Pol2 observation is conserved in mouse heart/kidney/ MEF (**Fig.4.4**). However, a number of cMYC family TFBS ChIP-seq signals align close to or overlap with the 4 candidate SNPs or the known COL4 silencer e.g. cMYC, MAX, MXI1 & MAZ. Strong repressor TFBS ChIP-seq signals are observed for SIN3A and MXI1 in H1ES aligned with rs4773144 (N.B. SIN3A is known to be part of a repressor complex with MXD1 and MAX (Swanson et al. 2004)). Known repressor E2F6 and MAX ChIP-seq TFBS signals both overlap with the known COL4 silencer. A 100bp strong signal density for Pol2 (HUVEC) & a moderate signal for TBP (H1ES) aligns with rs7986871 (**Fig.4.3B**). 11) RNA-seq (Caltech) dense signals for HeLa, HUVEC and NHLF cells. In particular, a strong HUVEC signal is seen overlying the SNP rs7986871, however the repeat signal below it is not dense at all, suggesting this result could be spurious. Additional RNA-seq (CSHL) data shows some raw signal beyond *COL4A2* exon 3, in HUVEC and HVMF, but it is most likely background noise (**Fig.4.3C**). 12) The common non-risk allele (G) of rs3809346 creates a CpG dinucleotide, and sits within a short 219bp CpG island (Fig.4.3C). The CpG island still exists even after the risk allele (A) removes a CpG dinucleotide (Obs/Exp ratio only drops from 0.85 to 0.82). The other 3 SNPs, do not sit in a CpG island.

Type of Analysis	Methods and Resources	Main Findings
TFBS analysis: Transcription factor binding site (TFBS) predictions using web based tools. **Rationale:** The purpose of assessing TFBS was previously described in full (see Chapter 3, **Table 3.6**). As regards putative function of the 4 candidate 13q34 locus SNPs, it is important to see whether any of the SNPs are located with a TFBS, and whether the major or minor allele creates or disrupts the binding of a particular *cis*-regulatory transcription factor.	**Resource:** JASPAR website: http://jaspar.genereg.net/ (Portales-Casamar et al. 2010). MatInspector (version 8.06, Aug 2012 released) accessed via the Genomatrix software suite, GmbH website: http://www.genomatix.de/index.html TRANSFAC® Professional (v10.2, Jul 2007) (Gotea & Ovcharenko 2008) (sourced from BIOBASE (http://www.biobase-international.com/ product/transcription-factor-binding-sites). **Methods**: All three databases generate and utilize position weighted matrix (PWM) scores that are analogous to binding energy affinity (Stormo 2000; Maerkl & Quake 2007). ***JASPAR*** is an open access database resource for eukaryote DNA TFBS profiles. The JASPAR CORE *Vertebrata* (2010 release) database consists of a non-redundant set of 129 PWM profiles, derived from literature published experimentally proven, nucleotide sequence binding to transcription factors in eukaryotes. The JASPAR database resource is complemented by a web interface for browsing, searching, subset selection and *fasta* sequence analysis utility that is ideal for assessing regulatory regions of genomic sequence. The output score results are based on differences in similarity between the PWM model binding affinity scores for each SNP allele (i.e. major *vs* minor) for the same TFBS on the same orientation DNA strand.	**Visualisation of TFBS analysis results:** See **Tables 4.4A/B** for details **Key findings:** 1) Several putative TFBS predictions are made by all three web-based tools for the lead candidate SNP rs4773144. JASPAR predictions showed high binding affinity differences (>3.8; 12-18%) for several ETS family TFBS, i.e. SPIB, FEV, GABPA, ETS1 and SPI1 (These bind to a PU box in particular GGAA/T). However, MatInspector/TRANSFAC®/RegulomeDB identified STAT1/STAT3 and STAT6; and MatInspector/ TRANSFAC® identified MIF1. The risk allele (G) disrupted all of these putative TFBS. 2) The 2nd CARDIoGRAM proxy SNP rs3809346 shows putative high differential binding scores (>4.0; 19%) for 2 JASPAR TFBS – BRCA1 & SOX10. MatInspector/ TRANSFAC® identified EVI-1, and MatInspector only identified MYT1. In all cases the risk allele (A) created a putative TFBS. 3) Proxy SNP rs4773143 risk allele (C) disrupts 6 different TFBS families. 4 were identified by JASPAR with high binding affinity scores (>3.2, 12-19%) – SOX10, FOXO3, CEBPA, & NFATC2. MatInspector identified FOXP1 & FAC1, whereas TRANSFAC® / RegulomeDB identified CEBPB. 4) The final proxy SNP rs7986871 risk allele (A) disrupts 7 putative TFBS. JASPAR identifies 6 putative TFBS – TEAD1, RORA_2, SPI1, ZEB1, GATA3 & AP1, with high binding affinity differences (>3.1; 9-18%).

Type of Analysis	Methods and Resources	Main Findings
TFBS analysis *(Continued):*	**Methods** *(Continued):* ***MatInspector*** is a commercial well-established, web-based tool (first released in 1995). It utilises a matrix database called MatBase (Matrix family library version 9.0, released Aug 2012) that consists of a large library of 1381 PWMs within 411 family descriptions for TFBS. A sequence analysis utility enables assessment of specific genomic sequence for TFBS at the four candidate SNPs of interest. Specificity is improved by incorporating a conservation profile and four base core region, so that a matrix similarity score can be calculated as a measure of conservation at any given nucleotide (Cartharius et al. 2005; Quandt et al. 1995). ***TRANSFAC® Professional*** is another commercial, well-established, web-based tool. The publicly available version I have used is v10.2 (released Jul 2007) with a library of 584 vertebrate PWMs from ~400 families. This resource was accessed through the multiple sequence local alignment and visualization tool (MULAN) (http://mulan.dcode.org). This allows the identification of TFBS that are evolutionarily conserved across multiple species. It does this by aligning multiple genomic sequences, and then making use of another NCBI dcode.org web tool, called MultiTF (http://multitf.dcode.org) that utilises the TRANSFAC® v10.2, described earlier. Scoring is given as optimised for function (O4F) for a higher putative binding affinity, based on PWM similarity scores >0.8 and conservation.	**Key findings** *(Continued):* Both MatInspector/JASPAR identified RORA_2. Both TRANSFAC®/JASPAR identified AP1. However MatInspector alone identified BACH2. 5) Some of the TFBS predictions were a potential match to the TFBS ChIP-seq data mentioned earlier under ENCODE results, and shown in **Figure 4.3*B***. E.g. a weak hESC - GABP ChIP-seq signal is a match for the rs4773144 GABPA putative TFBS, and hESC - BACH1 ChIP-seq is a weak match for the rs7986871 predicted BACH2 TFBS. 6) E2F6 and MAX ChIP-seq data in H1ES and A549 (epithelial lung carcinoma) cells show high density signal alignments with the Haniel *et al.* COL4 silencer, but also with or very near to all 4 candidate SNPs. A highly conserved TRANSFAC® putative TFBS (see ***Appendix 6.3, Table S4.1***) is E2F6 (**TTGCGCGC**), and it aligns with the known COL4 silencer sequence (Haniel et al. 1995). In addition a potential non-canonical E-box (CGCTTG) is also present within the reported 25 nt sequence (JASPAR binding affinity score for Arnt::Ahr is 5.7; 0.84 matrix similarity). 7) A MatInspector predicted TFBS for rs3809346, EVI-1 is known to be part of a multi-component complex involving SP1 (a GC-box). There is a known canonical SP1 site situated 16bp upstream of rs3809346. 8) The strongest putative TFBS predictions discovered by all four *in silico* prediction tools used, but without ChIP-seq data support, were for a STAT transcription factor disruption by rs4773144 risk allele (G), and a C/EBP disruption by rs4773143 risk allele (C).

Type of Analysis	Methods and Resources	Main Findings
RegulomeDB SNP annotation curated database **Rationale:** RegulomeDB is a database that annotates SNPs with known and predicted regulatory elements in the human non-coding regions. The four lead candidate SNPs at locus 13q34 are assessed.	**Resource:** RegulomeDB curated database with a web-based software interface - http://regulome.stanford.edu/ (Boyle et al. 2012) **Methods:** Source data from public datasets such as GEO (gene expression omnibus), ENCODE project, in silico prediction web-based tools, and published literature are utilised to identify known and predicted genomic *cis*-regulatory elements. For example, DNaseI HS sites, TFBS, and histone mark signatures. An overall likelihood of *cis*-regulatory element binding of each SNP, is given a putative functional evidentiary weighted simple score. A separate visual RegulomeDB SNP annotation data output was generated for each of the 4 lead candidate SNPs (rs4773143, rs4773144, rs7986871 and rs3809346). TFBS scores are provided as for TRANSFAC® Professional (v10.2), i.e. O4F.	**Visualisation RegulomeDB output results:** See **Figure 4.5*A-D*** for details **Key findings:** 1) The RegulomeDB output for each SNP are shown in order of most likely putative *cis*-regulatory binding element in **Figures 4.5*A-D*** for rs4773144, rs3809346, rs7986871 and rs4773143 respectively. 2) The candidate SNP with the best score for influencing regulatory element binding was the lead SNP rs4773144 (Score:2b) described as likely to affect binding; whereas the other three SNPs were considered to have minimal evidence of binding: rs3809346 and rs7986871 (Score:4) and rs4773143 (Score:5). 3) The regulatory elements driving these results were a Pol2 ChIP-seq peak and a STAT3 predicted motif for rs4773144 (Score:2b), a Pol2 ChIP-seq peak aligned with SNPs rs7986871 and rs3809346 (Score:4), and a CEBPB predicted motif for rs4773143 (Score 5). 4) All SNPs were in an active open chromatin promoter region as signified by a strong signal for DNaseI HS sites and H3K27Ac ENCODE tracks. 5) RegulomeDB reported TFBS identified using TRANSFAC® are included in **Tables 4.4A/B.** For instance, rs4773144 is a TFBS for STAT3 and rs4773143 is a TFBS for CEBPB.

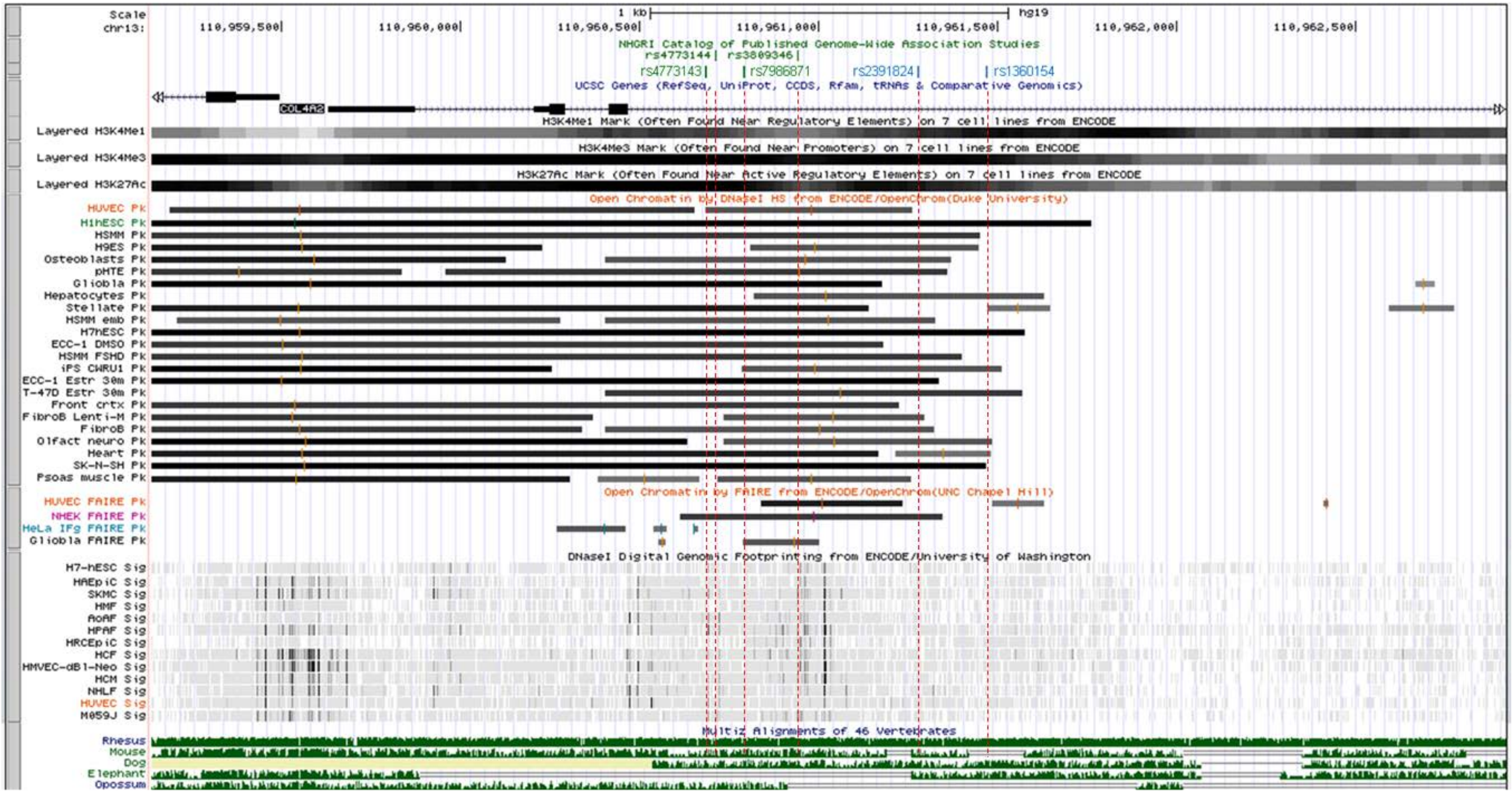


Figure 4.3A. Adapted ENCODE alignment tracks for Histone modfications, DNaseI HS, FAIRE, DNaseI DGF and Multiz alignment conservation from the UCSC Genome Browser (HG19) for the rs4773144 CAD-associated Chr13q34 locus. Track 1) shows the lead CAD GWA study SNPs rs4773144 and rs3809346, and 2 further SNPs in strong pairwise LD (r^2≈1) in green, and 2 extra SNPs with LD r^2≥0.5<0.8 in blue (Red vertical dashed lines indicate the SNP locations throughout the other alignment tracks); Track 2) shows UCSC genes for the promoter region of head-to-head genes *COL4A1* and *COL4A2*; Track 3) shows dense ChIP-seq signals for H3K4me1, H3K4me3 and H3K27Ac histone marks; Track 4) shows open chromatin DNaseI HS; Track 5) shows open chromatin by FAIRE; Track 6) shows single nucleotide resolution DNaseI digital genomic footprint (DGF) signals indicative of DNA-bound regulatory proteins; Track 7) shows Multiz alignments for the most conserved vertebrate species. (Tracks with the strongest peaks and signals are indicated by darker grey, short vertical lines when present indicate the peak summits).

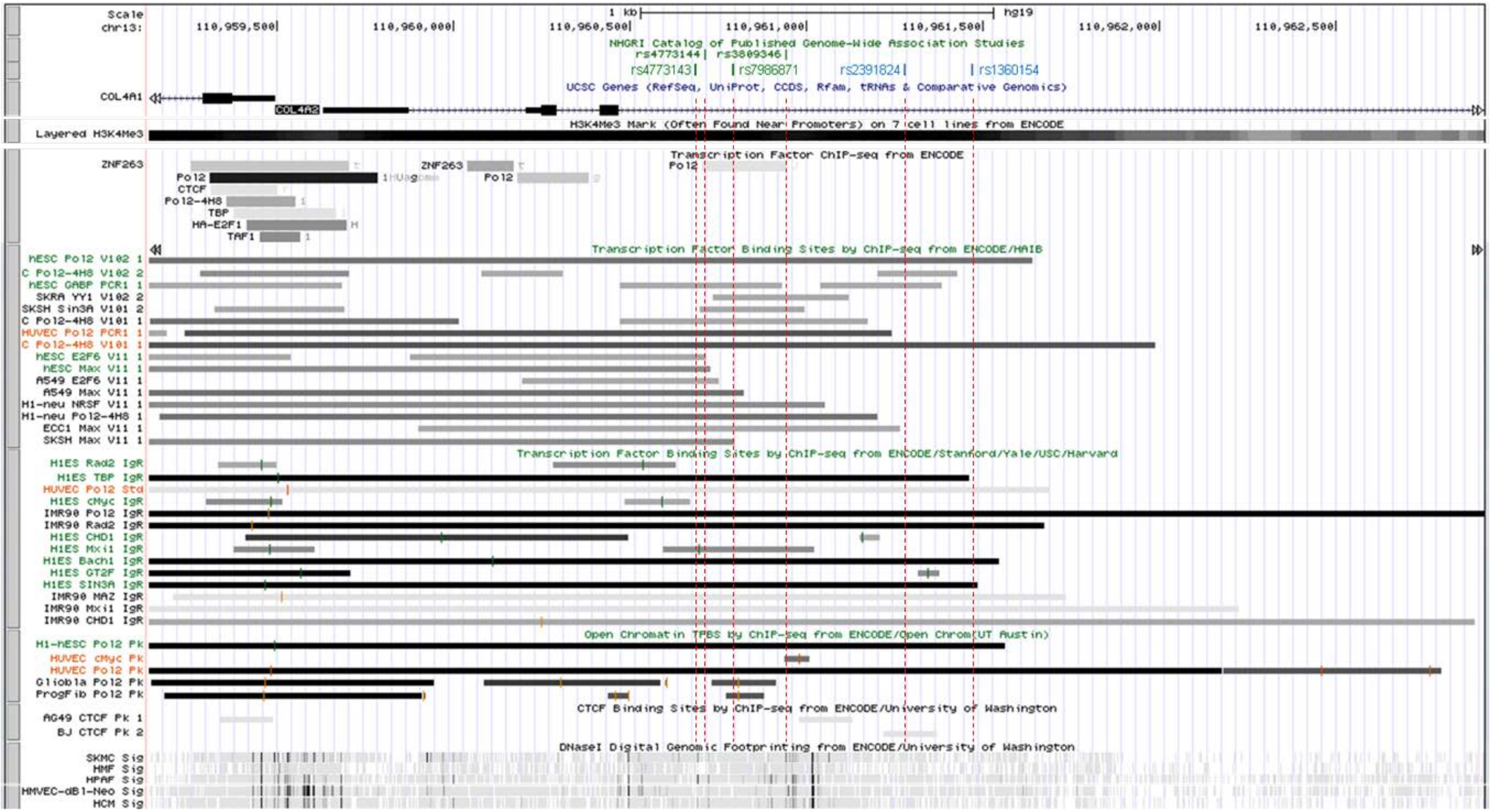


Figure 4.3B. Adapted ENCODE alignment tracks focussed on TFBS ChIP-seq from the UCSC Genome Browser (HG19) for the rs4773144 CAD-associated Chr13q34 locus. Track 1) shows the lead CAD GWA study SNPs rs4773144 and rs3809346, plus 2 further SNPs in strong pairwise LD (r^2≈1) in green, and 2 extra SNPs with LD r^2≥0.5<0.8 in blue (Red vertical dashed lines indicate the SNP locations throughout the other alignment tracks); Track 2) shows UCSC genes for the promoter region of head-to-head genes *COL4A1* and *COL4A2*; Track 3) shows dense ChIP-seq signal for the promoter H3K4me3 histone mark; Track 4) shows ENCODE TFBS ChIP-seq (multiple centres); Track 5) shows TFBS by ChIP-seq (HAIB); Track 6) shows TFBS ChIP-seq (SYDH); Track 7) shows the TFBS ChIP-seq (UT Austin); Track 8) shows the strongest signals for DNaseI DGF. (Tracks with the strongest peaks and signals are indicated by darker shading, short vertical lines when present indicate the peak summits).

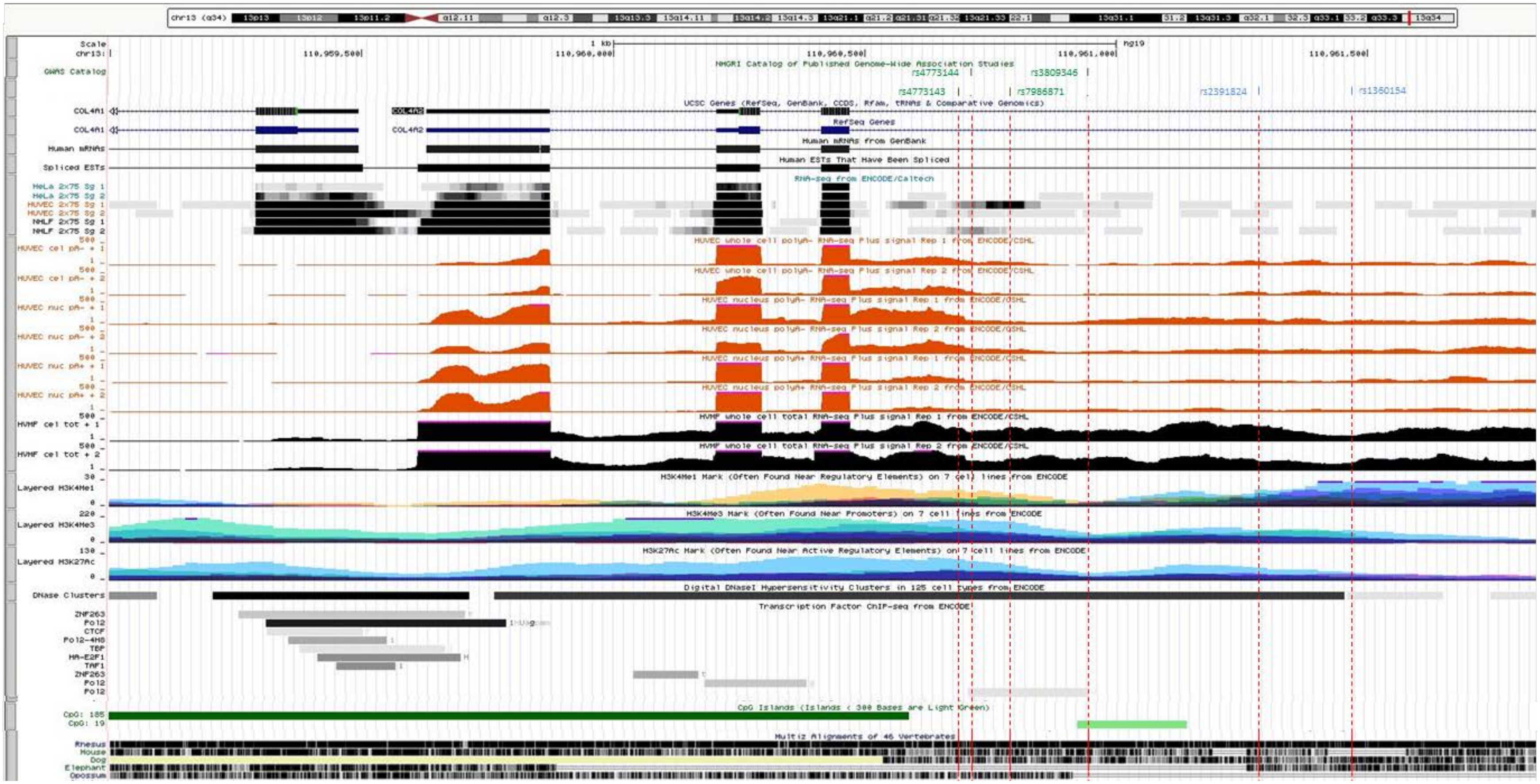

chr13 (q34)
13p13
13p12
13p11.2
q12.11
q12.3
13q13.3
13q14.11
13q14.2
13q14.3
13q21.1
q21.2
q21.31
q21.32
13q21.33
22.1
13q31.1
31.2
13q31.3
q32.1
32.3
q33.1
33.2
q33.3
13q34
Scale
chr13:
1 kb
hg19
110,959,500
110,960,000
110,960,500
110,961,000
110,961,500
GWAS Catalog
NHGRI Catalog of Published Genome-Wide Association Studies
rs4773144
rs3809346
rs4773143
rs7986871
rs2391824
rs1360154
UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)
COL4A1
COL4A2
RefSeq Genes
COL4A1
COL4A2
Human mRNAs
Human mRNAs from GenBank
Spliced ESTs
Human ESTs That Have Been Spliced
RNA-seq from ENCODE/Caltech
HeLa 2x75 Sg 1
HeLa 2x75 Sg 2
HUVEC 2x75 Sg 1
HUVEC 2x75 Sg 2
NHLF 2x75 Sg 1
NHLF 2x75 Sg 2
500
HUVEC cel pA- + 1
1
HUVEC whole cell polyA- RNA-seq Plus signal Rep 1 from ENCODE/CSHL
500
HUVEC cel pA- + 2
1
HUVEC whole cell polyA- RNA-seq Plus signal Rep 2 from ENCODE/CSHL
500
HUVEC nuc pA- + 1
1
HUVEC nucleus polyA- RNA-seq Plus signal Rep 1 from ENCODE/CSHL
500
HUVEC nuc pA- + 2
1
HUVEC nucleus polyA- RNA-seq Plus signal Rep 2 from ENCODE/CSHL
500
HUVEC nuc pA+ + 1
1
HUVEC nucleus polyA+ RNA-seq Plus signal Rep 1 from ENCODE/CSHL
500
HUVEC nuc pA+ + 2
1
HUVEC nucleus polyA+ RNA-seq Plus signal Rep 2 from ENCODE/CSHL
500
HVMF cel tot + 1
1
HVMF whole cell total RNA-seq Plus signal Rep 1 from ENCODE/CSHL
500
HVMF cel tot + 2
1
HVMF whole cell total RNA-seq Plus signal Rep 2 from ENCODE/CSHL
30
Layered H3K4Me1
0
H3K4Me1 Mark (Often Found Near Regulatory Elements) on 7 cell lines from ENCODE
220
Layered H3K4Me3
0
H3K4Me3 Mark (Often Found Near Promoters) on 7 cell lines from ENCODE
130
Layered H3K27Ac
0
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
DNase Clusters
Digital DNaseI Hypersensitivity Clusters in 125 cell types from ENCODE
Transcription Factor ChIP-seq from ENCODE
ZNF263
Pol2
CTCF
Pol2-4H8
TBP
HA-E2F1
TAF1
ZNF263
Pol2
Pol2
CpG: 185
CpG: 19
CpG Islands (Islands < 300 Bases are Light Green)
Multiz Alignments of 46 Vertebrates
Rhesus
Mouse
Dog
Elephant
Opossum

Figure 4.3C. Adapted ENCODE alignment tracks for RNA-seq, histone modifications, TFBS ChIP-seq, Multiz alignment conservation and CpG islands from the UCSC Genome Browser (HG19) for the rs4773144 CAD-associated Chr13q34 locus. Track 1) shows the lead CAD GWA study SNPs rs4773144 and rs3809346, plus 2 further SNPs in strong pairwise LD (r^2≈1) in green, and 2 extra SNPs with LD $r^2 \geq 0.5 < 0.8$ in blue (Red vertical dashed lines indicate the SNP locations throughout the other alignment tracks); Track 2) shows UCSC genes for the promoter region of head-to-head genes *COL4A1* and *COL4A2*; Track 3) shows Poly-A RNA-seq dense paired-ends signal (Caltech); Track 4) shows Poly-A and total RNA-seq raw signals (CSHL); Track 5) shows raw ChIP-seq signal for H3K4me1, H3K4me3 & H3K27Ac histone marks on 7 cell lines (GM12878=red, K562=Navy blue, H1-HESC=yellow, HSMM=green, HUVEC=light blue, NHEK=purple, NHLF=pink) ; Track 6) shows ENCODE TFBS ChIP-seq (multiple centres); Track 7) shows CpG islands>300bp are dark green and <300bp are light green; Track 8) shows Multiz alignments for conserved vertebrate species. (Tracks with the strongest peaks and signals are indicated by darker shading, short vertical lines when present indicate the peak summits).

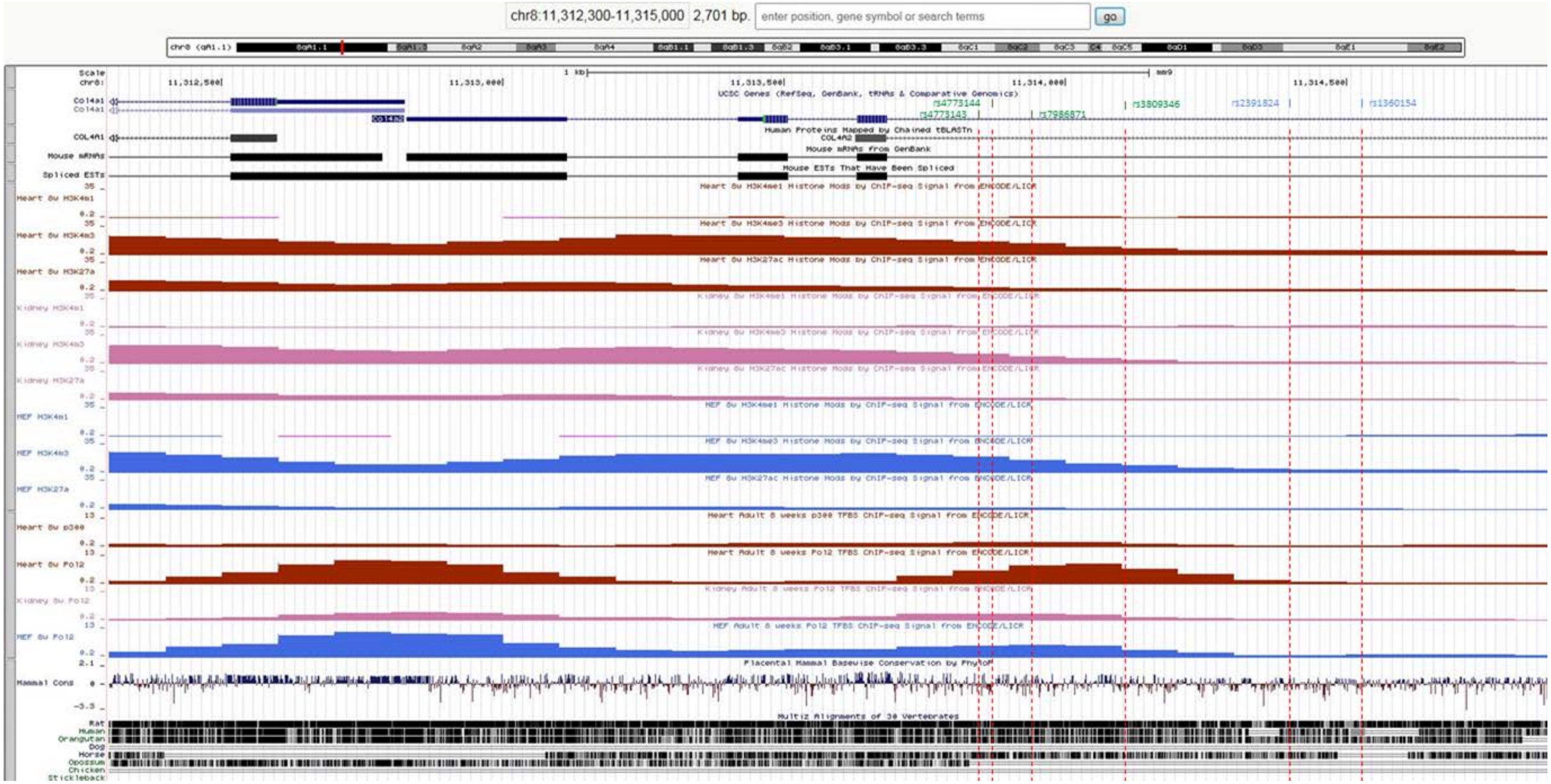
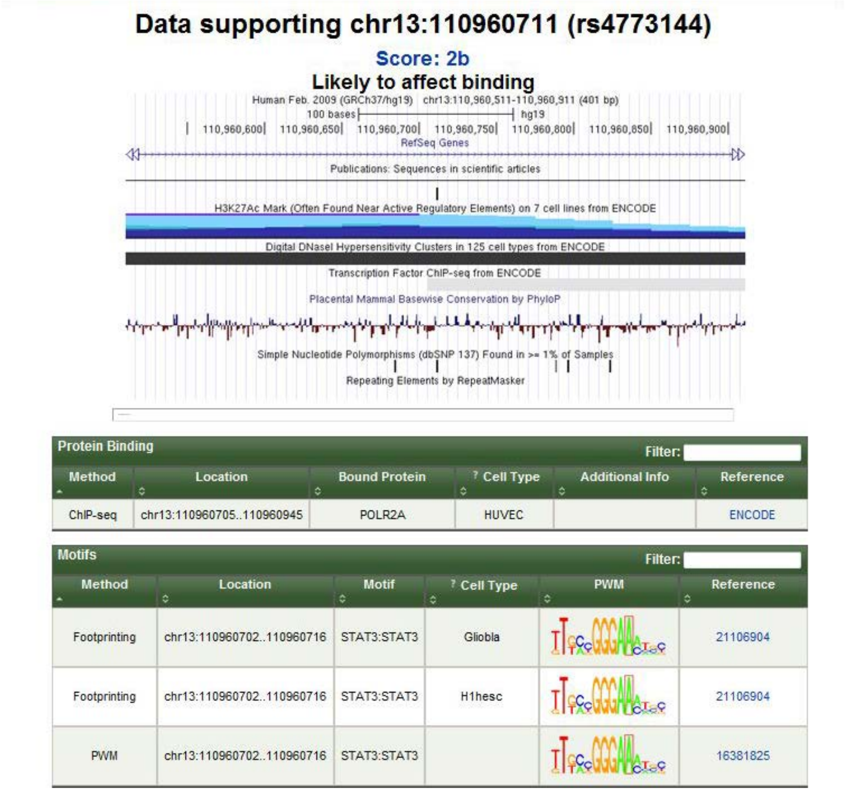

chr8:11,312,300-11,315,000
2,701 bp.
enter position, gene symbol or search terms
go
chr8 (qA1.1)
8qA1.1
8qA1.3
8qA2
8qA3
8qA4
8qB1.1
8qB1.3
8qB2
8qB3.1
8qB3.3
8qC1
8qC2
8qC3
8qC4
8qC5
8qD1
8qD3
8qE1
8qE2
Scale
chr8:
1 kb
mm9
11,312,500
11,313,000
11,313,500
11,314,000
11,314,500
UCSC Genes (RefSeq, GenBank, tRNAs & Comparative Genomics)
Col4a1
Col4a1
Col4a2
rs4773144
rs4773143
rs7986871
rs3809346
rs2391824
rs1360154
Human Proteins Mapped by Chained tBLASTn
COL4A1
COL4A2
Mouse mRNAs
Mouse mRNAs from GenBank
Spliced ESTs
Mouse ESTs That Have Been Spliced
35
0.2
Heart 8w H3K4m1
Heart 8w H3K4me1 Histone Mods by ChIP-seq Signal from ENCODE/LICR
Heart 8w H3K4m3
Heart 8w H3K4me3 Histone Mods by ChIP-seq Signal from ENCODE/LICR
Heart 8w H3K27a
Heart 8w H3K27ac Histone Mods by ChIP-seq Signal from ENCODE/LICR
Kidney H3K4m1
Kidney 8w H3K4me1 Histone Mods by ChIP-seq Signal from ENCODE/LICR
Kidney H3K4m3
Kidney 8w H3K4me3 Histone Mods by ChIP-seq Signal from ENCODE/LICR
Kidney H3K27a
Kidney 8w H3K27ac Histone Mods by ChIP-seq Signal from ENCODE/LICR
MEF H3K4m1
MEF 8w H3K4me1 Histone Mods by ChIP-seq Signal from ENCODE/LICR
MEF H3K4m3
MEF 8w H3K4me3 Histone Mods by ChIP-seq Signal from ENCODE/LICR
MEF H3K27a
MEF 8w H3K27ac Histone Mods by ChIP-seq Signal from ENCODE/LICR
13
Heart 8w p300
Heart Adult 8 weeks p300 TFBS ChIP-seq Signal from ENCODE/LICR
Heart 8w Pol2
Heart Adult 8 weeks Pol2 TFBS ChIP-seq Signal from ENCODE/LICR
Kidney 8w Pol2
Kidney Adult 8 weeks Pol2 TFBS ChIP-seq Signal from ENCODE/LICR
MEF 8w Pol2
MEF Adult 8 weeks Pol2 TFBS ChIP-seq Signal from ENCODE/LICR
2.1
0
-3.3
Mammal Cons
Placental Mammal Basewise Conservation by PhyloP
Multiz Alignments of 30 Vertebrates
Rat
Human
Orangutan
Dog
Horse
Opossum
Chicken
Stickleback

Figure 4.4. Adapted ENCODE alignment tracks for human and mouse histone modifications, TFBS ChIP-seq and Multiz alignment conservation from the UCSC Genome Browser (July 2007 NCBI37/mm9 mouse strain C57BL/6) for human aligned rs4773144 CAD-associated Chr13q34 locus. Track 1) shows the lead CAD GWA study SNPs rs4773144 and rs3809346, plus 2 further SNPs in strong pairwise LD ($r^2 \approx 1$) in green, and 2 extra SNPs with LD $r^2 \geq 0.5 < 0.8$ in blue (Red vertical dashed lines indicate the SNP locations throughout the other alignment tracks); Track 2) shows UCSC genes for the promoter region of head-to-head genes *col4a1* and *col4a2*; Track 3) shows raw ChIP-seq signal for H3K4me1, H3K4me3 and H3K27Ac histone marks in 8 week mouse embryonic fibroblasts (MEF), plus embryonic heart and kidney tissue; Track 4) shows raw ChIP-seq signal for Pol2 (p300) in 8 week mouse embryonic fibroblasts (MEF), plus embryonic heart and kidney tissue; Track 5) shows Multiz alignments for conserved vertebrate species. (Tracks with the strongest peaks and signals are indicated by darker shading).



