

The cellular response to atherogenic lipid



Anil Chalisey
Oriel College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy
September, 2019

I declare that this thesis has been composed solely by myself and that it has not been submitted, in whole or in part, in any previous application for a degree. Except where stated otherwise, by reference or acknowledgement, the work presented is entirely my own.

Abstract

The cellular response to atherogenic lipid

Atherosclerotic vascular disease is the leading cause of disease worldwide. It is an inflammatory condition affecting the arterial wall, and lesions are composed of vascular and immune cells. The purpose of this study is to understand how oxidised low density lipoprotein, a key risk factor for the development of atherosclerosis, affects immune and vascular cells involved in formation of the atherosclerotic lesion. Recent insights into the pathogenesis of atherosclerosis have highlighted changes in expression and activity of gene regulatory proteins (transcription factors) and epigenetic modifications, such as alterations in chromatin accessibility and histone marks, as important mechanisms of atherogenesis. To understand how these changes come about in atherosclerosis, genome-wide approaches have been used to investigate changes in gene expression, chromatin accessibility, and histone marks following exposure of primary human aortic endothelial cells, aortic vascular smooth muscle cells, CD4⁺ lymphocytes and CD14⁺ monocytes to oxidised low density lipoprotein.

Keywords: atherosclerosis, epigenetics, inflammation, chromatin, histone modifications, endothelial, vascular smooth muscle, monocytes

Acknowledgements

I would like to express my gratitude to my supervisor, Professor Chris O’Callaghan, whose expertise, understanding, and patience have been invaluable towards my graduate experience. I would also like to thank the other members of his group, Lucy Davison, Da Lin, Michael Reschen, and Tom Hiron, who in their own ways have provided essential aid over the last three years.

I would also like to thank my family for the support they provided me throughout this entire process (and indeed, my entire life!). I don’t say it often enough, but they know how important they have been to me. In particular, because I want them to be able to say their names can be found in the Bodleian Library at Oxford, I list them all here: Amar, Anju, Anup, Anisha, Shulochana, Prabesh, Sophia and Arya.

Finally, I recognise that none of this would have been possible without funding and financial assistance from the Wellcome Trust - thank you.

If we knew what we were doing, it
would not be called research, would
it?

Albert Einstein

Contents

Abstract	i
Acknowledgements	ii
List of Figures	v
List of Tables	vii
Abbreviations	ix
1 Background	1
1.1 Introduction	2
1.2 Mechanisms of atherosclerotic plaque progression	3
1.3 Biology of oxidised low density lipoprotein	5
1.4 Effects and atherogenicity of oxLDL	9
1.5 Epigenetic regulation of atherosclerosis	18
1.6 Aims and objectives	23
2 Materials and Methods	26
2.1 Basic cell culture techniques and cell types used	27
2.2 Isolation and purification of CD14 ⁺ and CD4 ⁺ cells	29
2.3 Cell treatments	31
2.4 Angiogenesis assays	31
2.5 LDL isolation	32
2.6 LDL oxidation	35
2.7 DiI-labelling of LDL	36
2.8 Flow cytometry	36
2.9 RNA-seq sample preparation	38
2.10 ChIP-seq sample preparation	40
2.11 ATAC-seq sample preparation	42
3 Phenotypic and functional responses of vascular and immune cells to oxLDL	45
3.1 Introduction	46
3.2 LDL isolation and measurement of its oxidation	49
3.3 Oxidised LDL induces changes in cellular appearance	52
3.4 Oxidised LDL is taken up by cells in a concentration and time dependent manner	54

3.5	Oxidised LDL induces activation of vascular endothelial and smooth muscle cells.	59
3.6	Oxidised LDL induces cell death at high concentrations	62
3.7	Discussion and Summary	67
4	Development of custom bioinformatic and statistical analysis tools	75
4.1	Introduction	76
4.2	RNA-seq analysis: the <i>rseqR</i> package	81
4.3	ChIP- and ATAC-seq analysis: the <i>chompR</i> package	87
4.4	Trait-associated SNP enrichment: the <i>snpR</i> package	94
4.5	Discussion	97
4.6	Conclusions	102
5	Transcriptomic Changes in Vascular and Immune Cells in Response to oxLDL	103
5.1	Introduction	104
5.2	Total RNA isolation and quality	106
5.3	Quality control of raw reads	107
5.4	Abundance estimation and quality control of count data	114
5.5	Differential gene expression and pathway analysis	117
5.6	Genes differentially regulated by oxLDL are associated with cardiovascular risk variants	126
5.7	Discussion	126
6	Epigenetic Changes in Vascular and Immune Cells in Response to oxLDL	139
6.1	Introduction	140
6.2	Chromatin accessibility: ATAC-seq	144
6.3	Oxidised LDL-induced changes in enhancer activity: H3K27ac CHIP-seq analysis	157
6.4	Dynamic enhancer sites show strong enrichment for ETS family of transcription factors	164
6.5	Coronary artery disease associated variants overlap with dynamic accessible and enhancer elements	165
6.6	Discussion	167
7	General Discussion	172
8	Appendix	179
8.1	Published work	179
	Bibliography	205

List of Figures

1.1	Sequence of events that lead to atherosclerosis and its progression . . .	6
2.1	Lipoprotein isolation	34
2.2	Amplification of ATAC-seq libraries.	44
3.1	Western blot analysis of lipoprotein samples.	50
3.2	Measurement of the oxidation of LDL	52
3.3	Measurement of batch-to-batch variation in oxLDL.	53
3.4	Oil Red O staining and phase contrast microscopy of cells treated with oxidised LDL.	55
3.5	Flow cytometric analysis of HAECs treated with DiI-labelled oxLDL.	56
3.6	Oxidised LDL is taken up by cells in a concentration-dependent manner.	58
3.7	Oxidised LDL is taken up by cells in a time-dependent manner. . . .	60
3.8	Flow cytometric analysis of intercellular cell adhesion molecule 1 (ICAM1) expression in human aortic endothelial cells (HAECs) treated with oxLDL.	61
3.9	Oxidised LDL induces cell surface expression of ICAM1 in HAECs and VSMCs in a concentration-dependent manner.	63
3.10	Oxidised LDL induces cell surface expression of ICAM1 in HAECs and VSMCs in a time-dependent manner.	64
3.11	Oxidised LDL induces cell death at high concentrations.	66
3.12	Oxidised LDL at high concentrations induces cell death.	68
4.1	Library preparation for high-throughput sequencing	78
4.2	Generalised workflow for HTS data analysis	80
4.3	Workflow of the rseqR package.	82
4.4	Workflow of the chompR package.	88
4.5	Quality control plots of ChIP-seq and ATAC-seq experiments.	91
4.6	Workflow of the snpR package.	95
5.1	RNA quality.	107
5.2	Representative QC plots of raw reads as produced by the rseqR pipeline	108
5.3	Summary of PASS/WARN/FAIL results of FASTQC modules.	109
5.4	Heatmap plot of per tile sequence quality for a failing sample.	111
5.5	Per sequence GC content of RNA-seq samples.	112