Data supporting chr13:110960711 (rs4773144)

Score: 2b

Likely to affect binding

Human Feb. 2009 (GRCh37/hg19) chr13:110,960,511-110,960,911 (401 bp)
100 bases
hg19
110,960,600
110,960,650
110,960,700
110,960,750
110,960,800
110,960,850
110,960,900
RefSeq Genes
Publications: Sequences in scientific articles
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
Digital DNaseI Hypersensitivity Clusters in 125 cell types from ENCODE
Transcription Factor ChIP-seq from ENCODE
Placental Mammal Basewise Conservation by PhyloP
Simple Nucleotide Polymorphisms (dbSNP 137) Found in >= 1% of Samples
Repeating Elements by RepeatMasker

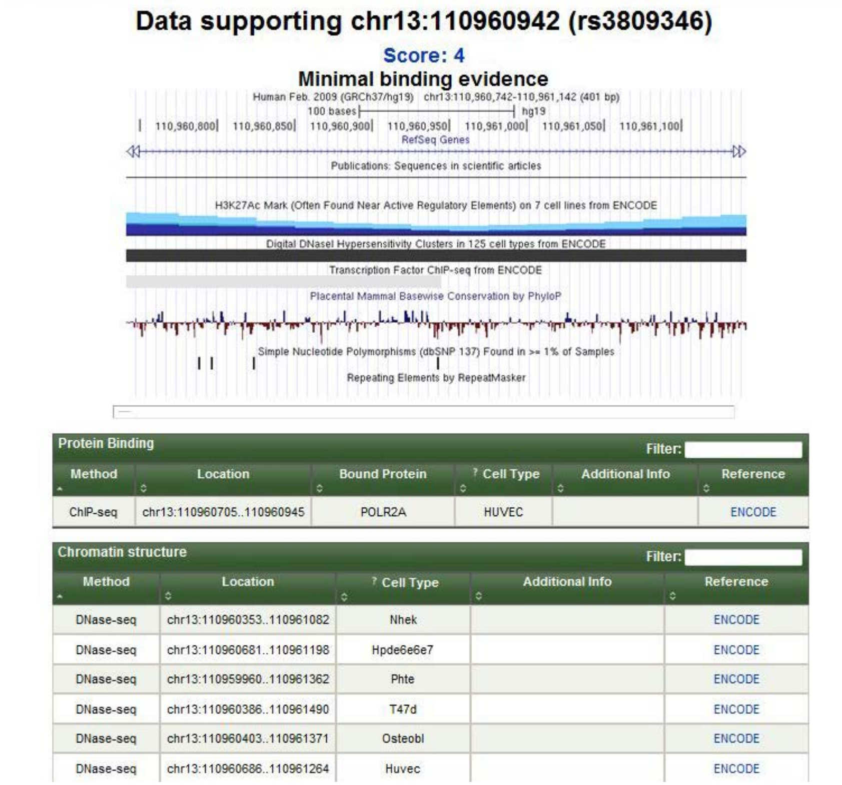
Protein Binding Filter:

Method	Location	Bound Protein	? Cell Type	Additional Info	Reference
ChIP-seq	chr13:110960705..110960945	POLR2A	HUVEC		ENCODE

Motifs Filter:

Method	Location	Motif	? Cell Type	PWM	Reference
Footprinting	chr13:110960702..110960716	STAT3:STAT3	Gliobla		21106904
Footprinting	chr13:110960702..110960716	STAT3:STAT3	H1hesc		21106904
PWM	chr13:110960702..110960716	STAT3:STAT3			16381825

Figure 4.5*A* Regulome DB results for rs4773144



Data supporting chr13:110960942 (rs3809346)

Score: 4

Minimal binding evidence

Human Feb. 2009 (GRCh37/hg19) chr13:110,960,742-110,961,142 (401 bp)
100 bases
hg19
110,960,800
110,960,850
110,960,900
110,960,950
110,961,000
110,961,050
110,961,100
RefSeq Genes
Publications: Sequences in scientific articles
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
Digital DNaseI Hypersensitivity Clusters in 125 cell types from ENCODE
Transcription Factor ChIP-seq from ENCODE
Placental Mammal Basewise Conservation by PhyloP
Simple Nucleotide Polymorphisms (dbSNP 137) Found in >= 1% of Samples
Repeating Elements by RepeatMasker

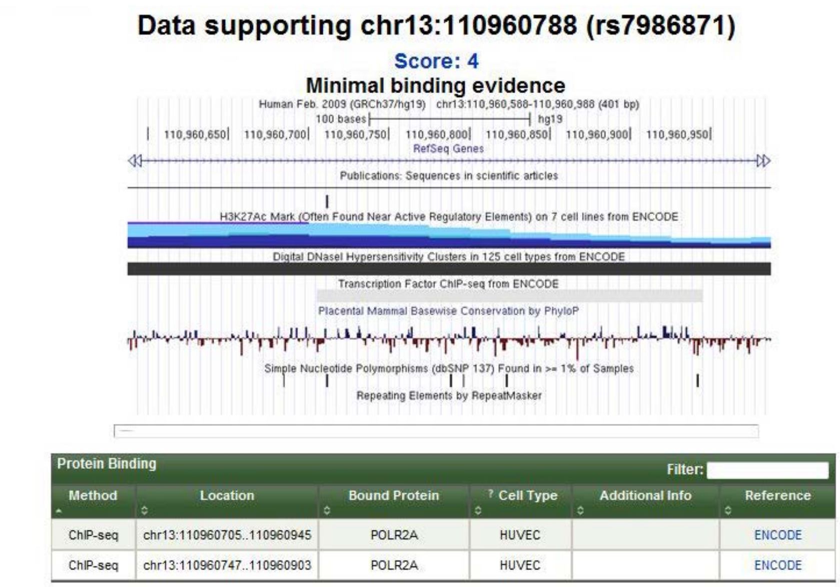
Protein Binding Filter:

Method	Location	Bound Protein	? Cell Type	Additional Info	Reference
ChIP-seq	chr13:110960705..110960945	POLR2A	HUVEC		ENCODE

Chromatin structure Filter:

Method	Location	? Cell Type	Additional Info	Reference
DNase-seq	chr13:110960353..110961082	Nhek		ENCODE
DNase-seq	chr13:110960681..110961198	Hpde6e6e7		ENCODE
DNase-seq	chr13:110959960..110961362	Phte		ENCODE
DNase-seq	chr13:110960386..110961490	T47d		ENCODE
DNase-seq	chr13:110960403..110961371	Osteobl		ENCODE
DNase-seq	chr13:110960686..110961264	Huvec		ENCODE

Figure 4.5*B* Regulome DB results for rs3809346



Data supporting chr13:110960788 (rs7986871)

Score: 4

Minimal binding evidence

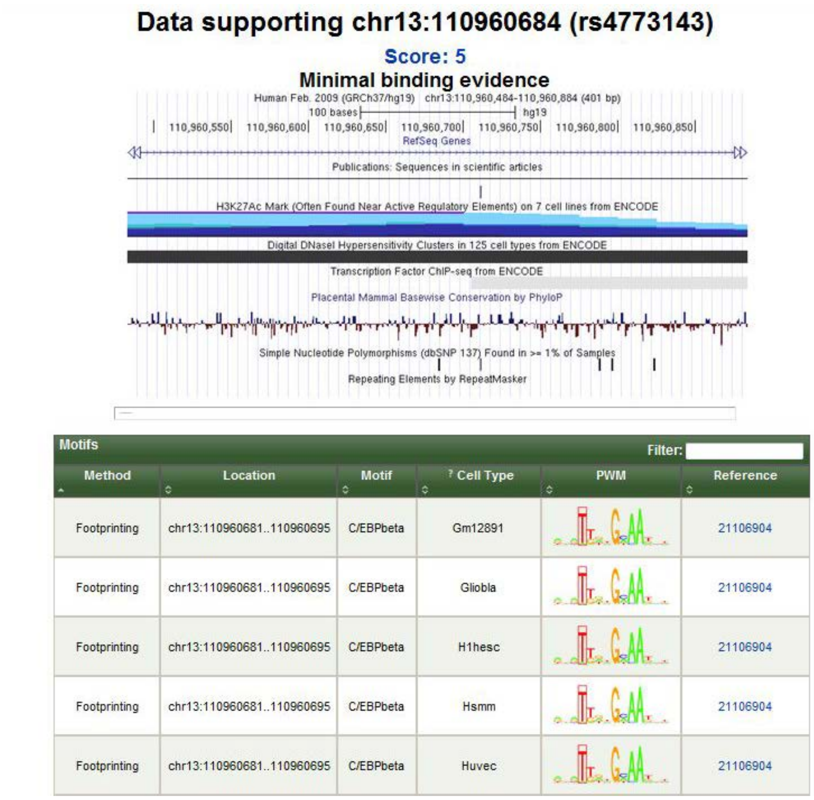
Human Feb. 2009 (GRCh37/hg19) chr13:110,960,588-110,960,988 (401 bp)
100 bases
hg19
110,960,650
110,960,700
110,960,750
110,960,800
110,960,850
110,960,900
110,960,950
RefSeq Genes
Publications: Sequences in scientific articles
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
Digital DNaseI Hypersensitivity Clusters in 125 cell types from ENCODE
Transcription Factor ChIP-seq from ENCODE
Placental Mammal Basewise Conservation by PhyloP
Simple Nucleotide Polymorphisms (dbSNP 137) Found in >= 1% of Samples
Repeating Elements by RepeatMasker

Protein Binding

Filter:

Method	Location	Bound Protein	? Cell Type	Additional Info	Reference
ChIP-seq	chr13:110960705..110960945	POLR2A	HUVEC		ENCODE
ChIP-seq	chr13:110960747..110960903	POLR2A	HUVEC		ENCODE

Figure 4.5*C* Regulome DB results for rs7986871



Data supporting chr13:110960684 (rs4773143)

Score: 5

Minimal binding evidence

Human Feb. 2009 (GRCh37/hg19) chr13:110,960,484-110,960,884 (401 bp)
100 bases
hg19
110,960,550
110,960,600
110,960,650
110,960,700
110,960,750
110,960,800
110,960,850
RefSeq Genes
Publications: Sequences in scientific articles
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
Digital DNaseI Hypersensitivity Clusters in 125 cell types from ENCODE
Transcription Factor ChIP-seq from ENCODE
Placental Mammal Basewise Conservation by PhyloP
Simple Nucleotide Polymorphisms (dbSNP 137) Found in >= 1% of Samples
Repeating Elements by RepeatMasker

Motifs

Filter:

Method	Location	Motif	? Cell Type	PWM	Reference
Footprinting	chr13:110960681..110960695	C/EBPbeta	Gm12891		21106904
Footprinting	chr13:110960681..110960695	C/EBPbeta	Gliobla		21106904
Footprinting	chr13:110960681..110960695	C/EBPbeta	H1hesc		21106904
Footprinting	chr13:110960681..110960695	C/EBPbeta	Hsmm		21106904
Footprinting	chr13:110960681..110960695	C/EBPbeta	Huvec		21106904

Figure 4.5*D* Regulome DB results for rs4773143

Figure 4.5*A-D*. Shows the RegulomeDB binding scores based upon experimental ChIP-seq protein binding from ENCODE and predicted binding motifs.

Table 4.4A. Putative transcription factor binding sites for SNPs rs4773144 and rs3809346, predicted using TFBS web-based prediction tools.

SNPs	Transcription Factor	Strand	Major Allele		Minor Allele		Binding Score; PWM Sim. Difference	Prediction Program PWM
			Matrix Sim.	Sequence	Matrix Sim.	Sequence		
rs4773144	SPIB	(+)	0.91	acggga**a**	0.75	acggga**a**	4.8; 0.16	JASPAR
	FEV	(+)	0.81	cggga**a**ct	0.67	cggga**g**ct	3.8; 0.14	JASPAR
	GABPA	(+)	0.8	cggga**a**ctggg	0.68	cggga**g**ctggg	7.0; 0.12	JASPAR
	ETS1	(-)	0.83	g**t**tccc	0.65	g**c**tccc	4.0; 0.18	JASPAR
	SPI1	(+)	0.83	ggga**a**ct	0.7	ggga**g**ct	4.0; 0.13	JASPAR
	STAT6	(-)	0.91	ccag**T**TCCcgtgaaaggcc	<0.75	ccag**c**TCCcgtgaaaggcc	Na; >0.16	MatInspector
	STAT6	(+)	0.91	gcctTTCAcggga**a**ctggg	0.78	gcctTTCAcggga**g**ctggg	Na; 0.13	MatInspector
	V$STAT1_01	(+)	O4F	ggcctttcacggga**A**ctggga	Na	ggcctttcacggga**g**ctggga	Na; O4F	TRANSFAC
	V$STAT1_02	(-)	O4F	cag**T**tccc	Na	cag**c**tccc	Na; O4F	TRANSFAC
	STAT3	(+)	O4F	ttcacggga**A**ctgg	Na	ttcacggga**g**ctgg	na	RegulomeDB
	V$MIF1_01	(+)	0.8	ggcctttcacgGGA**A**ctgg	<0.75	ggcctttcacgGGA**g**ctgg	Na; >0.05	MatInspector
	V$MIF1_01	(+)	O4F	gcctttcacggga**A**ctgg	Na	gcctttcacggga**g**ctgg	Na; O4F	TRANSFAC
rs3809346	BRCA1	(+)	0.71	ac**g**agac	0.9	ac**a**agac	-4.0; 0.19	JASPAR
	SOX10	(-)	0.63	tct**c**gt	0.82	tct**t**gt	-4.3; 0.19	JASPAR
	NFE2L1::MafG	(+)	0.68	c**g**agac	0.8	c**a**agac	-2.8; 0.12	JASPAR
	V$EVI1	(+)	0.78	cgaac**g**AGACaacctta	0.88	cgaac**a**AGACaacctta	Na; -0.10	MatInspector
	V$EVI1_06	(+)	na	ac**g**agacaa	O4F	ac**A**agacaa	Na; O4F	TRANSFAC
	V$MYT1	(-)	0.66	ctaAGGTtgtct**c**	0.76	ctaAGGTtgtct**t**	Na; -0.10	MatInspector

Legend for Tables 4.4A & 4.4B. Putative TFBS for the 4 lead SNPs, using different TFBS web-based prediction tools. Matrix similarity refers to how similar the test sequence is to the position weighted matrix (PWM). Sequence: 1) JASPAR – bold nucleotides indicate the SNP allele; 2) MatInspector – bold nucleotides indicate the SNP allele, 4 capitalised nucleotides indicate the core of the TFBS, red nucleotides are conserved; 3) TRANSFAC - bold nucleotides indicate the SNP allele and capitalized nucleotides are the lead allele with the higher % matrix similarity; RegulomeDB – provides TRANSFAC matrix hits optimized for function (O4F) and again the lead allele nucleotides are capitalized. Binding affinity score difference: 1) JASPAR –scores genomic sequence with the major or minor allele sequence for each matrix, here are shown binding affinity difference scores of >2.5, differences in matrix similarity are also provided; MatInspector – uses a matrix similarity threshold of 0.75 and red nucleotide conservation index (ci)-value >0.6, and individually O4F each matrix; TRANSFAC – each matrix is individually O4F based on matrix similarity >0.8 and conservation.

Table 4.4B. Putative transcription factor binding sites for SNPs rs4773143 and rs7986871, predicted using TFBS web-based prediction tools.

SNPs	Transcription Factor	Strand	Major Allele		Minor Allele		Binding Score; PWM Sim Difference	Prediction Program PWM
			Matrix Sim.	Sequence	Matrix Sim.	Sequence		
rs4774143	SOX10	(+)	0.92	ctgtg**t**	0.73	ctgtg**c**	4.2; 0.19	JASPAR
	BRCA1	(-)	0.87	aaa**a**cac	0.74	aaa**g**cac	2.7; 0.13	JASPAR
	FOXO3	(-)	0.87	gcaaa**a**ca	0.74	gcaaa**g**ca	3.9; 0.13	JASPAR
	FOXP1	(-)	1	cttgcaaA**A**CAcaggat	<0.75	cttgcaaA**g**CAcaggat	Na; <0.25	MatInspector
	CEBPA	(-)	0.84	ttgcaaa**a**c	0.69	ttgcaaa**g**c	4.2; 0.15	JASPAR
	CEBPA	(+)	0.91	**t**tttgcaag	0.77	**c**tttgcaag	3.9; 0.14	JASPAR
	CEBPB_01	(+)	O4F	gtg**T**tttgcaagcg	Na	gtg**c**tttgcaagcg	Na; O4F	TRANSFAC
	CEBPB_02	(-)	O4F	cgcttgcaaa**A**cac	Na	cgcttgcaaa**g**cac	Na; O4F	TRANSFAC
	CEBPB	(+)	O4F	gtg**T**tttgcaagcg	Na	gtg**c**tttgcaagcg	Na; O4F	RegulomeDB
	NFATC2	(+)	0.84	**t**tttgca	0.72	**c**tttgca	3.2; 0.12	JASPAR
	SOX10	(+)	0.71	**t**tttgc	0.81	**c**tttgc	-2.3; -0.10	JASPAR
	V$FAC1.01	(-)	0.96	tgcaaA**A**CAca	<0.75	tgcaaA**g**CAca	Na; >0.21	MatInspector
rs7986871	TEAD1	(-)	0.83	c**a**ctttcctggc	0.74	c**t**ctttcctggc	3.5; 0.09	JASPAR
	RORA_2	(+)	0.83	ggaaag**t**gtgtcag	0.74	ggaaag**a**gtgtcag	4.6; 0.09	JASPAR
	RORA_2	(+)	0.82	aggaaag**t**gTGTCagtcactagg	<0.75	aggaaag**a**gTGTCagtcactagg	Na; >0.07	MatInspector
	SPI1	(+)	0.83	ggaaag**t**	0.74	ggaaag**a**	2.9; 0.09	JASPAR
	ZEB1	(-)	0.81	c**a**cttt	0.65	c**t**cttt	3.9; 0.16	JASPAR
	GATA3	(+)	0.65	ag**t**gtg	0.83	ag**a**gtg	-4.4; -0.18	JASPAR
	AP1	(-)	0.93	tgacac**a**	0.82	tgacac**t**	3.05; 0.11	JASPAR
	V$AP1_Q2	(-)	O4F	ag**t**gtgtcagt	Na	ag**a**gtgtcagt	Na; O4F	TRANSFAC
	BACH2	(+)	0.91	ggaaag**T**GTGtcagtcactag	<0.75	ggaaag**a**GTGtcagtcactag	Na; >0.16	MatInspector

4.3.1.4 Overall conclusions from the in silico bioinformatics refinement

All four candidate CAD-associated 13q34 locus SNPs (rs4773143, rs4773144, rs7986871 and rs3809346) are situated within a bidirectional promoter, as identified by human and mouse ENCODE histone mark signatures, i.e. strong H3K4Me3, moderate H3K27Ac and low H3K4Me1 (**Fig.4.3A/C, Fig.4.4**). A multi-centre RNA Pol II ChIP-seq signal in HUVECs overlays rs4773144, rs7986871 and rs3809346 and is shortly downstream of rs4773143 (**Fig.4.3B**). Given the 200bp resolution of ChIP-seq, and the fact that ChIP-seq does not necessarily represent direct DNA-protein binding, any one of these SNPs, or indeed all of these SNPs could be involved in a transcriptional multi-component complex. Thus the 4-SNP haplotype as a whole could be of relevance to CAD risk; indeed three of the four SNPs (i.e. all except rs7986871) show moderately strong signals for bp resolution DNaseI footprints for DNA-protein binding (**Fig.4.3A**). Also, several family members of the homo- or hetero-dimer cMYC TFBS family, (i.e. cMYC, MAX, MXI1 and MAZ) align with or near to, all of the four candidate SNPs (**Fig.4.3B**). Moreover, another interesting observation was seen with E2F6 and MAX ChIP-seq signals for H1ES and A549 cells (**Fig.4.3B**). These two transcription factors showed strong signals with, or near to, all four candidate SNPs, and particularly with the known COL4 silencer (5'-CGCGCTTGGAC**TTGCGCGC**CCGAGA-3') sequence (Haniel et al. 1995). Furthermore, putative TFBS analyses identified: 1) a highly conserved TRANSFAC® TFBS prediction (see **Appendix 6.3, Table S4.1**) for a consensus E2F6 (TTGCGCGC); and 2) via JASPAR, a possible non-canonical E-box (CGCTTG) that could bind MAX (JASPAR binding affinity score for Arnt::Ahr = 5.7; matrix similarity score = 0.84). Interestingly, E2F6 is known to form a transcriptional repressor complex with a MAX-MGA heterodimer (see **Appendix 6.3, Figure S4.1**). Nonetheless, the E2F6 and MAX ChIP-seq signal, at or near to, all four candidate SNPs, could be indicative of an *in vivo* multi-component complex, which includes the COL4 silencer, and one or more of the four (13q34 locus), CAD-associated candidate SNPs.

All candidate 13q34 SNPs show evidence for being the causal SNP, but the SNP showing the strongest evidentiary support is rs3809346.

Th evidence supporting rs3809346 comes from several sources (**Table 4.3**). Firstly, FAIRE-seq signal peaks align very near to rs3809346, but not the other SNPs (**Fig.4.3A**). Secondly, DNaseI HS sites show peak density signals just beyond rs3809346, with no peaks aligning near the other SNPs (**Fig.4.3A**). Thirdly, the strongest bp resolution DNaseI footprints, align most strongly (in many cell types) with rs3809346, and a region ~100bp downstream that shows a very strong footprint signal. Fourthly, a mouse conserved 'dip' in all three histone mark signatures occurs over a distance of ~150-200bp (**Fig.4.3C, Fig.4.4**). This 'dip' is inclusive of rs3809346, the DNase HS site peaks, and the strong DNaseI footprint ~100bp beyond rs3809346. The 'dip' is also inclusive of the canonical SP1 GC box (16bp upstream of rs3809346). Fifthly, the rs3809346 risk allele (A) diminishes a CpG island by one CpG dinucleotide, i.e. reduces methylation of the adjacent cytosine. Sixthly, rs3809346 CAD risk allele (A) creates an EVI-1 putative TFBS (**Table 4.4A**). Of interest, EVI-1 is known to interact with SMAD3 and repress the TGFβ1-induced transcription of anti-growth genes, such as cyclin dependent kinase inhibitor P15, a.k.a. *CDKN2B* (which inhibits CDK4 or CDK6) – whereas SMAD3 interacts with SP1 and then binds to CBP/p300 for coactivation of TGFβ1-induced transcription (see **Appendix 6.3, Figure S4.2**) (Alliston et al. 2005; Derynck & Zhang 2003). It just so happens that a SP1 canonical site exists, just 16bp upstream of rs3809346.

Evidentiary support for the next most likely causal candidate SNP rs4773144 (the lead from CARDIoGRAM), comes from several sources also. Firstly, ENCODE alignments show bp high resolution DNaseI footprints very near to rs4773144. Secondly, strong signals for ChIP-seq TFBS SIN3A and MXI1 in H1ES cells are aligned with rs4773144. This is of interest because SIN3A is known to form a repressor complex with MXD1 (functionally similar to MXI1) and MAX (Swanson et al. 2004). Thirdly, RegulomeDB identified rs4773144 as being the SNP with the best evidence for putative binding by giving the best score (2b) (**Fig. 4.5A**).

Fourthly, three of four TFBS software prediction packages identified the rs4773144 CAD risk allele (G) disruption of a STAT TFBS: STAT3 for *COL4A2* (RegulomeDB); STAT6 for *COL4A1/A2* (MatInspector) and STAT1 for COL4A1/A2 (TRANSFAC®) (**Table 4.4A**). Fifthly, multiple ETS family putative TFBS were disrupted by the rs4773144 CAD risk allele (G) through the DNA binding motif (PU box) GGAA/T (**Table 4.4A**), a known PDGF and TNF-α induced ECM transcription factor expressed by endothelial cells, vascular SMC, mesangial cells and renal fibroblasts (Okano et al. 2012). The ETS1 transcription factor is involved in numerous pathways, including strong repression of TGFβ1-induced transcription of collagen type 1 alpha 2 (*COL1A2*), probably by interfering with the CBP/p300 and SMAD3-SP1 multi-component interaction (see **Appendix 6.3, Figure S4.3**) (Czuwara-Ladykowska et al. 2002).

The two almost perfect proxy SNPs ($r^2 \approx 1$) rs4773143 and rs7986871, also have some source evidence that points towards plausible DNA-protein binding. This evidence although weaker overall, does not rule out the possibility that either one of these SNPs, is causal of the CAD risk.

The evidence for rs4773143 comes from: 1) multiple cell, high bp resolution DNaseI footprints that align very near to rs4773143 (**Fig.4.3A**); 2) the ChIP-seq E2F6 and MAX signal intensity (**Fig.4.3B**); and 3) the putative TFBS disruption of six difference TFBS families, including FOX and C/EBP (**Table 4.4B**).

The evidence for rs7986871 comes from: 1) the BACH2 ChIP-seq signal that is weakly matched with the rs7986871 CAD risk allele (A) disruption of putative TFBS BACH1 (**Fig.4.3B, Table 4.4B**); 2) FAIRE-seq peak signal is nearest to rs7986871 after rs3809346 (**Fig.4.3A**); 3) the ChIP-seq E2F6 and MAX signal intensity (**Fig.4.3B**); 4) the CAD risk allele (A) disruption of six putative TFBS with high binding affinity differences (**Table 4.4B**); 5) a strong 100bp ChIP-seq signal for RNA Pol II (HUVECs) and a moderate ChIP-seq signal for TBP (H1ES) that align with rs7986871 (**Fig.4.4B**); 6) A strong signal for RNA-seq alignment with rs7986871 in one of two reads from CALTECH in HUVECs, but this was not seen in RNA-seq data from another source (CSHL), so it could be a spurious result (**Fig.4.3C**).

4.2.2 Assessment of rs4773144 as a COL4A1/A2 eQTL in the kidney

4.2.2.1 Polish kidney project subjects

The 130 Polish kidney project (PKP) subjects were recruited as part of an on-going study to collect healthy kidney tissue from renal cancer patients undergoing unilateral nephrectomy, from three centres in Poland (Silesian Renal Tissue Bank (SRTB), Poznań and Zabrze), for the purpose of extracting RNA to investigate cardiovascular gene expression. About 1cm^3 of tissue was collected from the healthy (unaffected by cancer) pole of the kidney immediately post-surgery, preserved in RNAlater® (Ambion®) and stored at -70^{0}C ready for future total RNA extraction. Recruited subjects were phenotyped and classified as either hypertensive or normotensive according to a set of criteria used in the Silesian hypertension study (SHS) (Tomaszewski et al. 2002; Tomaszewski et al. 2007).

4.2.2.2 RNA Extraction, DNaseI treatment and quantification

Total RNA was extracted from 130 human kidney samples using the commercially available assay RNeasy® Plus Mini Kit (Qiagen®), according to the manufacturer's handbook. The extracted total RNA was then subjected to recombinant DNaseI treatment using the commercially available DNA-*free*™ Kit- DNase treatment and removal reagents (Ambion®), for the purpose of removing genomic DNA. Finally, the total RNA extraction concentration and purity were measured using a NanoDrop™ 8000 (Thermo Fisher scientific, Inc.) spectrophotometer, good quality total RNA showed a 260/280 ratio ≥ 1.8.

4.2.2.3 cDNA synthesis by reverse transcription PCR (RT-PCR)

First strand cDNA was synthesised by preparing 200ng of total RNA in a maximum volume of 8.6µL sterile MilliQ (Milli-Q® system, EMD Millipore Corp.) water and adding this to a master mix prepared on ice with all components (sourced from: TaqMan® reverse transcription reagents (Applied Biosystems®), and 0.1M DTT (Invitrogen™) listed in **Table 4.5**.

Table 4.5. TaqMan® Reverse Transcription Reagents used in RT assay

Component	Volume (µl)	Final Conc.
TaqMan® RT Buffer (5X)	4	1x
$MgCl_2$ (25mM)	2	2.5mM
dNTPs (10mM)	2	1mM
DTT (100µM)	2	10µM
Oligo dT_{16} (50µM)	0.3	750nM
Random Hexamers (50µM)	0.3	750nM
RNase Inhibitor (20U/µL)	0.5	0.5U/µL
MultiScribe™ RTase (50U/µL)	0.3	750nM
Milli-Q Water	7.6	-
Total RNA (200ng/µL)	1	10ng/µL
Total	20	-

Legend. Components of the TaqMan® Reverse Transcription Reagents (Applied Biosystems®), except for 0.1M DTT (Invitrogen™).

To a 384-well PCR plate, 8.6µL of total RNA in MilliQ water was added followed by 11.4µL of master mix, the plate was then sealed and centrifuged briefly. A three-step RT-PCR was performed in a G-Storm GS4 thermal cycler (G-Storm, Somerton Biotechnology Centre) for: 1) 10 min @ 25^0C; 2) 12 min @ 42^0C; and 3) 5 min @ 85^0C; and the cDNA first strand PCR product stored at -70^0C.

4.2.2.4 Kidney cDNA TaqMan® genotyping of rs4773144

SNP rs4773144 was genotyped using a TaqMan® SNP genotyping assay (Applied Biosystems®), (for a description of the methods see ***Chapter 2, Section: 2.2.2***). Hardy-Weinberg equilibrium (HWE) was calculated in STATA using the HWE equation ($p^2 + 2pq + q^2 = 1$) and a chi-squared test (χ^2) with one degree of freedom (as we are assessing 2 alleles p and q), under the *null* hypothesis assumption that breeding is random and the genotyping data will not deviate from HWE.

4.2.2.5 TaqMan® real-time quantitative PCR (real-time qPCR)

Gene expression (mRNA) of the two target genes *COL4A1* and *COL4A2* were quantified using two inventoried TaqMan® FAM™ labelled *COL4A1* and *COL4A2 3'* MGB probes, and a VIC® labelled *β-2-microglobulin* (*B2M*) housekeeping gene 3'

MGB probe that was used to normalise the expression of the target genes (see **Table 4.6)**. The real-time qPCR fluorescence-based assay was performed using a ViiA™ 7 real-time PCR system instrument (Applied Biosystems®).

Table 4.6. TaqMan® labelled MGB probes used for quantitative real time PCR

Probe Target gene	Species	TaqMan assay ID	Fluorescent labelled dye
COL4A1	Human	Hs00266237_m1	FAM™
COL4A2	Human	Hs00300500_m1	FAM™
Beta-2-microglobulin (*B2M*)	Human	RefSeq: NM_004048.2	VIC®

A 20µL (TaqMan® gene expression assay) real-time qPCR total reaction volume was prepared per sample using 10µL of 2X TaqMan® gene expression master mix (Applied Biosystems), 1µL of 20X TaqMan® gene expression inventoried (250mM FAM™ /VIC®-dye labelled TaqMan® MGB probe / 9mM forward and reverse primers) assay mix (Applied Biosystems), 1µL of cDNA RT-PCR template, and 8µL of MilliQ water. All samples were run on the same 384-well PCR plate and each sample was run in duplicate for the two target genes (*COL4A1* and *COL4A2*) and the reference gene (*B2M*) on the same experimental plate for direct comparison to avoid batch effects. Real-time qPCR reaction conditions were as follows: 1) 2 min @ 50^{0}C; 2) 10 min @ 95^{0}C; and 3) 40 (two-step) cycles of 15 seconds @ 95^{0}C (denaturation) followed by 1 min @ 60^{0}C (annealing and extension). Importantly, prior to processing the full cohort of cDNA samples, an initial optimisation of the cDNA template input concentration was determined, by performing a serial dilution of a cDNA template of known concentration. This was used to generate a standard curve by plotting cycle threshold (Ct) values on the y-axis *vs* the logarithmic DNA concentrations (log[DNA]) on the x-axis. [**N.B.** The Ct value is a user defined early exponential phase cycle number threshold that is precisely proportional to the input cDNA amount, under the assumption that amplification should be doubling at 100% efficiency]. The standard curve plot provides an equation for a linear regression straight line (y = mx + b, where Ct = slope x log[DNA] + y-axis intercept), and the coefficient of determination (R^2) shows how well the experimental data fit the regression line; i.e. how linear the data are (N.B. R^2>0.98 shows good linearity).

The amplification efficiency (E) is calculated from the slope of the standard curve by the formula: $E = 10^{-1/slope}$, where E=2 represents 100% efficient amplification. Of note, a slope of -3.3219 gives a perfect 100% doubling; an acceptable slope parameter should ideally be between -3.6 to -3.1, i.e. 90-110%.

4.2.2.6 Statistical analysis of COL4A1/A2 real-time qPCR in kidney

The relative quantification measured is a comparison of both *COL4A1* and *COL4A2* gene expression by genotype to identify an eQTL. For directionality of fold change, the analysis initially used the comparative ΔCt method, commonly known as the ΔΔCt (delta delta Ct) method (Livak & Schmittgen 2001) with the formula: Fold change (R) = $2^{-\Delta\Delta Ct}$, where ΔCt = mean $Ct_{(\textit{target gene})}$ – mean $Ct_{(\textit{B2M}\_\text{ref gene})}$, and $\Delta\Delta Ct$ = mean $\Delta Ct_{(test)}$ - mean $\Delta Ct_{(calibrator)}$. (The calibrator is typically considered the untreated or basal control and the $2^{-\Delta\Delta Ct}$ fold change is calculated for the test (unknown) sample mean ΔCt and normalised to that of the calibrator). Therefore for rs4773144, the mean $\Delta Ct_{(calibrator)}$ is the non-risk major homozygote (AA), and the mean $\Delta Ct_{(test)}$ is either of two genotype risk allele (G) carriers, i.e. heterozygote (AG), or the minor homozygote (GG).

However, in order to get a more precise covariate adjusted comparative ΔCt relative gene expression, an alternative analysis was performed that is primarily equivalent to the standard ΔΔCt method formula: Fold change (R) = $2^{-\Delta\Delta Ct}$; here the relative difference in genotype mean ΔCt was covariate adjusted in a multiple linear regression model resulting in the formula: Fold change (R) = $2^{-(\beta\text{-coefficient})}$. Statistical analysis of the ΔCt values was performed in STATA® 12.1 (StataCorp LP) using a multiple linear regression, so as to be able to regress the dependent variable ΔCt, and adjust for relevant covariate independent variables age, sex, BMI, hypertensive status, centre and genotype under an additive, dominant or recessive model of genetic inheritance.

The *null* hypothesis is that the rs4773144 risk allele (G) will have no effect on gene expression in *COL4A1* or *COL4A2*.

4.3.2 Findings from the assessment of rs4773144 as a COL4A1/A2 eQTL in the kidney

4.3.2.1 Polish kidney project demographics

The demographics and clinical characteristics for the 133 Polish kidney project (PKP) subjects are summarised in **Table 4.7**. The majority of the subjects were hypertensive (67.7%) and most of these subjects were taking anti-hypertensive medication.

Table 4.7. Clinical phenotypic characteristics of subjects from PKP

Demographics	Polish kidney project
Subjects	133
Age (years), mean ±SD	61.1 ±10.6
Gender (M/F), n (%)	77 (57.9) / 56 (42.1)
Body mass index (kg/m^2), mean ±SD	27.7 ±4.4
Systolic BP (mmHg), mean ±SD	136.6 ±13.5
Diastolic BP (mmHg), mean ±SD	83.3 ±7.9
Hypertension, n (%)	90 (67.7)
Anti-hypertensive medication, n (%)	82 (61.7)

Legend. Data are given as means and standard deviations (SD), or counts and percentages.

4.3.2.2 COL4A1 and COL4A2 kidney genotyping analysis

The rs4773144 genotypes for the PKP subjects showed distinct clusters for the three genotypes AA, AG and GG (See **Figure 4.6**) and no deviation from Hardy-Weinberg Equilibrium (n=133, p=0.58).

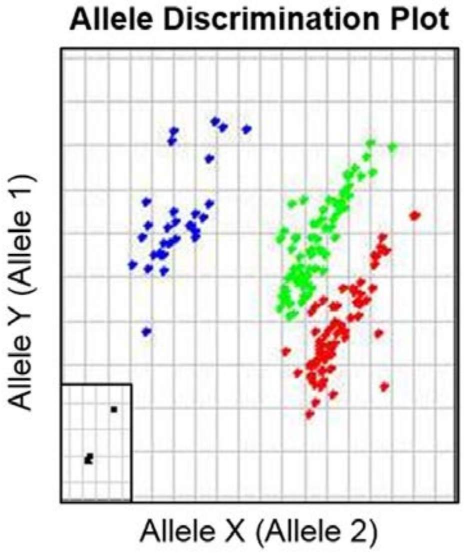


Figure 4.6. TaqMan® genotype cluster plot for rs4773144. Allele 1 on the Y-axis discriminated by fluorescent dye FAM™ are the minor homozygotes, GG (coloured blue); Allele 2 on the X-axis discriminated by fluorescent dye VIC® are the major homozygotes, AA (coloured red). Heterozygotes are the central cluster, AG (coloured green). No template control (NTC) samples are depicted as a black solid box, undetermined samples are depicted as are a cross 'X'.

4.3.2.3 COL4A1 and COL4A2 kidney gene expression analysis

Prior to running the PKP samples, a standard curve was plotted for the target (*COL4A1* and *COL4A2*) and reference (*B2M*) genes (see **Figure 4.7**), the amplification efficiencies (E) were all within the 90%-110% range and within 5% of each other and the coefficient of determination (R^2) was >0.98.

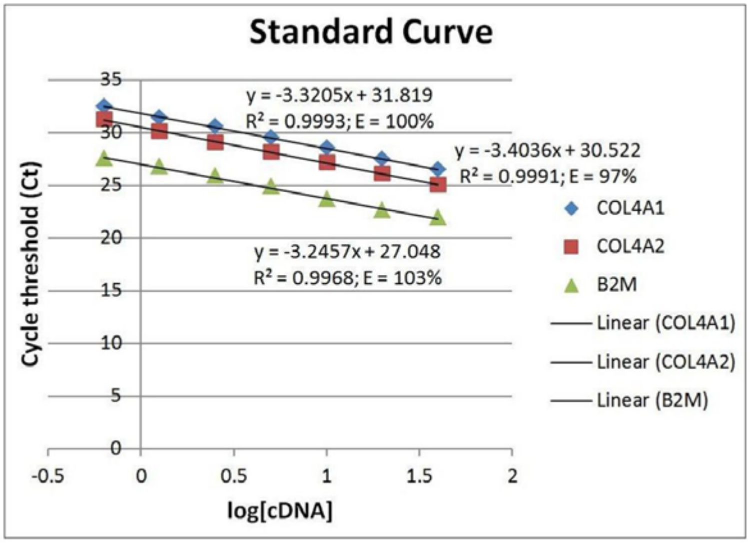


Figure 4.7. Standard curves for target (*COL4A1*, *COL4A2*) and reference (*B2M*) genes to measure linearity and amplification efficiencies.

TaqMan® gene expression analysis was carried out on n=133 samples. However, prior to ΔCt analyses, samples were excluded if: i) samples had a Ct difference between (technical replicate) duplicate readings >0.5 cycles; ii) samples were outliers beyond the linear range of the upper and lower points of the standard curve; and iii) samples had missing genotype data. Exclusions for *COL4A1*: i) technical replicates >0.5 cycles, n=8; ii) outliers beyond the standard curve linear range, n=16; and iii) missing genotypes, n=8. Exclusions for *COL4A2*: i) technical replicates >0.5 cycles, n=6; ii) outliers beyond the standard curve linear range, n=18); and iii) missing genotypes, n=6. This meant that for *COL4A1*, n=101 and for *COL4A2*, n=103 ΔCt values were taken forward for analysis.

A correlation plot of kidney mRNA gene expression *COL4A1* ΔCt values versus *COL4A2* ΔCt values (n=94) shows strong correlation, r=0.90 (see **Figure 4.8**).

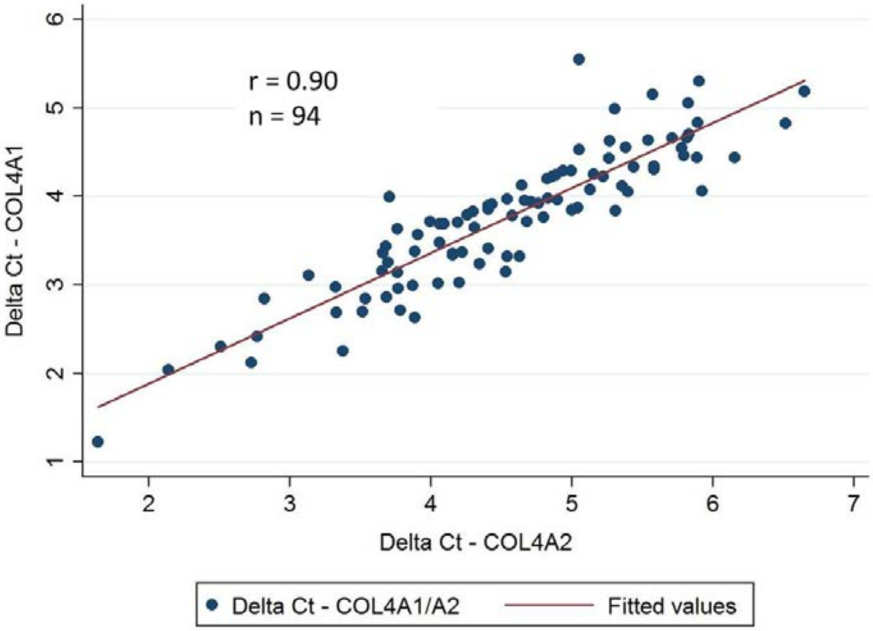


Figure 4.8. Pearson’s pairwise correlation (r) between *COL4A1* and *COL4A2* indicates that both genes are expressed to the same extent for the same kidney mRNA.

Following multiple linear regression analyses in STATA using the QC passed *COL4A1*, n=101 and *COL4A2*, n=103 ΔCt values after exclusions, I plotted residuals vs fitted samples as a means of checking the quality of the data for unusual patterns and potentially problematic outliers that lay well beyond a residual threshold of ±2.0 standard deviations. The data showed no unusual patterns that would cause concern, but did identify outliers for *COL4A1*, n=4 and *COL4A2*, n=5 ΔCt values (see **Figure 4.9*A*-*B***). As a consequence, the data were re-analysed as a sensitivity analysis, the results of which concluded that these samples should indeed be excluded, leaving *COL4A1* with n=97 ΔCt values and *COL4A2* with n=98 ΔCt values in the final analyses. **Figure 4.9*C*-*D*** show the plotted residuals vs fitted data used in the final analyses.

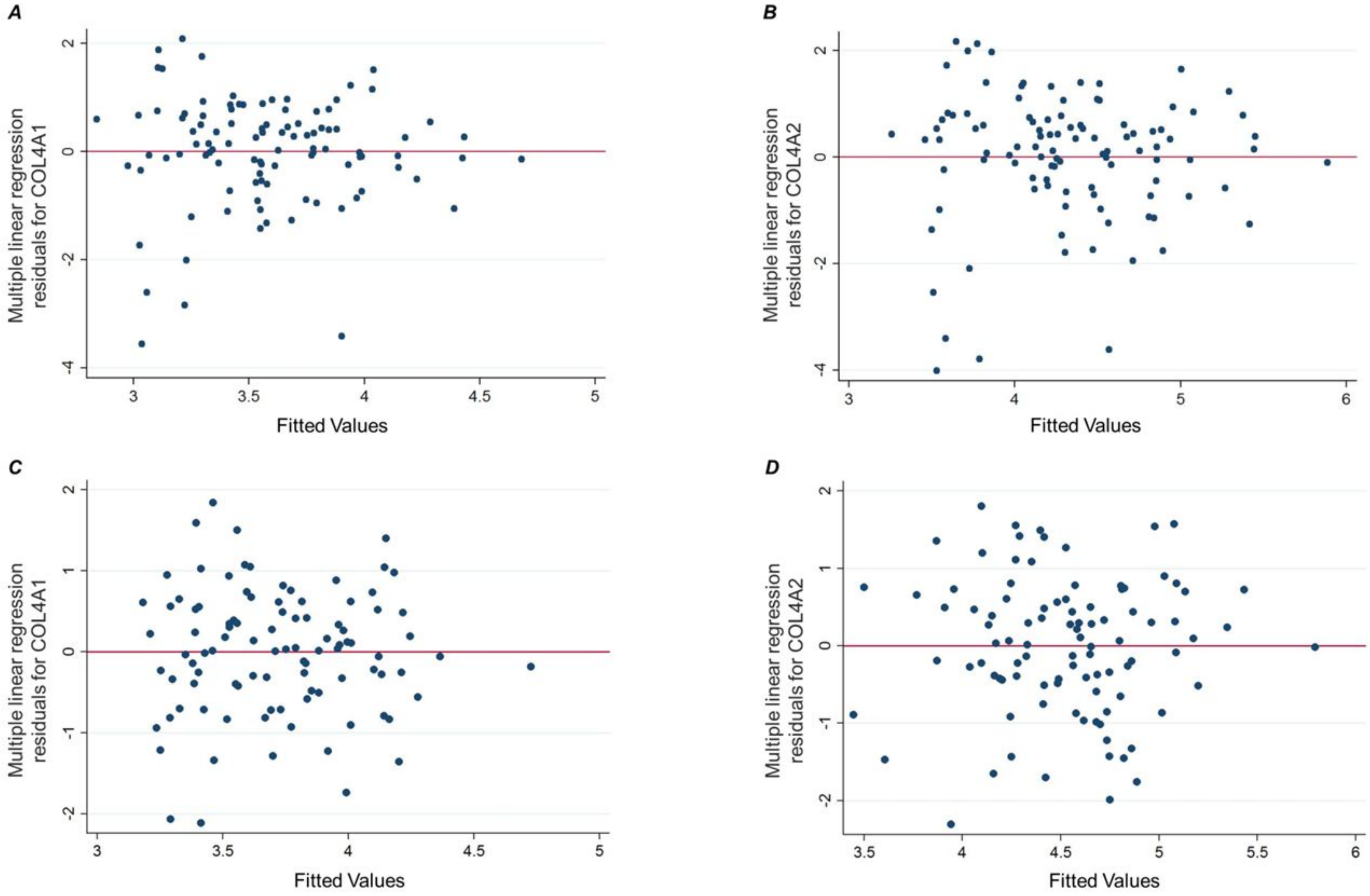


Figure 4.9. Scatter plots showing multiple linear regression residuals vs fitted data for *COL4A1* and *COL4A2* gene expression data. **Fig.** ***A*** and ***B*** show scatter plots before outlier exclusion for *COL4A1* and *COL4A2* respectively. **Fig.** ***C*** and ***D*** show scatter plots after outlier exclusion for *COL4A1* and *COL4A2* respectively.

The two bar graphs in **Figure 4.10*A*/*B*** show by multiple linear regression analysis adjusted for age, sex, BMI, centre and hypertension, a suggestive trend towards a relative fold decrease in renal *COL4A1* and *COL4A2* for those subjects that carry the CAD risk allele (G) compared to the subjects that carry the (normalised to 1) non-risk homozygous (AA) genotype. Under an additive model the relative fold decrease for *COL4A1* (see **Figure 4.10*A***) was - 0.847 (95% CI: 0.655-1.097) when comparing AA (non-risk) homozygotes to AG heterozygotes, and 0.754 (95% CI: 0.526-1.079) when comparing AA homozygotes to GG homozygotes; *P*=0.094 (for a dominant model (i.e. AA vs. AG+GG) - *P*=0.123, and a recessive model (i.e. GG vs. AG+AA) – *P*=0.269). A similar trend was also observed for *COL4A2* (see **Figure 4.10*B***), where under an additive model the relative fold decrease was - 0.813 (95% CI: 0.604-1.094) when comparing AA homozygotes to AG heterozygotes, and 0.732 (95% CI: 0.486-1.102) when comparing AA homozygotes to GG homozygotes; *P*=0.099 (for a dominant model - *P*=0.110, and a recessive model – *P*=0.330).

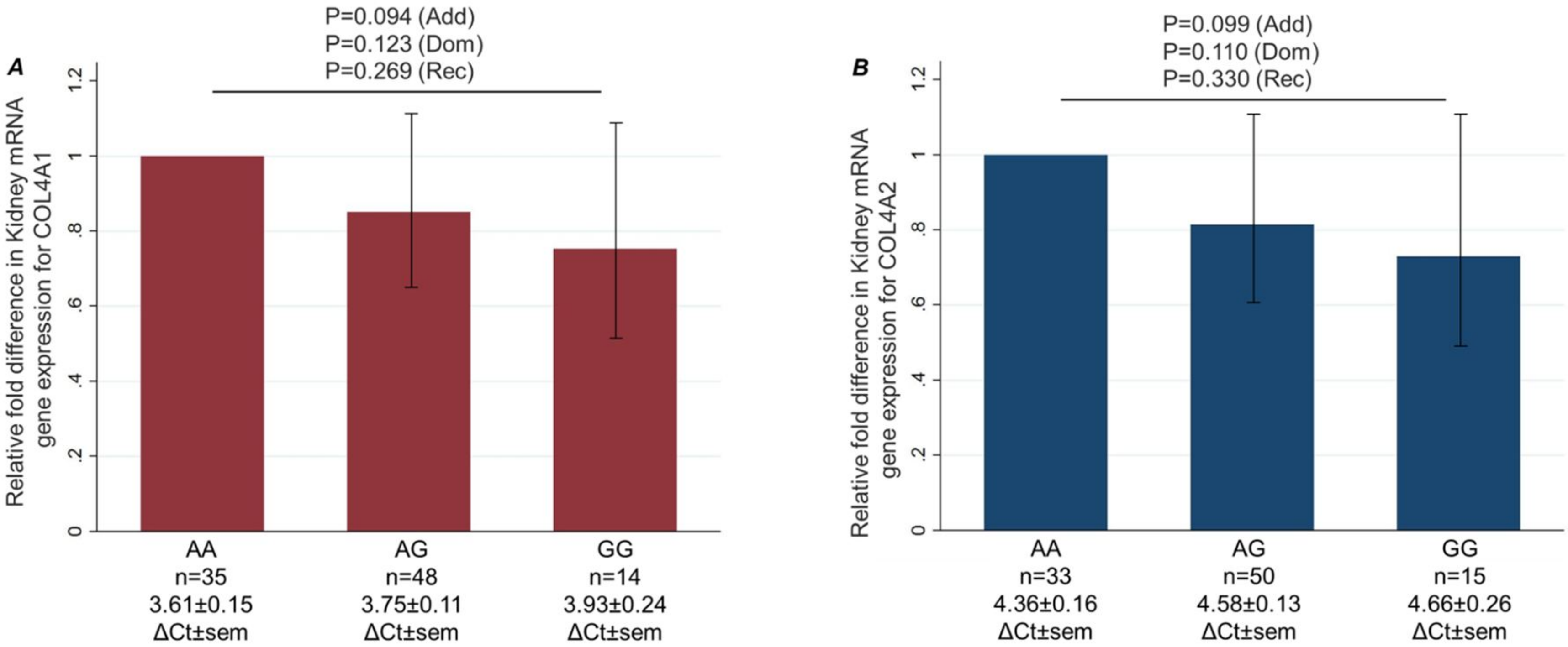


Figure 4.10A/B. Bar graphs show multiple linear regression adjusted relative fold change in renal COL4A1 (red bars) and COL4A2 (blue bars) gene expression mRNA levels (with 95% confidence intervals shown as error bars). **Fig.*A*** and ***B*** show the relative fold change in gene expression for rs4773144 risk allele (G) under an additive model, where AG and GG genotypes are compared with AA genotypes (normalised to 1); *P*-values are given for additive, dominant and recessive models of genetic inheritance.

4.2.3 Functional experiments using EMSA and Luciferase reporter assay

4.2.3.1 Cell culture

4.2.3.1.1 Materials and equipment

Primary human aorta smooth muscle cells (HASMC) (Gibco® Invitrogen™); HeLa (Henrietta Lacks cervical carcinoma epithelial-like cell-line) cells; medium 231 (Gibco® Invitrogen™); 20X smooth muscle growth supplement (SMGS) (Gibco® Invitrogen™); Dulbecco's modified eagles medium (DMEM), high glucose, GlutaMAX™, pyruvate (Gibco® Invitrogen™), foetal calf serum (FBS) (Gibco® Invitrogen™); 1% penicillin/streptomycin antibiotics; 0.05% trypsin/EDTA (0.2% trypsin/1mM EDTA) ; tissue culture flasks: T-75 cm^2 (NUNC), tissue culture plates: 6-well, 24-well (NUNC); sterile pipettes: 5mL, 10mL, 25mL (Sarstedt); Gilson™ pipettes and sterile tips; multichannel pipette and sterile tips; 10) pasteur pipettes, sterile, unplugged; 2% Trigene® disinfectant; 70% industrial methylated spirit (IMS) i.e. alcohol; sterile 50mL and 15 mL conical tubes (Fisher), laminar flow hood, 37^{0}C & 5% CO$_2$ Incubator, mechanical pipettor (Pipetboy), inverted phase contrast microscope and 37^{0}C water bath.

4.2.3.1.2 Maintenance of cell lines

HASMC and HeLa cells were both sub-cultured as monolayer adherent dependent cells in T-75 cm^2 culture flasks. HASMC (~4x10^5) were seeded in 12mL of medium 231 supplemented with SMGS, whereas HeLa (~8x10^5) cells were seeded in 12mL of DMEM with 10% (v/v) FCS and both cell types were supplemented with 1% penicillin/streptomycin to avoid contamination. Typically the cells were grown to 70-80% confluency (in 6-7 days for HASMC and 3-4 days for HeLa cells) and then passaged. Cells were passaged by removing the cell growth medium, quickly washing with 3mL of 0.05% Trypsin/EDTA and immediately removing, and then incubating with 3mL of 0.05% Trypsin/EDTA for 3-5 minutes at room temperature. After incubation, rap the flask very gently to dislodge cells from the surface and neutralise the cells in 12mL of fresh pre-warmed media containing SMGS for HASMC or 10% (v/v) FCS for HeLa cells. Finally, the volume of the cells was adjusted to enable a 1:3 dilution for HASMC and a 1:4 dilution for HeLa cells.

4.2.3.1.3 Freezing cells for storage

The long-term storage of cells was achieved by using Synth-a-Freeze® - Defined, Cryopreservation Medium (Invitrogen™). The manufacturer's instructions were followed and cells stored initially at 80^{0}C in a slow freezing insulation box containing propan-2-ol, and then subsequently transferred to liquid nitrogen after 24 hours.

4.2.3.2 Electrophoresis mobility (gel) shift assay (EMSA)

4.2.3.2.1 Preparation of whole cell lysate protein extracts

CelLytic™ M, mammalian cell lysis extraction reagent (Sigma-Aldrich) was used to extract whole cell lysate protein from endogenous unstimulated and 2µg/mL transforming growth factor beta 1 (TGFβ1) 3 hour stimulated HASMC (Gibco® Invitrogen) using the manufacturer's instruction leaflet. As recommended to avoid endogenous protease activity, a 1X final concentration supplement of c0mplete, mini, protease inhibitor cocktail (Roche) was added to the CelLytic M reagent.

4.2.3.2.2 Measurement of whole cell lysate protein concentration

The protein concentration was measured using Bradford Reagent (Sigma-Aldrich) according to manufacturer's instructions for the Standard 3.1 mL assay protocol, but with the recommended adaptation of halving the reaction volumes (i.e. protein sample used was 0.05mL, and the Bradford reagent used was 1.5mL). In brief, protein standards of bovine serum albumin (BSA; Sigma-Aldrich) at 0.25, 0.5, 1.0 and 1.4mg/mL were prepared and for the unknown sample of interest, the whole cell lysate protein extract was diluted 10X to ensure it was within the linear range of 0.1 to 1.4mg/mL. The 50µL protein samples were added to cuvettes and 1.5mL of 1X room temperature Bradford reagent was added and mixed prior to an absorbance reading at 595nm a fixed wavelength on a spectrophotometer.

The protein concentration was then calculated by preparing a graph of the standards (the standard curve) with the dependent variable (conc. mg/mL) on the X-axis and the independent variable (abs 595 nm) on the Y-axis and generating a 'best fit' linear equation straight line, y = mx + c, where (where y=absorbance, m=slope, c=constant and x=concentration). Therefore, the final concentration was achieved by rearrangement of the linear equation and multiplying by the dilution factor of 10.

4.2.3.2.3 Preparation of EMSA double-stranded labelled oligonucleotides

The genomic DNA sequences for complimentary sense and antisense oligonucleotides that were used as either labelled probes or unlabelled competitor for EMSA are shown in **Table 4.8**. Firstly, the lyophilised oligonucleotides were resuspended in TEN buffer (Tris-HCl pH8.0, 150mM NaCl and 5mM EDTA) at a

Table 4.8. Double stranded oligonucleotide sequences used in EMSA

Oligonucleotide Name	Sense/ Antisense	Sequence (5' to 3')	Length (nt)
rs4773143_(T)_Fwd	Sense	CAGCGTCAATCCTGTG**T**TTTGCAAGCGTCGGC	32
rs4773143_(T)_Rev	Antisense	GCCGACGCTTGCAAA**A**CACAGGATTGACGCTG	32
rs4773143_(C)_Fwd	Sense	CAGCGTCAATCCTGTG**C**TTTGCAAGCGTCGGC	32
rs4773143_(C)_Rev	Antisense	GCCGACGCTTGCAAA**G**CACAGGATTGACGCTG	32
rs4773144_(A)_Fwd	Sense	CCTTTCACGGGA**A**CTGGGAACTTA	24
rs4773144_(A)_Rev	Antisense	TAAGTTCCCAG**T**TCCCGTGAAAGG	24
rs4773144_(G)_Fwd	Sense	CCTTTCACGGGA**G**CTGGGAACTTA	24
rs4773144_(G)_Rev	Antisense	TAAGTTCCCAG**C**TCCCGTGAAAGG	24
rs7986871_(A)_Fwd	Sense	GCCAGGAAAG**A**GTGTCAGTCA	21
rs7986871_(A)_Rev	Antisense	TGACTGACAC**T**CTTTCCTGGC	21
rs7986871_(T)_Fwd	Sense	GCCAGGAAAG**T**GTGTCAGTCA	21
rs7986871_(T)_Rev	Antisense	TGACTGACAC**A**CTTTCCTGGC	21
rs3809346_(G)_Fwd	Sense	CGAGGAGGCGAAC**G**AGACAACCTTAGTA	28
rs3809346_(G)_Rev	Antisense	TACTAAGGTTGTCT**C**GTTCGCCTCCTCG	28
rs3809346_(A)_Fwd	Sense	CGAGGAGGCGAAC**A**AGACAACCTTAGTA	28
rs3809346_(A)_Rev	Antisense	TACTAAGGTTGTCT**T**GTTCGCCTCCTCG	28

final concentration of 200ng/µL. Annealing was achieved by mixing the sense and antisense oligonucleotides in a 1:1 molar ratio, incubating the samples for 10 minutes at 95^{0}C in a heat block, and slowly cooling within the heat block to 15-25^{0}C. The double-stranded (ds) oligonucleotides were then further diluted in TEN buffer to 25ng/µL prior to the digoxigenin (DIG) labelling reaction, so that a close approximation of 3.885pmol of ds oligonucleotide starting material could be added to the labelling reaction (for the purpose of controlling the downstream EMSA reaction labelled probe concentration to ~15-30fmol). The labelling reaction was prepared on ice and contained: 1) 3.885pmol of ds oligonucleotide, made up to a final volume of 10µL with MilliQ water; and 2) final concentrations of 1x labelling buffer (Roche), 5mM $CoCl_2$ (Roche), 0.05mM DIG-ddUTP (Roche) and 20U/µL terminal transferase (Roche), totalling 20µL in the reaction. The labelling reaction was then mixed and centrifuged briefly, incubated at 37^{0}C for 15 minutes, placed on ice, and the reaction stopped by adding 2µL of 0.2M EDTA (pH8.0). Finally, 3µL of MilliQ water was added to make a final volume of 25µL, and generate a final concentration of 0.1554pmol/µL.

The labelling reaction efficiency of was determined by direct nylon membrane spot blotting comparison of a 10-fold dilution series (i.e. 0, 0.1554fmol/µL, 1.554fmol/µL, 15.54fmol/µL, 155.4fmol/µL) to a known concentration control. The spot blot dilutions were visualised by chemiluminescence and autoradiography (as described in section 4.2.3.2.4)

4.2.3.2.4 EMSA (Gel shift)

On ice, DIG-labelled ds oligonucleotides (20fmol), 1x binding buffer (Roche), 1µg poly [d(I-C)], 0.1µg poly L-lysine, MilliQ water and 5µg of whole cell lysate protein extract were gently mixed in a standard gel shift reaction and incubated for 15 minutes at room temperature (15-25^{0}C). Simultaneously, four additional control gel shift reactions were also prepared by either: 1) not adding any protein (negative control); 2) adding a 200x unlabelled identical probe competitor; 3) adding a 200x unlabelled alternate allele for the same SNP competitor; and 4) adding a 200x unlabelled non-specific ds oligonucleotide that recognises a different protein consensus site (i.e. a consensus site that is essentially mutated from that of the tested DIG-labelled ds oligonucleotide). All reactions were made up to a 20uL total volume by altering the MilliQ water volumes accordingly. After incubation, the reactions were placed on ice and 5µL of 5X Novex® high-density TBE sample buffer (Invitrogen™) was added to each reaction tube. Samples were then, immediately resolved on a pre-electrophoresed (≥5 min), pre-cast, Novex® 6% DNA retardation polyacrylamide gel 1.0mm, 10 well (Invitrogen™), in 0.5X Novex® TBE (prepared from a 10x conc.: 890mM Tris, 890mM boric acid, 20mM EDTA, pH 8.0) running buffer (Invitrogen™) at 100V for 90 minutes at 4^0C,(i.e. run sample loading dye to two-thirds only), within the XCell SureLock™ mini-cell apparatus (Invitrogen™).

Post-electrophoresis, the migrated samples were transferred by electro-blotting, to a Nylon membrane, positively charged (Roche) in a XCell II™ blot module, run for 60 minutes at 300mA. After electro-blotting, the DNA-Protein complexes are cross-linked to the nylon membrane by first placing the membrane on a Whatman 3mm blotting paper pre-soaked in 2x saline-sodium citrate (SSC) buffer, and then cross-linking at 120mJ for 2 minutes in a Stratalinker (Stratgene). DIG-labelled oligonucleotides were visualised via an enzyme immunoassay following the manufacturer's instructions by incubating the nylon membrane with 75mU/mL anti-DIG, F(ab) fragments conjugated with alkaline phosphatase (Roche) and a subsequent chemiluminescence reaction with 100µg/mL CPSD substrate (Roche), followed by Amersham Hyperfilm ECL (GE Healthcare Life Sciences) exposure (~15 to 30 minutes).

4.3.3 Findings from the functional experiments using EMSA and Luciferase reporter assay

4.3.3.1 Gel shift DNA-Protein binding assay (EMSA)

The **Figures 4.11*A*** and 4.**11*B*** show rs4773144 specific DIG-labelled ds oligonucleotide EMSA gel shift experiments for unstimulated and TGFβ1 3hr stimulated HASMC whole cell protein extract, respectively. There is no major difference in binding affinity between the labelled ds oligonucleotides for the major and minor alleles for rs4773144 with or without TGFβ1 stimulation (see lanes 2 and 7 for **Figures 4.11*A*** and 4.**11*B***). The major and minor allele specific DIG-labelled probes are both out-competed by their unlabelled same allele-probe or their unlabelled alternative-allele-probe (i.e. the major or minor allele of rs4773144 for lanes 3 and 4 and the minor and major allele of rs4773144 for lanes 8 and 9 in **Figures 4.11*A*** and 4.**11*B***), indicating the shift position of specific DIG-labelled probe DNA-protein binding, but with no allele specific binding difference seen. Whilst the non-specific positive control is unable to compete with the DIG-labelled probe and has no effect on the labelled probe shift. Therefore, as there is no allele specific differential binding it is unlikely that a SNP allele change is affecting a sequence specific transcription factor binding site with endogenous unstimulated or TGFβ1-3hr-stimulated HASMC cellular lysate proteomes. The same result was found for the three other SNPs (rs4773143, rs7986871 and rs3809346) in almost perfect pairwise LD ($r^2 \approx 1$), [data not shown].

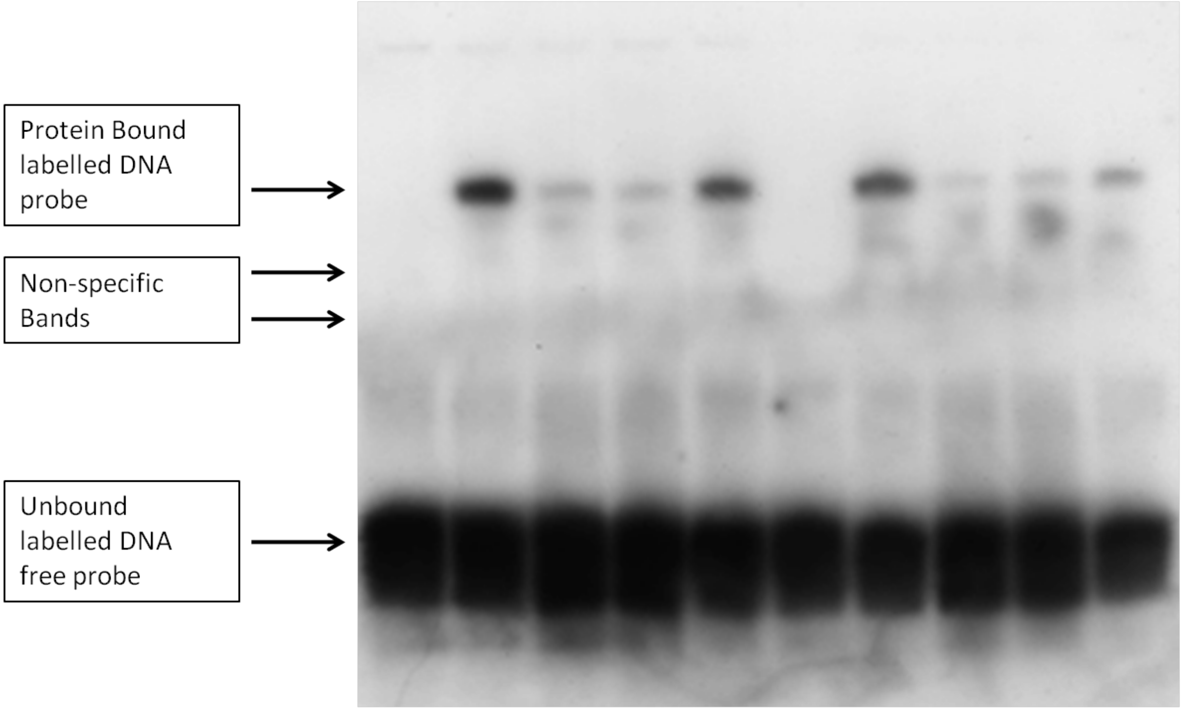

Protein Bound labelled DNA probe
Non-specific Bands
Unbound labelled DNA free probe

Figure 4.11*A*. EMSA (gel shift) showing rs4773144 specific DIG-labelled DNA probe with HASMC unstimulated whole cell protein lysate.

The EMSA gel image (26.5min exposure) compares the differential binding affinities of two 20fmol DIG-labelled ds oligonucleotide (24bp) probes for rs4773144 major (A) and minor (G) alleles with a 5µg protein lysate extracted from endogenous unstimulated HASMCs.

Lanes 1-5, all contain rs4773144_A allele DIG-labelled probe: Lane1 = No lysate ctrl; Lane 2 = gel shift; Lane 3 = 125x unlabelled rs4773144_A competitor; Lane 4 = 125x unlabelled rs4773144_G competitor; Lane5 = non-specific competitor (PPARG);
Lanes 6-10, all contain rs4773144_G allele DIG-labelled probe: Lane6 = No lysate ctrl, Lane7 = gel shift, Lane8 = 125x unlabelled rs4773144_G competitor, Lane9 = 125x unlabelled rs4773144_A competitor, Lane10 = non-specific competitor (PPARG).

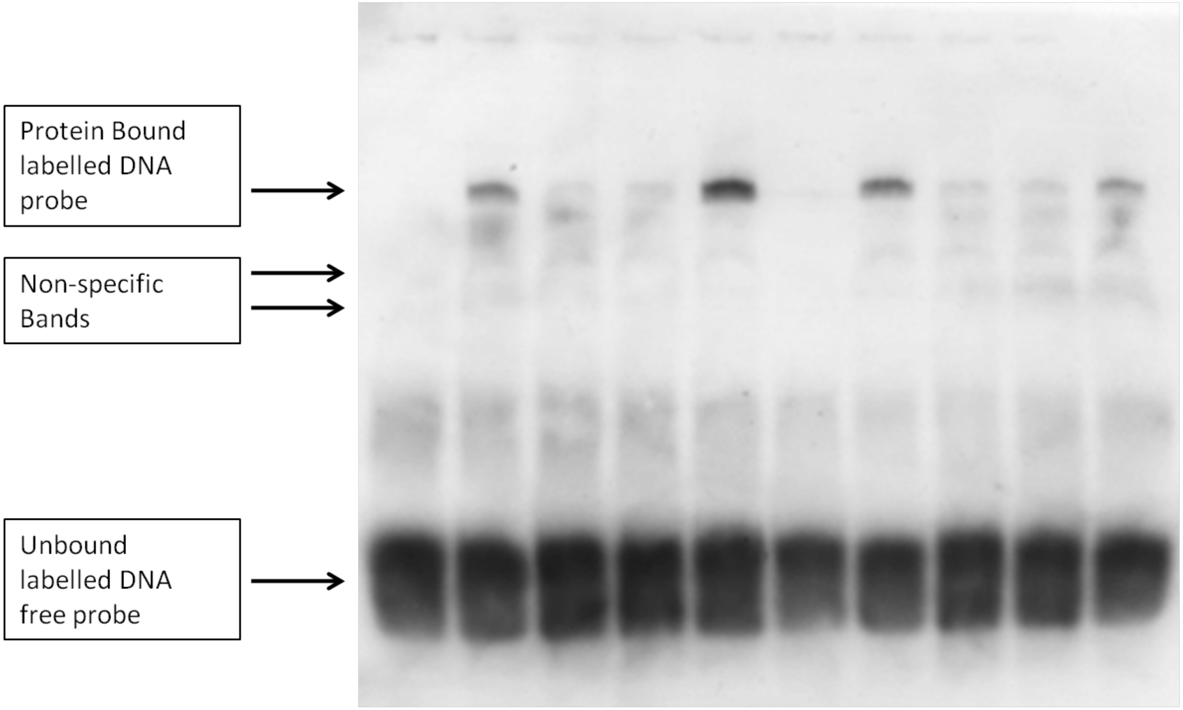

Protein Bound labelled DNA probe
Non-specific Bands
Unbound labelled DNA free probe

Figure 4.11*B*. EMSA (gel shift) showing rs4773144 specific DIG-labelled DNA probe with HASMC TGFβ1 stimulated whole cell protein lysate.

The EMSA gel image (18 min exposure) compares the differential binding affinities of two 20fmol DIG-labelled ds oligonucleotide (24bp) probes for rs4773144 major (A) and minor (G) alleles with a 5µg protein lysate extracted from TGFβ1, 3 hour stimulated HASMC.

Lanes 1-5, all contain rs4773144_A allele DIG-labelled probe: Lane1 = No lysate ctrl; Lane 2 = gel shift; Lane 3 = 125x unlabelled rs4773144_A competitor; Lane 4 = 125x unlabelled rs4773144_G competitor; Lane5 = non-specific competitor (PPARG);
Lanes 6-10, all contain rs4773144_G allele DIG-labelled probe: Lane6 = No lysate ctrl, Lane7 = gel shift, Lane8 = 125x unlabelled rs4773144_G competitor, Lane9 = 125x unlabelled rs4773144_A competitor, Lane10 = non-specific competitor (PPARG).

4.2.3.3 Cloning by DNA manipulation and Luciferase reporter gene assay

4.2.3.3.1 Promega cloning strategy to enable the Luciferase reporter expression assay

In order to perform the luciferase reporter expression assay, I first cloned a DNA PCR product 695bp in length, using forward and reverse genomic context primers, but with 12bp restriction-enzyme tails (selected for the purpose of sub-cloning later into the luciferase reporter vector pGL3-promoter) into a cloning vector plasmid called pGEM®-T Easy. The cloning site for this vector is very simple, as the vector has two 3'-T overhangs to enable greater ligation efficiency, by preventing re-circularisation of the vector, and providing a compatible overhang for PCR product insertions. A single deoxyadenosine is added to the 3'-ends of PCR amplicons by incubating for 20 minutes at 72^{0}C with a standard Taq polymerase (by making use of its non-template-dependent terminal transferase activity). Ligation reactions were set-up with pGEM®-T Easy vector and compatible PCR amplicons using T4 DNA Ligase in a 3:1 vector:insert ratio (as described in section 4.2.3.3.6) at 4^{0}C for ~16hr. Clean ethanol precipitated plasmid ligation product was then heat shock transformed into DH5α™ (*E.coli*) competent cells, and subsequently allowed to recover in pre-warmed LB media for an hour, before being plated-out onto LB-agar-ampicillin-Xgal plates. The Xgal was used to make use of the beta-galactosidase selection marker carried by the vector, which allows selection of positively transformed colonies after overnight growth at 37^{0}C – white colonies are successful transformants with DNA inserted into the plasmid and blue colonies are unsuccessful transformants with no insert. The plasmid containing the PCR-amplicon DNA of interest was then amplified by picking a single positive white colony and spiking a larger volume of LB media with the same ampicillin selection marker. The resultant culture was then extracted and purified using a plasmid extraction kit. The purified plasmid was then cut with restriction enzymes to identify whether the insert is the expected size and in the expected orientation. If an agarose gel provides expected band sizes, the plasmid is also sequenced to check the insert has no unexpected errors in comparison to the genomic sequence. The pGEM®-T Easy-719bp-insert construct and pGL3-promoter luciferase reporter cloning vector can then be digested with restriction enzymes corresponding to the 12bp restriction enzymes that flank the genomic 695bp insert and the MCS of the pGL3-promoter vector, ready for repeating the whole cloning strategy again with an ampicillin section marker - to sub-clone the 695bp insert of interest into pGL3-promoter vector. The successful pGL3-promoter-695bp-insert construct will then be ready for transfection into a mammalian cell culture for the purpose of measuring luciferase reporter expression after 24-48 hr. A schematic of the cloning strategy described above is shown in **Figure 4.12**.

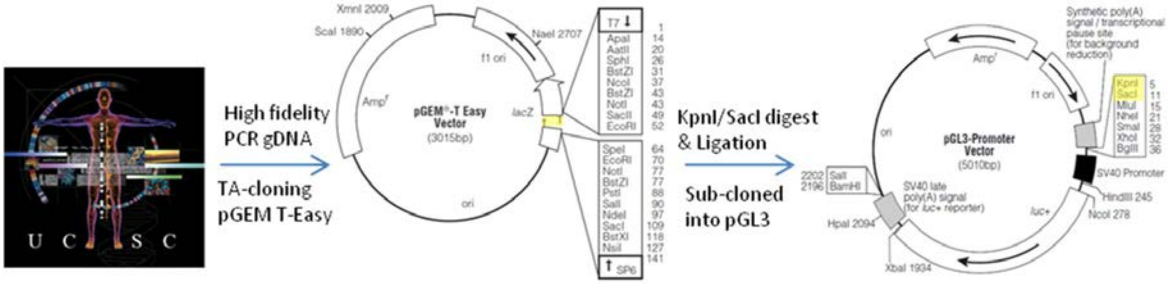


Figure 4.12. Cloning strategy flow chart (images from Center for Biomolecular Science & Engineering, UCSC and Promega).

4.2.3.3.2 High fidelity polymerase chain reaction (PCR) for cloning

Intron 3 *COL4A2* human genomic DNA (gDNA) sequence 500bp upstream and 500bp downstream of rs4773144 was identified using the UCSC genome browser and used to design primers (via Primer 3 software: http://biotools.umassmed.edu/bioapps/primer3_www.cgi) for a putative regulatory genomic region that would incorporate not only the CARDIoGRAM reported SNP rs4773144, but also the three SNPs (rs4773143, rs7986871 and rs3809346) in strong pairwise LD ($r^2 \geq 0.8$) and the Haniel *et al.* reported silencer sequence (see section 4.1 Introduction) (Haniel et al. 1995). The primers were designed with KpnI (forward primer) and SacI (reverse primer) restriction enzyme tails (see **Table 4.9**) for subsequent orientation specific sub-cloning into pGL3 luciferase reporter vectors (see **Table 4.10** and **Figure 4.12**).

PCR amplification of a human gDNA sequence (i.e. a 695bp *COL4A1*/*COL4A2* bidirectional promoter region) was achieved by adding 100ng of human gDNA to a master mix containing final concentration Phusion® high fidelity (HF) buffer

(Finnzymes), 200µM dNTPs (Promega), 0.5µM orientation-specific paired (**Table 4.9**) forward and reverse primers (Fisher), 0.02U/µL Phusion® Hot Start HF DNA Polymerase (Finnzymes) made up to a final volume of 50µL. The PCR master mix was then added to 200µL PCR tubes and placed in a GeneStorm PCR thermal cycler where the following cycling conditions were adhered to: a 30s initial denaturation at 98^{0}C, followed by a 3-step PCR cycle (10s at 98^{0}C, 30s at 63^{0}C and 30s at 72^{0}C) for 35 cycles and a final extension of 5 min at 72^{0}C.

Table 4.9. Oligonucleotides used to clone a 695bp genomic DNA fragment into pGEM®-T Easy vector and sub-clone into pGL3 luciferase reporter vectors.

Tailed Oligonucleotide name	Multicloning region Restriction enzyme within Tail for pGL3	sequence (5' to 3')
COL4A1_F_695	KpnI	ATATGGTACCATAGCTGCACCCACGGTATC
COL4A1_R_695	SacI	ATATGAGCTCATAGTGACCGTGGGGTTCCT
COL4A2_F_695	KpnI	ATATGGTACCATAGTGACCGTGGGGTTCCT
COL4A2_R_695	SacI	ATATGAGCTCATAGCTGCACCCACGGTATC

Table 4.10. Cloning, reporter and expression vectors.

Vectors	Type	Source
pGEM®-T Easy	cloning	Promega
pGL3_basic	reporter	Promega
pGL3_promoter	reporter	Promega
pRL-TK	reporter	Promega
pRL-SV40	reporter	Promega
pCMV6-XL4_STAT3	expression	Origene
pCMV6-Entry	expression	Origene
P611E4	Known STAT3 (NF-κB) intronic enhancer construct	Gift from Dr Tim Mckeithan, University of Nebraska†

†Described by (Brocke-Heidrich et al. 2006; Ge et al. 2003).

4.2.3.3.3 Agarose gel electrophoresis, extraction and purification

PCR products were diluted and mixed in a 1X bromophenol blue loading dye and loaded along with 100bp and 1kb ladders (bioline), (i.e. to identify the correctly sized migration band) to a 1% (w/v) agarose gel (prepared in a 500mL Duran from agarose powder melted in 1X Tris-Acetate ETDA (TAE) buffer, cooled to hand hot before adding [0.035µg/mL] ethidium bromide), submerged in 1X TAE running buffer. The gel was then electrophoresed in a gel tank for 30 minutes to 1 hour at

5V/cm. The gels migrated bands were then visualised under ultra-violet (UV) light (Syngene gel documentation and imaging system), excised with a razor blade and purified using an agarose gel extraction kit (Roche) following the manufacturer's instructions.

4.2.3.3.4 DNA quantification

DNA concentration and purity were measured using a NanoDrop™ 8000 (Thermo Fisher scientific, Inc.) spectrophotometer, good quality DNA showed a 260/280 ratio ≥ 1.8.

4.2.3.3.5 Restriction enzyme digest of pGL3 vectors

Prior to ligation pGL3-promoter plasmid DNA was double digested at 37^{0}C for 2 hours with KpnI (10U) and SacI (10U) restriction enzymes in NEBuffer 1 and 10µg/mL BSA (All reagents were from New England Biolabs® Inc.), in a 20µL final volume. The digest was then run on a 1% agarose gel and extracted using the excised with a razor blade and purified using an agarose gel extraction kit (Roche) following the manufacturer's instructions. Post cloning digests were also performed to ensure transformant clones had successfully subcloned gDNA inserts in the correct orientation, prior to sequencing.

4.2.3.3.6 Ligation of DNA insert into vector

The DNA fragment insert was placed into the ligation reaction with the cloning vector (50ng) (i.e. pGEM®-T Easy or pGL3-based vector) typically in a 3:1 insert:vector ratio, the amount of insert was calculated using the following formula: ng of insert = (vector (ng) x insert size (kb) ÷ vector size (kb)) x insert:vector ratio, along with 1µL of T4 DNA ligase (3U/µL), 5µL of 2X rapid ligation buffer (T4 DNA ligase buffer) and finally made up to a final volume of 10µL with MilliQ water. In addition, a no insert ligation control was also prepared and both ligation reactions left for 16 hr at 4^{0}C.

4.2.3.3.7 Transformation of bacteria with plasmid DNA

5µL of ethanol precipitation (tRNA carrier) purified ligation reaction was added to 50µL of subcloning efficiency™ DH5α™ competent cells (Invitrogen) and incubated on ice for 30 min. The cells were then heat shocked at 42^{0}C for 20 sec in a pre-heated water bath and then allowed to recover on ice for 2 min, prior to adding 950µL of pre-warmed Luria broth (LB) media. The cells were then placed at 37^0C for 1 hr in a shaking incubator at 225 RPM. Generally, 100µL of cells were spread on to already prepared LB-agar plates containing 100µg/mL ampillicin to enable transformed cell selection. In the case of pGEM®-T Easy subcloning a blue/white selection was achieved by adding 40µL of Xgal (20mg/mL) to the LB-agar-ampicillin plates. The plates were then inverted and placed in a 37^0C incubator overnight. The next morning plates were checked for successful transformation and stored at 4^0C. Single isolated transformant colonies were then picked with a sterile pipette tip and grown in 10mL of pre-warmed LB containing ampicillin overnight for 16 hr in a shaking incubator (225RPM) at 37^0C. Potential insert containing plasmid DNA was then extracted from 5mL of the growth culture using the GenElute™ plasmid miniprep kit (Sigma-aldrich®) according to manufacturer's instructions. The resultant plasmid DNA was measured for concentration using the ND-8000 Nanodrop spectrophotometer and diagnostically restriction enzyme digested to see if positive insert clones of the correct size and orientation had been achieved. A 50% (v/v) glycerol stock was prepared with some of the remaining growth culture for long term storage at -80^0C. In addition, another 40mL LB-ampicillin culture was spiked to extract more plasmid DNA for sequencing or transfection using a GenElute™ endotoxin-free plasmid midiprep kit (Sigma-aldrich®).

4.2.3.3.8 DNA sequencing

Cloned plasmid constructs were sequenced to ensure no insert sequencing errors were introduced when using the high fidelity Phusion® Taq by sending plasmid miniprep purified template and appropriate sequencing primers (see **Table 4.11**) to PNACL (Protein nucleotide acid chemistry laboratory, University of Leicester). PNACL used ABI Prism BigDye™ terminator ready reaction mix (v3.1) for cycle sequencing and an automated 3730xl DNA sequence analyzer (Applied biosytems®). The resultant (*.abi) sequencing data files were converted into fasta sequence using ABI 2 FASTA converter free software (Heracle) and aligned with wild type genomic DNA with WClustal v2.0.12 (http://www.ebi.ac.uk/Tools/msa/clustalw2/; EMBL-EBI) free software for analysis.

Table 4.11. Sequencing primers for cloning and sub-cloning vectors to ensure the genomic DNA inserts are correct.

Oligonucleotide name	Plasmid vector or gDNA	sequence (5' to 3')	Length (nt)
pUC/M13_F	pGEM® T Easy	CGCCAGGGTTTTCCCAGTCACGAC	24
pUC/M13_R	pGEM® T Easy	TCACACAGGAAACAGCTATGAC	22
F1	gDNA	AGCTTGCTCTTCCCTCATCA	20
F2	gDNA	GGAGTCTCTGGGTAAAGGGG	20
R1	gDNA	TGATGAGGGAAGAGCAAGCT	20
R2	gDNA	CCCCTTTACCCAGAGACTCC	20
GLprimer2	pGL3	CTTTATGTTTTTGGCGTCTTCCA	23
RVprimer3	pGL3	CTAGCAAAATAGGCTGTCCC	20

4.2.3.3.9 Transient transfection

Initially, [hard to transfect] primary HASMC were transfected in 6-well plates using the (electroporation) Amaxa® HASMC Nucleofector® kit optimised protocol (LONZA) according to manufacturer's instructions. 1x10^6 HASMC were used per sample (i.e. 1 x 75cm^2 flask at 70-80% confluency). Cells were co-transfected with 1-5µg/well of plasmid DNA, that included *Firefly* luciferase reporter *COL4A1*/*COL4A2* bidirectional promoter constructs (see **Table 4.12**), or an empty control vector (see **Table 4.10**) along with 50-250ng/well of *Renilla* luciferase pRL-SV40 (see **Table 4.10**) to normalise for transfection efficiency (i.e. 20:1 ratio of Firefly to Renilla luciferase). The total amount of DNA transfected was kept constant per sample well across all directly comparable transfections.

Additionally, [easier to transfect] HeLa cells were transfected in 24-well plates using the GeneJuice® transfection reagent (Novagen), according to manufacturer's instructions. About 12,000 HeLa cells/well were seeded in 24-well plates and incubated at 37^{0}C/5% CO_2 for 24hr, (cells reached 70-80% confluent) before transfection. The Genejuice® transfection reagent to plasmid DNA ratio used was as recommended at 3:1. Cells were co-transfected with 250ng/well of plasmid DNA, that included *Firefly* luciferase reporter *COL4A1*/*COL4A2* bidirectional promoter constructs (see **Table 4.12**), an empty pGL3-Promoter negative (-ve) control vector and 'P611E4' a known STAT3 (NF-κB) (*BCL3*) intronic enhancer construct, which contains the full length *BCL3* promoter and pGL3_basic vector (a gift from Dr McKeithan, University of Nebraska) as a positive (+ve) control (see **Table 4.10**) along with 12.5ng/well of *Renilla* luciferase pRL-TK (see **Table 4.10**) to normalise for transfection efficiency; these transfections were either left unstimulated or stimulated with cytokines (i.e. IL-6, IL-4, IFN-γ, or TNF-α), or the growth factor TGF-β1.

Table 4.12. *COL4A1* and *COL4A2* bidirectional promoter luciferase reporter constructs for wild type human genomic DNA.

Luciferase reporter constructs	pGL3 vector	Insert Size (bp)	4-SNP Haplotypes
-1632/-955_*COL4A1*_Major-Hap-SV40-*Luc*	Promoter	695	CA**T**A
-1632/-955_*COL4A1*_Minor-Hap-SV40-*Luc*	Promoter	695	TT**C**G
+804/+1498_*COL4A2*_Major-Hap-SV40-*Luc*	Promoter	695	T**A**TG
+804/+1498_*COL4A2*_Minor-Hap-SV40-*Luc*	Promoter	695	C**G**AA

Legend. The luciferase reporter constructs indicate the genomic position in relation to the *COL4A1* and *COL4A2* TSS(+1). The 4 SNP haplotypes are ordered rs4773143, rs4773144, rs7986871, rs3809346 for the plus strand (i.e. *COL4A2*) and rs3809346, rs7986871, rs4773144 and rs4773143 for the minus strand (i.e. *COL4A1*). The lead SNP rs4773144 is highlighted in bold text.

As mentioned above, transient transfection of HeLa cells took place once the cells had reached 70-80% confluency, after 24h. 6h post transfection the DMEM media was replaced by serum starve DMEM media with 0.5% FCS. After a further 24h the media was either left unstimulated or a cytokine/growth factor stimulus was added: 1) 10ng/µL of IL-6 (Brocke-Heidrich et al. 2006); 2) 10ng/µL of IL-4 (Pesu et al. 2002), 10ng/µL of IFN-γ (Yang et al. 2012); 2ng/µL of TGF-β1 (Czuwara-Ladykowska et al. 2002); and 20ng/µl of TNF-α (Brasier et al. 2001). After another 18hr (i.e. 48h from the initial transfection), the transfection was halted by cell lysis and luciferase activity measured by luciferase assay (see section 4.2.3.3.10). There was however one exception in that TNF-α was only stimulated for 4hr, before cell lysis. All cytokines were purchased from Peprotech EC Ltd., London, UK, except TGF-β1 which came from R&D systems Europe Ltd., UK.

Additional co-transfections were also specifically performed for STAT3, with 250ng/well pGL3_promoter constructs and either 250ng/well STAT3-pCMV6-XL4 (see **Table 4.10**) expression vector or an empty 250ng/well pCMV6-Entry (see **Table 4.10**) vector along with 25ng of *Renilla* pRL-TK. The empty pCMV6-Entry vector transfections were unstimulated or IL-6 stimulated, whereas the STAT3-pCMV6-XL4 expression vector transfection was unstimulated only. Media changes and stimulation with IL-6 were as decribed in the previous paragraph.

The activities of *Firefly* and *Renilla* luciferases were measured 48 hr post transfection using the Dual-Luciferase® Reporter (DLR™) Assay System (Promega), for both transfection methods, as described in section 4.2.3.3.10.

4.2.3.3.10 Dual Luciferase reporter gene assay

Firefly and *Renilla* luciferase activities were measured 48 hours post transfection using the DLR™ Assay System (Promega) and the Luminat LB9507 luminometer (Berthold Technologies). The assay was performed as described in the DLR™ Assay System technical manual (Promega). Briefly, phosphate buffered saline (PBS) washed transfected cells in 6-well or 24-well plates were lysed using 500µL/well or 100µL/well of passive lysis buffer (PLB) (Promega), respectively. The standard protocol was then adhered to by measuring the first luminescence for the *Firefly* luciferase (i.e. the bidirectional promoter reporter activity) by adding 20µL of PLB lysate to 100µL of LARII substrate pre-dispensed in a luminometer tube, mixing 3 times with a pipette, placing in the luminometer and reading. The second luminescence for the *Renilla* luciferase (i.e. the internal control activity) was measured by adding 100µL of Stop and Glo® substrate, vortexing briefly to mix, placing back in the luminometer and reading. (**N.B.** The Stop and Glo® substrate acts by quenching the first reaction and simultaneously initiating the *Renilla* luciferase reaction). Data were then calculated as relative luciferase activity, i.e. the ratio of the first *Firefly* luminescence divided by the second *Renilla* luminescence.