List of Figures

5.6	Gene counts statistics	116
5.7	Heirarchical clustering and inter-correlation analysis of samples	118
5.8	Upset plots displaying the number of differentially expressed genes detected by edgeR, DESeq2, and limma+voom between oxLDL-treated and untreated cells	119
5.9	Differentially regulaed genes in response to oxLDL	120
5.10	Overlap of differential genes.	125
5.11	Genes differentially expressed in each cell type that are also associated with coronary artery disease (CAD) loci.	127
5.12	Heatmap of enriched pathway terms for differentially expressed genes across all four cell types.	138
6.1	Quality control of ATAC-seq DNA samples.	144
6.2	Summary of PASS/WARN/FAIL results of FastQC modules for raw untrimmed reads.	145
6.3	Summary of PASS/WARN/FAIL results of FastQC modules for trimmed reads.	147
6.4	QC plots for ATAC-seq alignment.	152
6.5	Changes in chromatin structure at promoters in response to oxLDL .	153
6.6	Representative chromatin accessibility changes at the CD36 gene locus for macrophages.	154
6.7	Dynamic accessibility sites are enriched for non-coding regulatory ele- ments.	156
6.8	Characterisation of dynamic enhancer sites.	161
6.9	OxLDL-induced changes at dynamic enhancer sites is associated with a concordant fold-change in nearby genes.	162
6.10	Enrichment of transcription factor motifs in dynamic enhancer sites. .	165

List of Tables

1.1	Properties of lipoproteins	7
1.2	Oxidative modification of oxLDL	10
1.3	Anti- and pro-inflammatory factors induced by oxLDL	12
1.4	Common histone methylation and acetylation modifications	20
1.5	Studies investigating the epigenetic effects of atherogenic lipids	25
2.1	ATAC-seq PCR recipe	43
2.2	ATAC-seq oligonucleotides used for PCR	44
3.1	Important atherogenic effects of oxLDL	48
4.1	Comparison of features of rseqR with other RNA-seq pipelines	99
4.2	Comparison of features of chompR with other ChIP-seq and ATAC-seq pipelines	101
5.1	Microarray studies of oxLDL-treated vascular and immune cells	106
5.2	Basic summary statistics of raw RNA-seq reads	110
5.3	Alignment of VSMC RNA-seq reads to ribosomal RNA database	113
5.4	Summary statistics of read mapping	115
5.5	Top ten most up- and down-regulated genes in CD4+ cells following treatment with oxLDL	134
5.6	Top ten most up- and down-regulated genes in HAECs following treatment with oxLDL	135
5.7	Top ten most up- and down-regulated genes in macrophages following treatment with oxLDL	135
5.8	Top ten most up- and down-regulated genes in VSMCs following treatment with oxLDL	136
5.9	Enriched pathway terms for differentially expressed genes	137
6.1	Genome-wide high throughput chromatin accessibility assays	143
6.2	Basic summary statistics of raw ATAC-seq reads	146
6.3	General and quality metrics of ATAC-seq alignment	149
6.4	Genomic distribution of open chromatin sites	151
6.5	Changes in chromatin accessibility following oxLDL treatment	155
6.6	General and quality metrics of ChIP-seq alignment	158

List of Tables

6.7	Genomic distribution of H3K27ac sites	160
6.8	Co-localisation of open chromatin and H3K27ac marked sites	160
6.9	Enriched pathway terms for dynamic enhancer sites	163
6.10	CAD-associated variants overlapping regulatory elements in untreated and oxLDL-treated vascular and immune cells	166
6.11	Annotation of CAD-associated variants overlapping with dynamic ac- cessible sites	168

Abbreviations

Abbreviation	Term
7-AAD	7-amino-actinomycin D
ABC	ATP-binding cassette
ACAT	acyl-CoA:cholesterol acyltransferase
APC	allophycocyanin
ATAC	assay for transposase-accessible chromatin
BAM	binary alignment map
BCA	bicinchoninic acid
BSA	bovine serum albumin
CAD	coronary artery disease
cAMP	cyclic adenosine monophosphate
cDNA	complementary DNA
ChIP	chromatin immunoprecipitation
DAS	dynamic accessible sites
DEG	differentially expressed gene
DEPC	diethylpyrocarbonate
DES	dynamic enhancer sites
DiI	1,1'-dioctadecyl-3,3',3'-tetramethylindocarbocyanine perchlorate
DMSO	dimethyl sulfoxide
DNA	deoxyribonucleic acid
EC	endothelial cell
ECM	extracellular matrix
EDTA	ethylenediaminetetraacetic acid
EGF	endothelial growth factor
eNOS	endothelial nitric oxide synthase
ER	endoplasmic reticulum
ET-1	endothelin 1
FCS	foetal calf serum
FDR	false discovery rate
FGF	fibroblast growth factor
FRiP	fraction of reads in peaks
GAGE	Generally Applicable Gene-Set Enrichment
GO	gene ontology
GREAT	Genomic Regions Enrichment of Annotation Tool
GSA	gene set analysis
GWAS	genome wide association study

Abbreviations

(continued)

Abbreviation	Term
HAEC	human aortic endothelial cell
HDAC	histone deacetylase
HDL	high density lipoprotein
HIF	hypoxia-inducible factor
HMOX1	haemoxygenase-1
HRP	horse radish peroxidase
HTS	high throughput sequencing
ICAM1	intercellular cell adhesion molecule 1
IDL	intermediate density lipoprotein
IFN γ	interferon γ
IGF	insulin-like growth factor
IL-1	interleukin-1
IL-2	interleukin-2
IL-8	interleukin-8
KEGG	Kyoto Encyclopaedia of genes and genomes
KLF	kruppel-like factor
LD	linkage disequilibrium
LDL	low density lipoprotein
LDLR	low density lipoprotein receptor
LFA1	lymphocyte function-associated antigen-1
lncRNA	long non-coding RNA
LPS	lipopolysaccharide
MACS	magnetic cell separation
MAPK	mitogen activated protein kinase
MCP-1	monocyte chemoattractant protein-1
MCSF	monocyte colony stimulating factor
MDA	malondialdehyde
MFI	mean fluorescence intensity
MI	myocardial infarction
miRNA	microRNA
MM-LDL	minimally-modified LDL
MMP	matrix metalloproteinase
MOPS	3-(N-morpholino)propanesulfonic acid
MPO	myeloperoxidase
mRNA	messenger ribonucleic acid
NB	negative binomial