4.2.3.3.11 Statistical analysis of Luciferase reporter assay

All transfection experiments were performed in triplicate, and each experiment typically performed on at least three separate occasions. Measurements of luciferase activity were made for each luciferase reporter construct transfection experiment, and are expressed as means ± standard error of mean (s.e.m.). Differences between *COL4A1* and *COL4A2* constructs and the empty pGL3_promoter negative control were detected using a two sample student's t-test, with the alpha level of statistical significance set at α=0.01 (and nominal significance set at α=0.05).

4.3.3.2 Luciferase gene expression assay

The four *COL4A1*/*COL4A2* bidirectional promoter 695bp isoforms (see **Table 4.12**), were successfully sub-cloned (inserted) into the pGL3_promoter (SV40) vector, and these isoforms were confirmed by HindIII restriction enzyme digest producing two distinctly sized bands from a single cut within the insert and a single cut within the pGL3_promoter vector as shown on a 1% agarose gel (see **Figure 4.13**). Subsequently, these agarose gel correctly sized isoform pGL3_promoter constructs were also found to have no sequencing errors.

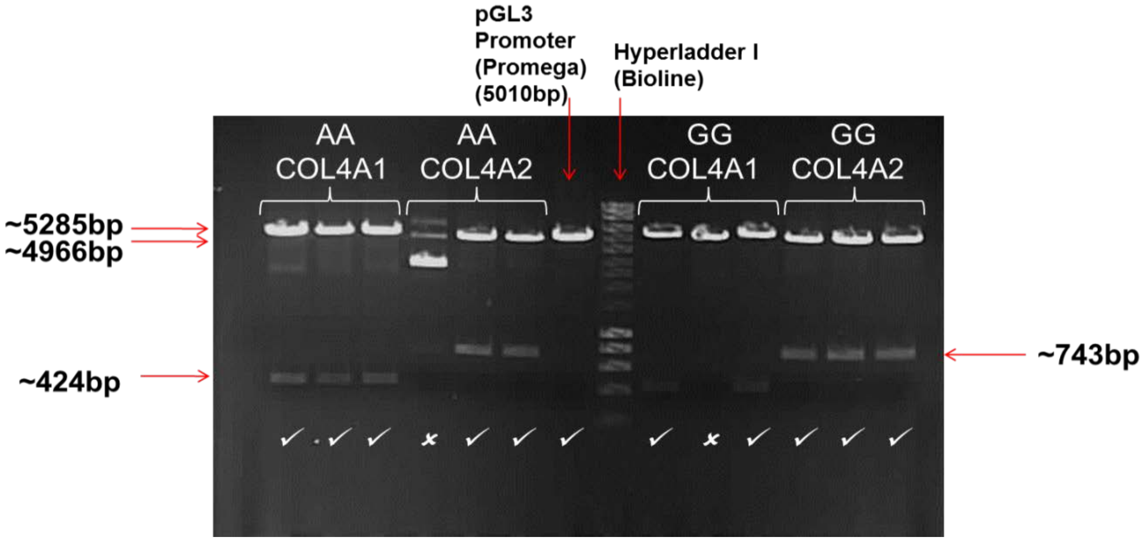


Figure 4.13. Agarose gel (1%) showing successful HindIII restriction enzyme digestion of four 695bp isoform pGL3_promoter constructs (Lanes 1-3: -1632/-955_*COL4A1*_Maj; Lanes 5-6: +804/+1498_*COL4A2*_Maj; Lanes 9&11: -1632/-955_*COL4A1*_Min; and Lanes 12-14: +804/+1498_*COL4A2*_Min, described in **Table 4.12**).

Initial transient transfection experiments using electroporation Nucleofector® experiments involving primary HASMC failed to transfect efficiently across a 1-5µg range of plasmid DNA input, and so no useful dual luciferase expression data was generated. Subsequently, transfections were performed with GeneJuice® transfection reagent in (easier to transfect) HeLa cells. Four pGL3_promoter constructs described in **Table 4.12** were successfully transfected with usable dual luciferase assay results.

The genetic architectural context of the distal 695bp *COL4A1*/*COL4A2* bidirectional promoter genomic fragment, which contains a 4-SNP CAD-associated haplotype and a known silencer *cis*-regulatory element (Haniel et al. 1995), is shown as a schematic diagram in **Figure 4.14*A***.

Luciferase assay results in **Figure 4.14*B*** show three statistical experiment comparisons. In the first instance, mean relative luciferase reporter gene expression were compared by two sample t-test for each of the four 695bp insert isoform pGL3_promoter constructs, and a STAT3 (NF-κB) intronic (*BCL3*) enhancer construct (+ve control) against an empty pGL3_promoter vector (-ve control) normalised to 1.0, under endogenous unstimulated or cytokine (IL-4, IL-6, IFN-γ and TNF-α) and TGFβ1 growth factor stimulated HeLa cell transient transfection conditions. Secondly, mean relative luciferase activities were compared by two sample t-test for unstimulated versus stimulated transfection conditions for each construct individually. Thirdly, there was a statistical comparison between the minor and major haplotype of each haplotype orientation, i.e. Haplotype 1 (-ve strand) and Haplotype 2 (+ve strand). Only the expected IL-6 (P<0.05) and TNF-α (P<0.01) stimulated STAT3 (NF-κB) intronic (*BCL3*) enhancer construct (+ve control) showed a statistically significant increase in expression above their unstimulated construct equivalent (P<0.01). In addition, IL-6 and TNF-α showed a nominal significant increase in expression above the empty pGL3_promoter -ve control (P<0.05). However, none of the four 695bp insert isoform pGL3_promoter constructs showed a significant difference in relative luciferase activity when compared with the empty pGL3_promoter –ve control

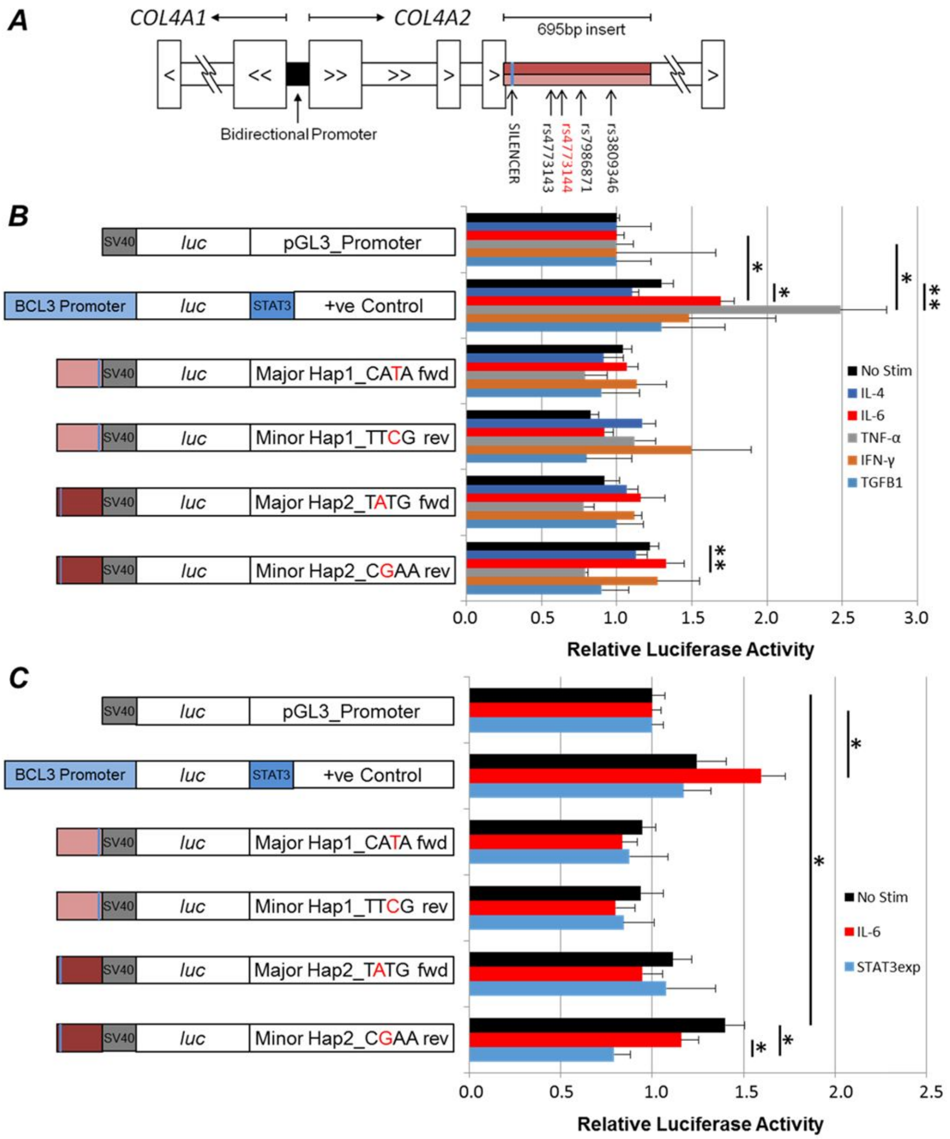


Figure 4.14A-C. Luciferase assay results for *COL4A1*/*COL4A2* gene expression assessing 695bp human genomic DNA 4-SNP haplotype inserts sub-cloned into the pGL3_promoter vector (Promega).

A. Shows a schematic of the *COL4A1*/*COL4A2* shared bidirectional promoter, a 4-SNP haplotype (lead SNP rs4773144 is highlighted red) and a known silencer (Haniel et al. 1995) as a narrow blue box.

B. Shows luciferase assay gene expression results following HeLa cell transfections with a -ve control empty pGL3_promoter (SV40) vector, four 695bp haplotype specific pGL3_promoter *COL4A1*/*COL4A2* plasmid constructs, and a known STAT3 (NF-κB) *cis*-regulatory intronic (*BCL3*) enhancer construct containing the *BCL3* promoter and pGL3_basic vector (Gift from Dr McKeithan, University of Nebraska) +ve control using no stimulation or stimulation with cytokines or growth factors.

C. Shows the same luciferase assay gene expression construct transfections as for ***B.,*** except this time the focus is specifically assessing a putative STAT3 TFBS, and involves additional co-transfection with either a STAT3-pCMV6-XL4 (Origene) expression vector or an empty pCMV6-Entry (Origene) promoter vector. The empty pCMV6-Entry (Origene) vector transfections are unstimulated or IL-6 stimulated, whereas the STAT3-pCMV6-XL4 (Origene) expression vector transfection is unstimulated only.

All experiments were carried out on at least three occasions in triplicate (i.e. 3 technical replicates) and data are represented as mean ± s.e.m., *P < 0.05, **P < 0.01, t-test, between haplotypes or versus empty vector. (**N.B.** As *n* < 10, approximate 95% CI can be created by multiplying the s.e.m. by 4).

under endogenous unstimulated or IL-4, IL-6, IFN-γ, TNF-α and TGFβ1 stimulated transfection conditions. Interestingly, TNF-α stimulated HeLa cells have a significant decrease in luciferase activity for the +804/+1498_*COL4A2*_minor haplotype 2 (C**G**AA) construct (rs4773144 risk allele (G)) compared to the endogenous unstimulated transfection (*P*<0.01). However, the level of gene expression for this construct does not show a significant difference in gene expression with its TNF-α stimulated counterpart, the +804/+1498_*COL4A2*_major haplotype 2 (T**A**TG) construct (rs4773144 non-risk allele (A)), as shown in **Figure 4.14*B***.

Finally, **Figure 4.14*C*** also shows mean relative luciferase reporter gene expression for four 695bp insert isoform pGL3_promoter constructs, and a STAT3 (NF-κB) intronic (*BCL3*) enhancer construct (+ve control) against an empty pGL3_promoter vector (-ve control) normalised to 1.0. Except this time, the transfection experiments were co-transfected with either STAT3-pCMV6-XL4 expression vector under endogenous unstimulated conditions, or an equimolar amount of empty pCMV6-Entry vector under endogenous unstimulated, and IL-6 cytokine stimulated HeLa cell transient transfection conditions.

Again as for the results under Figure 4.14*B*, there was a statistical comparison between stimulated and unstimulated transfection conditions for each construct individually, and a comparison between the major (Hap2) and minor (Hap1) haplotypes for each haplotype orientation.

There were three main findings: Firstly, the IL-6 stimulated STAT3 (NF-κB) intronic (*BCL3*) enhancer construct (+ve control) showed a nominally significant (*P*<0.05) increased gene expression when compared to its unstimulated construct equivalent, but none of the 695bp insert test constructs showed a difference in expression when stimulated with IL-6. In fact, for the most part experiments with the 695bp insert test constructs co-transfected with STAT3-pCMV6-XL4 expression vector, showed lower relative luciferase activity than the empty pGL3_promoter (-ve control), but none were significantly different. Secondly, the +804/+1498_*COL4A2*_minor haplotype 2 (C**G**AA) construct (rs4773144 risk allele

(G)) showed a nominally significant (*P*<0.05) increase in relative luciferase activity compared to the empty pGL3_promoter (-ve control) under endogenous unstimulated conditions. But, there was no significant difference in gene expression between the +804/+1498_*COL4A2*_minor haplotype 2 (C**G**AA) construct (rs4773144 risk allele (G)) and its endogenous unstimulated counterpart, i.e. construct +804/+1498_*COL4A2*_major haplotype 2 (T**A**TG) (rs4773144 non-risk allele (A)). Thirdly, the +804/+1498_*COL4A2* _minor haplotype 2 (C**G**AA) construct (rs4773144 risk allele (G)) showed a nominally significant decrease in relative luciferase activity when co-transfected with the STAT3-pCMV6-XL4 expression vector (*P*<0.05), when compared with either IL-6 or endogenous (unstimulated) conditions; but there was no significant difference between the STAT3-pCMV6-XL4 expression vector co-transfections for the +804/+1498_*COL4A2* _minor haplotype 2 (C**G**AA) construct (rs4773144 risk allele (G)) and its counterpart +804/+1498_*COL4A2* _major haplotype 2 (T**A**TG) construct (rs4773144 non-risk allele (A)). A final observation was that all four co-transfected empty pCMV6-Entry and 695bp insert constructs that were stimulated with IL-6, showed a slight (non-significant) decrease in relative luciferase activity when compared with the endogenous unstimulated co-transfections. This was suggestive of an inhibitory effect by IL-6, in the presence of an empty expression vector, and no insert enhancer (i.e. STAT3) sequence effect.

4.2.4 Assessment to see if the CAD-associated variant rs4773144 affects restenosis rates during coronary angioplasty (Methods)

4.2.4.1 Subjects and SHARP trial procedure

The subjects used for this study are from the Subcutaneous Heparin and Angioplasty Restenosis Prevention (SHARP) trial (Brack et al. 1995). This randomised clinical trial investigated subcutaneous unfractionated heparin on angiographic restenosis after PTCA. Patients were selected for randomised administration of no heparin or 12,500 IU of heparin subcutaneously twice daily for 4 months, if they had a successful PTCA procedure (i.e. post-PTCA, a visually assessed diameter stenosis of less than 50%). The study was undertaken in the pre-stent era, where simple balloon angioplasty was the norm and restenosis rates due to intimal hyperplasia (a smooth muscle and fibrotic reaction to the vessel wall injury) were quite high. The study therefore provided a valuable model to investigate whether the CAD-associated COL(IV) locus SNP rs4773144 affected vessel wall behaviour to injury, directly in humans. For this PTCA restenosis study, quantitative angiography, i.e. arterial lumen diameter measurements were collected by analysing identical orthogonal angiograms on three separate occasions: i) pre-PTCA; 2) immediately post-PTCA; and iii) at 4-months follow-up of the PTCA procedure. DNA was extracted from the blood sampling of single-vessel only PTCA subjects (n=233), which had complete clinical and angiographic follow-up (Samani et al. 1995).

4.2.4.2 Genotyping

TaqMan® genotyping of rs4773144 (*COL4A2*), was performed as described previously (Chapter 2).

4.2.4.3 Statistical analysis

The quantitative angiographic percentage (%) diameter stenosis data are tabulated as mean (s.e.m), and graphically represented as cumulative distribution curves for pre-PTCA, post-PTCA and follow-up PTCA. Differences between genotype groups for % diameter stenosis, were compared by (non-parametric) Kruskal-Wallis one-way analysis of variance (ANOVA), assuming a non-normal distribution. Two restenosis categories were measured: 1) subjects with >50% loss of acute luminal gain, and 2) subjects with >50% luminal stenosis at follow-up, and tabulated as counts and percentages. The restenosis differences between genotype groups were analysed by Pearson's χ^2 test, and relative risk ratios calculated for the risk allele, under a dominant and recessive mode of genetic inheritance. A level of statistical significance of α=0.05, was used for both statistical tests.

4.3.4 Findings from the assessment to see if the CAD-associated variant rs4773144 affects restenosis rates during coronary angioplasty

The genotype frequencies for rs4773144 (AA [39.7%], AG [42.1%], GG [18.2%], MAF=39.2%), did not deviate from Hardy-Weinberg equilibrium. The pre-PTCA percentage diameter stenosis was similar across all genotype groups (**Table 4.13**).

Table 4.13. Quantitative angiographic findings.

rs4773144 - *COL4A2*			
Percentage diameter stenosis	**Genotypes**		
	AA (n=83)	AG (n=88)	GG (n=38)†
Pre-angioplasty (%)	71.1 (1.3)	70.7 (1.3)	71.1 (1.9)
Post-angioplasty (%)	32.1 (1.2)	32.9 (1.4)	34.3 (2.2)
Follow-up (%)	50.1 (2.2)	51.1 (2.2)	50.8 (3.5)

Legend. Percentage diameter stenosis data are given as mean (standard error), †CAD risk allele.

The cumulative frequency curves of percentage diameter stenosis for each of the three genotype group, i.e. pre-PTCA, post-PTCA and 4-month follow-up are shown for rs4773144 in **Figure 4.15**.

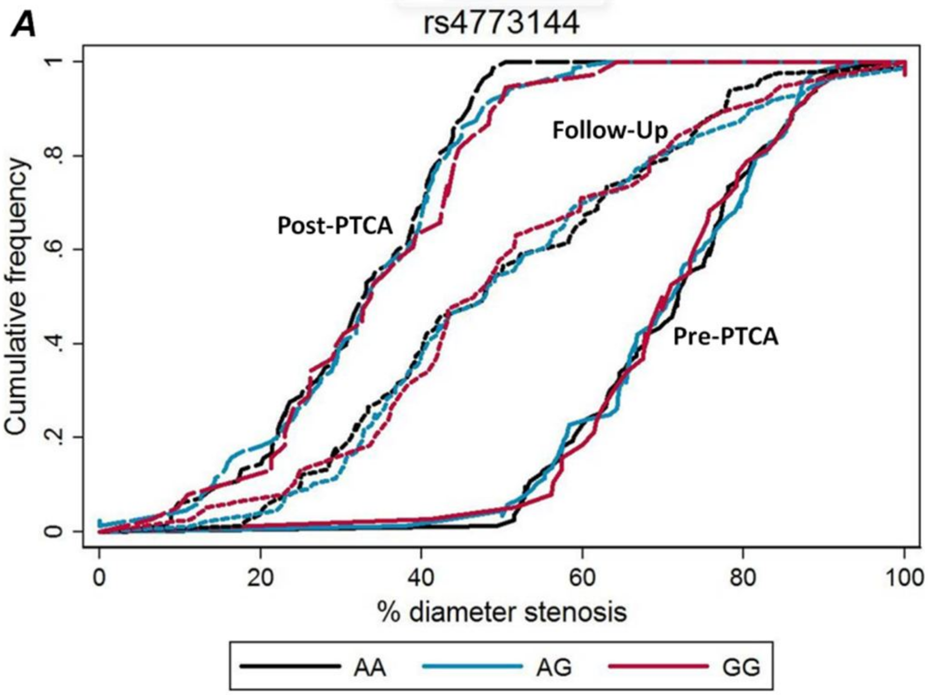


Figure 4.15. Cumulative frequency curves of percentage diameter stenosis pre-PTCA, post-PTCA and at 4-month follow-up in subjects grouped by genotypes for rs4773144.

The restenosis rates at 4-months follow-up are given for rs4773144, using two different criteria measurements (see **Table 4.14**).

Table 4.14. Restenosis rates at follow-up.

rs4773144 - *COL4A2*			
Measure of restenosis	**Genotypes**		
	AA (n=83)	AG (n=88)	GG (n=38)†
>50% loss of acute gain in MLD	37 (45)	43 (49)	13 (34)
>50% diameter stenosis	38 (46)	40 (45)	16 (42)

Legend. Restenosis data are given as counts (percentages), MLD=minimal luminal diameter, †CAD risk allele.

There was no significant difference between genotype groups for subjects with greater than 50% loss of acute luminal gain (L50%) - rs4773144: χ^2=2.31, *P*=0.32; or greater than 50% luminal diameter stenosis (>50%) at follow-up - rs4773144: χ^2=0.16, *P*=0.93.

The relative risk of restenosis for carriers of the CAD risk allele (G) rs4773144 under a dominant genetic model (GG/AG vs AA) was 1.00 (95% CI: 0.80 to 1.25) and 0.98 (95% CI: 0.78 to 1.22) for L50% and >50% respectively; whereas the relative risk of restenosis for carriers of the CAD risk allele (G) rs4773144 under a recessive genetic model (GG vs AG/AA) was 0.65 (95% CI: 0.35 to 1.20) and 0.89 (95% CI: 0.50 to 1.59) for L50% and >50% respectively.

Retrospective *post-hoc* power calculations were calculated for >50% and L50%, and there was 80% power to detect a difference of 22.5% between carriers and non-carriers of the risk allele. The actual difference was 6.63%, for which there was less than 10% power to detect at an alpha=0.05, significant level. Hence, it should be noted that the power to detect a genotypic effect for restenosis in this study is very low, and so it might mean that my hypothesis might not shift from the *null* – and so this raises the likelihood of observing a false negative (Type II error) effect.

4.4 Discussion:

This chapter has yet again shown how difficult it is, to move from a multi-factorial common disease GWA study signal, to a truly causal SNP and mechanism. The identification of the non-coding 13q34 locus *COL4A1*/*COL4A2* bidirectional promoter (within intron 3 of *COL4A2*, and distal promoter of *COL4A1*) signal, required a meta-analysis of 14 Caucasian ancestry GWA studies (22,233 cases and 64,762 controls), and replication in 58,623 subjects (29,894 cases and 28,729 controls) - in order to have enough statistical power, for its detection at a genome wide level of significance. Therefore, the signal for 13q34 is considerably weaker than for single GWA studies, such as the two strongest loci identified for CAD risk - 9p21.3 and 1p13.3.

Although the GWA studies, were this time imputed with HapMap phase II genotypes (i.e. ~2.2 million SNPs), a level far greater than for the early GWA studies, refinement of the locus is still required. This time I was able to firstly, assess the 1000 genomes pilot study data, via the Broad Institute’s web-based tool SNAP. The purpose of this was to search for additional SNPs that were in strong pairwise LD, with the two lead CARDIoGRAM discovery SNPs - rs4773144 and rs3809346. These two lead CARDIoGRAM SNPs are inseparable, in terms of which is more likely to be causal, because based on HapMap phase II data, these two SNPs are in perfect pairwise LD r^2=1, in subjects of European ancestry. Yet again, as with the CAD-associated locus 1p13.3, the locus 13q34 acquired two additional almost perfect proxy candidate SNPs (r^2≈1) - rs4773143 and rs7986871 from the more detailed analysis. The four SNPs form a haplotype that spans a narrow 258bp of genomic DNA sequence (**Table 4.2**).

Unfortunately, I was not privy to ‘wet-lab’ genotyped CAD GWA study meta-analysis data, on these *in silico* acquired 1000 genomes identified proxy SNPs, at the inception of the functional assessment of the 13q34 locus. However, subsequently, the Metabochip study through a combined consortia meta-analysis called ‘CARDIoGRAMplusC4D’ did identify the four SNPs in question as being associated with CAD risk, and that these SNPs following conditional analysis were representative of a single independent locus (coincidently an additional genome-wide significant SNP rs9515203 was discovered for the 13q34 locus, within the *COL4A2* gene, but located ~90kb downstream, and not in pairwise LD r^2=0.01, with the lead CARDIoGRAM discovery SNP rs4773144; this SNP was found to be independent following conditional analysis) (CARDIoGRAMplusC4D Consortium et al. 2013).

Furthermore, I also checked the trans-ethnic fine-mapping of these 4 candidate SNPs. This showed that none, or very few recombination events over evolutionary time have occurred, even for the African 1000 genomes pilot study YRI population. Indeed, three of the four SNPs in YRI have a perfect pairwise LD (r^2=1), and only one SNP rs4773143 has a slightly lower pairwise LD (r^2=0.923). Therefore, trans-ethnic fine-mapping is unlikely to be able to easily further refine the rs4773144 lead 13q34 locus.

Additionally as mentioned in the previous chapter, it can be easier to identify a particular casual SNP and mechanism, if there is a putative, quantitative, intermediate phenotype. Unfortunately to this end, CARDIoGRAM and the IBC 50K chip CAD association meta-analyses, did not identify any CAD risk phenotype association with the chromosome 13q34 locus (Schunkert et al. 2011; IBC 50K CAD Consortium 2011).

Following on from these refinement strategies, the next steps I took to identify a definitive causal SNP(s), at the 4-SNP candidate 13q34 locus was through *in silico* bioinformatics. The first step was to assess whether any of the 4-SNP candidates had been identified previously, as having a *cis*- or *trans*-acting regulatory effect on gene expression. I looked at the eQTL Chicago browser, the Sanger Institute's Genevar database, and Cardiogenics (as described in the previous chapter), but none of them showed a *cis*-regulatory gene expression effect on *COL4A1* and/or *COL4A2* - nor indeed were there any form of *trans*-acting regulatory gene expression effect. Therefore, based on their genomic location, and proximity to a known COL4A1/A2 bidirectional transcription repressor element called the COL4 silencer (Haniel et al. 1995), the subsequent *in silico* bioinformatics steps taken were to assess the 4 candidate SNPs for putative *cis*-regulatory DNA elements. The web-based tools used to perform in silico analysis were ENCODE, the curated RegulomeDB SNP annotation database, and TFBS prediction software JASPAR, MatInspector and TRANFAC®.

ENCODE is a very powerful tool for identifying putative *cis*-regulatory elements, because it uses a multitude of complementary, confirmatory, *in vivo* experimentally captured DNA-protein binding, alongside next generation sequencing - which enables genomic alignment, across a library of multiple cell types. The key alignment tools are low resolution (~200bp), ChIP-seq identified histone modification signatures and TFBS, FAIRE-seq identified nucleosome depletion, and DNaseI HS sites, and high single bp resolution DNaseI footprints (described in detail in Chapter 3). Each ENCODE alignment tool has its limitations, but as a combination they are a very powerful tool for identify putative *cis*-regulatory elements, which in turn can help inform EMSA and luciferase reporter assay functional experimental design, by providing important clues.

Yet again, as was the case for 1p13.3, each of the 4 candidate SNPs, shows clear evidence of DNA-protein binding at *cis*-regulatory elements. Firstly, all but rs7986871 clearly show nearby single bp, high resolution DNaseI footprints.

The second important finding as regards the *COL4A1-COL4A2* bidirectional promoter, is that the histone mark signatures indicate through strong H3K4Me3, moderate H3K27Ac and weak H3K4Me1 signals that all four candidate SNP, co-exist within the distal location of an active, or poised bidirectional promoter.

The third discovery is that a RNA Pol II ChIP-seq spans all four SNPs except rs4773143, in more than one cell type. However, given the low resolution of ChIP-seq signalling and the close proximity of rs4773143 to the RNA Pol II ChIP-seq signal, it too may be considered a RNA Pol II target.

The fourthly key finding is that there strong E2F6 and MAX ChIP-seq TFBS signals that align with a known 25 nucleotide sequence COL4 silencer. Additionally, the E2F6 and MAX ChIp-seq signal is also seen at or near to all 4 candidate SNPs.

Moreover, a mouse conserved TRANSFAC® canonical E2F6 TFBS (**TTGCGCGC**) and a JASPAR identified non-canonical E-box (CGCTTG) can be found within the narrowed 25 nucleotide sequence identified by Haniel *et al.* 5’-CGCGCTTGGAC**TTGCGCGC**CCGAGA-3’ (Haniel et al. 1995). The E2F6 isoform of E2F proteins is known to form a repression complex with a MAX-MGA heterodimer, see **Appendix 6.3, Fig. S4.1** (Ogawa et al. 2002; La Thangue 2002). If these ChIP-seq signals are non-spurious, they could be indicative of a 4 SNP haplotype multi-component protein complex with the known COL4 silencer (Haniel et al. 1995).

The strongest likelihood for an individual causal SNP is rs3809346. The evidence for this notion, comes from the strongest evidence for DNA-protein binding being situated at, or just beyond rs3809346. This observation is signified in multiple cell types, by a ‘dip’ in all three ChIP-seq identified histone mark signatures (in a fashion akin to that seen at the 127bp common core promoter TSS(+1)) (Poschl, Pollner & Kuhn 1988), a CpG island alignment, FAIRE-seq, DNaseI HS sites and single bp resolution DNaseI footprints. Taken together, these alignements could be indicative of a multi-component transcriptional complex. Indeed this proposition, is all the more plausible, when one considers the *in silico* TFBS prediction tool findings for rs3809346, and an assessment of nearby genomic sequence human-mouse conserved TFBS by TRANSFAC®. Two TFBS tools identify a CAD risk allele (A) created EVI-1 TFBS (5’-GAC**A**AGATAA-3’; where the bold nucleotide is the risk allele and the underlined nucleotides match the genomic DNA sequence), and TRANSFAC® confirms a mouse conserved canonical SP1 GC-box (5’-GGGGCGGGGC-3’), previously described in the literature (Poschl, Pollner & Kuhn 1988), just 16bp upstream of rs3809346. TGFβ1 regulates cell cycle genes, e.g. *CDKN2B* (Derynck & Zhang 2003), but it is also one of the most significant mediators for the production of the ECM, and is responsible for the TGFβ1/SMAD3 induced expression of COL(I), COL(III) and COL(IV) (Okano et al. 2012). However, the TGFβ1 effector SMAD3 regulates transcription by interacting with coactivators, or corepressors (Derynck & Zhang 2003). For example, SMAD3 interacts with SP1 and CBP/p300 to coactivate TGFβ1-induced transcription of *COL1A2*, but if

SMAD3 interacts with EVI-1, the required interaction with coactivator CBP/p300 is disrupted and causes a strong repression of *COL1A2* transcription (Alliston et al. 2005; Czuwara-Ladykowska et al. 2002).

Although there is strong evidence for rs3809346 being the causal SNP, it is just as likely that the causal SNP could be, the lead CARDIoGRAM SNP rs4773144. Firstly, there is evidence of DNaseI footprints very near to the rs4773144 gene. Secondly, there are ChIP-seq TFBS alignments with two transcription factors SIN3A (SIN3 transcription regulator family member A) and MXI1 (MAX interactor 1, dimerization protein; aka MAD2) in H1ES cells, which in turn are likely to form a repressor complex with histone deacetylase (HDAC) and MAX (MYC associated factor X) respectively, i.e. a HDAC-SIN3A::MXI1-MAX complex (Swanson et al. 2004), (given that MXD1 (MAX dimerization protein 1; aka MAD1) and MXI1 are both repressors that compete with cMYC to form heterodimer complexes with MAX). Thirdly, there are ChIP-seq TFBS signals for E2F6 and MAX in H1ES and A549 cells, again as mentioned earlier these proteins are known to form a repressor complex (Ogawa et al. 2002; La Thangue 2002). Fourthly, the RegulomeDB SNP annotation database gave rs4773144 the best score (2b), in terms of overall evidence of being a likely transcriptional binding site; Fifthly, TFBS predictions disrupted two important TFBS families, when the CAD risk allele (G) was present, either a STAT (signal transducer and activator of transcription) protein via the consensus motif 5’-TTC(N2-4)GA**A**-3’, or an ETS1 protein via the consensus motif 5’-GGA**A/T**-3’ (bold nucleotides in the consensus motifs are the non-risk allele (A)). STAT proteins are strongly influenced by pro- and anti-inflammatory cytokines, such as those within the atheroma, i.e. pro-inflammatory such as - INF-γ induces STAT1, IL-6 induces STAT3, or anti-inflammatory such as - IL-4 inducing STAT6. Also, ETS1 proteins are expressed by endothelial cells and vascular SMC in the atherosclerotic plaque, induced by PDGF, TNF-α and IL-1β pro-inflammatory cytokines (Okano et al. 2012; Peng et al. 2008). Like EVI-1 mentioned earlier ETS1 is also a corepressor of ECM TGFβ1-induced genes, e.g. COL1A2, when SMAD3 interacts with ETS1 (Czuwara-Ladykowska et al. 2002).

As for the other two candidate SNPs rs4773143 and rs7986871 (in almost perfect pairwise LD ($r^2 \approx 1$)), although the evidence is weaker for their putative causal involvement - neither can be fully ruled out. In the case of rs4773143 there is some evidence of its involvement through: 1) multiple cell DNaseI footprints that align via near to rs4773143; 2) the aforementioned E2F6 and MAX TFBS ChIP-seq signal intensities; and 3) a number of putative TFBS disruptions identified through prediction web-based tools, including important transcription factor families, such as FOX (Forkhead Box) and C/EBP proteins. As for rs7986871, there is also some evidence for *cis*-regulatory binding, even though there are no obvious ENCODE observed DNaseI footprints near to this SNP. Firstly, BACH2 (BTB and CNC homology 1, basic leucine zipper transcription factor 2) shows weak ChIP-seq signal alignment with rs7986871 in H1ES cells, which is in turn supported by the CAD risk allele (A) disruption of a putative BACH1 TFBS. Of interest, BACH1 and BACH2 are each known to form a heterodimer with MAFK (v-maf avian musculoaponeurotic fibrosarcoma oncogene homolog K) and repress transcription, by binding to the NF-E2 consensus site 5’-TGC**TG**A(G/C)TCA(T/C)-3’ (bold text is the non-CAD risk allele, underlined are the genomic DNA matches to the consensus site) (Oyake et al. 1996). Secondly, the FAIRE-seq signal that aligns best for rs3809346, is also aligned with rs7986871. Thirdly, the aforementioned E2F6 and MAX ChiP-seq signals include rs7986871. Fourthly, there is a strong ChIP-seq signal for RNA Pol II in HUVECs and a moderate ChIP-seq signal for TBP in H1ES cells that align with rs7986871. Fifthly, there is a strong RNA-seq signal aligned with rs7986871 in HUVECs, although this finding was not found in a second read for same experiment, or in RNA-seq from another source, so this finding could be spurious. Finally, the CAD risk allele (A) disrupts six putative TFBS, including AP-1 (activator protein 1).

Given the location of the four candidate SNPs within the bidirectional *COL4A1/A2* promoter, and the multitude of potentially relevant, putative *cis*-regulatory elements identified through *in silico* bioinformatics analyses, it seemed possible that there is likely to be an involvement of one or more of these four candidate SNPs, in the control of transcriptional gene expression in one or both *COL4A1/A2* genes. These putative findings helped mold future, focussed functional experiments at the 13q34 locus.

To date there has been no identification of a *cis*-acting eQTL relating to these four candidate SNPs, or any SNPs even in weak pairwise LD (r^2>0.2) at the 13q34 locus, or indeed within the vicinity of the bidirectonal *COL4A1*/*COL4A2* promoter. On this basis, I thought it would be useful to see if healthy kidney tissue extracted RNA, (in tissue/cell types that hadn't been assessed previously on a sizable scale) could affect gene expression in *COL4A1* and/or *COL4A2*, and thus generate an eQTL for the 13q34 locus.

To achieve this, the PKP study (a unique resource of healthy human renal tissue) was used to investigate whether rs4773144 is a *cis*-acting eQTL that significantly alters *COL4A1*, and *COL4A2* gene expression.

In the first instance, the *COL4A1* ΔCt *vs. COL4A2* ΔCt plotted readings, showed a strong correlation with a Pearson's correlation coefficient of r=0.90, which means the steady state mRNA gene expression levels are most likely co-regulated, and expressed to a similar degree for both *COL4A1* and *COL4A2* across all samples. The mean ΔCt ratio of *COL4A1*:*COL4A2* mRNA gene expression is approximately 1:1.2, which is similar to the ranges seen previously in the literature, although mRNA levels have been shown to vary between cell types (Schmidt et al. 1992). Nonetheless, it should be noted that these ubiquitously expressed genes are subsequently translated into *COL4A1* and *COL4A2* peptide alpha chains under highly regulated post-transcriptional processes, so that they are always secreted into the ECM in a 2:1 ratio (i.e. α1α1α2) as a heterotrimeric protomer (i.e. the main building block that makes up 50% of the BM, and gives it structural strength and integrity).

The primary PKP study analysis results show that rs4773144 locus genotypes carrying the CAD risk allele (G) have an additive (i.e. per risk allele) mRNA gene expression reduction after adjusting for age, gender, hypertension, and centre in both *COL4A1* (AA vs AG: 0.847 (95% CI: 0.655-1.097) and AA vs GG: 0.754 (95% CI: 0.526-1.079)), and *COL4A2* (AA vs AG: 0.813 (95% CI: 0.604-1.094) and AA vs GG: 0.732 (95% CI: 0.486-1.102)), with a trend towards statistical significance (*P*=0.094 and *P*=0.099 respectively). In conclusion, we must therefore accept the *null* hypothesis that the PKP study did not provide an allele-specific *cis*-acting eQTL, which reached statistical significance, but there was a clear trend towards a decrease in both *COL4A1* and *COL4A2* gene expression for those subjects carrying the CAD-associated rs4773144 risk allele (G).

So the question remains, is the rs4773144 identified 13q34 locus truly non-functional? It is possible, given these trends towards significance for the CAD risk allele (G) carriage that if we analysed additional samples, or used a more sensitive gene expression measuring apparatus, to reduce sample variation, such as the Rotor-Gene Q instrument (Qiagen), where technical replicates can be read far more accurately (i.e. to within a Ct threshold of 0.2, rather than 0.5) - we may have seen a statistically acceptable result at the 5% level of statistical significance. However, although the results of the PKP study did not reach statistical significance, the directional trend of the results are still of interest, in terms of biological mechanistic plausibility. For instance, carriers of the rs4773144 CAD risk allele (G) that show an additive reduction in both *COL4A1* and *COL4A2* mRNA gene expression, fit well with the notion that a decrease in COL(IV) BM can lead to CAD risk.

Firstly, a hastened medial SMC migration and proliferation could result in an advancement of atherosclerosis. Secondly, an endothelial desquamation (via MMP2 breakdown of COL(IV)) of the atheroma fibrous cap, which results in plaque erosion, and which in turn leads to luminal thrombi and a possible MI (Kolodgie et al. 1998). Thirdly, all the rare Mendelian genetic mutations described in the literature for either *COL4A1* or *COL4A2* tend to be as a result of dysregulated secretion of the α1α1α2 heterotrimeric protomers, and hence reduced COL(IV) protein and BM, which leads to cerebrovascular diseases and HANAC (Plaisier & Ronco 1993; Plaisier et al. 2007). Additionally, the recently identified *COL4A1* non-synonymous SNP rs3742207 for CAD/MI risk (Tarasov et al. 2009; Yamada et al. 2008), may also through an amino acid change plausibly have reduced secretion of COL(IV) to its ECM destination. Therefore, it is reasonable to assume that a down regulation in mRNA gene expression of these ubiquitously expressed genes will have potentially detrimental effects on vascular health that could cause CAD (see **Appendix 6.3, Figure S4.4 -** for a possible mechanism of genetic variant rs4773144 CAD risk).

As a follow-up to these studies it might be worth considering the isolation of specific cell types from human kidney tissue prior to RNA extraction, in case the genotype specific effect is driven by particular cell types. This can be achieved following renal dissection by using standard serial sieving techniques or immunomagnetic cell sorting separation (MACS) techniques, to isolate for example, mesangial cells (renal glomerular vSMC) (Pawluczyk, Patel & Harris 2004), renal microvascular endothelial cells (RMEC) (Muczynski et al. 2003), or proximal and distal tubule separated epithelial cells (Baer et al. 1997).

The suggestive trend towards a reduced *COL4A1/COL4A2* gene expression, in the presence of the rs4773144 CAD risk allele (G), led on to the next functional genomics approach, which was to perform DIG-labelled EMSA - DNA-protein binding assays.

The purpose of the EMSA, is to compare the DNA-protein binding affinity, between two short DIG-labelled ds oligonucleotides of genomic DNA context, where one oligonucleotide contains centrally the major (non-risk) allele, and another oligonucleotide contains centrally the minor (risk) allele. All four candidate SNPs of interest, were assessed for differential DNA binding affinity to protein from HASMC - whole cell lysate protein extracts.

The findings were that all four SNPs within the short DIG-labelled ds oligonucleotides (21bp to 32bp) were capable of binding to a protein, or rather more likely, a multi-component protein complex. This was because the DNA-protein binding, gel band shift, travelled the same equal distance for each SNP examined. However, there were no observable DNA-protein binding differences identifiable, between the major and minor alleles of each SNP for the gel shift experiments. This observation stood firm, even when comparing competition assays. For example, where a given labelled candidate SNP non-risk allele containing oligonucleotide, competes for binding with a 100x excess unlabelled same sequence non-risk allele oligonucleotide vs. a 100x excess unlabelled same sequence risk allele oligonucleotide. These experiments were initially performed using endogenous unstimulated primary HASMC, and subsequently with a known inducer of COL(IV) gene expression by 3hr TGFβ1 stimulation, but still no SNP allele differential binding affinity was observed.

The hope was that under endogenous or stimulated cell culture conditions, there would be a genotype specific binding difference in HASMC, a cell type that is known to express *COL4A1* and *COL4A2*. It is interesting that all four SNP containing short ds oligonucleotides, bind to a DNA-protein complex of the same size. This suggests several possible DNA-protein binding scenarios: 1) all four SNPs bind to one large protein with multiple DNA binding domains; 2) each ds oligonucleotide binds to only one separate protein, but each bound protein happens to be the same size; or 3) that each ds oligonucleotide binds to one or more separate different sized proteins, but that these proteins all form part of a large multi-component protein complex, which means each separate DNA-protein gel shift ends up the same size. The third scenario is perhaps the most likely, if the ENCODE RNA Pol II ChIP-seq signal is to be believed, because multiple transcription factor proteins are known to be involved in the pre-initiation complex (PIC).

These results were somewhat disappointing, because putative data from the *in silico* assessment of TFBS prediction tools, were suggestive of a rs3809346 CAD risk allele (A) (TGFβ1-induced and SMAD-dependent) repression of transcription in the presence of a putative DNA-bound EVI-1 transcription factor. Additionally, the rs4773144 CAD risk allele (G) is known to disrupt both putative STAT and ETS1 bound proteins, and so one might have expected a reduction in binding affinity for the risk allele.

Although the EMSA results suggest that there is no binding affinity difference, it does not necessarily mean that some of the putative TFBS could be binding under normal physiological or pathological conditions. Firstly, the binding affinity difference may be too subtle to detect using this method. Secondly, the non-risk and risk allele containing oligonucleotide may be binding two different proteins or protein complexes of a similar size. However, it may also be that under the *in vitro* cultured cell conditions used, not all the necessary coactivators or corepressors for a given multi-component complex were expressed; such that potentially important DNA-protein binding that would show binding differences cannot be seen. It simply shows that functional experiments may be 'hit and miss', and perhaps depend on a bit of luck, in terms of picking the appropriate cell type and culturing conditions for your experiment.

As described above, the kidney tissue mRNA showed an additive trend towards a relative decrease in mRNA gene expression for human individuals that carry the rs4773144 CAD risk allele (G). A differential binding affinity test was performed by using the EMSA, and importantly in terms of cis-acting transcriptional regulation, showed definitive protein binding to genomic DNA that contained individually, any one of the four candidate SNPs. However, there were no binding affinity differences under the conditions tested, and so one cannot distinguish between any of the SNP in terms of causality regarding CAD risk, or the reduction in gene expression seen in the kidney for the rs4773144 risk allele (G). We are merely left with a CAD risk 4-SNP haplotype.

Therefore, the next functional genomics approach used was to perform luciferase reporter gene expression assays. The purpose of using this method is to identify, a perhaps more subtle, and specific difference in gene expression. This was achieved by creating pGL3_promoter luciferase reporter gene vector constructs that contain either a *COL4A1 minus*-strand (-1632/-955) or a *COL4A2 plus*-strand (+804/+1498) orientated 695bp intron 3 putative *cis*-regulatory genomic sequence, inclusive of either the major or the minor 4-SNP candidate haplotype, as well as the known *cis*-regulatory Haniel *et al.* reported 25bp COL4 silencer element (Haniel et al. 1995), so they can be transiently transfected into human cells (primary HASMC and HeLa cells).

Unfortunately, after numerous attempts the known to be hard to tranfect primary HASMC, failed to transfect - even when using a HASMC cell-type specific electroporation Nucleofector® kit protocol. Therefore, I subsequently used easier to transfect HeLa cells, for my luciferase reporter assay experiments. These were successfully performed under endogenous unstimulated, and cytokine or growth factor stimulated conditions. The stimuli used were selected based upon *in silico* Regulome DB and TFBS predictions, and represent the likely pro-inflammatory adaptive immune reponse to an *in vivo* atherosclerotic plaque environment. This was to induce a statistically significant difference in gene expression that would support, and hopefully enhance the kidney gene expression results.

The first set of dual luciferase experiments (see **Fig.4.14*B***) showed that none of the four intronic 695bp insert isoform pGL3_promoter constructs, had a significant difference in relative luciferase activity, when compared with the empty pGL3_promoter –ve control. This was the case under all conditions tested, i.e. endogenous unstimulated, or IL-4, IL-6, IFN-γ, TNF-α and TGFβ1 individually stimulated, transient transfection conditions. Only the expected IL-6 and TNF-α stimulated STAT3 (NF-κB) intronic (*BCL3*) enhancer construct positive control, showed a statistically significant increase in expression, in comparison with their unstimulated construct equivalent (*P*<0.01). This positively proved IL-6 stimulation of STAT3, and TNF-α stimulation of NF-κB, increased expression of the *BCL3* gene in HeLa cells. Another finding was that TNF-α stimulated HeLa cells show a significant decrease in luciferase activity for the +804/+1498_*COL4A2*_minor haplotype 2 (C**G**AA) construct (rs4773144 risk allele (G)) when compared to the endogenous unstimulated transfection (*P*<0.01). However, the level of gene expression for this construct does not show a significant difference in gene expression with its TNF-α stimulated counterpart, containing the alternative +804/+1498_*COL4A2*_major haplotype 2 (T**A**TG) construct (rs4773144 non-risk allele (A)), and so this result would seem to be spurious. Indeed, if anything the real difference is a slight non-significant increase in the unstimulated +804/+1498_*COL4A2*_minor haplotype 2 (C**G**AA) construct (rs4773144 risk allele (G)) compared with its counterpart, the +804/+1498_*COL4A2*_major haplotype 2 (T**A**TG) construct (rs4773144 non-risk allele (A)).

The second set of dual luciferase experiments (see **Fig.4.14*C***), involved co-transfection with the same four intronic (4-SNP candidate haplotypes) 695 bp insert isoform pGL3-promoter constructs, and either a STAT3-pCMV6-XL4 expression vector, or an empty pCMV6-Entry vector. This experiment was performed off the back of a STAT3 Regulome DB TFBS prediction, which uses DNA sequence PWM and chromatin accessible ENCODE experimental data to make an accurate putative TFBS inference (Pique-Regi et al. 2011). The only positive result was that the +804/+1498_*COL4A2* orientated construct containing rs4773144 risk allele (G) showed a nominally significant (*P*<0.05) increase in relative luciferase activity compared to the empty pGL3_promoter vector under endogenous (unstimulated) conditions. In addition, it would seem that the +804/+1498_*COL4A2* orientated construct containing rs4773144 risk allele (G) has a same direction effect, small increase in relative gene expression, compared with the +804/+1498_*COL4A2* orientated construct containing rs4773144 non-risk allele (A), under endogenous conditions - but this difference is non-significant, which was also seen in **Fig4.14*B***. The same non-significant directional trend is also seen for IL-6, in the +804/+1498_*COL4A2* orientated constructs, for the rs4773144 risk allele (G) over the rs4773144 non-risk allele (A). However, when the +804/+1498_*COL4A2* orientated construct containing rs4773144 risk allele (G) is co-transfected with the STAT3-pCMV6-XL4 expression vector, there is a nominally significant decrease in relative luciferase expression compared to co-transfection with the empty pCMV6-Entry vector, under both endogenous unstimulated and IL-6 stimulated conditions (*P*<0.05); but the same was not seen for the +804/+1498_*COL4A2* orientated construct containing rs4773144 non-risk allele (A). This would suggest that the STAT3-pCMV6-XL4 expression vector cannot increase relative luciferase expression due to the rs4773144 risk allele (G) disrupted STAT3 TFBS. But then again, it should be noted that the

STAT3-pCMV6-XL4 expression vector co-transfection relative luciferase expression for the rs4773144 non-risk allele (A) has a large s.e.m., and is not significantly different to the STAT3-pCMV6-XL4 co-transfection with rs4773144 risk allele (G).

Henceforth, the only result of any consequence was the slight increase in +804/+1498_*COL4A2* orientated construct containing rs4773144 risk allele (G), compared to it's counterpart +804/+1498_*COL4A2* orientated construct containing rs4773144 non-risk allele (A) - under endogenous conditions. Unfortunately, this gene expression effect is in the opposite direction to the previously described kidney gene expression results. It should be noted at this point that Promega technical information warn of known CMV promoter cross-talk with other promoters, and that the CMV promoter will drain the pool of transcription factors and affect luciferase expression levels. Another concern mentioned by Promega is the number of cryptic TFBS with pGL3 vectors that can lead to misleading results, compared with their newer pGL4 vectors (Chapter 12: Transfections. Co-Transfection and Dual-Reporter Assays (page 6) © 2004–2011 Promega Corporation). Also, a recent report described that the whole plasmid sequence is transcribed at different levels, and that some plasmids when co-transfected affect luciferase reporter expression in a dose-dependent manner. Indeed, on one occasion, a Neomycin/Kanomycin antibiotic selection cassette influenced transcription, by generating a unique population of sense and anti-sense small RNAs (Nejepinska et al. 2012). These considerations may go some way to explaining the gene expression differences, particularly those involving co-transfection, but it is perhaps more likely that the differences are tissue and cell-type specific. It would be useful to repeat the luciferase assay using a kidney specific cell-type, and see if the earlier TaqMan® probe kidney gene expression finding of a trend towards a decrease in kidney gene expression for the rs4773144 risk allele (G) can be repeated. Nonetheless, it is important to mention that the luciferase assay has its limitations, and may not provide the answer one might hope to discover - due to the fact that it is an *in vitro* artificial system.

All things being considered the luciferase assay findings would suggest that the rs4773144 putative transcription factor STAT3 is unlikely to alter gene expression in HeLa cells. In addition, given the array of different cytokine and growth factor stimuli used in the initial luciferase assay, with the potential to increase the pool of other transcription factors of relevance to the putative TFBS predictions given in **Table 4.4A/B**, perhaps HeLa cells themselves are the main reason why no allele specific luciferase gene expression difference are seen. Indeed, from the *in silico* ENCODE data other cell types showed a much stronger raw ChIP-seq signal for promoter specific H3K4Me3 histone marks suggesting that these cell types (i.e. HUVEC, HVMF, HCF, HCM or HRE cells) are worth considering beyond the use of the easier to transfect HeLa cells. In addition, if HUVEC or fibroblasts, etc. are again troublesome to transfect, it may be worth considering a viral transduction method, such as lentivirus, so that the the most biologically plausible human cells can be fully tested and properly assessed.

Overall the gel shift and luciferase reporter assays were inconclusive, in their ascertainment of whether the lead CARDIoGRAM candidate SNP rs4773144, or another member of the strongly linked (pairwise LD r^2≈1) 4-SNP candidate haplotype are causal of CAD at locus chr13q34. However, carriers of the rs4773144 risk allele (G) did show a trend (albeit non-significant) towards a decrease in kidney gene expression suggesting that there is a potential eQTL for the rs4773144 locus. This information along with the previously discussed *in silico* bioinformatics would certainly suggest that one or all of the 4-SNP haplotype SNPs are within a *cis*-regulatory element(s), which will in all likelihood play a binding role as part of a multi-component protein complex (as was indeed suggested by the DNA-protein gel shift assay).

The final assessment of function for the CAD-associated candidate SNP rs4773144 (or one of its proxies), was to see whether restenosis after a single-vessel PTCA procedure was influenced by genotype. The PTCA procedure opens up the lumen of the coronary arteries, but in doing so causes injury to the vessel wall, which forms a neo-intimal layer from proliferating vascular medial migrated SMCs, which in turn plays an important role in restenosis. The PTCA procedure is successful for the most part, but 30-40% of patients result in restenosis at 1-3 months follow-up. It is likely that individuals undergo restenosis at varying rates depending upon their genetic predisposition. The SHARP study was a randomised clinical trial that assessed subcutaneous unfractionated calcium heparin on angiographic restenosis, for subjects at three PTCA procedure time points pre-PTCA, immediately post-PTCA, and at 4 months follow-up of PTCA, and DNA was collected via blood sampling of single-vessel PTCA patients only.

Given that *COL4A1* and *COL4A2* encode for heterotrimeric protomer building blocks for the BM, and the BM surrounds vascular media SMC, these genes are prime targets to test for restenosis genetic predisposition. I therefore genotyped rs4773144 in the PTCA SHARP trial samples, to see if the CAD risk allele was also associated with restenosis, using two restenosis criteria measurements.

The restenosis analysis was non-significant for rs4773144 genotypes. Therefore, as regards the levels of COL(IV) protein within the BM, it seems to have no genotype specific effect on stenosis through *COL4A1* or *COL4A2*, but it should be noted that the power to see an effect in this exploratory study was low.

Another GWA study meta-analysis and replication association for this same 13q34 locus, is for rs3809346 with coronary artery calcification (CAC) via the CHARGE consortium (Discovery meta-analysis: P=1.25x10^{-6}; Discovery meta-analysis and replication: P=8.64x10^{-7}) (O'Donnell et al. 2011). CAC is a known risk factor for CAD, and this observation fits very nicely with the immunohistochemistry studies from the 1980s and 1990s, which mentioned that COL(IV) was present within atherosclerotic plaque calcified deposits (Katsuda et al. 1992; Tanimura, McGregor & Anderson 1986a; Tanimura, McGregor & Anderson 1986b). Furthermore, another SNP rs13260 located within the 3'UTR of *COL4A1* was very recently shown to have suggestive CAC association (P=8.7x10^{-5}) in the chronic renal insufficiency cohort (CRIC), and association with MI in the South Asian PROMIS replication cohort with an odds ratio of 1.13 (95% CI: 1.06-1.20), P=9.7x10^{-4} (Ferguson et al. 2013).

With this information in mind, a closer assessment of the CARDIoGRAM and CARDIoGRAMplusC4D meta-analysis data was made, for SNP rs13260, and the *P*-values are seen to be just below the *a priori* threshold required for subsequent replication in CARDIoGRAM (OR=1.12 (95% CI:1.06-1.19), *P*=2.3x10^{-4} for CAD; OR=1.23 (95% CI:1.12-1.34), p=5.1x10^{-6} for MI) and only nominally significant in Stage 2: CARDIoGRAMplusC4D (OR=1.05 (95% CI:1.01-1.09), *P*=0.01 for CAD; OR=1.05 (95% CI:1.00-1.10), *P*=0.036 for MI). Interestingly, the European only meta-analysis CARDIoGRAM and the PROMIS South Asian cohort both gave strong CAD/MI associations for rs13260, but surprisingly the Stage 2: CARDIoGRAMplusC4D meta-analysis that includes both European and South Asian (including PROMIS) cohort ethnicities, only gives a weak association. This is most likely due to issues with heterogeneity caused when meta-analysing such a large number of different studies (CARDIoGRAMplusC4D Consortium et al. 2013; Schunkert et al. 2011).

This means there are now two *COL4A1* moderately significant CAD-associated exonic SNPs – rs13260 (3'UTR) (Ferguson et al. 2013) and rs3742207 (Gln1334His) (Tarasov et al. 2009; Yamada et al. 2008), and two *COL4A2* non-coding, genome wide significant CAD-associated loci - bidirectional promoter 4-SNP haplotype, and an intronic SNP rs9515203, ~90kb downstream of rs4773144 within *COL4A2* (CARDIoGRAMplusC4D Consortium et al. 2013). Therefore, the *COL4A1* and *COL4A2* genes are surely going to be future targets for therapeutic intervention.

5. Chapter 5

General discussion and perspective

5.1 Summary of the main findings

There were several key findings made in this thesis at two CAD-associated GWA study discovery loci.

Firstly, a CAD GWA study SNP signal rs599839 at locus 1p13.3 located just beyond the 3'UTR of *PSRC1*, was found to be associated with an intermediate CV risk phenotype trait - LDL-C; and in particular with two small dense subclasses of LDL-C – (B) and (D). Fine-mapping CAD and LDL-C association analyses of the 1p13.3 locus identified four alternate candidate SNPs (rs7528419, rs12740374, rs629301 and rs646776) in almost perfect pairwise LD (r^2≈1) within, or just beyond the 3'UTR of *CELSR2*. *In silico* analyses through ENCODE suggest that rs12740374 is the most likely causal SNP based on DNase-seq, FAIRE-seq and ChIP-seq DNA-protein binding data, across multiple cell types, but rs646776 and another adjacent SNP rs660240, also in strong pairwise LD (r^2=0.95), also show evidence of DNA-protein binding.

Liver expression analysis by Schadt *et al.* identified an eQTL for CAD and LDL-C associated SNP rs646776 risk allele (A), which was strongly associated with the reduced expression of three clustered genes (Schadt et al. 2008). The 1p13.3 eQTL acted in *cis* upon *CELSR2* and *PSRC1*, and *SORT1*. In relation to the liver eQTL findings, *in silico* analyses for TFBS predictions show that the rs12740374 risk allele (G) disrupted two liver specific TFBS - HLF and CEBPA.

Functional genomics experiments by Musunuru *et al.* subsequently proved that rs12740374 was indeed the causal SNP, due to the risk allele (G) TFBS disruption of CEBPA in hepatocytes (Musunuru et al. 2010). Musunuru and colleagues than went on to elucidate the mechanism by which the causal SNP rs12740374 at locus 1p13.3 resulted in elevated LDL-C and CAD risk, through humanised mouse model experiments. Essentially, disruption of the C/EBP alpha TFBS resulted in the decreased expression of *SORT1*, which dysregulates intracellular nascent VLDL secretion, which in turn increases serum LDL-C, and henceforth also increases CAD risk (Musunuru et al. 2010).

Secondly, a CAD GWA meta-analysis SNP signal rs4773144 at locus 13q34, was narrowed down to a 258bp haplotype of four candidate SNPs located within intron 3 of the *COL4A2* gene. This four SNP signal was shown by *in silico* ChIP-seq histone mark signatures, to exist within a shared common bidirectional promoter by the paired genes *COL4A1* and *COL4A2*. These two genes are ubiquitously expressed and encode for two alpha chain subunits that form heterotrimeric protomers α1α1α2, an ECM secreted protein, which makes up the main structural protein building blocks of all BM. This four SNP candidate 13q34 locus was shown to be a putative eQTL in the kidney, i.e. the rs4773144 risk allele (G) was associated with a suggestive trend towards a (additive model) decrease in both *COL4A1* and *COL4A2*. This 13q34 locus has no intermediate cardiovascular trait phenotype. EMSA and luciferase reporter assay *in vitro* studies, showed there is DNA-protein binding, but no differential binding by genotype in vascular HASMC, or epithelial-like HeLa cells, two cell types known to be associated with BM. An assessment of restenosis levels showed no difference by rs4773144 genotypes, perhaps suggestive that the CAD-association is not driven through medial vascular SMC.

5.2 Implications

5.2.1 Challenges of moving from a GWA signal to a causal mechanism

This thesis has shown for locus 1p13.3 that it is possible to move from a novel SNP signal, identified as a GWA study genetic marker of CAD (a common, multifactorial disease) to a causal SNP, and its mechanistic pathway elucidated. However, of the 46 novel loci identified to date (CARDIoGRAMplusC4D Consortium et al. 2013), the 1p13.3 locus is the only one that has been properly pinpointed in terms of a causal SNP and mechanism of effect.

In general, there are a number of key challenges that need to be faced and overcome, in order to move from a CAD GWA study signal, to an identified causal SNP(s), and eventual elucidated mechanistic pathway (see **Fig. 5.1**) - so that targeted and preventative pharmacological therapeutics can be administered.

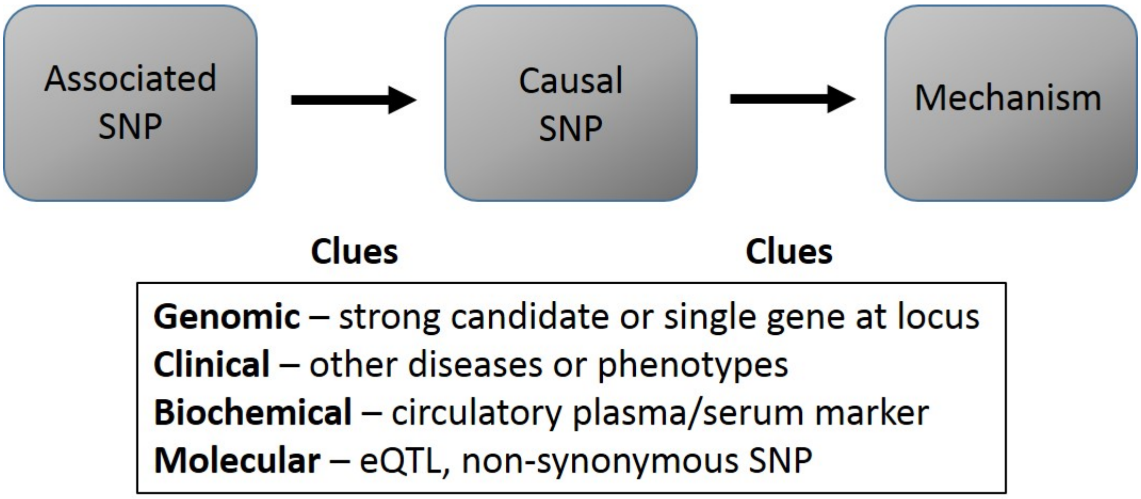

Associated SNP
Causal SNP
Mechanism
Clues
Clues
Genomic – strong candidate or single gene at locus
Clinical – other diseases or phenotypes
Biochemical – circulatory plasma/serum marker
Molecular – eQTL, non-synonymous SNP

Figure 5.1. The key challenge emerging from GWA studies.

The ease with which this challenge can be met and undertaken, is dependent upon the number of clues that can be garnered from: 1) genomic architecture, 2) other known clinical disease and phenotypes, 3) biochemical circulatory markers, and 4) molecular level effects.

In the case of genomic structure, it would be of great benefit if there was an obvious *cis*-acting candidate gene at the novel discovery locus; with annotation and published data relevant to a CAD related pathway mechanism. Certainly, this was the case for the 13q34 locus, studied in this thesis. For instance, mutations in the 13q34 locus genes *COL4A1* and *COL4A2* have both been implicated in diseases of blood vessel structures - including cerebral microvascular haemorrhage, i.e. stroke, porencephaly, HANAC, and more recently CAD and MI. The role of the two *COL4A1* and *COL4A2* encoded peptides are to form the major structural protein building blocks of the BM. The BM is responsible for structural integrity by forming an anchorage protein network lattice to maintain endothelial and vascular medial SMC stability, and to enable substrate permeability. Therefore, the COL(IV) α1α1α2 heterotrimeric protomers play a potentially critical and poignant role, in atheroma formation. In contrast, the 1p13.3 locus studied in this thesis was first identified as a SNP signal intergenically located between two *cis*-located genes with no known affiliation to CAD.

On the other hand, should there be no obvious biologically plausible candidate gene, the next best alternative would be that there is only one gene at the locus.

However, if there are multiple genes within the locus, in close proximity to the lead CAD-associated SNP, then it would require two alternative scenarios in order to pinpoint the most likely primary status candidate gene. Firstly, that the lead SNP, or another SNP in strong pairwise LD is a coding SNP. Secondly, that the lead SNP in whatever guise, i.e. intergenic, intronic, or non-coding, must have a cluster of supporting SNPs in strong pairwise LD. That being said if the lead SNP, and its supporting SNPs in weaker association are located intergenically, it will be very difficult to ascertain what candidate gene is most likely affected, without further clues.

Interestingly, in the case of 1p13.3 the lead SNP was intergenic, but strong local pairwise LD enabled fine-mapping to locate the truly causal SNP within the non-coding exomic 3'UTR of a nearby gene *CELSR2*, but this was not the nearest gene to the lead SNP, i.e. *PSRC1*. However, in the case of 13q34 locus the lead SNP and strong pairwise LD SNPs, were all located within a biologically plausible candidate gene *COL4A2* within intron 3, but near enough distally, to be part of the bidirectional common promoter for both *COL4A1* and *COL4A2* genes. Furthermore, it should be noted that although these two loci 1p13.3 and 13q34, seem to be located in their respective candidate genes, their non-coding status makes them likely to play a regulatory role. Henceforth, it should not be assumed that the effect will be *cis*-acting on local genes, indeed the association could very well end up being cis-acting, but affect genes much further away, i.e. up to 1Mb.

On a molecular level there are other alternative clues that can be assessed. One such clue is to see whether there is a genotype specific effect on gene expression that is either *cis*- or *trans*-acting - this is known as an eQTL. If there is an eQTL in a particular cell or tissue type, this will also provide clues as to mechanism and of course candidate gene(s). A prime example of this was a proxy to the lead SNP, i.e. rs646776 of locus 1p13.3, which was identified in liver cells to express a cluster of three nearby genes (*CELSR2*, *PSRC1* and *SORT1*). Indeed, here we see a *cis*-acting eQTL for *CELSR2*, *PSRC1,* and *SORT1*, whose promoter region is ~120kb downstream of the *CELSR2* 3'UTR, which harbours the causal SNP that regulates the CEBPA transcription factor activation. In contrast, 13q34 had no previous evidence of an eQTL for the lead

SNP, or any other SNPs either in close proximity or in weak pairwise LD (r^2>0.2), prior to this thesis. However, a putative kidney eQTL was identified in this thesis, suggesting a potential role for COL(IV) α1α1α2 reduction, via the rs4773144 SNP identified 13q34 locus in the kidney, which could influence CAD risk.

Moreover, as was mentioned earlier for the 13q34 locus, further clues could be ascertained by observing whether the CAD GWA study signal, is also seen in other clinically relevant diseases. Indeed, the 13q34 locus is associated with a CAD surrogate coronary artery calcification (CAC) (O'Donnell et al. 2011). Interesting the chronic renal insufficiency cohort (CRIC) identified a SNP association with rs13260, located in the 3'UTR of *COL4A1*, and this same SNP was found to be associated with MI in the PROMIS study (Ferguson et al. 2013). Another good example would be that of locus 9p21.3, where the same SNP signal rs1333049 is associated with a number of cardiovascular diseases (CVD) - abdominal aortic aneurysms and carotid intima-media thickness (a vascular surrogate for CAD), type 2 diabetes, and a non-atheroma CVD intracranial aneurysm; and an array of different cancer disease forms, due to deletions (Samani & Schunkert 2008). This indicates that the association in CAD maybe due to a general effect on blood vessel structure, and perhaps through cell cycle and cell proliferation, as is indicated by the nearby candidate gene cyclin dependent kinase inhibitors - *CDKN2A*, *CDKN2B*, and *CDKN2B-AS1* (Samani & Schunkert 2008).

Another alternative clue to help elucidate the mechanism of effect at a given locus is to see if the SNP has a genotypic effect on an intermediate quantitative trait of known risk to CAD. This could be a relevant biochemical marker from within the circulation - such as elevated cholesterol, physiological - such as high blood pressure, or behavioural - such as smoking. To this end, this functional characterisation procedure was successfully employed for locus 1p13.3 for LDL-C. Indeed, it turned out that this association was the strongest identified of any lipid trait GWA study (Teslovich et al. 2010), and was one of the largest for MI risk association at ~40% (Samani et al. 2007; Myocardial Infarction Genetics Consortium et al. 2009).

As one can see the accumulation of evidence from the specific types of clues described above (i.e. genomic, clinical, molecular and biochemical), give strength to predictions of causal SNPs, and help functionally characterise the potential mechanistic pathways that lead to CAD risk association. Indeed, the 1p13.3 locus is a clear cut success story, mainly due to the fact that it has very strongly significant associations with: a) CAD risk, b) LDL-C - which is one of the main risk factors for CAD, and c) liver gene expression QTL - which is the main tissue involved in the regulation of LDL-C metabolism and catabolism. Moreover, the *cis*-acting eQTL identified *SORT1* gene, had previously characterised evidence of acting as a multi-ligand receptor, and internalising LPL and apoAV in to liver cells (Nielsen et al. 1999; Nilsson et al. 2008) – making this gene an extremely plausible candidate gene for focussed functional genomics experimentation. In fact, the DNA manipulated EMSAs and luciferase assays using hepatocytes, which included site-directed mutagenesis were very successful at identifying a liver specific TFBS prediction; and the overexpression and knockdown assays were very successful at pinpointing the involvement of *SORT1*, in the liver.

For the most part, two-thirds of the 46 novel CAD-associated loci identified to date (CARDIoGRAMplusC4D Consortium et al. 2013), are not known to be associated with an intermediate CAD risk quantitative trait - such as lipids or blood pressure. Therefore, as was seen for locus 13q34 it is harder to discover the underlying mechanisms of less well characterised loci. Nonetheless, locus 13q34 has two ideal candidate genes in *COL4A1* and *COL4A2*, and did indicate a potential eQTL in the kidney. Unfortunately, the follow-up functional studies were unable to identify a genotype specific effect, which might best explain the CAD risk at this 4-SNP narrowed down, bidirectional promoter 13q34 locus. Vascular SMCs are an excellent choice for assessing genotype specific effects on the binding affinity to TFBS or differential expression levels; because the BM anchors media SMC of blood vessels, and atheroma extremely elongated migratory SMC secrete a thick layer of COL(IV) when deep in the fibrous cap. However, the role of media derived SMCs may have been ruled out on the basis of EMSA assays and the restenosis study analysis.

Overall these evidentiary clues indicate that finding the right cell types, for downstream functional characterisation studies is essential. In addition, even if you strike ‘gold’ with the correct cell type, you still have to find the right stimuli conditions, to identify a genotypic effect relating to DNA-protein binding affinity, and gene expression. It you don’t you may discard a cell type that plays an important mechanistic pathway role. It will also help if each functional experiment executed, provide strong and clear results, to give confidence to subsequent studies.

The key aim, is to be able to mimic the *in vivo* human pathological conditions, at any one, given time and space. Interestingly, even with a wonderful new library of *in silico* DNA binding elements, as identified through the ENCODE project - pinpointing the causal event is a challenge.

5.2.2 Pros and cons of functional genomics approaches used

ENCODE project *in silico* data provide a very powerful resource, for potentially relevant *cis*-regulatory element discoveries, which can be related to genetic variation, such as SNPs. By using a multitude of both low and high resolution experimental techniques, such as ChIP-seq, DNase-seq and FAIRE-seq in cultured cells - it is possible to align where upon the genomic assembly, DNA-protein binding is likely to be taking place. These techniques also indicate, what regions of the genome are open and actively able to engage with DNA-protein binding.

However, there are some limitations. In particular, many of the experiments performed for ENCODE (and referred to in this thesis) utilise cultured human cell lines, and there are three potential problems with the experiments performed, using cultured cell lines.

Firstly, cultured cell lines are often described as *in vivo*, but are not *in vivo* in the truest sense (i.e. within the animal). That is to say, cultured human cells are not necessarily a satisfactory substitute for the truly *in vivo* environment. This could be of particular relevance when it comes to *cis*-regulatory elements, as whole organism compensatory mechanisms maybe missed, due to perhaps hormonal or environmental changes.

Secondly, each experiment is for a given ‘snap-shot’ in time, you are not getting a truly dynamic profile of what proteins maybe binding to open/active chromatin, under different stages of the cell cycle, and different environmental conditions (such as normal physiological vs. pathological conditions).

Thirdly, the binding identified in a ChIP-seq TFBS experiment is often only for one cell line, i.e. one individual’s genomic profile. This means that the DNA-protein binding identified at a site where a polymorphism exists, will be dependent on the allele present for that one particular cell-line genomic profile. Therefore, you are missing any potential binding difference that would occur for an alternate, genomic profile. Hence, just because a non-ENCODE identified TFBS is discovered by a PSSM prediction web-based tool, but not identified by ChIP-seq - it does not mean it couldn’t still be relevant in the same cell type or even another cell type, with an alternate genomic profile. However, false positive TFBS PSSM predictions are most likely in the majority of cases, due to the short sequence specific recognition sites, and the evolutionary need for flexibility. Indeed, certain ENCODE track alignments, and this includes ChIP-seq for TFBS, show the overall data, based on several different cell-lines; and currently in the case of ChIP-seq TFBS prediction, in as many as 125 cell-lines.

Although ENCODE has some limitations overall, you would be hard pushed to find a better methodology, for the mimicry of the true *in vivo* condition. It is more that ENCODE, is still in its infancy for certain aspects of DNA-protein binding identification, such as ChIP-seq TFBS. Nonetheless, overtime, there will be many more individuals (with unique genomic profiles) used for any one given cell-type, and cell culture performed DNA-protein capture will take place under multiple conditions (i.e. more representative of true *in vivo* conditions). The collection of such data, will enable the construction of far larger reference libraries, for the purposes of genomic assembly alignment. Hence, ENCODE will become all the more beneficial. Indeed, the clues provided by ENCODE at present, are already a great resource for the development of functional experiments, for hypotheses testing.

Another limitation of the TFBS ChIP-seq methodology is that signals are low resolution (~200bp), and can only test one protein at a time, dependent on the 'snapshot' of transcriptional regulation in your gene(s) of interest at the time of DNA-protein binding capture, i.e. repressed or activated.

The human cell culture functional genomics DNA manipulation experiments, using EMSAs and luciferase reporter assays were far from convincing. This was in terms of supporting the potentially relevant findings that the rs4773144 CAD risk allele (G), or one of three proxy SNPs - decreased kidney *COL4A1* and *COL4A2* gene expression. This could be because the most appropriate cell type was not used and/or the cell culture conditions were not optimised to identify a genotype specific effect. For instance, it may be important to use a different vascular cell type, such as endothelial cells rather than HASMC or HeLa cells. It may be necessary to stimulate the cell culture conditions with a stronger stimulus, perhaps increasing the concentration of cytokines or growth factor, or adding different transcription factor expression vectors, or adding recombinant human forms of transcription factor proteins.

In the case of the luciferase reporter assays, another potential limitation is the use of pGL3 Promega vectors, rather than the newer pGL4 vectors. The newer vectors have fewer cryptic TFBS, so the vectors will mop-up less of the endogenous proteins within the cell culture environment. For instance, a loss of important transcription factors within the cell culture environment, could detrimentally affect the formation of a particular putative multi-component complex, under hypothesis.

5.2.3 GWA studies and the future

It is now well-established from a strong family history of CAD that it is important to study the genetic component of the CAD phenotype (i.e. the population phenotypic variance attributed to genetic variation). Evidentiary support for this notion, has been backed-up by over 30 years of follow-up data measurements of twin and family study CAD heritability estimates that vary between 38% and 57% (Zdravkovic et al. 2002).

To date, even after 30 years of studying the underlying genetic aetiology of CAD, a common and complex multifactorial, polygenic disease phenotype, it is still poorly understood. Nonetheless, since 2007, as part of the post-genomics era, technological advances have enabled the introduction of a large scale GWA study. The results produced from GWA studies have enabled a rapid and exciting progression in the genetics of CAD and other complex diseases. Indeed, the GWA study approach, along with improved statistical power by use of meta-analysis - has proved to be the most successful at garnering reproducible polymorphism disease associations. Moreover, recently as a consequence of the collaborative CARDIoGRAMplusC4D consortia, the number of CAD-associated discovery loci has been enhanced by meta-analysis of studies using the Metabochip, to 46 loci that have reached genome wide significance (i.e. $P < 5 \times 10^{-8}$) (CARDIoGRAMplusC4D Consortium et al. 2013).

As expected many of the 46 loci were associated with known metabolic pathways, involving traditional CAD risk factors. Indeed, twelve loci were significantly associated with dyslipidaemia traits, eight of these loci were most strongly associated with LDL-C (*APOB, ABCG5-ABCG8, PCSK9, SORT1, ABO, LDLR, APOE and LPA*) indicating the importance of LDL-C as a causative risk factor in CAD. Two loci were most strongly associated with TG (*TRIB1* and *APOA5*). One locus was associated with HDL-C (*ANKS1A*), and another locus was strongly associated with both HDL-C and TG (*LPL*). Of note, all but two loci *LPA* and *ANKS1A* reached genome wide significance for a lipid trait (CARDIoGRAMplusC4D Consortium et al. 2013; Kathiresan et al. 2009; Chasman et al. 2009). Five loci were significantly related to BP (*CYP17A1-NT5C2, SH2B3, ZC3HC1, GUCY1A3* and *FES*) (International Consortium for Blood Pressure Genome-Wide Association Studies et al. 2011). One locus each showed suggestive association with WHR (*CYP17A1-NT5C2*) and BMI (*RAI1-PEMT-RASD1*), respectively (CARDIoGRAMplusC4D Consortium et al. 2013). Nevertheless the vast majority (two-thirds), have no known affiliation with traditional CAD risk factors and indicate that, as yet unknown mechanisms are likely to play an important role in CAD aetiology.

In spite of this great leap in CAD genetics, to date the 46 novel discovery loci explain only a small proportion (approx. 6%) of the considerable heritability estimates. Indeed, even if we considered the additional CARDIoGRAMplusC4D identified 104 SNPs with likely independent association at a FDR of 5% and pairwise LD at $r^2 < 0.2$ (i.e. very weak to no LD) - the heritability estimate only advances to an approximate 10.6%. Moreover, it has also been bemoaned that the GWA study era discoveries, have provided little improvement in our biological understanding of the complexities, underlying the CAD phenotype. In reality, such negativity is unwarranted as these findings merely signify how complex multifactorial conditions, such as CAD truly are. Indeed, the recent ENCODE revelations show that most of the genome is transcribed (~75%), and contains regulatory elements involved in biochemical function (~80%). This just indicates how naive geneticists' were to believe that most of the human genome was inactive ancestral 'junk DNA' (ENCODE Project Consortium et al. 2012; Djebali et al. 2012). The reality is that we are just starting to unpeel the outer onion layers of genetic complexity, in relation to CAD and other polygenic diseases.

So a key question that arises from GWA studies, is what are the reasons that best explain the shortfall in heritability estimates? To this end, there have been numerous hypotheses proposed, to explain the term "missing heritability" (Maher 2008). These include for instance, gene-gene interactions, gene-environment interactions, structural variation, epigenetics, as well as unmeasured common variants with smaller effects, unmeasured rare variants, and allelic heterogeneity (Manolio et al. 2009; Eichler et al. 2010; Li & Leal 2008; Zhang et al. 2012). The notion given most credence and highest priority by researchers right now, is the "common disease, rare variant" (CD/RV) hypothesis as a means of investigating whether - rare variants (i.e. those with MAF < 0.05) explain a large portion of the missing heritability (Manolio et al. 2009). If rare variants truly do explain the missing heritability, it is in part due to the fact that GWA studies lack the statistical power to detect most rare variant associations, because of the low pairwise LD (r^2) between the HapMap common tagged SNPs and rare potentially causative variants. Indeed, there is strong evidence supporting the CD/RV hypothesis, which suggests that low-frequency (MAF 1% to 5%) and rare variants (MAF < 1%) are involved in the causality of complex disease (Bodmer & Bonilla 2008; Gibson 2012). For example, 11 of the 30 genes, including *LDLR*, *PCSK9* and *ABCA1*, with common variants discovered via GWA studies, are also known to carry rare variants that cause Mendelian dyslipidaemia, indicating that genes can carry both common variants of small or moderate effect size and rare variants of large effect size (Kathiresan et al. 2009; Lusis & Pajukanta 2008). Now with the realisation of next-generation sequencing technology, it is possible to carry out large scale sequencing projects even going as far as whole genome sequencing (with the promise from Illumina, Inc. of a $1000 human genome, just around the corner). Thus, the validity of the CD/RV hypothesis and the proportion of missing heritability that is attributable to rare variants can now be investigated. In the first instance, due to cost constraints sequencing projects of this kind have been aimed at the exome, for example the exome array (Illumina Inc.). The idea being that the exomic amino acid changes, will generate a number of large effect size associations that geneticists' will have the statistical power to detect. In addition, the 1000 genomes project recently completed 1092 haplotype resolved genomes, for the purpose of generating a genome-wide integrated map inclusive of 38 million SNPs. This map can be utilised for whole genome imputation that will include rare SNPs (1000 Genomes Project Consortium et al. 2012). However, it will most likely require whole genome sequencing studies, in order to scan the genome and detect mutations that affect the ~80% ENCODE predicted, regulatory functional elements.

In contrast, it should not be assumed that the "common disease, common variant" (CD/CV) hypothesis (i.e. MAF ≥ 0.05) driven GWA studies, do not explain more of the missing heritability. For instance, it is likely that many of the common variants that do not reach the stringent level of genome wide significance, will actually be genuine CAD risk variants of small effect size. Unfortunately, they are discarded as statistical type II errors (i.e. false negatives) due to the lack of statistical power. Evidentiary support for this reasoning, has been shown in height GWA study meta-analysis, using maximum likelihood statistical methods that utilise the *P*-values for all SNPs, regardless of significance. Additive genetic variance for height using these methods on common genetic variants, explained between 29% and 45% of the total phenotypic variance (Yang et al. 2010; Kutalik et al. 2011). This was a 3-fold to 4.5-fold increase in the phenotypic variance explained by the 180 loci, which reached genome wide significance (Lango Allen et al. 2010). Although, these methods are yet to be standardised, they do indicate in due course that GWA studies are likely to make a larger contribution to the twin and family study heritability estimates than first expected, but that the remaining genetic variation lies elsewhere, perhaps through the CD/RV hypothesis.

5.2.4 Translational benefits to medicine

As mentioned earlier in the introduction chapter the financial and social cost of CAD is vast, in 2006, the estimated NHS cost of CAD to the UK was £3.3 billion, with an approximate £54 per capita (Scarborough et al. 2010). However, the scope of the financial and social burden in the UK goes far beyond this cost. Firstly, in terms of productive years lost through premature deaths <75 years (i.e. 1 in 3); and secondly by the productive years lost through a morbidity prevalence of 1 million men and 500,000 women who suffer from an MI, and 2.7 million registered as having CAD. Surprisingly, these great economic and social burdens exist in spite of vast medical advancements reducing the death rate from CAD in the UK by 65%, between 1980 and 2007 (Scarborough et al. 2010).

However, the good news is that since the earliest GWA studies for CAD, an impressive and robust 46 novel loci have been discovered, and this means that each locus has the potential to advance therapeutic preventative intervention, which will in turn help to reduce both the economic and social burden. Furthermore, these 46 loci only represent approximately 6% of the CAD heritability estimates. Henceforth, if we can bridge the remaining gap in missing heritability estimates, there could be yet further reductions in death rate for CAD, perhaps akin to those seen over the past 30 years, through advancements in modern medicine.

It is essential therefore that we are able to harness the GWA study signals of CAD risk, for the development of new therapies, and to enhance prevention strategies through predictions of CAD risk scores, such as the Framingham risk score.

It is possible to obtain genomic profiling of patients or indeed all individuals, for purposes of generating a predictive score for all genotypic risk that can be added to the current Framingham risk score parameters. However, with each individual having hundreds of potentially causative and preventative genotypes in equal measure, it will be a difficult task to decipher who is truly at greater risk statistically. However, pharmaceutical therapies can be very specific in their targeting of each potentially causative novel locus in turn, and will perhaps achieve a more additive protective benefit for all, over time. Of course, in order for this to be achieved we must identify very specific molecular targets for drug therapy, through understanding the mechanistic pathways involved.

To this end, as a proof of principle, there is one example of a drug therapy molecular target currently undergoing phase III clinical trials - which has been developed as a consequence of a known genetic association that pre-dates the GWA study era. The example in question is a drug therapy for targeting the action of serine protease PCSK9. Gain-of-function monogenic polymorphisms in *PCSK9* were first discovered by Abifadel *et al.*, in subjects with autosomal familial hypercholesterolaemia, i.e. high circulating LDL-C and premature CAD (Abifadel et al. 2003). Moreover, the MI genetics consortium identified an intronic gain-of-function common risk variant rs11206510, through GWA study combined analyses (Myocardial Infarction Genetics Consortium et al. 2009). A molecular elucidated role for PCSK9 was described in the following year using animal models, where PCSK9 post-translationally regulates LDLR activity (Park, Moon & Horton 2004; Maxwell & Breslow 2004). A schematic diagram showing the role of PCSK9 and its inhibition is shown in **Fig. 5.2**.