Abbreviations

(continued)

Abbreviation	Term
ncRNA	non-coding RNA
NGS	next generation sequencing
NO	nitric oxide
NRF	Non-Redundant Fraction
NSC	normalised strand cross-correlation coefficient
oxLDL	oxidised LDL
PBC1	PCR bottlenecking coefficient 1
PBC2	PCR bottlenecking coefficient 2
PBMC	peripheral blood mononuclear cell
PBS	phosphate-buffered saline
PDGF	platelet-derived growth factor
PE	phycoerythrin
PPAR	peroxisome proliferator-activated receptor
PS	phosphatidylserine
PTM	post-translational modification
PUFA	polyunsaturated fatty acid
PVDF	polyvinylidene fluoride
REM	relative electrophoretic mobility
RNA	ribonucleic acid
ROS	reactive oxygen species
rpm	revolutions per minute
RSC	relative strand cross-correlation coefficient
SAM	sequence alignment map
SDS	sodium dodecyl sulphate
siRNA	short interfering RNA
SNP	single nucleotide polymorphism
SR	scavenger receptor
taSNPs	trait-associated SNPs
TBARS	Thiobarbituric acid reacting substances
TGF β	transforming growth factor β
TNF α	tumour necrosis factor α
TSA	trichostatin A
TSS	transcriptional start site
VCAM1	vascular cell adhesion molecule 1
VEGF	vascular endothelial growth factor
VEGFR	vascular endothelial growth factor receptor

Abbreviations

(continued)

Abbreviation	Term
VLDL	very low density lipoprotein
VSMC	vascular smooth muscle cell

Chapter 1

Background

Contents

1.1	Introduction	2
1.2	Mechanisms of atherosclerotic plaque progression	3
1.3	Biology of oxidised low density lipoprotein	5
1.3.1	Overview of lipoproteins	5
1.3.2	Mechanisms of LDL oxidation	7
1.3.3	Types of oxLDL	9
1.4	Effects and atherogenicity of oxLDL	9
1.4.1	Cytotoxicity of oxLDL	9
1.4.2	Inflammatory effects of oxLDL	11
1.4.3	Effects of oxLDL on endothelial cells	12
1.4.4	Effects of oxLDL on macrophages	15
1.4.5	Vascular smooth muscle cells and oxLDL	17
1.4.6	T cells and oxLDL	18
1.5	Epigenetic regulation of atherosclerosis	18
1.5.1	Epigenetic mechanisms	18
1.6	Aims and objectives	23

1.1 Introduction

Atherosclerotic vascular disease is the leading cause of morbidity and mortality worldwide^{1,2}. It is a disease of medium-to-large sized vessels arising from focal accumulation of lipid in the arterial wall. It is a chronic and progressive process that generally begins in childhood as ‘fatty streaks’ and manifests clinically in late adulthood with plaque formation and rupture leading to occlusion of vessels with consequences such as myocardial infarction (MI) and ischaemic stroke³. One of the key risk factors for atherosclerosis is high circulating plasma levels of low density lipoprotein (LDL), the oxidised form of which is extremely atherogenic. Atherosclerosis affects

mainly medium-to-large-sized arteries, and manifests predominantly at sites of arterial injury and at sites where there is non-laminar flow of blood (e.g. bifurcations). The earliest lesion of atherosclerosis is the ‘fatty streak’, which is an asymptomatic lesion characterised by the accumulation of foam cells⁴. It typically develops in childhood and is potentially reversible. Foam cells arise mainly from circulating monocytes that enter the sub-endothelial space, transform into macrophages and then take up and accumulate excess lipid⁵. Progression of fatty streaks leads to intimal thickening and formation of a fibrous plaque, a lesion primarily composed of layers of vascular smooth muscle cells (VSMCs) in a proteoglycan-collagen matrix with an underlying lipid pool consisting of an acellular area, rich in hyaluronan and proteoglycans. Outside this lipid pool, there is accumulation of macrophage and VSMC-derived foam cells. In more advanced disease, the lesions are described as fibro-atheromatous and consist of an acellular necrotic core covered by a thick fibrous cap composed of endothelial cells (ECs) and VSMCs and extracellular matrix (ECM). Eventually, the fibrous cap may thin and become vulnerable to rupture. Once this happens, the necrotic core contents are exposed to circulating blood, resulting in activation of coagulation and thrombotic cascades, with subsequent ischaemia and infarction of downstream tissues and organs. This results in the clinical presentation of diseases such as angina, myocardial infarction, transient ischaemic events, cerebrovascular accidents, peripheral vascular disease and mesenteric ischaemia.

1.2 Mechanisms of atherosclerotic plaque progression

It is well established that high levels of circulating LDL lead to atherosclerotic vascular disease⁶. Several mechanisms by which LDL might lead to and cause progression of atherosclerosis have been proposed. Whilst the details regarding the initiating step of

atherosclerosis differ between these explanations, what is accepted is that oxidative modification of LDL by components of the vascular wall, in particular by ECs⁷, is an essential step in atherosclerosis and results in EC injury and activation (Figure 1.1). The consequence of EC activation by oxidised LDL (oxLDL) is the induction of cellular adhesion molecules on the surface of ECs and expression of chemokines, which stimulate the transmigration of monocytes and T lymphocytes to the sub-endothelial space⁸.

Within the sub-endothelial space, under the influence of local growth factors, monocytes differentiate into macrophages which express several scavenger receptors (SRs) that facilitate the uptake of modified lipoproteins⁹, leading to formation of foam cells¹⁰. These foam cells, in turn, produce cytokines that increase the activation of ECs and modulate the function of other cells involved in atherosclerosis¹¹. This migration of monocytes and activation continues even in established lesions and not just in the initial stages of atherosclerosis¹² and is a key driver for the progression of fatty streaks into more advanced lesions.

The prominent subclass of T cells recruited into the sub-endothelial lesions of atherosclerosis are CD4⁺ T cells, which are activated by presentation of oxLDL-derived antigens. The exact location of antigen presentation is not clear, but there are dendritic cells located both in the periphery of and within atherosclerotic lesions¹³. The predominant subtype of CD4⁺ T cells in atherosclerotic lesions is the Th1 subtype. These cells secrete pro-inflammatory cytokines, which augment the production of other inflammatory cytokines by macrophages and vascular cells, thus perpetuating the inflammatory process. Th2 cells are rarely found in atherosclerotic lesions and their role is unclear. Th17 cells have been identified in lesions and have been reported to have both pro-atherogenic and anti-atherogenic roles¹⁴.