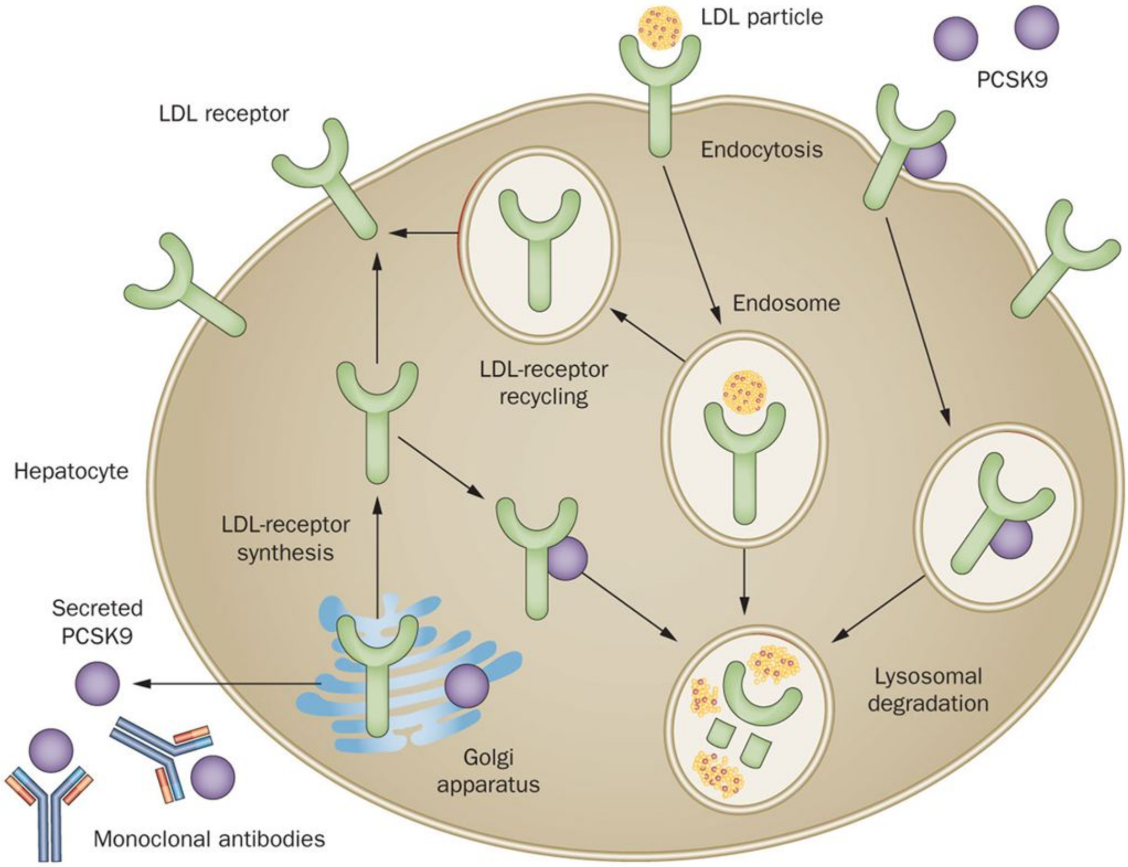


Figure 5.2. Mechanism of PCSK9 and its inhibition with monoclonal antibodies (Image taken from (Brautbar & Ballantyne 2011)).

The mechanistic pathway is one where PCSK9 is synthesised and secreted from the liver into the circulation. The catalytic domain of PCSK9 then binds to the epidermal growth factor-like repeat A (EGFA) domain of hepatic LDLR, and targets them for lysosomal degradation. One of the main roles of the LDLR is to bind and internalise LDL-C into endosomes, and then for LDLR to be recycled back to the hepatocyte cell surface, ready to bind and internalise more LDL-C. However, PCSK9 reduces the ability of the liver to remove LDL-C from the circulation and results in elevated LDL-C in the circulation. The therapeutic potential of this mechanistic pathway was made all the more plausible, following the discovery of loss-of-function monogenic mutations (Cohen et al. 2005; Kotowski et al. 2006), which resulted in reduced LDL-C levels and decreased incidence of CAD (Cohen et al. 2006). As a consequence of these findings, drugs or to be more specific human monoclonal antibodies, have been developed that recognise and bind to PCSK9 near the catalytic domain, e.g.

evolocumab (Amgen), ID05-IgG2 (Merck & Co.) and alirocumab (Aventis/Regeneron), and block its interaction with LDLR. In fact, several phase III clinical trials are now under way, to see if the drug PCSK9 inhibitor can lower disease incidence, with acceptable levels of side effects. Phase I clinical trials in healthy and heterozygous familial hypercholesterolaemia subjects - have already shown that PCSK9 inhibitors (REGN727, aka alirocumab) can reduce LDL-C levels by 50% (Stein et al. 2012); with no resistance to therapy caused by antibody associated immunogenicity. Furthermore, the actions of PCSK9 inhibitors are enhanced still further in the presence of statins. Statins on their own cannot reduce LDL-C as much as PCSK9, because although they target and reduce HMGCR, a key enzyme in cholesterol synthesis, there is a feedback compensatory loop that means LDLR and PCSK9 levels are increased (Sheridan 2013). Hence, a ‘polypill’ including both PCSK9 inhibitors and statins is more effective at reducing LDL-C levels, and hopefully will prove to reduce the clinical incidence of CAD too.

5.3 Future studies

There are two possible follow-up experiments that could be undertaken, to see if the suggestive putative eQTL identified at locus 13q34, in kidney tissue *COL4A1* and *COL4A2* mRNA levels, is valid.

Firstly, it may be possible to perform another eQTL assessment in another dataset with gene expression measurements for *COL4A1* and *COL4A2* from kidney tissue RNA, and measure rs4773144 genotypes in cDNA synthesised from this alternative RNA. To this end, another group Wheeler *et al.* performed an eQTL mapping assessment to find genes associated with aging in kidney tissue, in a manner similar to that of the PKP study. Indeed, Wheeler *et al.* made use of whole genome microarray (inclusive of probes for both *COL4A1* and *COL4A2*) gene expression measurements of kidney tissue RNA, from another group Rodwell *et al.* that had performed a transcriptional profile of aging (29 to 92 years) (Rodwell et al. 2004). Wheeler *et al.*, then genotyped ~1400 SNPs from 630 age-regulated genes in n=96 archived DNA (a similar sample size to the PKP study, n=101) that had kidney gene expression microarray data (Wheeler et al. 2009). Of note, the *COL4A1* and *COL4A2* genes did not show an expression change with age (Rodwell et al. 2004). Therefore, it may be feasible to perform an experiment similar to Wheeler *et al.* to genotype rs4773144 in archived DNA that corresponds to the individuals used to generate eQTL data in the Rodwell *et al.* microarray dataset (or another similar dataset), and compare the findings reported in this thesis for a putative kidney eQTL (Rodwell et al. 2004; Wheeler et al. 2009).

Secondly, another possible notion, as described previously in chapter 4, is that the kidney tissue decreased *COL4A1* and *COL4A2* gene expression (putative eQTL) effect is driven primarily through a specific cell type. If so, it might be worth performing kidney tissue sieving or immunomagnetic cell sorting separation (MACS) techniques, to isolate for example, mesangial cells (renal glomerular vSMC) (Pawluczyk, Patel & Harris 2004), or renal microvascular endothelial cells (RMEC) (Muczynski et al. 2003), prior to RNA extraction.

In chapter 4, there were some interesting ENCODE project ChIP-seq TFBS alignments seen at the 13q34 locus. In particular, E2F6 and MAX showed dense signals aligned with the known COL4 silencer, 25 nucleotide region (Haniel et al. 1995), just beyond exon 3 of *COL4A2*; and with or near to the 13q34 locus candidate SNPs. The scientific literature suggests that E2F6 binds to a MAX-MGA or MAX-MXI1 heterodimers, as part of a multi-component polycomb repressor complex (Ogawa et al. 2002; La Thangue 2002).

Given that the polycomb repressor complex is destabilised by other E2F family members, such as E2F1-3 to enable transcription - it would be very useful to be able to capture ChIP-seq under the ENCODE project umbrella, at different stages of the cell cycle, and under stimulated conditions. Under these circumstances, you could prove the existence of a regulatory complex, and indeed whether the 4-SNP candidate haplotype plays a role.

In addition, it is quite plausible that there is another putative *in silico* derived multi-component DNA-protein binding complex, which could control transcriptional regulation. This region certainly requires a deeper assessment functionally. Several strong lines of ENCODE evidence and TFBS software predictions, point to a putative TGFβ1/SMAD3 mediated binding complex near to and possibly including rs3809346, as well as transcription factors: SP1, EVI-1 and ETS.

As for the functional genomics experiments used in chapter 4, namely EMSAs and luciferase reporter assays. It would be useful to repeat these assays using an alternative vascular cell type, for instance HUVECs. There are two reasons for this selection. Firstly, because endothelial cells are a potentially relevant cell type, due to its anchorage to the BM. Secondly, because HUVECs show a strong (ENCODE) ChIP-seq H3K4Me3 signature signal for an active promoter.

Yet another interesting idea, would be to test for a genotype specific effect of the four CAD-associated candidate SNPs at locus 13q34, on measurements of circulatory endothelial cells (CEC), a known CAD biomarker. It might be possible to attribute a CEC-QTL using immunomagnetic factors and flow cytometry in subjects of known genotype. CEC are indicative of EC shedding due to apoptosis, and EC denudation by subendothelial matrix proteolysis. Therefore, it is feasible that a loss of endothelial anchorage and structural integrity, enhance the future risk of an atheroma fibrous cap superficial erosion, and a subsequent thrombotic event.

Finally a technically feasible recent breakthrough in terms of disease modelling and functional genomics is human genome editing. This method could be used for instance, for point mutagenesis of the 1p13.3 locus CAD causal SNP rs12740374 in hepatocytes differentiated from embryonic stem cells (ESCs), induced pluripotent stem cells (iPSCs), or mesenchymal stem cells (MSCs). This would be a useful future step for assessing *SORT1* gene expression and VLDL secretion before moving on to drug targeted therapy. Indeed, this technique could even ultimately, be used for corrective gene therapy therapeutic intervention.

Genome editing is a type of genetic engineering that enables the insertion, replacement, or removal of DNA from the genome, using artificially engineered nucleases. The role of nucleases is to create specific double-stranded break (DSBs) at targeted locations within the genome, and then exploit the cell’s own natural biology via endogenous repair mechanisms, such as homologous recombination (HR) and non-homologous end-joining (NHEJ). At present there are currently four types of engineered nucleases: zinc finger nucleases (ZFNs), transcription activator-like effector nucleases (TALENs), the CRISPR/Cas system and engineered meganucleases. There is also an additional genome editing proprietary recombination (vector) adeno-associated virus (rAAV) technique that boasts 1,000 times greater targeting efficiency within the nucleus than plasmid-based methods. This rAAV technique has no off-target sequence alteration elsewhere in the genome, and single base precision editing - because it only makes use of the precise HR natural repair mechanisms. This rAAV technique is only available through a precision genome editing company called Horizon (www.horizondiscovery.com). In fact, this same company has recently introduced a complementary modified CRISPR/Cas plasmid delivery technique, which makes use of an engineered Cas9 bacterial protein that introduces single-stranded 'nicks' rather than DSBs, and targets nucleus genomic DNA using two complementary guide-RNAs for precise genomic DNA hybridisation. The key utility of genome editing for GWA study discovery loci, is being able to use SNP mutagenesis on putative causal SNPs, to see what phenotypic effects this will have on a human cell line (including those derived from differentiated stem cells (such as ESCs and iPSCs or MSCs). This is of particular use when other functional genomics techniques have failed to detect a subtle phenotypic change, i.e. locus 13q34. This is perhaps, because the EMSA and luciferase reporter assay experiments are unable to represent a scenario that has resemblance, to a truly *in vivo* human condition. Indeed, if this genome-editing technique were to be harnessed by the ENCODE project, or individual research groups employing ENCODE type analyses, the possibilities for mechanistic pathway discoveries could be potentially very impressive.

5.4 Final concluding remarks

This thesis has shown that it is possible to move from a CAD GWA study novel SNP signal, to a causal SNP and its mechanistic pathway elucidation.

In particular, it was a great success that the 1p13.3 CAD-associated locus first identified through GWA study SNP rs599839, could be attributed to the causal SNP rs12740374 risk allele (G); and that the mechanistic pathway has been proven to decrease *SORT1* liver gene expression, which increases intracellular nascent VLDL secretion, which in turn increases serum levels of LDL-C, and hence increases CAD risk. Henceforth the 1p13.3 CAD-associated locus is a beacon of hope for the future of new therapeutic strategies. Particularly, given the success of statins and potential success of PCSK9 inhibitors.

As for the 13q34 CAD-associated locus studied, overall the implications are that no mechanism has been discovered. However, the PKP study did produce a putative kidney tissue eQTL, which suggests an additive trend towards a reduced expression of both genes *COL4A1* and *COL4A2*. Prior to this thesis there had been no previous eQTL reported for genes *COL4A1* and *COL4A2* at locus 13q34, for rs4773144 or one of its almost perfect proxies ($r^2 \approx 1$), or indeed for any correlated SNP in even weak LD $r^2 > 0.2$. Therefore, the putative eQTL in kidney, provides evidence that there really could be an effect on CAD at the 13q34 locus perhaps through the kidney, as is strongly suggested by *in silico* ENCODE data, and two large meta-analyses studies reported by two consortia (CARDIoGRAMplusC4D Consortium et al. 2013; IBC 50K CAD Consortium 2011).

Furthermore, in agreement with the putative eQTL, there are potential mechanistic pathways through which CAD risk could be enhanced, and these are mostly likely to be found through reduced levels of BM at particular sites, within the atherosclerotic plaque: 1) anchorage of the endothelium, a weakened BM may lead to an erosion prone fibrous cap via desquamation of the endothelial layer; 2) reduced BM surrounding media SMC may speed up the progression of the atheroma, through an enhanced release of migratory SMC into the intima, where they become proliferative (although the EMSAs and restenosis study perhaps indicates this effect is unlikely); and 3) reduced BM may result in a higher level of apoptosis in SMC, and release of calcified nodules not removed via efferocytosis, which could explain the CAC association (O'Donnell et al. 2011).

In conclusion, the unbiased non-hypothesis driven GWA studies have provided the field of CAD genetics with exciting novel loci associations - with great potential for future clinical preventative therapies. However, moving from the initial enthusiasm of an association to a definitive function has proven, in most cases to be a slow and laborious process. In addition, as regards any new loci identified through exome and whole genome sequencing, they are likely to have even smaller effect sizes, which could potentially make discovery of function even more difficult, due to the likely nuances of the association effect, alongside unforeseen compensatory mechanisms (that are essential for the adaptable survival of all living things). However, with constant and ever growing technological advancements, including innovative endeavours, such as the ENCODE and the 1000 Genomes projects and genome editing – the constraints of today will inevitably be overcome in the future.

6. Appendix

6.1 Supplementary Tables for Chapter 2

Table S2.1. Association findings for all quantitative traits in the GRAPHIC study for the seven analysed SNPs

Chr.	SNP	Phenotype	β-Coefficient	SE	Lower 95% CI	Upper 95% CI	*P*-value
1	rs17465637	mean 24 SBP	0.395	0.3853	-0.36	1.151	0.3
1	rs17465637	mean 24 DBP	0.278	0.2552	-0.222	0.778	0.28
1	rs17465637	mean 24 PP	0.107	0.2454	-0.374	0.587	0.66
1	rs17465637	BMI	0.157	0.1608	-0.159	0.472	0.33
1	rs17465637	WHR	0.003	0.0025	-0.002	0.008	0.25
1	rs17465637	Total cholesterol	0.053	0.0331	-0.012	0.118	0.11
1	rs17465637	HDL cholesterol	-0.009	0.013	-0.034	0.017	0.5
1	rs17465637	Blood glucose	0.056	0.0415	-0.025	0.137	0.18
1	rs17465637	Urate	1.312	2.2374	-3.073	5.698	0.56
1	rs17465637	CCR	0.379	1.3823	-2.33	3.088	0.78
1	rs599839	mean 24 SBP	-0.673	0.4111	-1.479	0.133	0.1
1	rs599839	mean 24 DBP	-0.498	0.3027	-1.091	0.095	0.1
1	rs599839	mean 24 PP	-0.173	0.2643	-0.691	0.345	0.51
1	rs599839	BMI	-0.265	0.1656	-0.589	0.06	0.11
1	rs599839	WHR	-0.001	0.0026	-0.006	0.004	0.66
1	rs599839	Total cholesterol	0.168	0.0363	0.096	0.239	**3.84×10^{-6}**
1	rs599839	HDL cholesterol	-0.018	0.0143	-0.046	0.01	0.21
1	rs599839	Blood glucose	0.019	0.0454	-0.07	0.108	0.68
1	rs599839	Urate	-2.475	2.3963	-7.172	2.222	0.3
1	rs599839	CCR	-0.379	1.3583	-3.041	2.283	0.78
2	rs2943634	mean 24 SBP	-0.064	0.3697	-0.788	0.661	0.86
2	rs2943634	mean 24 DBP	-0.07	0.25	-0.56	0.419	0.78
2	rs2943634	mean 24 PP	0.011	0.2522	-0.505	0.483	0.96
2	rs2943634	BMI	-0.093	0.1512	-0.389	0.203	0.54
2	rs2943634	WHR	-0.003	0.0022	-0.008	0.001	0.14
2	rs2943634	Total cholesterol	0.011	0.031	-0.049	0.072	0.71
2	rs2943634	HDL cholesterol	-0.006	0.0119	-0.03	0.017	0.6
2	rs2943634	Blood glucose	-0.051	0.0391	-0.127	0.026	0.2
2	rs2943634	Urate	0.586	2.1993	-3.724	4.897	0.79
2	rs2943634	CCR	-0.081	1.3232	-2.674	2.512	0.95
6	rs6922269	mean 24 SBP	-0.639	0.372	-1.368	0.09	0.086
6	rs6922269	mean 24 DBP	-0.127	0.2612	-0.639	0.385	0.63
6	rs6922269	mean 24 PP	-0.504	0.2517	-0.997	-0.011	**0.045**
6	rs6922269	BMI	0.087	0.1645	-0.236	0.409	0.6
6	rs6922269	WHR	0	0.0025	-0.005	0.005	0.91
6	rs6922269	Total cholesterol	0.011	0.0353	-0.058	0.08	0.76
6	rs6922269	HDL cholesterol	-0.017	0.0126	-0.042	0.008	0.18
6	rs6922269	Blood glucose	0.092	0.0454	0.003	0.181	**0.042**

6	rs6922269	Urate	-0.389	2.2612	-4.821	4.043	0.86
6	rs6922269	CCR	0.892	1.3669	-1.787	3.572	0.51
9	rs1333049	mean 24 SBP	0.028	0.3691	-0.696	0.751	0.94
9	rs1333049	mean 24 DBP	-0.288	0.2453	-0.769	0.193	0.24
9	rs1333049	mean 24 PP	0.313	0.2327	-0.143	0.769	0.18
9	rs1333049	BMI	0.304	0.1503	0.009	0.598	**0.043**
9	rs1333049	WHR	0.005	0.0023	0.001	0.01	**0.023**
9	rs1333049	Total cholesterol	-0.002	0.0301	-0.061	0.057	0.95
9	rs1333049	HDL cholesterol	-0.011	0.0112	-0.033	0.011	0.34
9	rs1333049	Blood glucose	0.027	0.0401	-0.052	0.105	0.51
9	rs1333049	Urate	1.396	2.0374	-2.597	5.39	0.49
9	rs1333049	CCR	1.367	1.2639	-1.11	3.845	0.28
10	rs501120	mean 24 SBP	0.118	0.5607	-0.981	1.216	0.83
10	rs501120	mean 24 DBP	0.287	0.3417	-0.382	0.957	0.4
10	rs501120	mean 24 PP	-0.176	0.3773	-0.915	0.563	0.64
10	rs501120	BMI	-0.029	0.2231	-0.467	0.408	0.9
10	rs501120	WHR	-0.005	0.0033	-0.011	0.002	0.17
10	rs501120	Total cholesterol	-0.012	0.043	-0.096	0.072	0.78
10	rs501120	HDL cholesterol	0.017	0.0184	-0.02	0.053	0.37
10	rs501120	Blood glucose	-0.057	0.0498	-0.155	0.041	0.25
10	rs501120	Urate	-0.451	3.4415	-7.196	6.295	0.9
10	rs501120	CCR	1.034	1.6982	-2.294	4.363	0.54
15	rs17228212	mean 24 SBP	-0.191	0.3669	-0.91	0.528	0.6
15	rs17228212	mean 24 DBP	-0.302	0.2535	-0.798	0.195	0.23
15	rs17228212	mean 24 PP	0.091	0.2461	-0.391	0.574	0.71
15	rs17228212	BMI	-0.023	0.1635	-0.343	0.298	0.89
15	rs17228212	WHR	0.003	0.0024	-0.002	0.007	0.3
15	rs17228212	Total cholesterol	-0.016	0.0343	-0.083	0.052	0.65
15	rs17228212	HDL cholesterol	-0.019	0.0119	-0.042	0.004	0.11
15	rs17228212	Blood glucose	0.055	0.0484	-0.04	0.15	0.26
15	rs17228212	Urate	2.622	2.176	-1.643	6.887	0.23
15	rs17228212	CCR	0.6	1.4444	-2.231	3.431	0.68

Legend. The results are presented by locus and for mean 24-hr systolic blood pressure (SBP), 24-hr diastolic blood pressure (DBP), 24-hr pulse pressure (PP), body mass index (BMI), waist-hip ratio (WHR), TC, HDL-C, urate and creatinine clearance (CCR). The coefficients are described with respect to the effect of the coronary artery disease (CAD) risk allele on the trait after adjustment for age, gender and familial correlations in the GRAPHIC Study (a minus sign indicates that the CAD risk allele is associated with a lower value for the trait). **N.B. Table 2.6** describes the CAD risk alleles for each SNP.

Table S2.2. Cardiovascular trait associations from six GWA study meta-analysis consortia.

SNP	EA	OA	*P*-value	Beta	SE	Trait
rs599839	A	G	0.6	0.0036	0.007	BMI
rs599839	G	A	$2.89\text{x}10^{-10}$	-0.1067	0.0169	CAD CARDIoGRAM
rs599839	G	A	0.061	0.0392	0.02	Diabetes
rs599839	G	A	0.6593	0.0327	0.0741	Diastolic blood pressure (DBP)
rs599839	A	G	0.1018	-0.0063	0.0038	Glucose
rs599839	A	G	$5.56\text{x}10^{-07}$	-0.0303	0.006	HDL
rs599839	A	G	0.3672	-0.0029	0.0032	Insulin
rs599839	A	G	$2.94\text{x}10^{-168}$	0.1718	0.0063	LDL
rs599839	G	A	0.2047	0.0899	0.0709	Mean arterial pressure (MAP)
rs599839	G	A	0.7873	0.0192	0.0711	Pulse Pressure (PP)
rs599839	G	A	0.6609	0.0053	0.0121	Smoking
rs599839	G	A	0.8912	0.016	0.1171	Systolic Blood Pressure (SBP)
rs599839	A	G	$4.12\text{x}10^{-130}$	0.1483	0.0062	TC
rs599839	A	G	0.0443	0.0113	0.0058	TG
rs2943634	A	C	0.89	0.0009	0.006	BMI
rs2943634	C	A	$9.63\text{x}10^{-05}$	0.0566	0.0145	CAD CARDIoGRAM
rs2943634	C	A	$1.10\text{x}10^{-05}$	0.077	0.0193	Diabetes
rs2943634	C	A	0.0482	0.1279	0.0648	Diastolic blood pressure (DBP)
rs2943634	A	C	0.1373	-0.005	0.0034	Glucose
rs2943634	A	C	$2.33\text{x}10^{-09}$	0.0317	0.0051	HDL
rs2943634	A	C	$1.49\text{x}10^{-13}$	-0.021	0.0028	Insulin
rs2943634	A	C	0.1358	-0.0085	0.0055	LDL
rs2943634	C	A	0.0133	0.1537	0.0621	Mean arterial pressure (MAP)
rs2943634	C	A	0.0904	0.1055	0.0623	Pulse Pressure (PP)
rs2943634	C	A	0.9125	0.0012	0.0106	Smoking
rs2943634	C	A	0.0562	0.1949	0.1021	Systolic Blood Pressure (SBP)
rs2943634	A	C	0.9071	-0.0003	0.0054	TC
rs2943634	A	C	$5.17\text{x}10^{-08}$	-0.0271	0.0051	TG
rs6922269	G	A	0.76	0.002	0.0064	BMI
rs6922269	G	A	$7.38\text{x}10^{-05}$	-0.061	0.0154	CAD CARDIoGRAM
rs6922269	A	G	0.084	0.0296	0.0151	Diabetes
rs6922269	G	A	0.2993	0.0723	0.0697	Diastolic blood pressure (DBP)
rs6922269	A	G	0.1958	-0.0046	0.0036	Glucose
rs6922269	A	G	0.8526	0.0002	0.0055	HDL
rs6922269	A	G	0.1696	-0.0042	0.0031	Insulin
rs6922269	A	G	0.734	-0.0028	0.0058	LDL
rs6922269	G	A	0.4318	0.0524	0.0666	Mean arterial pressure (MAP)
rs6922269	G	A	0.907	0.0078	0.0667	Pulse Pressure (PP)
rs6922269	G	A	0.7394	-0.0038	0.0113	Smoking
rs6922269	G	A	0.5871	0.0599	0.1103	Systolic Blood Pressure (SBP)
rs6922269	A	G	0.3572	-0.0061	0.0058	TC
rs6922269	A	G	0.8935	0.0004	0.0054	TG
rs1333049	G	C	0.75	0.0021	0.0066	BMI
rs1333049	C	G	$2.06\text{x}10^{-20}$	0.2552	0.0276	CAD CARDIoGRAM

rs1333049	C	G	0.021	0.0392	0.0149	Diabetes
rs1333049	G	C	0.1751	0.0856	0.0631	Diastolic blood pressure (DBP)
rs1333049	C	G	0.587	0.0018	0.0033	Glucose
rs1333049	C	G	0.2622	-0.0049	0.0052	HDL
rs1333049	C	G	0.2246	-0.0033	0.0028	Insulin
rs1333049	C	G	0.1564	-0.0095	0.0055	LDL
rs1333049	G	C	0.3636	0.0555	0.0611	Mean arterial pressure (MAP)
rs1333049	G	C	0.2214	0.0744	0.0609	Pulse Pressure (PP)
rs1333049	G	C	0.4475	0.0078	0.0103	Smoking
rs1333049	G	C	0.9928	0.0009	0.0993	Systolic Blood Pressure (SBP)
rs1333049	C	G	0.0081	-0.0154	0.0054	TC
rs1333049	C	G	0.711	-0.0036	0.0051	TG
rs501120	T	C	0.54	0.0052	0.0084	BMI
rs501120	C	T	2.29x10^{-05}	-0.0882	0.0208	CAD CARDIoGRAM
rs501120	C	T	0.74	0.01	0.0259	Diabetes
rs501120	T	C	0.1395	0.1309	0.0886	Diastolic blood pressure (DBP)
rs501120	T	C	0.173	-0.0063	0.0046	Glucose
rs501120	T	C	0.6731	-0.0055	0.0071	HDL
rs501120	T	C	0.89	-0.0005	0.0038	Insulin
rs501120	T	C	0.4793	-0.0051	0.0075	LDL
rs501120	T	C	0.0947	0.1416	0.0848	Mean arterial pressure (MAP)
rs501120	T	C	0.4611	0.0626	0.0849	Pulse Pressure (PP)
rs501120	T	C	0.9838	0.0003	0.0145	Smoking
rs501120	T	C	0.538	0.0858	0.1393	Systolic Blood Pressure (SBP)
rs501120	T	C	0.1754	-0.0105	0.0074	TC
rs501120	T	C	0.2388	-0.0057	0.007	TG
rs17228212	C	T	0.24	0.0075	0.0063	BMI
rs17228212	C	T	0.3825	0.0139	0.0159	CAD CARDIoGRAM
rs17228212	C	T	0.14	0.0296	0.0202	Diabetes
rs17228212	T	C	0.5974	0.0363	0.0686	Diastolic blood pressure (DBP)
rs17228212	T	C	0.7421	-0.0012	0.0036	Glucose
rs17228212	T	C	0.7498	0.0019	0.0055	HDL
rs17228212	T	C	0.0052	-0.0085	0.003	Insulin
rs17228212	T	C	0.0317	0.0129	0.0058	LDL
rs17228212	T	C	0.7541	0.0208	0.0663	Mean arterial pressure (MAP)
rs17228212	T	C	0.4286	0.0522	0.066	Pulse Pressure (PP)
rs17228212	T	C	0.1033	-0.0183	0.0112	Smoking
rs17228212	T	C	0.7383	0.0364	0.1089	Systolic Blood Pressure (SBP)
rs17228212	T	C	0.0876	0.0101	0.0058	TC
rs17228212	T	C	0.1252	-0.0064	0.0054	TG

Legend. Seven WTCCC GWA study novel loci assessed for subsequent association with cardiovascular traits in six large scale GWA study meta-analyses consortia: Type 2 diabetes (T2D) – DIAGRAM (DIAbetes Genetics Replication And Meta-analysis) (Zeggini et al. 2008), CAD – CARDIoGRAM (Coronary ARtery DIsease Genome wide Replication and Meta-analysis) (Schunkert et al. 2011), Fasting glucose and insulin – MAGIC

(Meta-Analyses of Glucose and Insulin-related traits Consortium) (Manning et al. 2012), Smoking – TAG (The Tobacco and Smoking Consortium) (Tobacco and Genetics Consortium 2010), Lipids – GLGC (Global Lipids Genetics Consortium) (Global Lipids Genetics Consortium et al. 2013), and BP – ICBP (International Consortium for Blood Pressure) (International Consortium for Blood Pressure Genome-Wide Association Studies et al. 2011). EA=Effect Allele; OA=Other Allele; Beta=β-coefficient; SE=standard error.

6.2 Supplementary Figures for Chapter 3

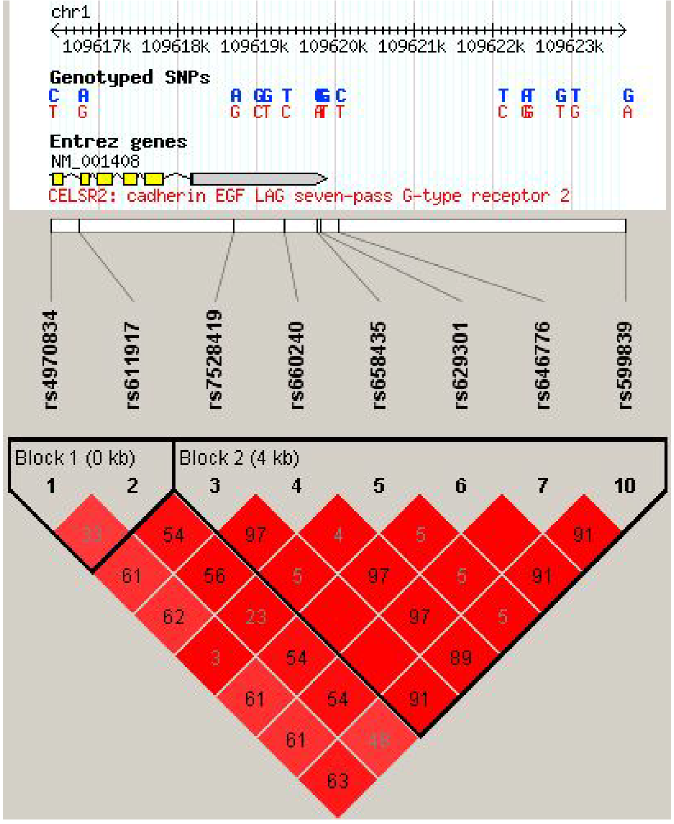


Figure S3.1. Haplotype Blocks for North & West European Ancestry, Utah.

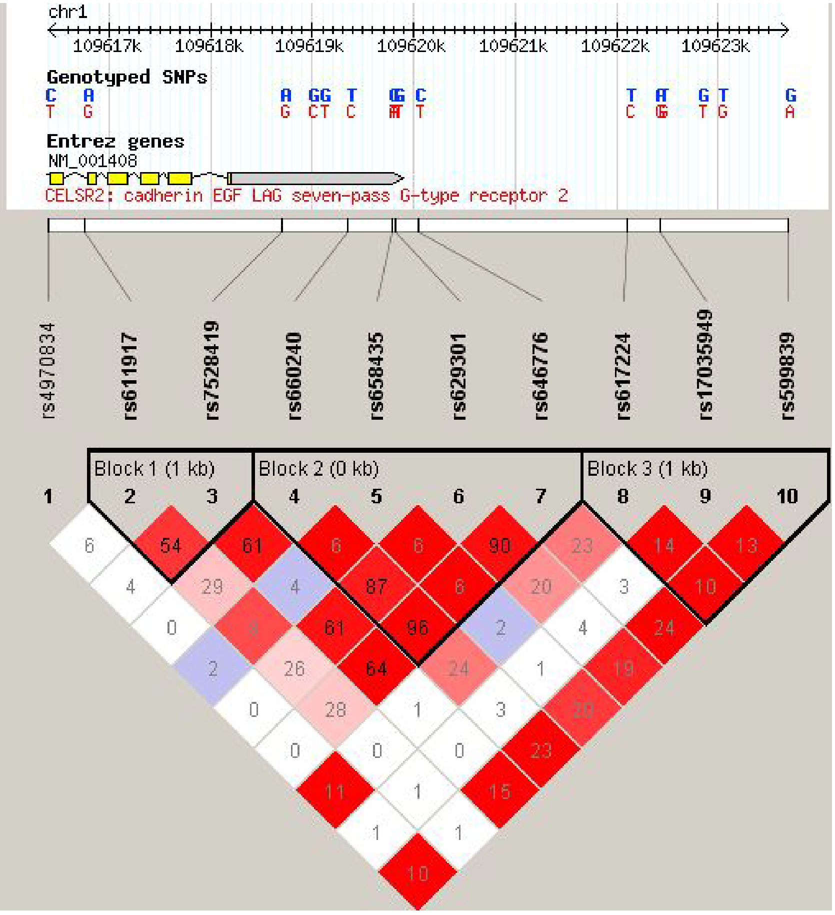


Figure S3.2. Haplotype Blocks for African ancestry in Southwest USA.

6.3 Supplementary Tables and Figures for Chapter 4

Table S4.1. Conserved major 4-SNP haplotype on 13q34 shows 48 PWM.

```
Human to Mouse
1 V$NERF_Q2 - 5-22 gcgactttCTGCCTGgtc - 784-801 gtcgccttCTGCCTGgtc
2 V$ER_Q6 - 26-44 gtgGGTCAcgcgcgcatgg - 804-822 ctaGGTGAaggggaccttg
3 V$GCM_Q2 + 78-89 cgggCCCGCACg + 850-861 ggggCCCGCATg
4 V$E47_01 + 115-129 catGCAGGTGgctgc + 895-909 tgtGCAGGTGgctgc
5 V$AP4_Q6_01 - 118-126 gCAGGTGgc - 898-906 gCAGGTGgc
6 V$USF_C - 118-125 gCAGGTGg - 898-905 gCAGGTGg
7 V$WHN_B - 163-173 agaGCGTcttg - 943-953 agaGCGTcttg
8 V$E2F1_Q6_01 + 202-211 cTTGCGCGcc + 989-998 gTTCCGCGcc
9 V$E2F_Q4_01 - 202-212 cTTGCGCGccc - 989-999 gTTCCGCGccc
10 V$E2F_Q6_01 - 202-213 ctTGCGCGcccg - 989-1000 gtTCCGCGcccg
11 V$E2F1DP1RB_01 + 203-210 TTGCGCGC + 990-997 TTCCGCGC
12 V$E2F1_Q4_01 + 203-211 TTGCGCGcc + 990-998 TTCCGCGcc
13 V$E2F4DP1_01 + 203-210 TTGCGCGC + 990-997 TTCCGCGC
14 V$E2F_02 + 203-210 TTGCGCGC + 990-997 TTCCGCGC
15 V$E2F_03 + 203-214 TTGCGCGCccga + 990-1001 TTCCGCGCccgt
16 V$E2F_Q3 + 203-210 TTGCGCGC + 990-997 TTCCGCGC
17 V$E2F_Q3_01 + 203-211 TTGCGCGcc + 990-998 TTCCGCGcc
18 V$E2F_Q4 + 203-210 TTGCGCGc + 990-997 TTCCGCGc
19 V$E2F_Q6 + 203-210 TTGCGCGc + 990-997 TTCCGCGc
20 V$ZF5_B - 203-215 ttGCGCGcccgag - 990-1002 ttCCGCGcccgtg
21 V$E2F_Q2 - 206-211 cgCGCC - 993-998 cgCGCC
22 V$NRF1_Q6 - 206-215 cGCGCCCGAG - 993-1002 cGCGCCCGTG
23 V$ZF5_01 - 258-265 ccGTGCGC - 1045-1052 tcGAGCCC
24 V$EGR2_01 + 274-285 gTGTGTGTGCGt + 1061-1072 gTGCGTGTGCAt
25 V$EGR3_01 + 274-285 gtGTGTGTGCGt + 1061-1072 gtGCGTGTGCAt
26 V$AP2ALPHA_02 + 434-448 gtaGCCTGAGGcacc + 1158-1172 gtaGCCTCAGGcacc
27 V$AP2ALPHA_03 + 434-448 gtaGCCTGAGGCacc + 1158-1172 gtaGCCTCAGGCacc
28 V$AP2ALPHA_02 - 434-448 gtagCCTGAGGCacc - 1158-1172 gtagCCTCAGGCacc
29 V$AP2ALPHA_03 - 434-448 gtaGCCTGAGGCacc - 1158-1172 gtaGCCTCAGGCacc
30 V$AP2ALPHA_01 + 437-445 GCCTGAGGc + 1161-1169 GCCTCAGGc
31 V$AP2GAMMA_01 + 437-445 GCCTGAGGc + 1161-1169 GCCTCAGGc
32 V$AP2ALPHA_01 - 437-445 gCCTGAGGC - 1161-1169 gCCTCAGGC
33 V$AP2GAMMA_01 - 437-445 gCCTGAGGC - 1161-1169 gCCTCAGGC
34 V$STAT5B_01 - 541-555 gttTTCCTGGAActc - 1263-1277 agtTTCCAGAAActc
35 V$GC_01 + 625-638 aagGGGCGGGGcga + 1340-1353 gagGGGCGGGGcaa
36 V$SP1_Q4_01 + 625-637 aagGGGCGGggcg + 1340-1352 gagGGGCGGggca
37 V$SP1_Q6 + 625-637 aagGGGCGGGgcg + 1340-1352 gagGGGCGGGgca
38 V$MAZR_01 + 626-638 aggGGCGGGGcga + 1341-1353 aggGGCGGGGcaa
39 V$SP1_Q2_01 - 626-635 agGGGCGGGg - 1341-1350 agGGGCGGGg
40 V$ETF_Q6 + 627-633 GGGGCGG + 1342-1348 GGGGCGG
41 V$MAZ_Q6 + 627-634 gGGGCGGG + 1342-1349 gGGGCGGG
42 V$SP1_01 + 627-636 ggGGCGGGgc + 1342-1351 ggGGCGGGgc
43 V$SP1_Q6_01 + 627-636 gGGGCGGGgc + 1342-1351 gGGGCGGGgc
44 V$ZF5_01 + 628-635 GGGCGGgg + 1343-1350 GGGCGGgg
45 V$CACD_01 - 628-635 GGGCGGGG - 1343-1350 GGGCGGGG
46 V$E2F_Q2 + 629-634 GGCGgg + 1344-1349 GGCGgg
47 V$VDR_Q3 + 632-646 GGGGCGAGGAGGcga + 1347-1361 GGGGCAAGGAGGaga
48 V$GABP_B + 735-746 cCTGGAAGAGct + 1433-1444 cCCTGAAGAGcc
```

Legend. Conserved MultiTF TRANSFAC Prof v10.2 data output from MULAN alignment of Human and Mouse orthologous sequence. Relevant silencer and candidate SNP positions: Silencer = 192-216 (Haniel et al. 1995); rs4773143 = 391; rs4773144 = 418; rs7986871 = 495; rs3809346 = 649.

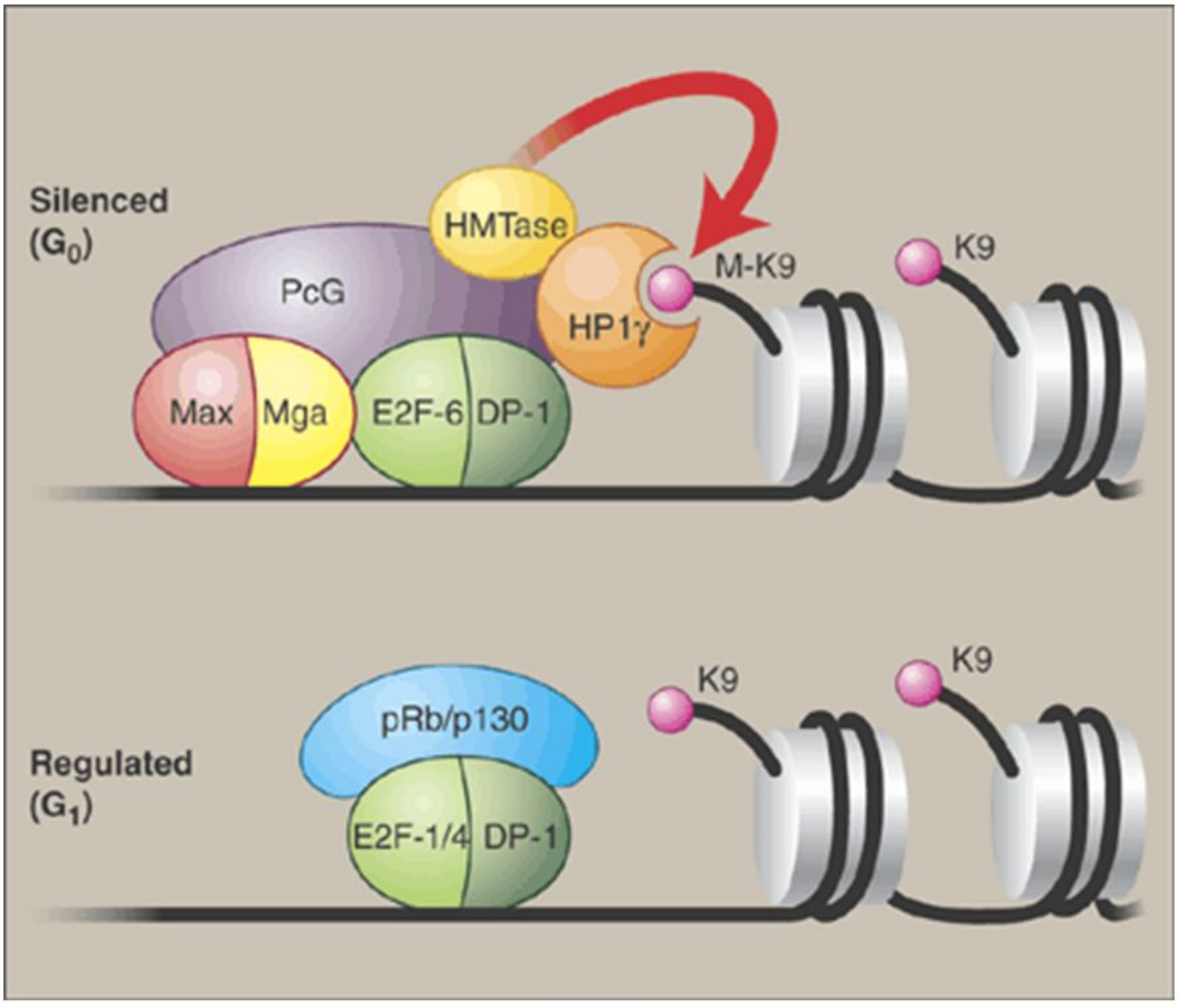


Figure S4.1. The dark horse of the E2F family.

Transcriptional silencing by E2F-6. In quiescent cells in G_0 of the cell cycle, E2F-6 nucleates the assembly of a chromatin-modifying complex. This complex contains HMTase, PcG proteins, and HP1γ, together with other sequence-specific transcription factors like Max and its partner Mga. By modifying the chromatin environment of target genes, this complex coordinates the long-term silencing of cell cycle-regulated genes. As the cell cycle ensues, the promoter binding sites occupied by E2F-6 become occupied by other E2F family proteins, such as E2F-1 and E2F-4, and by members of the pocket protein family, such as pRb and p130. This allows the regulated transcription of target genes that are involved in cell cycle progression. (K9, lysine 9; M-K9, methylated lysine 9) (Ogawa et al. 2002; La Thangue 2002).

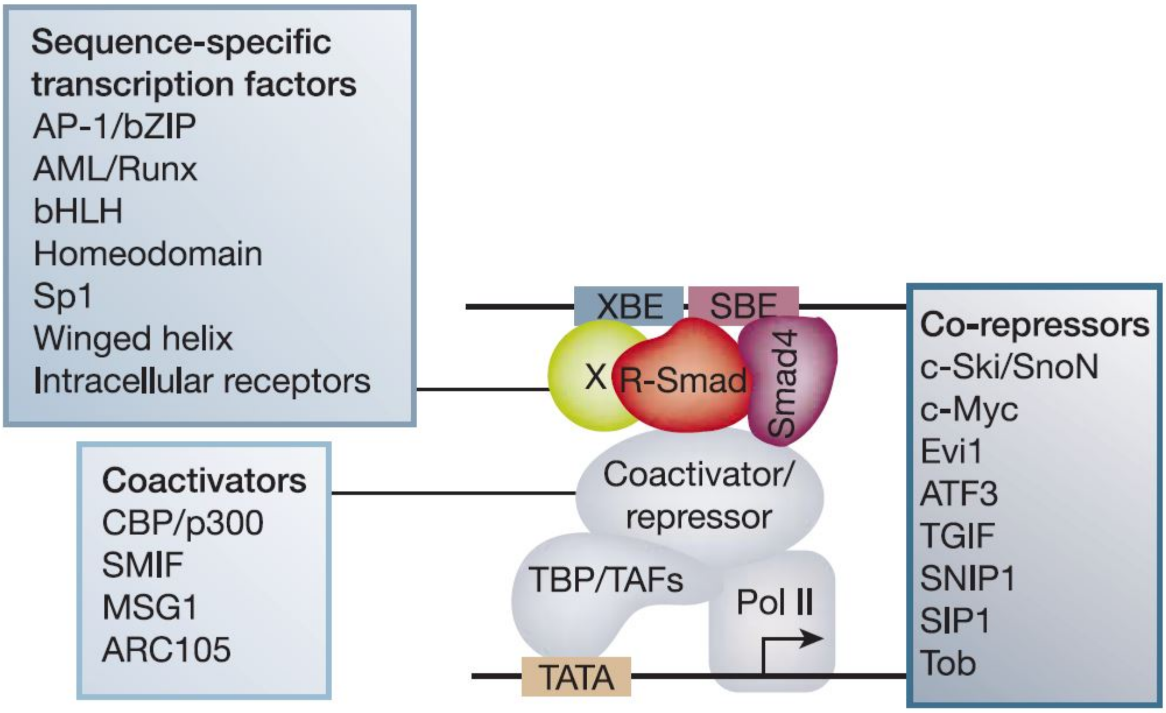


Figure S4.2. Figure showing the SP1 coactivation and EVI-1 corepression of SMAD3-dependent TGFβ1-induced transcription of cell cycle genes (Derynck & Zhang 2003).

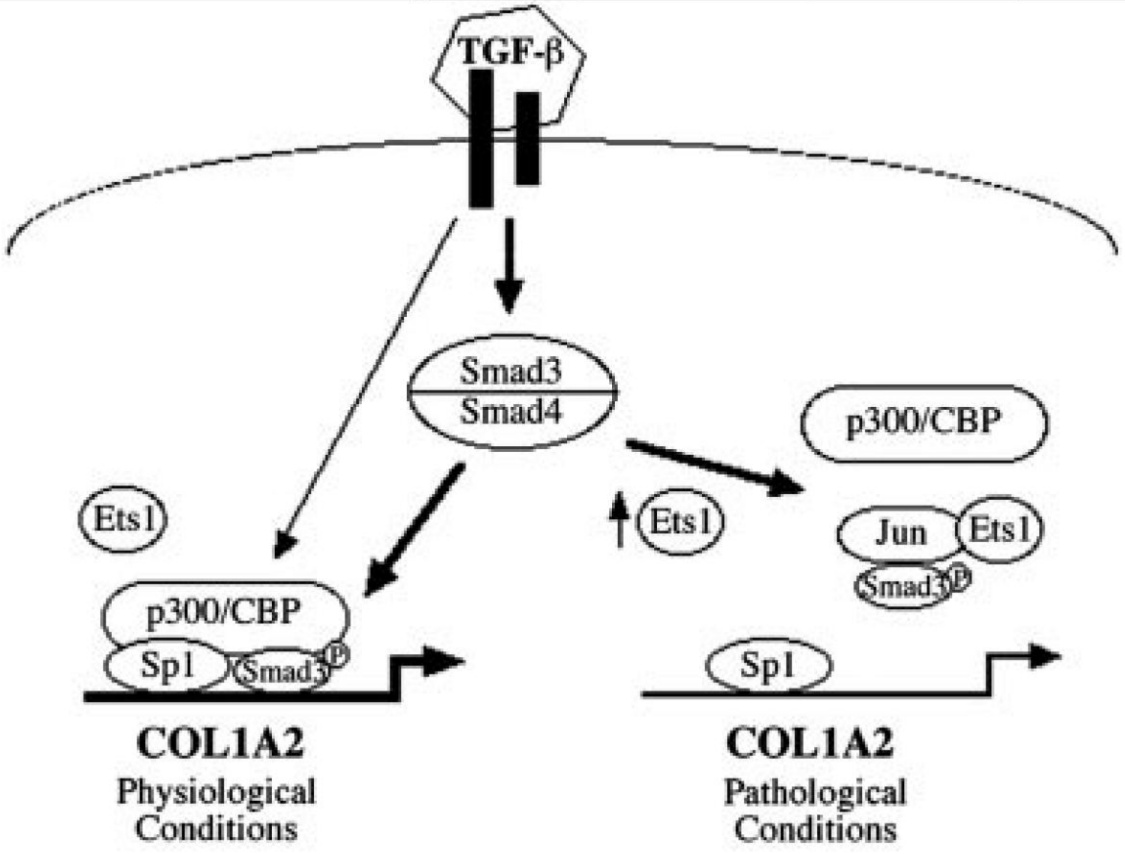


Figure S4.3. A hypothetical model for the role of Ets1 in TGFβ1-dependent response of the *COL1A2* gene under physiological and pathological conditions (Czuwara-Ladykowska et al. 2002).

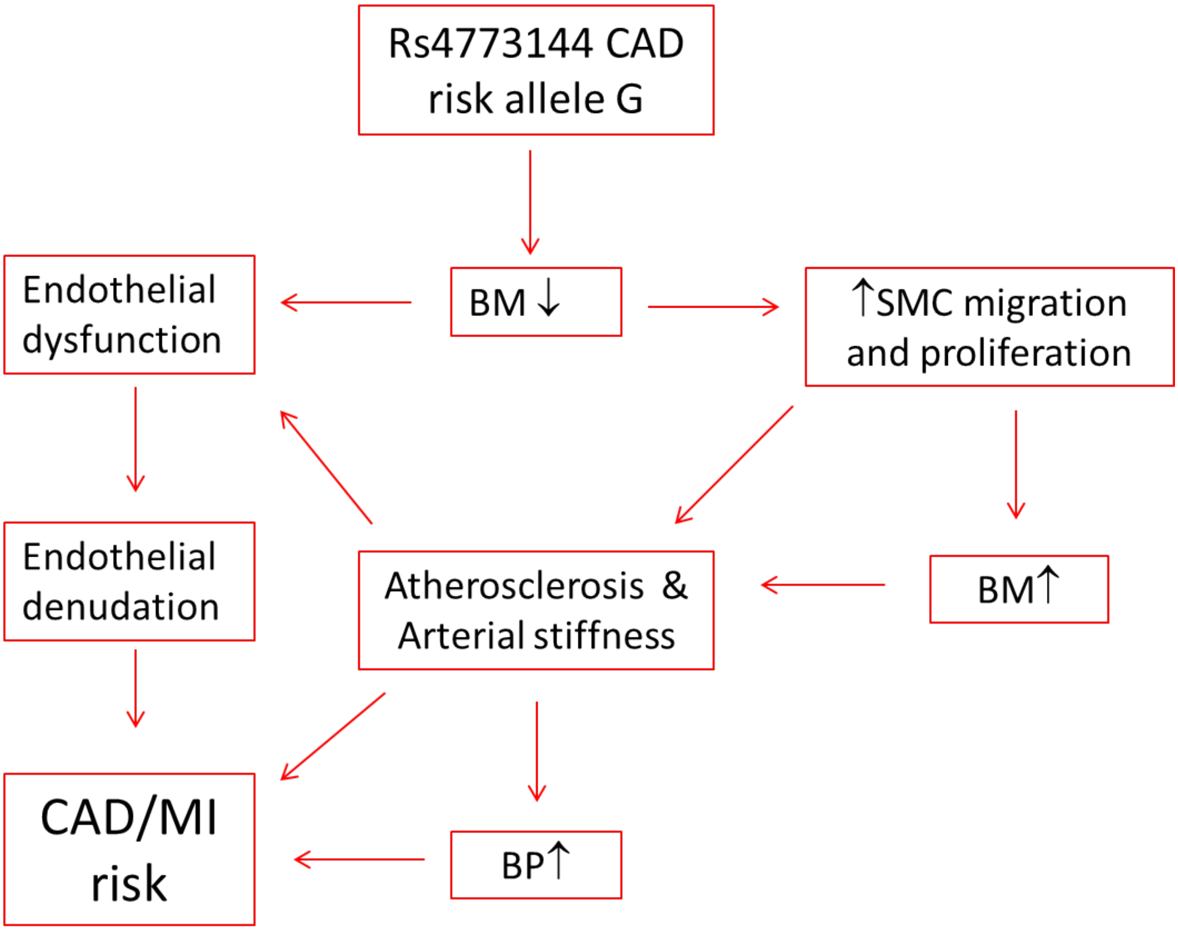

Rs4773144 CAD risk allele G
BM ↓
Endothelial dysfunction
↑SMC migration and proliferation
Endothelial denudation
Atherosclerosis & Arterial stiffness
BM↑
CAD/MI risk
BP↑

Figure S4.4. A possible mechanism for genetic variant rs4773144 CAD risk.

Bibliography

1000 Genomes Project Consortium, Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., et al. 2012. 'An integrated map of genetic variation from 1,092 human genomes'. *Nature.* 491 (7422): 56-65.

Abifadel, M., Varret, M., Rabes, J.P., Allard, D., Ouguerram, K., Devillers, M., et al. 2003. 'Mutations in PCSK9 cause autosomal dominant hypercholesterolemia'. *Nature Genetics.* 34 (2): 154-156.

Aguilera, C.M., George, S.J., Johnson, J.L. & Newby, A.C. 2003. 'Relationship between type IV collagen degradation, metalloproteinase activity and smooth muscle cell migration and proliferation in cultured human saphenous vein'. *Cardiovascular Research.* 58 (3): 679-688.

Alfakih, K., Brown, B., Lawrance, R.A., Warburton, P., Maqbool, A., Walters, K., et al. 2007. 'Effect of a common X-linked angiotensin II type 2-receptor gene polymorphism (-1332 G/A) on the occurrence of premature myocardial infarction and stenotic atherosclerosis requiring revascularization'. *Atherosclerosis.* 195 (1): e32-8.

Alliston, T., Ko, T.C., Cao, Y., Liang, Y.Y., Feng, X.H., Chang, C., et al. 2005. 'Repression of bone morphogenetic protein and activin-inducible transcription by Evi-1'. *The Journal of Biological Chemistry.* 280 (25): 24227-24237.

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

Baer, P.C., Nockher, W.A., Haase, W. & Scherberich, J.E. 1997. 'Isolation of proximal and distal tubule cells from human kidney by immunomagnetic separation. Technical note'. *Kidney International.* 52 (5): 1321-1331.

Barbagallo, C.M., Emmanuele, G., Cefalu, A.B., Fiore, B., Noto, D., Mazzarino, M.C., et al. 2003. 'Autosomal recessive hypercholesterolemia in a Sicilian kindred harboring the 432insA mutation of the ARH gene'. *Atherosclerosis.* 166 (2): 395-400.

Berneis, K.K. & Krauss, R.M. 2002. 'Metabolic origins and clinical significance of LDL heterogeneity'. *Journal of Lipid Research.* 43 (9): 1363-1379.

Bernstein, B.E., Kamal, M., Lindblad-Toh, K., Bekiranov, S., Bailey, D.K., Huebert, D.J., et al. 2005. 'Genomic maps and comparative analysis of histone modifications in human and mouse'. *Cell.* 120 (2): 169-181.

Betel, D., Koppal, A., Agius, P., Sander, C. & Leslie, C. 2010. 'Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites'. *Genome Biology.* 11 (8): R90.

Betel, D., Wilson, M., Gabow, A., Marks, D.S. & Sander, C. 2008. 'The microRNA.org resource: targets and expression'. *Nucleic Acids Research.* 36 (Database issue): D149-53.

Bhatia, S.K. 2010. 'Chapter 2 - coronary artery disease' in *Biomaterials for Clinical Applications.* (1st edn) Springer, 23-30.

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

Bodzioch, M., Orso, E., Klucken, J., Langmann, T., Bottcher, A., Diederich, W., et al. 1999. 'The gene encoding ATP-binding cassette transporter 1 is mutated in Tangier disease'. *Nature Genetics.* 22 (4): 347-351.

Bosserhoff, A.K. & Buettner, R. 2002. 'Expression, function and clinical relevance of MIA (melanoma inhibitory activity)'. *Histology and Histopathology.* 17 (1): 289-300.

Bourbon, M., Duarte, M.A., Alves, A.C., Medeiros, A.M., Marques, L. & Soutar, A.K. 2009. 'Genetic diagnosis of familial hypercholesterolaemia: the importance of functional analysis of potential splice-site mutations'. *Journal of Medical Genetics.* 46 (5): 352-357.

Boyle, A.P., Hong, E.L., Hariharan, M., Cheng, Y., Schaub, M.A., Kasowski, M., et al. 2012. 'Annotation of functional variation in personal genomes using RegulomeDB'. *Genome Research.* 22 (9): 1790-1797.

Brack, M.J., Ray, S., Chauhan, A., Fox, J., Hubner, P.J., Schofield, P., et al. 1995. 'The Subcutaneous Heparin and Angioplasty Restenosis Prevention (SHARP) trial. Results of a multicenter randomized trial investigating the effects of high dose unfractionated heparin on angiographic restenosis and clinical outcome'. *Journal of the American College of Cardiology.* 26 (4): 947-954.

Brasier, A.R., Lu, M., Hai, T., Lu, Y. & Boldogh, I. 2001. 'NF-kappa B-inducible BCL-3 expression is an autoregulatory loop controlling nuclear p50/NF-kappa B1 residence'. *The Journal of Biological Chemistry.* 276 (34): 32080-32093.

Brautbar, A. & Ballantyne, C.M. 2011. 'Pharmacological strategies for lowering LDL cholesterol: statins and beyond'. *Nature Reviews.Cardiology.* 8 (5): 253-265.

British Heart Foundation (BHF) 2010. *. Death by cause, sex and country, 2009, United Kingdom. Source: England and Wales, Office for National Statistics (2010); Scotland, General Register Office (2010); Northern Ireland, Statistics and Research Agency (2010). Personal communication.* [Homepage of British Heart Foundation (BHF)], [Online]. Available: http://www.bhf.org.uk/research/heart-statistics/mortality/numbers-dying.aspx.

Brocke-Heidrich, K., Ge, B., Cvijic, H., Pfeifer, G., Loffler, D., Henze, C., et al. 2006. 'BCL3 is induced by IL-6 via Stat3 binding to intronic enhancer HS4 and represses its own transcription'. *Oncogene.* 25 (55): 7297-7304.

Broeckel, U., Hengstenberg, C., Mayer, B., Holmer, S., Martin, L.J., Comuzzie, A.G., et al. 2002. 'A comprehensive linkage analysis for myocardial infarction and its related risk factors'. *Nature Genetics.* 30 (2): 210-214.

Broms, U., Silventoinen, K., Madden, P.A., Heath, A.C. & Kaprio, J. 2006. 'Genetic architecture of smoking behavior: a study of Finnish adult twins'. *Twin Research and Human Genetics : The Official Journal of the International Society for Twin Studies.* 9 (1): 64-72.

Brooks-Wilson, A., Marcil, M., Clee, S.M., Zhang, L.H., Roomp, K., van Dam, M., et al. 1999. 'Mutations in ABC1 in Tangier disease and familial high-density lipoprotein deficiency'. *Nature Genetics.* 22 (4): 336-345.

Brouilette, S., Singh, R.K., Thompson, J.R., Goodall, A.H. & Samani, N.J. 2003. 'White cell telomere length and risk of premature myocardial infarction'. *Arteriosclerosis, Thrombosis, and Vascular Biology.* 23 (5): 842-846.

Burton, P., Gurrin, L. & Sly, P. 1998. 'Extending the simple linear regression model to account for correlated responses: an introduction to generalized estimating equations and multi-level mixed modelling'. *Statistics in Medicine.* 17 (11): 1261-1291.

Businaro, R., Tagliani, A., Buttari, B., Profumo, E., Ippoliti, F., Di Cristofano, C., et al. 2012. 'Cellular and molecular players in the atherosclerotic plaque progression'. *Annals of the New York Academy of Sciences.* 1262 (1): 134-141.

Campos, H., Genest, J.J.,Jr, Blijlevens, E., McNamara, J.R., Jenner, J.L., Ordovas, J.M., et al. 1992. 'Low density lipoprotein particle size and coronary artery disease'. *Arteriosclerosis and Thrombosis : A Journal of Vascular Biology / American Heart Association.* 12 (2): 187-195.

CARDIoGRAMplusC4D Consortium, Deloukas, P., Kanoni, S., Willenborg, C., Farrall, M., Assimes, T.L., et al. 2013. 'Large-scale association analysis identifies new risk loci for coronary artery disease'. *Nature Genetics.* 45 (1): 25-33.

Cartharius, K., Frech, K., Grote, K., Klocke, B., Haltmeier, M., Klingenhoff, A., et al. 2005. 'MatInspector and beyond: promoter analysis based on transcription factor binding sites'. *Bioinformatics (Oxford, England).* 21 (13): 2933-2942.

Chasman, D.I., Pare, G., Mora, S., Hopewell, J.C., Peloso, G., Clarke, R., et al. 2009. 'Forty-three loci associated with plasma lipoprotein size, concentration, and cholesterol content in genome-wide analysis'. *PLoS Genetics.* 5 (11): e1000730.

Chen, Y., Rollins, J., Paigen, B. & Wang, X. 2007. 'Genetic and genomic insights into the molecular basis of atherosclerosis'. *Cell Metabolism.* 6 (3): 164-179.

Cohen, J., Pertsemlidis, A., Kotowski, I.K., Graham, R., Garcia, C.K. & Hobbs, H.H. 2005. 'Low LDL cholesterol in individuals of African descent resulting from frequent nonsense mutations in PCSK9'. *Nature Genetics.* 37 (2): 161-165.

Cohen, J.C., Boerwinkle, E., Mosley, T.H.,Jr & Hobbs, H.H. 2006. 'Sequence variations in PCSK9, low LDL, and protection against coronary heart disease'. *The New England Journal of Medicine.* 354 (12): 1264-1272.

Collins, F.S. 1999. 'The human genome project and the future of medicine'. *Annals of the New York Academy of Sciences.* 882: 42-55; discussion 56-65.

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

Coronary Artery Disease Consortium, Samani, N.J., Deloukas, P., Erdmann, J., Hengstenberg, C., Kuulasmaa, K., et al. 2009. 'Large scale association analysis of novel genetic loci for coronary artery disease'. *Arteriosclerosis, Thrombosis, and Vascular Biology.* 29 (5): 774-780.

Crawford, G.E., Davis, S., Scacheri, P.C., Renaud, G., Halawi, M.J., Erdos, M.R., et al. 2006. 'DNase-chip: a high-resolution method to identify DNase I hypersensitive sites using tiled microarrays'. *Nature Methods.* 3 (7): 503-509.

Czuwara-Ladykowska, J., Sementchenko, V.I., Watson, D.K. & Trojanowska, M. 2002. 'Ets1 is an effector of the transforming growth factor beta (TGF-beta) signaling pathway and an antagonist of the profibrotic effects of TGF-beta'. *The Journal of Biological Chemistry.* 277 (23): 20399-20408.

Daly, M.J., Rioux, J.D., Schaffner, S.F., Hudson, T.J. & Lander, E.S. 2001. 'High-resolution haplotype structure in the human genome'. *Nature Genetics.* 29 (2): 229-232.

Dawber, T.R. & Kannel, W.B. 1966. 'The Framingham Study An Epidemiological Approach to Coronary Heart Disease'. *Circulation.* 34 (4): 553-555.

Dawber, T.R., Moore, F.E. & Mann, G.V. 1957. ' II. Coronary Heart Disease in the Framingham Study '. *American Journal of Public Health.* 47 (4): 4-24.

de Vilhena e Santos, D.M., Katzmarzyk, P.T., Seabra, A.F. & Maia, J.A. 2012. 'Genetics of physical activity and physical inactivity in humans'. *Behavior Genetics.* 42 (4): 559-578.

Demer, L.L. 2002. 'Vascular calcification and osteoporosis: inflammatory responses to oxidized lipids'. *International Journal of Epidemiology.* 31 (4): 737-741.

Derynck, R. & Zhang, Y.E. 2003. 'Smad-dependent and Smad-independent pathways in TGF-beta family signalling'. *Nature.* 425 (6958): 577-584.

Djebali, S., Davis, C.A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., et al. 2012. 'Landscape of transcription in human cells'. *Nature.* 489 (7414): 101-108.

Donnelly, P. 2008. 'Progress and challenges in genome-wide association studies in humans'. *Nature.* 456 (7223): 728-731.

Drysdale, C.M., McGraw, D.W., Stack, C.B., Stephens, J.C., Judson, R.S., Nandabalan, K., et al. 2000. 'Complex promoter and coding region beta 2-adrenergic receptor haplotypes alter receptor expression and predict in vivo responsiveness'. *Proceedings of the National Academy of Sciences of the United States of America.* 97 (19): 10483-10488.

Eichler, E.E., Flint, J., Gibson, G., Kong, A., Leal, S.M., Moore, J.H., et al. 2010. 'Missing heritability and strategies for finding the underlying causes of complex disease'. *Nature Reviews.Genetics.* 11 (6): 446-450.

El Harchaoui, K., van der Steeg, W.A., Stroes, E.S., Kuivenhoven, J.A., Otvos, J.D., Wareham, N.J., et al. 2007. 'Value of low-density lipoprotein particle number and size as predictors of coronary artery disease in apparently healthy men and women: the EPIC-Norfolk Prospective Population Study'. *Journal of the American College of Cardiology.* 49 (5): 547-553.

ENCODE Project Consortium, Bernstein, B.E., Birney, E., Dunham, I., Green, E.D., Gunter, C., et al. 2012. 'An integrated encyclopedia of DNA elements in the human genome'. *Nature.* 489 (7414): 57-74.

Erdmann, J., Grosshennig, A., Braund, P.S., Konig, I.R., Hengstenberg, C., Hall, A.S., et al. 2009. 'New susceptibility locus for coronary artery disease on chromosome 3q22.3'. *Nature Genetics.* 41 (3): 280-282.

Esperon, P., Raggio, V., Stoll, M., Vital, M. & Alallon, W. 2008. 'A new APOA1 mutation with severe HDL-cholesterol deficiency and premature coronary artery disease'. *Clinica Chimica Acta; International Journal of Clinical Chemistry.* 388 (1-2): 222-224.

Euskirchen, G.M., Rozowsky, J.S., Wei, C.L., Lee, W.H., Zhang, Z.D., Hartman, S., et al. 2007. 'Mapping of transcription factor binding regions in mammalian cells by ChIP: comparison of array- and sequencing-based technologies'. *Genome Research.* 17 (6): 898-909.

Fagge, C.H. 1872. 'General xanthelasma or vitilogoidea.'. *Transactions of the Pathological Society of London.* 24: 242-250.

Farb, A., Tang, A.L., Burke, A.P., Sessums, L., Liang, Y. & Virmani, R. 1995. 'Sudden coronary death. Frequency of active coronary lesions, inactive coronary lesions, and myocardial infarction'. *Circulation.* 92 (7): 1701-1709.

Farrall, M., Green, F.R., Peden, J.F., Olsson, P.G., Clarke, R., Hellenius, M.L., et al. 2006. 'Genome-wide mapping of susceptibility to coronary artery disease identifies a novel replicated locus on chromosome 17'. *PLoS Genetics.* 2 (5): e72.

Ferguson, J.F., Matthews, G.J., Townsend, R.R., Raj, D.S., Kanetsky, P.A., Budoff, M., et al. 2013. 'Candidate Gene Association Study of Coronary Artery Calcification in Chronic Kidney Disease: Findings from the Chronic Renal Insufficiency Cohort Study'. *Journal of the American College of Cardiology.*

Fischer, G., Schmidt, C., Opitz, J., Cully, Z., Kuhn, K. & Poschl, E. 1993. 'Identification of a novel sequence element in the common promoter region of human collagen type IV genes, involved in the regulation of divergent transcription'. *The Biochemical Journal.* 292 (Pt 3) (Pt 3): 687-695.

Franchini, M., Peyvandi, F. & Mannucci, P.M. 2008. 'The genetic basis of coronary artery disease: from candidate genes to whole genome analysis'. *Trends in Cardiovascular Medicine.* 18 (5): 157-162.

Francke, S., Manraj, M., Lacquemant, C., Lecoeur, C., Lepretre, F., Passa, P., et al. 2001. 'A genome-wide scan for coronary heart disease suggests in Indo-Mauritians a susceptibility locus on chromosome 16p13 and replicates linkage with the metabolic syndrome on 3q27'. *Human Molecular Genetics.* 10 (24): 2751-2765.

Freeman, D.J., Samani, N.J., Wilson, V., McMahon, A.D., Braund, P.S., Cheng, S., et al. 2003. 'A polymorphism of the cholesteryl ester transfer protein gene predicts cardiovascular events in non-smokers in the West of Scotland Coronary Prevention Study'. *European Heart Journal.* 24 (20): 1833-1842.

Friedman, R.C., Farh, K.K., Burge, C.B. & Bartel, D.P. 2009. 'Most mammalian mRNAs are conserved targets of microRNAs'. *Genome Research.* 19 (1): 92-105.

Fruchart, J.C., Nierman, M.C., Stroes, E.S., Kastelein, J.J. & Duriez, P. 2004. 'New risk factors for atherosclerosis and patient risk assessment'. *Circulation.* 109 (23 Suppl 1): III15-9.

Gabriel, S.B., Schaffner, S.F., Nguyen, H., Moore, J.M., Roy, J., Blumenstiel, B., et al. 2002. 'The structure of haplotype blocks in the human genome'. *Science (New York, N.Y.).* 296 (5576): 2225-2229.

Garcia, C.K., Wilund, K., Arca, M., Zuliani, G., Fellin, R., Maioli, M., et al. 2001. 'Autosomal recessive hypercholesterolemia caused by mutations in a putative LDL receptor adaptor protein'. *Science (New York, N.Y.).* 292 (5520): 1394-1398.

Gardiner-Garden, M. & Frommer, M. 1987. 'CpG islands in vertebrate genomes'. *Journal of Molecular Biology.* 196 (2): 261-282.

Ge, B., Li, O., Wilder, P., Rizzino, A. & McKeithan, T.W. 2003. 'NF-kappa B regulates BCL3 transcription in T lymphocytes through an intronic enhancer'. *Journal of Immunology (Baltimore, Md.: 1950).* 171 (8): 4210-4218.

Giannoulatou, E., Yau, C., Colella, S., Ragoussis, J. & Holmes, C.C. 2008. 'GenoSNP: a variational Bayes within-sample SNP genotyping algorithm that does not require a reference population'. *Bioinformatics (Oxford, England).* 24 (19): 2209-2214.

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

Giresi, P.G., Kim, J., McDaniell, R.M., Iyer, V.R. & Lieb, J.D. 2007. 'FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin'. *Genome Research.* 17 (6): 877-885.

Global Lipids Genetics Consortium, Willer, C.J., Schmidt, E.M., Sengupta, S., Peloso, G.M., Gustafsson, S., et al. 2013. 'Discovery and refinement of loci associated with lipid levels'. *Nature Genetics.* 45 (11): 1274-1283.

Goldstein, J.L., Hobbs, H.H. & Brown, M.S. 2001. '
Familial hypercholesterolemia in' in *Volume2: The Metabolic &Molecular Bases of Inherited Disease.* , (eds) C.R. Scriver, A.L. Beaudet, W.S. Sly & D. Valle, (8th edn). New York: McGraw-Hill, 2863-2913.

Gonzalez, P., Garcia-Castro, M., Reguero, J.R., Batalla, A., Ordonez, A.G., Palop, R.L., et al. 2006. 'The Pro279Leu variant in the transcription factor MEF2A is associated with myocardial infarction'. *Journal of Medical Genetics.* 43 (2): 167-169.

Gotea, V. & Ovcharenko, I. 2008. 'DiRE: identifying distant regulatory elements of co-expressed genes'. *Nucleic Acids Research.* 36 (Web Server issue): W133-9.

Griffin, B.A., Caslake, M.J., Yip, B., Tait, G.W., Packard, C.J. & Shepherd, J. 1990. 'Rapid isolation of low density lipoprotein (LDL) subfractions from plasma by density gradient ultracentrifugation'. *Atherosclerosis.* 83 (1): 59-67.

Griffiths-Jones, S., Grocock, R.J., van Dongen, S., Bateman, A. & Enright, A.J. 2006. 'miRBase: microRNA sequences, targets and gene nomenclature'. *Nucleic Acids Research.* 34 (Database issue): D140-4.

Griffiths-Jones, S., Saini, H.K., van Dongen, S. & Enright, A.J. 2008. 'miRBase: tools for microRNA genomics'. *Nucleic Acids Research.* 36 (Database issue): D154-8.

Grimson, A., Farh, K.K., Johnston, W.K., Garrett-Engele, P., Lim, L.P. & Bartel, D.P. 2007. 'MicroRNA targeting specificity in mammals: determinants beyond seed pairing'. *Molecular Cell.* 27 (1): 91-105.

Hall, A.S., Jackson, B.M., Farrin, A.J., Efthymiou, M., Barth, J.H., Copeland, J., et al. 2009. 'A randomized, controlled trial of simvastatin versus rosuvastatin in patients with acute myocardial infarction: the Secondary Prevention of Acute Coronary Events--Reduction of Cholesterol to Key European Targets Trial'. *European Journal of Cardiovascular Prevention and Rehabilitation : Official Journal of the European Society of Cardiology, Working Groups on Epidemiology & Prevention and Cardiac Rehabilitation and Exercise Physiology.* 16 (6): 712-721.

Haniel, A., Welge-Lussen, U., Kuhn, K. & Poschl, E. 1995. 'Identification and characterization of a novel transcriptional silencer in the human collagen type IV gene COL4A2'. *The Journal of Biological Chemistry.* 270 (19): 11209-11215.

Hannon, G.J. & Beach, D. 1994. 'p15INK4B is a potential effector of TGF-beta-induced cell cycle arrest'. *Nature.* 371 (6494): 257-261.

Harrap, S.B., Zammit, K.S., Wong, Z.Y., Williams, F.M., Bahlo, M., Tonkin, A.M., et al. 2002. 'Genome-wide linkage analysis of the acute coronary syndrome suggests a locus on chromosome 2'. *Arteriosclerosis, Thrombosis, and Vascular Biology.* 22 (5): 874-878.

Hauser, E.R., Crossman, D.C., Granger, C.B., Haines, J.L., Jones, C.J., Mooser, V., et al. 2004. 'A genomewide scan for early-onset coronary artery disease in 438 families: the GENECARD Study'. *American Journal of Human Genetics.* 75 (3): 436-447.

Heikkila, P., Soininen, R. & Tryggvason, K. 1993. 'Directional regulatory activity of cis-acting elements in the bidirectional alpha 1(IV) and alpha 2(IV) collagen gene promoter'. *The Journal of Biological Chemistry.* 268 (33): 24677-24682.

Helgadottir, A., Gretarsdottir, S., St Clair, D., Manolescu, A., Cheung, J., Thorleifsson, G., et al. 2005. 'Association between the gene encoding 5-lipoxygenase-activating protein and stroke replicated in a Scottish population'. *American Journal of Human Genetics.* 76 (3): 505-509.

Helgadottir, A., Manolescu, A., Helgason, A., Thorleifsson, G., Thorsteinsdottir, U., Gudbjartsson, D.F., et al. 2006. 'A variant of the gene encoding leukotriene A4 hydrolase confers ethnicity-specific risk of myocardial infarction'. *Nature Genetics.* 38 (1): 68-74.

Helgadottir, A., Manolescu, A., Thorleifsson, G., Gretarsdottir, S., Jonsdottir, H., Thorsteinsdottir, U., et al. 2004. 'The gene encoding 5-lipoxygenase activating protein confers risk of myocardial infarction and stroke'. *Nature Genetics.* 36 (3): 233-239.

Helgadottir, A., Thorleifsson, G., Manolescu, A., Gretarsdottir, S., Blondal, T., Jonasdottir, A., et al. 2007. 'A common variant on chromosome 9p21 affects the risk of myocardial infarction'. *Science (New York, N.Y.).* 316 (5830): 1491-1493.

Hirose, M., Kosugi, H., Nakazato, K. & Hayashi, T. 1999. 'Restoration to a quiescent and contractile phenotype from a proliferative phenotype of myofibroblast-like human aortic smooth muscle cells by culture on type IV collagen gels'. *Journal of Biochemistry.* 125 (6): 991-1000.

Hokanson (ed) 2000. *In: Handbook of Lipoprotein Testing.* (2nd edn) AACC Press.

Holmer, S.R., Hengstenberg, C., Mayer, B., Doring, A., Lowel, H., Engel, S., et al. 2000. 'Lipoprotein lipase gene polymorphism, cholesterol subfractions and myocardial infarction in large samples of the general population'. *Cardiovascular Research.* 47 (4): 806-812.

Hou, J.C. & Pessin, J.E. 2007. 'Ins (endocytosis) and outs (exocytosis) of GLUT4 trafficking'. *Current Opinion in Cell Biology.* 19 (4): 466-473.

Hsieh, P.C., Chang, J.C., Sun, W.T., Hsieh, S.C., Wang, M.C. & Wang, F.F. 2007. 'p53 downstream target DDA3 is a novel microtubule-associated protein that interacts with end-binding protein EB3 and activates beta-catenin pathway'. *Oncogene.* 26 (34): 4928-4940.

Hsieh, S.C., Lo, P.K. & Wang, F.F. 2002. 'Mouse DDA3 gene is a direct transcriptional target of p53 and p73'. *Oncogene.* 21 (19): 3050-3057.

Iakoubova, O.A., Tong, C.H., Chokkalingam, A.P., Rowland, C.M., Kirchgessner, T.G., Louie, J.Z., et al. 2006. 'Asp92Asn polymorphism in the myeloid IgA Fc receptor is associated with myocardial infarction in two disparate populations: CARE and WOSCOPS'. *Arteriosclerosis, Thrombosis, and Vascular Biology.* 26 (12): 2763-2768.

Iakoubova, O.A., Tong, C.H., Rowland, C.M., Kirchgessner, T.G., Young, B.A., Arellano, A.R., et al. 2008. 'Association of the Trp719Arg polymorphism in kinesin-like protein 6 with myocardial infarction and coronary heart disease in 2 prospective trials: the CARE and WOSCOPS trials'. *Journal of the American College of Cardiology.* 51 (4): 435-443.

IBC 50K CAD Consortium 2011. 'Large-scale gene-centric analysis identifies novel variants for coronary artery disease'. *PLoS Genetics.* 7 (9): e1002260.

International Consortium for Blood Pressure Genome-Wide Association Studies, Ehret, G.B., Munroe, P.B., Rice, K.M., Bochud, M., Johnson, A.D., et al. 2011. 'Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk'. *Nature.* 478 (7367): 103-109.

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

International HapMap Consortium, Frazer, K.A., Ballinger, D.G., Cox, D.R., Hinds, D.A., Stuve, L.L., et al. 2007. 'A second generation human haplotype map of over 3.1 million SNPs'. *Nature.* 449 (7164): 851-861.

Jang, C.Y., Wong, J., Coppinger, J.A., Seki, A., Yates, J.R.,3rd & Fang, G. 2008. 'DDA3 recruits microtubule depolymerase Kif2a to spindle poles and controls spindle dynamics and mitotic chromosome movement'. *The Journal of Cell Biology.* 181 (2): 255-267.

Jansen, P., Giehl, K., Nyengaard, J.R., Teng, K., Lioubinski, O., Sjoegaard, S.S., et al. 2007. 'Roles for the pro-neurotrophin receptor sortilin in neuronal development, aging and brain injury'. *Nature Neuroscience.* 10 (11): 1449-1457.

Jee, S.H., Song, K.S., Shim, W.H., Kim, H.K., Suh, I., Park, J.Y., et al. 2002. 'Major gene evidence after MTHFR-segregation analysis of serum homocysteine in families of patients undergoing coronary arteriography'. *Human Genetics.* 111 (2): 128-135.

Johnson, G.C. & Todd, J.A. 2000. 'Strategies in complex disease mapping'. *Current Opinion in Genetics & Development.* 10 (3): 330-334.

Kaess, B., Fischer, M., Baessler, A., Stark, K., Huber, F., Kremer, W., et al. 2008. 'The lipoprotein subfraction profile: heritability and identification of quantitative trait loci'. *Journal of Lipid Research.* 49 (4): 715-723.

Kalinina, N., Agrotis, A., Antropova, Y., Ilyinskaya, O., Smirnov, V., Tararak, E., et al. 2004. 'Smad expression in human atherosclerotic lesions: evidence for impaired TGF-beta/Smad signaling in smooth muscle cells of fibrofatty lesions'. *Arteriosclerosis, Thrombosis, and Vascular Biology.* 24 (8): 1391-1396.

Kannel, W.B., Dawber, T.R., Kagan, A., Revotskie, N. & Stokes, J. 1961. 'Factors of Risk in the Development of Coronary Heart Disease—Six-Year Follow-up Experience: The Framingham Study'. *Annals of Internal Medicine.* 55 (1): 33-50.

Kathiresan, S., Melander, O., Guiducci, C., Surti, A., Burtt, N.P., Rieder, M.J., et al. 2008. 'Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans'. *Nature Genetics.* 40 (2): 189-197.

Kathiresan, S., Willer, C.J., Peloso, G.M., Demissie, S., Musunuru, K., Schadt, E.E., et al. 2009. 'Common variants at 30 loci contribute to polygenic dyslipidemia'. *Nature Genetics.* 41 (1): 56-65.

Katsuda, S., Okada, Y., Minamoto, T., Oda, Y., Matsui, Y. & Nakanishi, I. 1992. 'Collagens in human atherosclerosis. Immunohistochemical analysis using collagen type-specific antibodies'. *Arteriosclerosis and Thrombosis : A Journal of Vascular Biology / American Heart Association.* 12 (4): 494-502.

Keating, B.J., Tischfield, S., Murray, S.S., Bhangale, T., Price, T.S., Glessner, J.T., et al. 2008. 'Concept, design and implementation of a cardiovascular gene-centric 50 k SNP array for large-scale genomic association studies'. *PloS One.* 3 (10): e3583.

King, R.A., Rotter, J.I. & Motulsky, A.G. (eds) 2002.
the Genetic Basis of Common Diseases. . New York: Oxford University Press.

Kjolby, M., Andersen, O.M., Breiderhoff, T., Fjorback, A.W., Pedersen, K.M., Madsen, P., et al. 2010. 'Sort1, encoded by the cardiovascular risk locus 1p13.3, is a regulator of hepatic lipoprotein export'. *Cell Metabolism.* 12 (3): 213-223.

Kolodgie, F.D., Luna, R.E., Farb, A., Burke, A.P., Horiba, K., Ferrans, V.J., et al. 1998.
' Differential expression of matrix metalloproteinases in coronary thrombosis: plaque rupture and erosion. J Am Coll Cardiol. 1998;31:419A. Abstract. '. *J Am Coll Cardiol.* 31: 419A.

Kotowski, I.K., Pertsemlidis, A., Luke, A., Cooper, R.S., Vega, G.L., Cohen, J.C., et al. 2006. 'A spectrum of PCSK9 alleles contributes to plasma levels of low-density lipoprotein cholesterol'. *American Journal of Human Genetics.* 78 (3): 410-422.