Growth factors and chemokines produced by the activated endothelium and infiltrating inflammatory cells induce a change in VSMCs from a quiescent state to one where they migrate from the media to the intima and proliferate. Migration of VSMCs is facilitated by production of a broad range of matrix metalloproteinases (MMPs) by the VSMCs themselves; this is partly induced by inflammatory cytokines produced by activated ECs^{15–17}. Within the intima, VSMCs acquire a synthetic phenotype resulting in deposition of ECM proteins and formation of a fibrous cap¹⁸. VSMCs can also take up oxLDL and form foam cells. In more advanced lesions, VSMCs undergo apoptosis leading to plaque instability and rupture. This is partly through the direct toxic effects of oxLDL itself, but also via induction of surface death receptors by inflammatory cytokines.

1.3 Biology of oxidised low density lipoprotein

1.3.1 Overview of lipoproteins

Lipids are insoluble in aqueous solution and are therefore carried in plasma complexed with proteins. These lipoprotein particles contain a core of cholesterol esters and triglyceride covered by a surface mono-layer of phospholipids, free cholesterol and apoproteins. The major classes of lipoprotein found in the plasma are defined by their density and termed chylomicrons, very low density lipoprotein (VLDL), intermediate density lipoprotein (IDL), LDL and high density lipoprotein (HDL)¹⁹. Lipoproteins are synthesised in the liver or the intestine with a corresponding structural apoprotein. Table 1.1 lists the major properties of the different lipoproteins¹⁹.

Lipoproteins can be distinguished from one another by a variety of methods including lipid content, density, size, electrophoretic mobility and the proteins on their surfaces¹⁹.

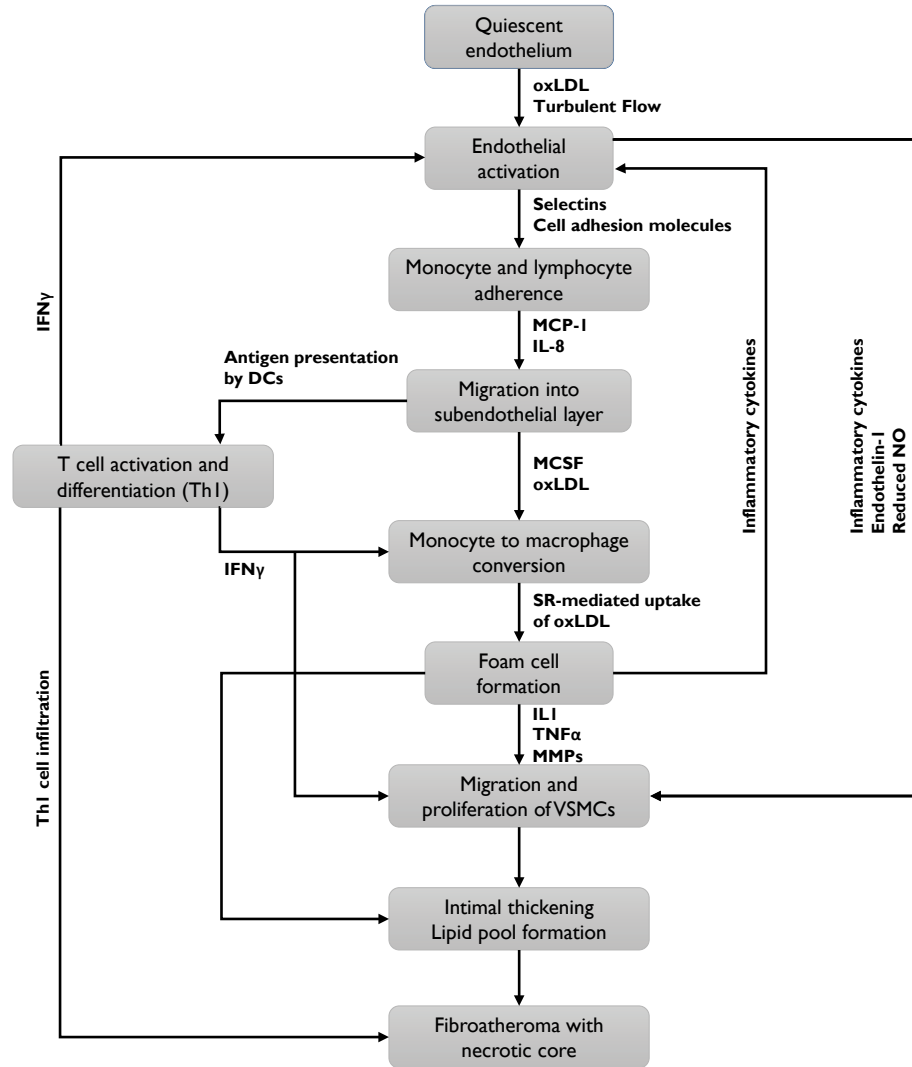


Figure 1.1: Sequence of events that lead to atherosclerosis and its progression. Activation of the endothelium is triggered by oxLDL. Monocyte attraction, rolling, and attachment is facilitated by secretion of monocyte chemoattractant protein-1 (MCP-1), and expression of cell adhesion molecules. In the arterial intima, the presence of MCP-1, monocyte colony stimulating factor (MCSF) and oxLDL trigger maturation of monocytes into macrophages and subsequent proliferation. Macrophages express SRs that facilitate the uptake of modified LDL and the conversion to foam cells. After antigen-specific activation, T cells enter the intima and secrete interferon γ ($\text{IFN}\gamma$) which send signals that exaggerate and sustain the inflammatory response. Activation of VSMCs promotes their migration and proliferation, which is facilitated by production of MMPs. Foam cell and VSMC migration and proliferation results in pathological intimal thickening, underlying which is a lipid pool. Subsequently, T cells infiltrate the lesion, and macrophages and VSMCs apoptose resulting in conversion of the lipid pool to a necrotic core. The lesion also develops a fibrous cap composed of ECs and VSMCs in a proteoglycan-collagen matrix.

Table 1.1: Properties of lipoproteins.

Lipoprotein	Density (g/ml)	Electrophoretic mobility	Source
chylomicrons	<0.95	origin	intestine
very low density lipoproteins (VLDL)	0.95-1.006	pre- β	liver
intermediate density lipoproteins (IDL)	1.006-1.009	broad β	catabolism of VLDL and chylomicrons
low density lipoproteins (LDL)	1.019-1.063	β	catabolism of VLDL
high density lipoproteins (HDL)	1.063-1.21	α	catabolism of VLDL and chylomicrons, liver and intestine

The different classes of lipoproteins vary in their contribution to atherosclerotic risk. LDL, chylomicron remnants and IDL are known to be pro-atherogenic, whereas HDL levels are thought to be inversely related to risk of atherosclerosis. The main cholesterol carrying lipoproteins are LDL and HDL. In a normal individual, LDL contains approximately 70% and HDL 20% of the total plasma cholesterol²⁰.