Kuhn, K. 1995. 'Basement membrane (type IV) collagen'. *Matrix Biology : Journal of the International Society for Matrix Biology.* 14 (6): 439-445.

Kuller, L., Arnold, A., Tracy, R., Otvos, J., Burke, G., Psaty, B., et al. 2002. 'Nuclear magnetic resonance spectroscopy of lipoproteins and risk of coronary heart disease in the cardiovascular health study'. *Arteriosclerosis, Thrombosis, and Vascular Biology.* 22 (7): 1175-1180.

Kumar, V., Abbas, A.K., Aster, J. & Fausto, N. 2009. 'Atherosclerosis' in *Pathologic Basis of Disease.* (Revised 8th edn). USA: W.B. Saunders Company, 496.

Kumar, V., Cotran, R.S. & Robbins, S.L. 1992. 'Atherosclerosis' in *Basic Pathology.* , (ed) J. Mitchell, (5th edn). USA: W.B. SAUNDERS COMPANY, 278-285.

Kutalik, Z., Whittaker, J., Waterworth, D., GIANT consortium, Beckmann, J.S. & Bergmann, S. 2011. 'Novel method to estimate the phenotypic variation explained by genome-wide association studies reveals large fraction of the missing heritability'. *Genetic Epidemiology.* 35 (5): 341-349.

La Thangue, N.B. 2002. 'Transcription. Chromatin control--a place for E2F and Myc to meet'. *Science (New York, N.Y.).* 296 (5570): 1034-1035.

Lander, E. & Kruglyak, L. 1995. 'Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results'. *Nature Genetics.* 11 (3): 241-247.

Lander, E.S. 1996. 'The new genomics: global views of biology'. *Science (New York, N.Y.).* 274 (5287): 536-539.

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

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

LeBleu, V.S., Macdonald, B. & Kalluri, R. 2007. 'Structure and function of basement membranes'. *Experimental Biology and Medicine (Maywood, N.J.).* 232 (9): 1121-1129.

Lee, M.H., Lu, K., Hazard, S., Yu, H., Shulenin, S., Hidaka, H., et al. 2001. 'Identification of a gene, ABCG5, important in the regulation of dietary cholesterol absorption'. *Nature Genetics.* 27 (1): 79-83.

Levy, D., DeStefano, A.L., Larson, M.G., O'Donnell, C.J., Lifton, R.P., Gavras, H., et al. 2000. 'Evidence for a gene influencing blood pressure on chromosome 17. Genome scan linkage results for longitudinal blood pressure phenotypes in subjects from the framingham heart study'. *Hypertension.* 36 (4): 477-483.

Li, B. & Leal, S.M. 2008. 'Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data'. *American Journal of Human Genetics.* 83 (3): 311-321.

Li, Y., Willer, C.J., Ding, J., Scheet, P. & Abecasis, G.R. 2010. 'MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes'. *Genetic Epidemiology.* 34 (8): 816-834.

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

Libby, P. & Theroux, P. 2005. 'Pathophysiology of Coronary Artery Disease'. *Circulation.* 111 (25): 3481-3488.

Linsel-Nitschke, P., Heeren, J., Aherrahrou, Z., Bruse, P., Gieger, C., Illig, T., et al. 2010. 'Genetic variation at chromosome 1p13.3 affects sortilin mRNA expression, cellular LDL-uptake and serum LDL levels which translates to the risk of coronary artery disease'. *Atherosclerosis.* 208 (1): 183-189.

Little, P.J., Chait, A. & Bobik, A. 2011. 'Cellular and cytokine-based inflammatory processes as novel therapeutic targets for the prevention and treatment of atherosclerosis'. *Pharmacology & Therapeutics.* 131 (3): 255-268.

Livak, K.J. & Schmittgen, T.D. 2001. 'Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method'. *Methods (San Diego, Calif.).* 25 (4): 402-408.

Lo, P.K., Chen, J.Y., Lo, W.C., Chen, B.F., Hsin, J.P., Tang, P.P., et al. 1999. 'Identification of a novel mouse p53 target gene DDA3.
'. *Oncogene.* 18 (54): 7765-7774.

Lowe, S.W. & Sherr, C.J. 2003. 'Tumor suppression by Ink4a-Arf: progress and puzzles'. *Current Opinion in Genetics & Development.* 13 (1): 77-83.

Luke, M.M., Kane, J.P., Liu, D.M., Rowland, C.M., Shiffman, D., Cassano, J., et al. 2007. 'A polymorphism in the protease-like domain of apolipoprotein(a) is associated with severe coronary artery disease'. *Arteriosclerosis, Thrombosis, and Vascular Biology.* 27 (9): 2030-2036.

Lusis, A.J. 2003. 'Genetic factors in cardiovascular disease. 10 questions'. *Trends in Cardiovascular Medicine.* 13 (8): 309-316.

Lusis, A.J., Mar, R. & Pajukanta, P. 2004. 'Genetics of atherosclerosis'. *Annual Review of Genomics and Human Genetics.* 5: 189-218.

Lusis, A.J. & Pajukanta, P. 2008. 'A treasure trove for lipoprotein biology'. *Nature Genetics.* 40 (2): 129-130.

Lusis, A.J., Weinreb, A., Drake, T.A. & Allayee, H. 2002. ' Genetics of atherosclerosis.' in *Textbook of Cardiovascular Medicine.* , (ed) E. Topol, (2nd edn). Philadelphia: Lippincott Williams and Wilkins.

Lyons, M.J., Schultz, M., Neale, M., Brady, K., Eisen, S., Toomey, R., et al. 2006. 'Specificity of familial vulnerability for alcoholism versus major depression in men'. *The Journal of Nervous and Mental Disease.* 194 (11): 809-817.

Maerkl, S.J. & Quake, S.R. 2007. 'A systems approach to measuring the binding energy landscapes of transcription factors'. *Science (New York, N.Y.).* 315 (5809): 233-237.

Maher, B. 2008. 'Personal genomes: The case of the missing heritability'. *Nature.* 456 (7218): 18-21.

Manning, A.K., Hivert, M.F., Scott, R.A., Grimsby, J.L., Bouatia-Naji, N., Chen, H., et al. 2012. 'A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance'. *Nature Genetics.* 44 (6): 659-669.

Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., et al. 2009. 'Finding the missing heritability of complex diseases'. *Nature.* 461 (7265): 747-753.

Maragkakis, M., Reczko, M., Simossis, V.A., Alexiou, P., Papadopoulos, G.L., Dalamagas, T., et al. 2009. 'DIANA-microT web server: elucidating microRNA functions through target prediction'. *Nucleic Acids Research.* 37 (Web Server issue): W273-6.

Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. 2007. 'A new multipoint method for genome-wide association studies by imputation of genotypes'. *Nature Genetics.* 39 (7): 906-913.

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

Martin, E.R., Lai, E.H., Gilbert, J.R., Rogala, A.R., Afshari, A.J., Riley, J., et al. 2000. 'SNPing away at complex diseases: analysis of single-nucleotide polymorphisms around APOE in Alzheimer disease'. *American Journal of Human Genetics.* 67 (2): 383-394.

Matys, V., Kel-Margoulis, O.V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., et al. 2006. 'TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes'. *Nucleic Acids Research.* 34 (Database issue): D108-10.

Maxwell, K.N. & Breslow, J.L. 2004. 'Adenoviral-mediated expression of Pcsk9 in mice results in a low-density lipoprotein receptor knockout phenotype'. *Proceedings of the National Academy of Sciences of the United States of America.* 101 (18): 7100-7105.

McCarthy, J.J., Parker, A., Salem, R., Moliterno, D.J., Wang, Q., Plow, E.F., et al. 2004. 'Large scale association analysis for identification of genes underlying premature coronary heart disease: cumulative perspective from analysis of 111 candidate genes'. *Journal of Medical Genetics.* 41 (5): 334-341.

McCully, K.S. 1993. 'Chemical pathology of homocysteine. I. Atherogenesis'. *Annals of Clinical and Laboratory Science.* 23 (6): 477-493.

McNally, A.K., Chisolm, G.M., Morel, D.W. & Cathcart, M.K. 1990. 'Activated human monocytes oxidize low-density lipoprotein by a lipoxygenase-dependent pathway.'. *The Journal of Immunology.* 145 (1): 254-259.

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

Miller, Y.I., Altamentova, S.M. & Shaklai, N. 1997. 'Oxidation of Low-Density Lipoprotein by Hemoglobin Stems from a Heme-Initiated Globin Radical: Antioxidant Role of Haptoglobin'. *Biochemistry.* 36 (40): 12189-12198.

Miyazono, K. 2000. 'TGF-beta signaling by Smad proteins'. *Cytokine & Growth Factor Reviews.* 11 (1-2): 15-22.

Morgan, T.M., Krumholz, H.M., Lifton, R.P. & Spertus, J.A. 2007. 'Nonvalidation of reported genetic risk factors for acute coronary syndrome in a large-scale replication study'. *JAMA : The Journal of the American Medical Association.* 297 (14): 1551-1561.

Morinville, A., Martin, S., Lavallee, M., Vincent, J.P., Beaudet, A. & Mazella, J. 2004. 'Internalization and trafficking of neurotensin via NTS3 receptors in HT29 cells'. *The International Journal of Biochemistry & Cell Biology.* 36 (11): 2153-2168.

Morozova, O., Hirst, M. & Marra, M.A. 2009. 'Applications of new sequencing technologies for transcriptome analysis'. *Annual Review of Genomics and Human Genetics.* 10: 135-151.

Motulsky, A.G. & Brunzell, J.D. 2003. '
Genetics of coronary atherosclerosis' in
the Genetic Basis of Common Disease. , (eds) R.A. King, J.I. Rotter & A.G. Motulsky, (2nd edn). New York: Oxford University Press, 105-126.

Muczynski, K.A., Ekle, D.M., Coder, D.M. & Anderson, S.K. 2003. 'Normal human kidney HLA-DR-expressing renal microvascular endothelial cells: characterization, isolation, and regulation of MHC class II expression'. *Journal of the American Society of Nephrology : JASN.* 14 (5): 1336-1348.

Munck Petersen, C., Nielsen, M.S., Jacobsen, C., Tauris, J., Jacobsen, L., Gliemann, J., et al. 1999. 'Propeptide cleavage conditions sortilin/neurotensin receptor-3 for ligand binding'. *The EMBO Journal.* 18 (3): 595-604.

Musunuru, K., Orho-Melander, M., Caulfield, M.P., Li, S., Salameh, W.A., Reitz, R.E., et al. 2009. 'Ion mobility analysis of lipoprotein subfractions identifies three independent axes of cardiovascular risk'. *Arteriosclerosis, Thrombosis, and Vascular Biology.* 29 (11): 1975-1980.

Musunuru, K., Strong, A., Frank-Kamenetsky, M., Lee, N.E., Ahfeldt, T., Sachs, K.V., et al. 2010. 'From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus'. *Nature.* 466 (7307): 714-719.

Myllyharju, J. & Kivirikko, K.I. 2004. 'Collagens, modifying enzymes and their mutations in humans, flies and worms'. *Trends in Genetics : TIG.* 20 (1): 33-43.

Myllyharju, J. & Kivirikko, K.I. 2001. 'Collagens and collagen-related diseases'. *Annals of Medicine.* 33 (1): 7-21.

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

Nejepinska, J., Malik, R., Moravec, M. & Svoboda, P. 2012. 'Deep sequencing reveals complex spurious transcription from transiently transfected plasmids'. *PloS One.* 7 (8): e43283.

Nickerson, D.A., Taylor, S.L., Weiss, K.M., Clark, A.G., Hutchinson, R.G., Stengard, J., et al. 1998. 'DNA sequence diversity in a 9.7-kb region of the human lipoprotein lipase gene'. *Nature Genetics.* 19 (3): 233-240.

Nielsen, M.S., Jacobsen, C., Olivecrona, G., Gliemann, J. & Petersen, C.M. 1999. 'Sortilin/neurotensin receptor-3 binds and mediates degradation of lipoprotein lipase'. *The Journal of Biological Chemistry.* 274 (13): 8832-8836.

Nilsson, S.K., Christensen, S., Raarup, M.K., Ryan, R.O., Nielsen, M.S. & Olivecrona, G. 2008. 'Endocytosis of apolipoprotein A-V by members of the low density lipoprotein receptor and the VPS10p domain receptor families'. *The Journal of Biological Chemistry.* 283 (38): 25920-25927.

O'Donnell, C.J., Kavousi, M., Smith, A.V., Kardia, S.L., Feitosa, M.F., Hwang, S.J., et al. 2011. 'Genome-wide association study for coronary artery calcification with follow-up in myocardial infarction'. *Circulation.* 124 (25): 2855-2864.

Ogawa, H., Ishiguro, K., Gaubatz, S., Livingston, D.M. & Nakatani, Y. 2002. 'A complex with chromatin modifiers that occupies E2F- and Myc-responsive genes in G0 cells'. *Science (New York, N.Y.).* 296 (5570): 1132-1136.

Okano, K., Hibi, A., Miyaoka, T., Inoue, T., Sugimoto, H., Tsuchiya, K., et al. 2012. 'Inhibitory effects of the transcription factor Ets-1 on the expression of type I collagen in TGF-beta1-stimulated renal epithelial cells'. *Molecular and Cellular Biochemistry.* 369 (1-2): 247-254.

Osler, W. 1897. *Lectures on Angina Pectoris and Allied States.* New York: D. Appleton & Company.

Oyake, T., Itoh, K., Motohashi, H., Hayashi, N., Hoshino, H., Nishizawa, M., et al. 1996. 'Bach proteins belong to a novel family of BTB-basic leucine zipper transcription factors that interact with MafK and regulate transcription through the NF-E2 site'. *Molecular and Cellular Biology.* 16 (11): 6083-6095.

Ozaki, K., Inoue, K., Sato, H., Iida, A., Ohnishi, Y., Sekine, A., et al. 2004. 'Functional variation in LGALS2 confers risk of myocardial infarction and regulates lymphotoxin-alpha secretion in vitro'. *Nature.* 429 (6987): 72-75.

Ozaki, K., Ohnishi, Y., Iida, A., Sekine, A., Yamada, R., Tsunoda, T., et al. 2002. 'Functional SNPs in the lymphotoxin-alpha gene that are associated with susceptibility to myocardial infarction'. *Nature Genetics.* 32 (4): 650-654.

Pajukanta, P., Cargill, M., Viitanen, L., Nuotio, I., Kareinen, A., Perola, M., et al. 2000. 'Two loci on chromosomes 2 and X for premature coronary heart disease identified in early- and late-settlement populations of Finland'. *American Journal of Human Genetics.* 67 (6): 1481-1493.

Parikh, H., Lyssenko, V. & Groop, L.C. 2009. 'Prioritizing genes for follow-up from genome wide association studies using information on gene expression in tissues relevant for type 2 diabetes mellitus'. *BMC Medical Genomics.* 2: 72-8794-2-72.

Park, S.W., Moon, Y.A. & Horton, J.D. 2004. 'Post-transcriptional regulation of low density lipoprotein receptor protein by proprotein convertase subtilisin/kexin type 9a in mouse liver'. *The Journal of Biological Chemistry.* 279 (48): 50630-50638.

Parthasarathy, S., Wieland, E. & Steinberg, D. 1989. 'A role for endothelial cell lipoxygenase in the oxidative modification of low density lipoprotein'. *Proceedings of the National Academy of Sciences.* 86 (3): 1046-1050.

Paulsson, M. 1992. 'Basement membrane proteins: structure, assembly, and cellular interactions'. *Critical Reviews in Biochemistry and Molecular Biology.* 27 (1-2): 93-127.

Pawluczyk, I.Z., Patel, S.R. & Harris, K.P. 2004. 'The role of bradykinin in the antifibrotic actions of perindoprilat on human mesangial cells'. *Kidney International.* 65 (4): 1240-1251.

Pearson, R., Sivananthan, U.M., Barth, J.H., Gale, C.P. & Hall, A.S. 2011. 'Comparison of 4-h heart fatty acid binding protein with 12-h troponin I to assess 6-month risk following percutaneous coronary intervention in acute coronary syndromes.'. *Heart.* 97 (Suppl1): A17.

Peng, H., Tan, L., Osaki, M., Zhan, Y., Ijiri, K., Tsuchimochi, K., et al. 2008. 'ESE-1 is a potent repressor of type II collagen gene (COL2A1) transcription in human chondrocytes'. *Journal of Cellular Physiology.* 215 (2): 562-573.

Pesu, M., Aittomaki, S., Takaluoma, K., Lagerstedt, A. & Silvennoinen, O. 2002. 'p38 Mitogen-activated protein kinase regulates interleukin-4-induced gene expression by stimulating STAT6-mediated transcription'. *The Journal of Biological Chemistry.* 277 (41): 38254-38261.

Petersen, A.K., Stark, K., Musameh, M.D., Nelson, C.P., Romisch-Margl, W., Kremer, W., et al. 2012. 'Genetic associations with lipoprotein subfractions provide information on their biological nature'. *Human Molecular Genetics.* 21 (6): 1433-1443.

Petersen, C.M., Nielsen, M.S., Nykjaer, A., Jacobsen, L., Tommerup, N., Rasmussen, H.H., et al. 1997. 'Molecular identification of a novel candidate sorting receptor purified from human brain by receptor-associated protein affinity chromatography'. *The Journal of Biological Chemistry.* 272 (6): 3599-3605.

Pique-Regi, R., Degner, J.F., Pai, A.A., Gaffney, D.J., Gilad, Y. & Pritchard, J.K. 2011. 'Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data'. *Genome Research.* 21 (3): 447-455.

Plaisier, E., Gribouval, O., Alamowitch, S., Mougenot, B., Prost, C., Verpont, M.C., et al. 2007. 'COL4A1 mutations and hereditary angiopathy, nephropathy, aneurysms, and muscle cramps'. *The New England Journal of Medicine.* 357 (26): 2687-2695.

Plaisier, E. & Ronco, P. 1993. 'COL4A1-related disorders' in *GeneReviews.* , (eds) R.A. Pagon, T.D. Bird, C.R. Dolan & K. Stephens,. Seattle (WA): University of Washington, Seattle. All rights reserved.

Pollner, R., Schmidt, C., Fischer, G., Kuhn, K. & Poschl, E. 1997. 'Cooperative and competitive interactions of regulatory elements are involved in the control of divergent transcription of human Col4A1 and Col4A2 genes'. *FEBS Letters.* 405 (1): 31-36.

Portales-Casamar, E., Thongjuea, S., Kwon, A.T., Arenillas, D., Zhao, X., Valen, E., et al. 2010. 'JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles'. *Nucleic Acids Research.* 38 (Database issue): D105-10.

Poschl, E., Pollner, R. & Kuhn, K. 1988. 'The genes for the alpha 1(IV) and alpha 2(IV) chains of human basement membrane collagen type IV are arranged head-to-head and separated by a bidirectional promoter of unique structure'. *The EMBO Journal.* 7 (9): 2687-2695.

Prasannan, P., Pike, S., Peng, K., Shane, B. & Appling, D.R. 2003. 'Human mitochondrial C1-tetrahydrofolate synthase: gene structure, tissue distribution of the mRNA, and immunolocalization in Chinese hamster ovary calls'. *The Journal of Biological Chemistry.* 278 (44): 43178-43187.

Pritchard, J.K. 2001. 'Are rare variants responsible for susceptibility to complex diseases?'. *American Journal of Human Genetics.* 69 (1): 124-137.

Prockop, D.J. & Kivirikko, K.I. 1995. 'Collagens: molecular biology, diseases, and potentials for therapy'. *Annual Review of Biochemistry.* 64: 403-434.

Prospective Studies Collaboration, Lewington, S., Whitlock, G., Clarke, R., Sherliker, P., Emberson, J., et al. 2007. 'Blood cholesterol and vascular mortality by age, sex, and blood pressure: a meta-analysis of individual data from 61 prospective studies with 55,000 vascular deaths'. *Lancet.* 370 (9602): 1829-1839.

Pruim, R.J., Welch, R.P., Sanna, S., Teslovich, T.M., Chines, P.S., Gliedt, T.P., et al. 2010. 'LocusZoom: regional visualization of genome-wide association scan results'. *Bioinformatics (Oxford, England).* 26 (18): 2336-2337.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., et al. 2007. 'PLINK: a tool set for whole-genome association and population-based linkage analyses'. *American Journal of Human Genetics.* 81 (3): 559-575.

Qin, B.Y., Lam, S.S., Correia, J.J. & Lin, K. 2002. 'Smad3 allostery links TGF-beta receptor kinase activation to transcriptional control'. *Genes & Development.* 16 (15): 1950-1963.

Quandt, K., Frech, K., Karas, H., Wingender, E. & Werner, T. 1995. 'MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data'. *Nucleic Acids Research.* 23 (23): 4878-4884.

Rader, D.J., Cohen, J. & Hobbs, H.H. 2003. 'Monogenic hypercholesterolemia: new insights in pathogenesis and treatment'. *The Journal of Clinical Investigation.* 111 (12): 1795-1803.

Randak, C., Roschinger, W., Rolinski, B., Hadorn, H.B., Applegarth, D.A. & Roscher, A.A. 2000. 'Three siblings with nonketotic hyperglycinaemia, mildly elevated plasma homocysteine concentrations and moderate methylmalonic aciduria'. *Journal of Inherited Metabolic Disease.* 23 (5): 520-522.

Regieli, J.J., Nathoe, H.M., Koerselman, J., van der Graaf, Y., Grobbee, D.E. & Doevendans, P.A. 2007. 'Coronary collaterals--insights in molecular determinants and prognostic relevance'. *International Journal of Cardiology.* 116 (2): 139-143.

Reich, D.E., Cargill, M., Bolk, S., Ireland, J., Sabeti, P.C., Richter, D.J., et al. 2001. 'Linkage disequilibrium in the human genome'. *Nature.* 411 (6834): 199-204.

Rieder, M.J. & Nickerson, D.A. 2000. 'Hypertension and single nucleotide polymorphisms'. *Current Hypertension Reports.* 2 (1): 44-49.

Rigler, R., Cronvall, E., Hirsch, R., Pachmann, U. & Zachau, H.G. 1970. 'Interactions of seryl-tRNA synthetase with serine and phenylalanine specific tRNA'. *FEBS Letters.* 11 (5): 320-323.

Rios, J., Stein, E., Shendure, J., Hobbs, H.H. & Cohen, J.C. 2010. 'Identification by whole-genome resequencing of gene defect responsible for severe hypercholesterolemia'. *Human Molecular Genetics.* 19 (22): 4313-4318.

Risch, N. 1990a. 'Linkage strategies for genetically complex traits. I. Multilocus models'. *American Journal of Human Genetics.* 46 (2): 222-228.

Risch, N. 1990b. 'Linkage strategies for genetically complex traits. II. The power of affected relative pairs'. *American Journal of Human Genetics.* 46 (2): 229-241.

Risch, N. 1990c. 'Linkage strategies for genetically complex traits. III. The effect of marker polymorphism on analysis of affected relative pairs'. *American Journal of Human Genetics.* 46 (2): 242-253.

Risch, N. & Merikangas, K. 1996. 'The future of genetic studies of complex human diseases'. *Science (New York, N.Y.).* 273 (5281): 1516-1517.

Roberts, R., Wells, G.A., Stewart, A.F., Dandona, S. & Chen, L. 2010. 'The genome-wide association study--a new era for common polygenic disorders'. *Journal of Cardiovascular Translational Research.* 3 (3): 173-182.

Rodwell, G.E., Sonu, R., Zahn, J.M., Lund, J., Wilhelmy, J., Wang, L., et al. 2004. 'A transcriptional profile of aging in the human kidney'. *PLoS Biology.* 2 (12): e427.

Rose, G. 1964. 'Familial Patterns in Ischaemic Heart Disease'. *British Journal of Preventive and Social Medicine.* 18 (2): 75-80.

Ross, R., Wight, T.N., Strandness, E. & Thiele, B. 1984. 'Human atherosclerosis. I. Cell constitution and characteristics of advanced lesions of the superficial femoral artery'. *The American Journal of Pathology.* 114 (1): 79-93.

Rust, S., Rosier, M., Funke, H., Real, J., Amoura, Z., Piette, J.C., et al. 1999. 'Tangier disease is caused by mutations in the gene encoding ATP-binding cassette transporter 1'. *Nature Genetics.* 22 (4): 352-355.

Sabo, P.J., Kuehn, M.S., Thurman, R., Johnson, B.E., Johnson, E.M., Cao, H., et al. 2006. 'Genome-scale mapping of DNase I sensitivity in vivo using tiling DNA microarrays'. *Nature Methods.* 3 (7): 511-518.

Sachidanandam, R., Weissman, D., Schmidt, S.C., Kakol, J.M., Stein, L.D., Marth, G., et al. 2001. 'A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms'. *Nature.* 409 (6822): 928-933.

Samani, N.J., Braund, P.S., Erdmann, J., Gotz, A., Tomaszewski, M., Linsel-Nitschke, P., et al. 2008. 'The novel genetic variant predisposing to coronary artery disease in the region of the PSRC1 and CELSR2 genes on chromosome 1 associates with serum cholesterol'. *Journal of Molecular Medicine (Berlin, Germany).* 86 (11): 1233-1241.

Samani, N.J., Burton, P., Mangino, M., Ball, S.G., Balmforth, A.J., Barrett, J., et al. 2005. 'A genomewide linkage study of 1,933 families affected by premature coronary artery disease: The British Heart Foundation (BHF) Family Heart Study'. *American Journal of Human Genetics.* 77 (6): 1011-1020.

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

Samani, N.J., Martin, D.S., Brack, M., Cullen, J., Chauhan, A., Lodwick, D., et al. 1995. 'Insertion/deletion polymorphism in the angiotensin-converting enzyme gene and risk of restenosis after coronary angioplasty'. *Lancet.* 345 (8956): 1013-1016.

Samani, N.J. & Schunkert, H. 2008. 'Chromosome 9p21 and cardiovascular disease: the story unfolds'. *Circulation.Cardiovascular Genetics.* 1 (2): 81-84.

Sandhu, M.S., Waterworth, D.M., Debenham, S.L., Wheeler, E., Papadakis, K., Zhao, J.H., et al. 2008. 'LDL-cholesterol concentrations: a genome-wide association study'. *Lancet.* 371 (9611): 483-491.

Savenkova, M.L., Mueller, D.M. & Heinecke, J.W. 1994. 'Tyrosyl radical generated by myeloperoxidase is a physiological catalyst for the initiation of lipid peroxidation in low density lipoprotein.'. *Journal of Biological Chemistry.* 269 (32): 20394-20400.

Sawadogo, M. & Sentenac, A. 1990. 'RNA polymerase B (II) and general transcription factors'. *Annual Review of Biochemistry.* 59: 711-754.

Scarborough, P., Bhatnagar, P., Wickramasinghe, K., Smolina, K., Mitchell, C. & Rayner, M. 2010, *Coronary Heart Disease Statistics 2010 Edition*, London: British Heart Foundation.

Schadt, E.E., Molony, C., Chudin, E., Hao, K., Yang, X., Lum, P.Y., et al. 2008. 'Mapping the genetic architecture of gene expression in human liver'. *PLoS Biology.* 6 (5): e107.

Schmidt, C., Fischer, G., Kadner, H., Genersch, E., Kuhn, K. & Poschl, E. 1993. 'Differential effects of DNA-binding proteins on bidirectional transcription from the common promoter region of human collagen type IV genes COL4A1 and COL4A2'. *Biochimica Et Biophysica Acta.* 1174 (1): 1-10.

Schmidt, C., Pollner, R., Poschl, E. & Kuhn, K. 1992. 'Expression of human collagen type IV genes is regulated by transcriptional and post-transcriptional mechanisms'. *FEBS Letters.* 312 (2-3): 174-178.

Schunkert, H., Konig, I.R., Kathiresan, S., Reilly, M.P., Assimes, T.L., Holm, H., et al. 2011. 'Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease'. *Nature Genetics.*

Sedlacek, K., Neureuther, K., Mueller, J.C., Stark, K., Fischer, M., Baessler, A., et al. 2007. 'Lymphotoxin-alpha and galectin-2 SNPs are not associated with myocardial infarction in two different German populations'. *Journal of Molecular Medicine (Berlin, Germany).* 85 (9): 997-1004.

Servin, B. & Stephens, M. 2007. 'Imputation-based analysis of association studies: candidate regions and quantitative traits'. *PLoS Genetics.* 3 (7): e114.

Shekhonin, B.V., Domogatsky, S.P., Idelson, G.L., Koteliansky, V.E. & Rukosuev, V.S. 1987. 'Relative distribution of fibronectin and type I, III, IV, V collagens in normal and atherosclerotic intima of human arteries'. *Atherosclerosis.* 67 (1): 9-16.

Shekhonin, B.V., Domogatsky, S.P., Muzykantov, V.R., Idelson, G.L. & Rukosuev, V.S. 1985. 'Distribution of type I, III, IV and V collagen in normal and atherosclerotic human arterial wall: immunomorphological characteristics'. *Collagen and Related Research.* 5 (4): 355-368.

Shepherd, J., Cobbe, S.M., Ford, I., Isles, C.G., Lorimer, A.R., MacFarlane, P.W., et al. 1995. 'Prevention of coronary heart disease with pravastatin in men with hypercholesterolemia. West of Scotland Coronary Prevention Study Group'. *The New England Journal of Medicine.* 333 (20): 1301-1307.

Sheridan, C. 2013. 'Phase 3 data for PCSK9 inhibitor wows'. *Nature Biotechnology.* 31 (12): 1057-1058.

Shiffman, D., Ellis, S.G., Rowland, C.M., Malloy, M.J., Luke, M.M., Iakoubova, O.A., et al. 2005. 'Identification of four gene variants associated with myocardial infarction'. *American Journal of Human Genetics.* 77 (4): 596-605.

Shiffman, D., O'Meara, E.S., Bare, L.A., Rowland, C.M., Louie, J.Z., Arellano, A.R., et al. 2008. 'Association of gene variants with incident myocardial infarction in the Cardiovascular Health Study'. *Arteriosclerosis, Thrombosis, and Vascular Biology.* 28 (1): 173-179.

Shiffman, D., Rowland, C.M., Louie, J.Z., Luke, M.M., Bare, L.A., Bolonick, J.I., et al. 2006. 'Gene variants of VAMP8 and HNRPUL1 are associated with early-onset myocardial infarction'. *Arteriosclerosis, Thrombosis, and Vascular Biology.* 26 (7): 1613-1618.

Shima, Y., Kawaguchi, S.Y., Kosaka, K., Nakayama, M., Hoshino, M., Nabeshima, Y., et al. 2007. 'Opposing roles in neurite growth control by two seven-pass transmembrane cadherins'. *Nature Neuroscience.* 10 (8): 963-969.

Siri-Tarino, P.W., Williams, P.T., Fernstrom, H.S., Rawlings, R.S. & Krauss, R.M. 2009. 'Reversal of small, dense LDL subclass phenotype by normalization of adiposity'. *Obesity (Silver Spring, Md.).* 17 (9): 1768-1775.

Slack, J. & Evans, K.A. 1966. 'The increased risk of death from ischaemic heart disease in first degree relatives of 121 men and 96 women with ischaemic heart disease'. *Journal of Medical Genetics.* 3 (4): 239-257.

Song, Y., Stampfer, M.J. & Liu, S. 2004. 'Meta-analysis: apolipoprotein E genotypes and risk for coronary heart disease'. *Annals of Internal Medicine.* 141 (2): 137-147.

Stein, E.A., Mellis, S., Yancopoulos, G.D., Stahl, N., Logan, D., Smith, W.B., et al. 2012. 'Effect of a monoclonal antibody to PCSK9 on LDL cholesterol'. *The New England Journal of Medicine.* 366 (12): 1108-1118.

Steinbrecher, U.P., Parthasarathy, S., Leake, D.S., Witztum, J.L. & Steinberg, D. 1984. 'Modification of low density lipoprotein by endothelial cells involves lipid peroxidation and degradation of low density lipoprotein phospholipids'. *Proceedings of the National Academy of Sciences.* 81 (12): 3883-3887.

Stormo, G.D. 2000. 'DNA binding sites: representation and discovery'. *Bioinformatics (Oxford, England).* 16 (1): 16-23.

Strachan, T. & Read, A.P. 2004. 'Mapping and identifying genes conferring susceptibility to complex diseases' in *Human Molecular Genetic 3.* (3rd edn). London & New York: Garland Science Taylor & Francis Group, 436-448.

Strachan, T. & Read, A.P. 1999. *Human Molecular Genetics 2.* (2nd edn). London &New York: Garland Science - Taylor & Francis Group.

Swanson, K.A., Knoepfler, P.S., Huang, K., Kang, R.S., Cowley, S.M., Laherty, C.D., et al. 2004. 'HBP1 and Mad1 repressors bind the Sin3 corepressor PAH2 domain with opposite helical orientations'. *Nature Structural & Molecular Biology.* 11 (8): 738-746.

Takeichi, M. 2007. 'The cadherin superfamily in neuronal connections and interactions'. *Nature Reviews.Neuroscience.* 8 (1): 11-20.

Tanimura, A., McGregor, D.H. & Anderson, H.C. 1986a. 'Calcification in atherosclerosis. I. Human studies'. *Journal of Experimental Pathology.* 2 (4): 261-273.

Tanimura, A., McGregor, D.H. & Anderson, H.C. 1986b. 'Calcification in atherosclerosis. II. Animal studies'. *Journal of Experimental Pathology.* 2 (4): 275-297.

Tarasov, K.V., Sanna, S., Scuteri, A., Strait, J.B., Orru, M., Parsa, A., et al. 2009. 'COL4A1 is associated with arterial stiffness by genome-wide association scan'. *Circulation.Cardiovascular Genetics.* 2 (2): 151-158.

Teo, Y.Y., Inouye, M., Small, K.S., Gwilliam, R., Deloukas, P., Kwiatkowski, D.P., et al. 2007. 'A genotype calling algorithm for the Illumina BeadArray platform'. *Bioinformatics (Oxford, England).* 23 (20): 2741-2746.

Teslovich, T.M., Musunuru, K., Smith, A.V., Edmondson, A.C., Stylianou, I.M., Koseki, M., et al. 2010. 'Biological, clinical and population relevance of 95 loci for blood lipids'. *Nature.* 466 (7307): 707-713.

Tobacco and Genetics Consortium 2010. 'Genome-wide meta-analyses identify multiple loci associated with smoking behavior'. *Nature Genetics.* 42 (5): 441-447.

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

Tobin, M.D., Tomaszewski, M., Braund, P.S., Hajat, C., Raleigh, S.M., Palmer, T.M., et al. 2008. 'Common variants in genes underlying monogenic hypertension and hypotension and blood pressure in the general population'. *Hypertension.* 51 (6): 1658-1664.

Tomaszewski, M., Brain, N.J., Charchar, F.J., Wang, W.Y., Lacka, B., Padmanabahn, S., et al. 2002. 'Essential hypertension and beta2-adrenergic receptor gene: linkage and association analysis'. *Hypertension.* 40 (3): 286-291.

Tomaszewski, M., Charchar, F.J., Lynch, M.D., Padmanabhan, S., Wang, W.Y., Miller, W.H., et al. 2007. 'Fibroblast growth factor 1 gene and hypertension: from the quantitative trait locus to positional analysis'. *Circulation.* 116 (17): 1915-1924.

Topol, E.J., McCarthy, J., Gabriel, S., Moliterno, D.J., Rogers, W.J., Newby, L.K., et al. 2001. 'Single nucleotide polymorphisms in multiple novel thrombospondin genes may be associated with familial premature myocardial infarction'. *Circulation.* 104 (22): 2641-2644.

Tournier-Lasserve, E. 2002. 'New players in the genetics of stroke'. *The New England Journal of Medicine.* 347 (21): 1711-1712.

Tregouet, D.A., Konig, I.R., Erdmann, J., Munteanu, A., Braund, P.S., Hall, A.S., et al. 2009. 'Genome-wide haplotype association study identifies the SLC22A3-LPAL2-LPA gene cluster as a risk locus for coronary artery disease'. *Nature Genetics.* 41 (3): 283-285.

Vincent, J.B., Skaug, J. & Scherer, S.W. 2000. 'The human homologue of flamingo, EGFL2, encodes a brain-expressed large cadherin-like protein with epidermal growth factor-like domains, and maps to chromosome 1p13.3-p21.1'. *DNA Research : An International Journal for Rapid Publication of Reports on Genes and Genomes.* 7 (3): 233-235.

Voetsch, B. & Loscalzo, J. 2004. 'Genetic determinants of arterial thrombosis'. *Arteriosclerosis, Thrombosis, and Vascular Biology.* 24 (2): 216-229.

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

Voss, B. & Rauterberg, J. 1986. 'Localization of collagen types I, III, IV and V, fibronectin and laminin in human arteries by the indirect immunofluorescence method'. *Pathology, Research and Practice.* 181 (5): 568-575.

Wald, N.J. & Law, M.R. 2003. 'A strategy to reduce cardiovascular disease by more than 80%'. *BMJ (Clinical Research Ed.).* 326 (7404): 1419.

Walkup, A.S. & Appling, D.R. 2005. 'Enzymatic characterization of human mitochondrial C1-tetrahydrofolate synthase'. *Archives of Biochemistry and Biophysics.* 442 (2): 196-205.

Wallace, C., Newhouse, S.J., Braund, P., Zhang, F., Tobin, M., Falchi, M., et al. 2008. 'Genome-wide association study identifies genes for biomarkers of cardiovascular disease: serum urate and dyslipidemia'. *American Journal of Human Genetics.* 82 (1): 139-149.

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

Wang, P., Dai, M., Xuan, W., McEachin, R.C., Jackson, A.U., Scott, L.J., et al. 2006. 'SNP Function Portal: a web database for exploring the function implication of SNP alleles'. *Bioinformatics (Oxford, England).* 22 (14): e523-9.

Wang, Q., Rao, S., Shen, G.Q., Li, L., Moliterno, D.J., Newby, L.K., et al. 2004. 'Premature myocardial infarction novel susceptibility locus on chromosome 1P34-36 identified by genomewide linkage analysis'. *American Journal of Human Genetics.* 74 (2): 262-271.

Wellcome Trust Case Control Consortium 2007. 'Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls'. *Nature.* 447 (7145): 661-678.

Wellcome Trust Case Control Consortium, Maller, J.B., McVean, G., Byrnes, J., Vukcevic, D., Palin, K., et al. 2012. 'Bayesian refinement of association signals for 14 loci in 3 common diseases'. *Nature Genetics.* 44 (12): 1294-1301.

West of Scotland Coronary Prevention Study Group 1998. 'Influence of pravastatin and plasma lipids on clinical events in the West of Scotland Coronary Prevention Study (WOSCOPS)'. *Circulation.* 97 (15): 1440-1445.

Wheeler, H.E., Metter, E.J., Tanaka, T., Absher, D., Higgins, J., Zahn, J.M., et al. 2009. 'Sequential use of transcriptional profiling, expression quantitative trait mapping, and gene association implicates MMP20 in human kidney aging'. *PLoS Genetics.* 5 (10): e1000685.

Willer, C.J., Sanna, S., Jackson, A.U., Scuteri, A., Bonnycastle, L.L., Clarke, R., et al. 2008. 'Newly identified loci that influence lipid concentrations and risk of coronary artery disease'. *Nature Genetics.* 40 (2): 161-169.

Williams, P.T., Vranizan, K.M. & Krauss, R.M. 1992. 'Correlations of plasma lipoproteins with LDL subfractions by particle size in men and women'. *Journal of Lipid Research.* 33 (5): 765-774.

Wilson, P.W.F., D'Agostino, R.B., Levy, D., Belanger, A.M., Silbershatz, H. & Kannel, W.B. 1998. 'Prediction of Coronary Heart Disease Using Risk Factor Categories'. *Circulation.* 97 (18): 1837-1847.

World Health Organisation (WHO) 2011. June 2011-last update*. **The top 10 causes of death***. Available: http://www.who.int/mediacentre/factsheets/fs310/en/index.html.

World Health Organisation (WHO) 2008. . *Global Health Observatory Data Repository - Noncommunicable diseases: Risk factors: Alcohol, Blood glucose, Blood pressure, Cholesterol, Physical inactivity, Overweight/Obesity, Tobacco (Smoking)* [Homepage of World Health Organisation (WHO)], [Online].
Available: http://apps.who.int/gho/data/node.main.A867?lang=en (September2012).

Wu, Q. & Maniatis, T. 1999. 'A striking organization of a large family of human neural cadherin-like cell adhesion genes'. *Cell.* 97 (6): 779-790.

Yamada, Y., Izawa, H., Ichihara, S., Takatsu, F., Ishihara, H., Hirayama, H., et al. 2002. 'Prediction of the risk of myocardial infarction from polymorphisms in candidate genes'. *The New England Journal of Medicine.* 347 (24): 1916-1923.

Yamada, Y., Kato, K., Oguri, M., Fujimaki, T., Yokoi, K., Matsuo, H., et al. 2008. 'Genetic risk for myocardial infarction determined by polymorphisms of candidate genes in a Japanese population'. *Journal of Medical Genetics.* 45 (4): 216-221.

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

Yang, W., Tan, J., Liu, R., Cui, X., Ma, Q., Geng, Y., et al. 2012. 'Interferon-gamma upregulates expression of IFP35 gene in HeLa cells via interferon regulatory factor-1'. *PloS One.* 7 (12): e50932.

Ye, Z., Liu, E.H., Higgins, J.P., Keavney, B.D., Lowe, G.D., Collins, R., et al. 2006. 'Seven haemostatic gene polymorphisms in coronary disease: meta-analysis of 66,155 cases and 91,307 controls'. *Lancet.* 367 (9511): 651-658.

Zdravkovic, S., Wienke, A., Pedersen, N.L., Marenberg, M.E., Yashin, A.I. & De Faire, U. 2002. 'Heritability of death from coronary heart disease: a 36-year follow-up of 20 966 Swedish twins'. *Journal of Internal Medicine.* 252 (3): 247-254.

Zeggini, E., Scott, L.J., Saxena, R., Voight, B.F., Marchini, J.L., Hu, T., et al. 2008. 'Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes'. *Nature Genetics.* 40 (5): 638-645.

Zhang, G., Karns, R., Sun, G., Indugula, S.R., Cheng, H., Havas-Augustin, D., et al. 2012. 'Finding missing heritability in less significant Loci and allelic heterogeneity: genetic variation in human height'. *PloS One.* 7 (12): e51211.

Zwicker, J.I., Peyvandi, F., Palla, R., Lombardi, R., Canciani, M.T., Cairo, A., et al. 2006. 'The thrombospondin-1 N700S polymorphism is associated with early myocardial infarction without altering von Willebrand factor multimer size'. *Blood.* 108 (4): 1280-1283.