1.3.2 Mechanisms of LDL oxidation

The critical factor determining the susceptibility of LDL to oxidation is the abundance of polyunsaturated fatty acids (PUFAs) in the particle. Other factors influencing the susceptibility of LDL to oxidation (at least *in vitro*) include particle size, composition of other lipids and presence of lipoprotein(a)²¹.

oxLDL has long been thought to be present only in tissues and not found in the circulating plasma. However, with the development of methods to measure oxLDL levels in plasma²²⁻²⁴, it is now clear that oxLDL is transferred between tissues and plasma and does not merely accumulate in lesions. Several groups have now shown

that plasma oxLDL levels in patients with or at high risk of coronary artery disease (CAD) are significantly greater than those measured in healthy individuals²⁵⁻²⁸. It has also been shown that plasma oxLDL levels increase in patients with acute MI or following acute cerebrovascular events.

Oxidative modification is thought to occur primarily within the intima of the arterial wall²⁹. Significant amounts of oxidation are not thought to occur in plasma due to the presence of antioxidants such as vitamin C and E and removal by SRs present on hepatic sinusoidal Kupffer cells.

The means by which cell mediated oxidative modification of LDL occurs is unclear, but *in vitro* experiments have identified several potential, non-exclusive, mechanisms²¹ by which this may occur. ECs, VSMCs and macrophages are all capable of oxidising LDL in cell culture³⁰. Release of superoxide anion, which can subsequently be broken down to form hydrogen peroxide, is thought to be the major pathway in cultured VSMCs^{30,31}. The generation of lipoperoxides by the action of 15-lipoxygenase is thought to be the major pathway of oxidation in macrophages and ECs³⁰. Macrophages also produce large amounts of superoxide and myeloperoxidase (MPO) as part of their inflammatory response³². MPO amplifies the oxidative potential of hydrogen peroxide by generating cytotoxic oxidants³³.

Nitric oxide (NO) also plays an important role in lipid oxidation. NO released by the action of endothelial nitric oxide synthase (eNOS) in ECs of the artery wall plays a critical role in normal physiology³⁴. In the normal state, NO reacts with superoxide in a protective manner preventing its oxidative effects³⁵. However, when NO production becomes dysregulated as occurs in chronic inflammatory states, the reaction of NO with superoxide results in excessive generation of peroxynitrite, a powerful oxidising intermediate³⁵. Peroxynitrite oxidises LDL lipids and converts the lipoprotein into a

form recognised by SRs.

Finally, transition metal ions have also been shown to oxidise LDL *in vitro*. However, there is insufficient evidence to show that this mechanism is also applicable *in vivo* and patients with haemochromatosis or Wilson's disease (in which there are elevated levels of iron and copper in tissues and the blood) are not at increased risk for atherosclerosis²⁹.

1.3.3 Types of oxLDL

The oxidation of LDL is a complex process during which both the protein and lipid components undergo modification. These modifications can range from minimal to extensive and the products formed may include just one, some, or all of those shown in Table 1.2. Oxidatively modified LDL also has an increased buoyant density, an increased negative charge, a loss of enzyme activity associated with LDL, and is not recognised by the low density lipoprotein receptor (LDLR)³⁶. Instead, oxLDL is recognised by specific receptors termed SRs, of which several have now been identified^{37,38}.

1.4 Effects and atherogenicity of oxLDL

1.4.1 Cytotoxicity of oxLDL

OxLDL is able to induce cell death via both apoptosis and necrosis in a wide range of cell types including ECs, VSMCs, and human monocyte derived macrophages^{39–42}. The degree of cytotoxicity of oxLDL depends on the nature and extent of oxidation. For example, lipoxygenases catalyse peroxidation of fatty acids, which generates a less

Table 1.2: Oxidative modification of oxLDL.

Oxidation Products
Fatty acid
Free and esterified fatty acid hydroxides
Free and esterified fatty acid peroxides
Aldehydes
Lipids
Lysophosphatidylcholine
Oxysterols
Internally modified phosphatidyl ethanolamine/serine products
Proteins
Protein carbonyls
Non-enzymatic proteolyzed fragments
Modified amino acids
Protein cross-linking due to modified tyrosines and aldehydes
Lipid-protein adducts
Conformational changes which may lead to antigenicity

cytotoxic oxLDL than that made with hypochlorite, which oxidises the apolipoprotein B100 protein⁴³. Mechanisms by which oxLDL may cause cell death include:

- induction of intracellular oxidant production by decreasing antioxidants such as glutathione or increasing production of reactive oxygen species (ROS)^{44–46};
- activation of caspases 3 and 9 and release of mitochondrial activators of caspases^{41,47–49} reduces expression of B-cell lymphoma 2 (Bcl-2) and cellular inhibitor of apoptosis protein 1 (c-IAP-1), both of which are anti-apoptotic proteins⁴⁷;
- upregulation of Fas surface expression, which triggers apoptosis when activated by its ligand, FasL⁴⁷;
- down-regulation of angiopoietin-2 in ECs, which results in apoptosis, possibly by reducing expression of Survivin⁵⁰.

1.4.2 Inflammatory effects of oxLDL

OxLDL is composed of a complex mixture of several bioactive compounds, each of which has independent and sometimes opposing cellular effects. Therefore, it is not surprising that oxLDL has been reported to have both pro- and anti-inflammatory effects (Table 1.3). The mode and extent of oxidation of LDL and the cell system being studied all play an important role in determining the overall pro- or anti-inflammatory outcome. The pro-inflammatory effects are primarily manifested through the transcription factors $\text{NF}\kappa\text{B}$, AP-1, STAT 1/3, NFAT, SP-1, and HF-1, while the anti-inflammatory effects are mediated by the activation of peroxisome proliferator-activated receptors (PPARs), Nrf2, and haemoxygenase-1 (HMOX1)⁵¹.

In addition to stimulating the expression of the transcription factors, the components of oxLDL exert their pro- and anti-inflammatory effects at other levels as well:

- in the extracellular space, oxidized phospholipids competitively inhibit the binding of bacterial lipopolysaccharide (LPS) to TLR-4, thus inhibiting LPS-mediated inflammatory reactions;
- oxidized fatty acids inhibit the tumour necrosis factor α ($\text{TNF}\alpha$)-induced upregulation of adhesion molecules and activate PPARs⁵²;
- in the nucleus, oxidised lipids have been shown to inhibit DNA binding and transactivation of $\text{NF}\kappa\text{B}$;
- oxidised phosphatidylcholines in minimally-modified LDL (MM-LDL) has been shown to induce HMOX1 in ECs and VSMCs⁵³. HMOX1 has both anti-oxidant and anti-inflammatory properties⁵⁴;
- oxidised phospholipids also elevate cyclic adenosine monophosphate (cAMP) levels in ECs, which leads to inhibition of adhesion molecule synthesis and

Table 1.3: Anti- and pro-inflammatory factors induced by oxLDL.

Transcription Factor	Active Component of OxLDL	Selected Target Genes and Cellular Effects
Anti-inflammatory		
PPAR α	9-HODE, 13-HODE, HETE, Free fatty acids	Increased lipolysis and fatty acid oxidation; decreased NF κ B activity, and VCAM1 and Tissue Factor expression
PPAR γ	Prostaglandin PGJ2, 9-HODE, 13-HODE, Arachidonate PAF-like products	SR-B1, CD-36, LPL, increased insulin sensitivity, adipocyte differentiation
Nrf2 Smad3	4-HNE Lipid hydroperoxides -	Decreased VCAM1, increased HMOX1 PAI-1
Pro-inflammatory		
NF κ B	LPC, 13-HODE	Inflammatory cytokines, immune receptors, adhesion molecules, impaired Glut4 function
AP-1	LPC, Oxidized PC	TNF, Osteopontin, Endothelin-1, ABCA1, TGF, MMP9
STAT 1/3	-	Cytokines, apoptotic pathway, sPLA2, cPLA2
NFAT	Oxidized PC, ROS	Cytokines, angiogenesis, tumorigenicity
Sp-1	7-keto cholesterol	VCAM1, ICAM1, TGF, PDGF β , Osteopontin, Tissue Factor
(HF-1)	-	VEGF1, PAI-1 COX-2, VCAM1, IL1 β

secretion of inflammatory cytokines, possibly through an inhibition of the NF κ B pathway⁵⁴.

1.4.3 Effects of oxLDL on endothelial cells

1.4.3.1 Endothelial cell activation and dysfunction

The vascular endothelium is responsible for regulation of vascular tone, angiogenic remodelling, thrombogenesis, inflammation, and platelet and leucocyte interaction with the vessel wall. Given this diversity of function, the definition of EC activation is not completely straightforward, but may be viewed as a switch from a quiescent phenotype that is anti-thrombogenic and anti-inflammatory, toward a phenotype that involves the host defence response, with increased expression of inflammatory cytokines,

chemokines and cell adhesion molecules⁵⁵. EC activation is part of the normal host defence and physiology; however, sustained activation leads to EC dysfunction, with dysregulation of vascular tone and promotion of a pro-atherogenic and pro-thrombotic state.

A switch from NO-mediated silencing of cellular processes to activation by redox signalling is a key feature of EC activation and dysfunction, and can occur in the face of a number of insults or following exposure to a number of stimuli, including oxLDL⁴⁷. In the quiescent state, basal production of NO by eNOS results in inhibition of pro-inflammatory and cell adhesion molecule gene expression⁵⁶, inhibition of fusion of secretory vesicles containing P-selectin and chemokines⁵⁷, and inhibition of leucocyte activation⁵⁸. This is partly achieved by s-nitrosylation of cysteine residues, resulting in reduced activity of a wide range of target proteins including NF κ B, cell-cycle proteins and proteins involved in generation of tissue factor⁵⁹, and partly by limitation of oxidative phosphorylation in mitochondria⁶⁰. In activated ECs, signalling via ROS predominates. Like NO, ROS are able to diffuse through the cell and alter protein function; however, by contrast, the consequence of this is phosphorylation of transcription factors, induction of genes involved in chromatin remodelling and transcription, and protease activation. This in turn induces expression of cellular adhesion molecules on the surface of ECs and expression of chemokines, such as MCP-1 and interleukin-8 (IL-8), which stimulate the transmigration of monocytes and T lymphocytes to the sub-endothelial space⁸.

1.4.3.2 Effects of oxLDL on NO signalling in endothelial cells

OxLDL reduces the activity of endothelial NO by several mechanisms⁴⁷:

- depletion of EC caveolae of cholesterol, resulting in displacement of eNOS;

- increased cellular production of ROS, thus inactivating NO;
- activation of endothelial arginase II by causing it to dissociate from the microtubule cytoskeleton; this competes with eNOS for the common substrate L-arginine.

OxLDL also stimulates the production of endothelin 1 (ET-1), a potent vasoconstrictor and proinflammatory peptide which antagonises the effects of NO.

1.4.3.3 Changes in expression of inflammatory cytokines, chemokines and cell-cell adhesion molecules

Incubation of ECs with oxLDL significantly increases MCP-1 expression via activation of a mitogen activated protein kinase (MAPK) pathway⁶¹. Other chemokines upregulated in response to oxLDL include IL-8, and chemokine (C-X-C motif) ligands 2 and 3 (CXCL2 and CXCL3)⁴⁷. OxLDL significantly increases expression of cellular adhesion molecules in ECs via activation of an NF κ B-dependent pathway that can be inhibited by pre-treatment of cells with statins^{62,63}. Incubation of human coronary artery ECs (HCAECs) with oxLDL also results in increased expression of CD40 and CD40L⁶⁴; signalling via the stimulation of CD40/CD40L is known to activate pro-atherogenic pathways⁶⁵.

1.4.3.4 Other effects of oxLDL on endothelial cells

MMP activity is important in atherosclerotic plaque stability, and expression of these enzymes is increased in atherosclerosis⁶⁶. OxLDL increases the expression of MMP-1 and MMP-3 mRNA and protein in human coronary arterial ECs⁶⁷ and MMP-9 expression in human aortic endothelial cells (HAECs)⁶⁸.

Angiogenesis is an important feature of late atherosclerosis and low concentrations of oxLDL appear to have a significant angiogenic effect, partly due to activation of the PI3K/Akt/eNOS pathway and by upregulation of vascular endothelial growth factor (VEGF). Microarray studies indicate that oxLDL enhances expression of multiple receptors involved in transmission of pro-angiogenic signalling, particularly via vascular endothelial growth factor receptor (VEGFR) 2 and Notch⁶³. OxLDL also stimulates tube formation by ECs *in vitro*⁶⁹.

1.4.4 Effects of oxLDL on macrophages

1.4.4.1 Normal cholesterol metabolism in macrophages

Once taken up by the cell, LDL particles are transferred to the lysosomal compartment and metabolised by an array of hydrolytic acid lipases and proteases, leading to degradation of the particle to its components, including cholesterol^{70,71}. The resulting free cholesterol is then transported to the endoplasmic reticulum (ER) where it is esterified by the enzyme acyl-CoA:cholesterol acyltransferase (ACAT), resulting in formation of cholesteryl ester that is stored in cytoplasmic inclusions. To prevent accumulation within cells, this excess cholesterol is cleared by efflux pathways; this requires the presence of an appropriate extracellular acceptor, mainly HDL and other apoA1 containing lipoproteins. Excess cholesterol can then be exported out of the cell via several pathways. Active cholesterol efflux is mediated by the membrane-bound lipid ATP-binding cassette (ABC) translocases, ABCA1 and ABCG1, and these comprise the major efflux pathways^{72,73}. Additional efflux capacity is provided by simple diffusion (aqueous diffusion pathway)⁷⁴ or facilitated diffusion, mediated by SR-B1⁷⁵. Cholesterol effluxed from the cells is accepted by nascent or mature HDL molecules and transported to the liver, where the cholesterol is de-esterified and

secreted into bile.

1.4.4.2 Foam cell formation

In macrophages located within atherosclerotic lesions, cholesterol efflux pathways are impaired leading to massive accumulation of intracellular cholesterol and foam cell formation. This impairment is a direct result of oxLDL, and not seen with cells exposed to native LDL^{76,77}. The exact mechanisms by which the efflux pathways are impaired is unclear, but the following are thought to contribute:

- scavenger receptors take up oxLDL 3-10x faster than native LDL is taken up by the LDLR; this may result in saturation of cholesterol efflux pathways;
- oxLDL is resistant to cathepsins and other lysosomal enzymes, resulting in retention and deposition of oxLDL components within lysosomal compartments⁷⁸;
- oxLDL (in particular the oxysterol components) causes dysfunction of ACAT and ABCA1 and ABCG1 transporters leading to impaired efflux of cholesterol from cells to HDL.

1.4.4.3 Activation of macrophages by oxLDL

Regardless of foam cell formation, oxLDL induces the expression of a number of genes associated with inflammation in macrophages, such as MCP-1, serum amyloid A, and HMOX1⁷⁹. Moreover, macrophage activation induces the release of pro-inflammatory cytokines (interleukin-1 (IL-1) β , TNF α), ROS and MMPs, all of which are associated with progression of inflammation¹⁰. Internalized oxLDL also provides oxidised lipids as ligands for the PPAR γ pathway, upregulating expression of the scavenger receptor CD36, facilitating, in turn, the internalization of more oxLDL^{80,81}. The PPAR γ

pathway also regulates the expression of the enzymes MPO and 12/15-lipoxygenase, which are thought to participate in the oxidization of new LDL, thus increasing the local pool of oxLDL^{82,83}.

1.4.4.4 Macrophage retention within atherosclerotic lesions

The internalization of oxLDL results in inhibition of macrophage migration, favouring cell spreading and the activation of focal adhesion kinase, in a process mediated by src-kinases and oxidative stress⁸⁴. OxLDL-CD36 interaction also induces the loss of cell polarization in macrophages, an essential process to cell migration⁸⁵. Thus, the evidence suggests that oxLDL not only participates in monocyte differentiation and macrophage activation, but also macrophage retention.

1.4.5 Vascular smooth muscle cells and oxLDL

The migration into and focal proliferation of VSMCs within the tunica intima are hallmarks of atherosclerosis. A key mechanism for this phenomenon by oxLDL is the induction of platelet-derived growth factor (PDGF), fibroblast growth factor (FGF), insulin-like growth factor (IGF)-1 and endothelial growth factor (EGF) expression by ECs and macrophages, and induction of their receptors by VSMCs⁸⁶⁻⁸⁸.

OxLDL has also been shown to induce changes directly in VSMCs. Incubation of VSMCs with oxLDL results in increased production of ECM⁸⁹. The production of interstitial collagen and elastin leads to the building of the fibrous cap that covers the developing atherosclerotic plaque⁹⁰. As already described, oxLDL can also stimulate cell death. In VSMCs, this results in formation of the necrotic core of the atherosclerotic plaque, and also contributes to plaque instability and rupture in the final stages of

atherosclerosis.

1.4.6 T cells and oxLDL

OxLDL presented by antigen presenting cells is a potent pro-inflammatory stimulus for T lymphocytes, and results in activation of effector T cell (in particular CD4⁺ Th1 cells)⁹¹. These T cells respond to oxLDL found in atherosclerotic plaques by proliferation and expression of inflammatory cytokines^{13,92,93}. Activated T cells also express CD40L (CD154), which binds to CD40 on the macrophage cell surface promoting expression of TNF receptors, lymphocyte function-associated antigen-1 (LFA1) and intercellular cell adhesion molecule 1 (ICAM1). On ECs, binding of CD40L triggers the increased expression of a number of cell adhesion molecules, MMPs, and VSMC-stimulating growth factors. OxLDL can also induce apoptosis in CD4⁺ cells via activation of a Fas-dependent pathway⁹⁴.

Evidence from animal models has shown that induction of neonatal tolerance to oxLDL not only reduces T cell immune responses, but also reduces atherosclerosis, suggesting that abrogating the response of T cells to oxLDL can modify the course of the disease⁹³.

1.5 Epigenetic regulation of atherosclerosis

1.5.1 Epigenetic mechanisms

There is increasing evidence that many of the effects of oxLDL are mediated by epigenetic mechanisms, including deoxyribonucleic acid (DNA) methylation, histone post-translational modification (PTM), and non-coding RNAs (ncRNAs). These

processes affect chromatin structure and gene expression without alteration of the underlying DNA sequence, and have been demonstrated to exert both short- and long-term effects, with the potential to be inherited. The basic structural unit of chromatin is the nucleosome, which consists of a histone octamer complex, comprising two each of histones H2A, H2B, H3 and H4, and which binds and wraps 146 base pairs of DNA^{95,96}. Linking of individual nucleosomes is achieved by additional histones such as H1, which also play an important role in stabilising higher order structures of chromatin⁹⁷. Chromatin is not uniformly distributed within the nucleus, and staining of DNA identifies weakly staining domains termed euchromatin, and strongly staining chromatin domains known as heterochromatin. At the molecular level, euchromatin is associated with relaxation of local chromatin structures and accessibility of DNA to gene regulatory proteins and ribonucleic acid (RNA) polymerase, thus favouring transcription, whereas heterochromatic DNA is compacted with reduced accessibility and thus associated with reduced gene expression⁹⁸. Many (but not all) of the epigenetic mechanisms that have been identified result in changes in gene expression by altering the accessibility of chromatin and driving a switch from a euchromatic to a heterochromatic state or vice versa.

DNA methylation is achieved through the addition of a methyl group to the 5' position of cytosine residues within CpG dinucleotides. It leads to transcriptional repression by acting as a site for recruitment of inhibitory chromatin remodelling factors and by preventing binding of transcription factors^{99–101}. DNA methylation modifications are typically stable and are involved in processes such as silencing of repetitive elements, X-inactivation, imprinting and oncogenesis¹⁰².

Histone PTMs modify their physical interaction and spatial distribution resulting in changes in chromatin structure and affinities for chromatin-binding proteins, thus

Table 1.4: Common histone methylation and acetylation modifications.

Histone modification	Putative function
H2AK5ac	transcriptional activation; found at both promoters and enhancers
H2BK5ac	transcriptional activation; found predominantly at promoters
H3K4me1	identifies active and primed enhancers and other distal elements; is also enriched downstream of transcription starts
H3K4me2	transcriptional activation; associated with both promoters and enhancers
H3K4me3	identifies active or poised promoters
H3K9ac	mark of active regulatory elements with preference for promoters
H3K9me1	transcriptional repression; located at 5' end of genes
H3K9me3	repressive mark associated with constitutive heterochromatin and repetitive elements; Rb-mediated silencing
H3K27ac	mark of active regulatory elements; distinguishes active enhancers and promoters from poised ones
H3K27me3	repressive mark established by polycomb complex activity associated with repressive domains and silent developmental genes
H3K36me3	elongation mark associated with transcribed portions of genes, with preference for 3' regions after intron 1
H3K79me2	telomeric silencing
H4K20me1	transcriptional silencing; mitotic condensation

affecting gene expression. Histone modifications include methylation, acetylation, phosphorylation, sumoylation, and ubiquitylation. Over a hundred PTMs have been identified, and whilst many are poorly understood, some of the more common modifications, in particular the acetylation and methylation modifications, have been well studied and characterised^{103,104} (Table 1.4). Even so, the specific outcome of a particular histone PTM is not always easy to predict and depends on a number of factors, including the degree of modification, the presence or absence of neighbouring modifications, and the cell type.

Non-coding RNA molecules are divided into two major families: long non-coding RNAs

(lncRNAs) and short ncRNAs. Long ncRNAs are able to regulate gene expression by complexing with chromatin-modifying proteins or transcription factors and recruiting them to specific sites in the genome; some lncRNAs also act as precursors for short ncRNAs. The short ncRNAs affect gene expression in a number of ways. Both microRNAs (miRNAs) and short interfering RNAs (siRNAs) bind to target messenger RNA and block its translation or induce cleavage and degradation¹⁰⁵. Small interfering RNAs can also induce formation of a RNA-induced transcriptional silencing complex which promotes H3K9 methylation and chromatin condensation. The effects of miRNAs are even more diverse: firstly, they are both regulated by, and regulators of epigenetic marks¹⁰⁶; and, secondly, epigenetic marks and miRNAs can co-operate to affect common targets¹⁰⁸.

Recent studies have shown these epigenetic mechanisms play important roles in the pathogenesis of atherosclerosis. The most studied epigenetic changes in atherosclerosis are those related to DNA methylation. Altered patterns in global DNA methylation are associated with cardiovascular disease and clinical outcomes. Measurement of global DNA methylation levels in peripheral blood mononuclear cells (PBMCs) showed that hypermethylation was found in patients with raised inflammatory markers and was associated with all-cause and cardiovascular mortality¹⁰⁹. Furthermore, patients with angiographically-proven CAD had higher levels of global DNA methylation in PBMCs compared to controls¹¹⁰. The degree of methylation was also positively correlated with plasma homocysteine levels, which is an independent risk factor for CAD¹¹¹. Subsequent to these findings, studies investigating endothelial dysfunction, the exposure of vascular cells to atherogenic lipids and homocysteine, and inflammation have identified changes in methylation status in a number of genes associated with atherosclerosis. For example, exposure of ECs to LDL results in hypomethylation of

the *p66shc* promoter¹¹²; in hypercholesterolaemic mice, this gene mediates endothelial dysfunction and fatty streak formation^{112,113}. LDL also causes hypermethylation of the transcription factor kruppel-like factor (KLF) 2 in endothelial cells, with consequent down-regulation of its target genes, eNOS, thrombomodulin, and plasminogen activating inhibitor-1¹¹⁴. As part of a homeostatic negative feedback loop, oxLDL has been demonstrated to induce hypermethylation of the promoter of the oxLDL receptor, LOX1, thus reducing its expression¹¹⁵.

The link between atherogenesis and histone modifications has also been investigated. For example, atherogenic shear-stress profiles have been shown to result in globally decreased histone H4 acetylation¹¹⁶. Histone deacetylase (HDAC) 3 has been shown to be upregulated in atherosclerosis-prone areas, and when knocked down in APOE^{-/-} mice severe atherosclerosis occurred¹¹⁷; conversely, in LDLR^{-/-} mice, knockdown of HDAC9, both systemically and in macrophages, reduced atherosclerosis. This was thought to be mediated by increased promoter acetylation of ABCA1, ABCG1, and PPAR γ ¹¹⁸. ECs incubated with oxLDL have reduced expression of HDAC1 and HDAC2, both of which are involved in transcriptional repression¹¹⁹. Administration of the HDAC inhibitor trichostatin A (TSA) to LDLR^{-/-} mice worsened neo-intimal lesions and increased expression of the oxLDL receptor CD36¹²⁰. Furthermore, TSA increased TNF α and vascular cell adhesion molecule 1 (VCAM1) expression and inhibited VSMC proliferation when administered to aortic plaques of rats¹²¹. Histone methylation changes are also important in atherogenesis. For example, a long-lasting pro-atherogenic macrophage phenotype can be induced in monocytes by stimulation with oxLDL via histone PTMs resulting in increased H3K4me3 marks at the promoter regions of TNF α , MMP2, MMP9, CD36, SRA, MCP-1, IL-6, and IL-18¹²².

Non-coding RNAs have also been identified as key regulators of atherogenesis. Recently,

the anti-sense RNA derived from the *INK4* locus (ANRIL) has been shown to be a lncRNA modulating epigenetic mechanisms. Levels of ANRIL are correlated with severity of atherosclerosis¹²³, and certain single nucleotide polymorphisms (SNPs) in the *INK4* locus result in differences in chromatin interactions with induction of pro-atherogenic genes^{124,125}. Several miRNAs have also been shown to regulate atherosclerosis. For example, miR-21 has been shown to be upregulated in rat atherosclerotic models¹²⁶, and in human studies of atherosclerosis obliterans, to modulate a hypoxia-inducible factor (HIF)-1 α /tropomyosin1 pathway essential for atherosclerosis¹²⁷. In response to atherogenic shear stress, miR-21 also inhibits PPAR γ translation and consequently reduces expression of cell adhesion molecules, which are essential for atherosclerosis¹²⁸. MiR-92 is pro-atherogenic miRNA that is induced by low shear stress and exerts its actions by inhibition of KLFs 2 and 4 translation. OxLDL strongly potentiates the induction of miR-92 by low shear stress, and inhibition of miR-92 reduces oxLDL-induced expression of inflammatory cytokines and cell adhesion molecules¹²⁹.

The above examples illustrate that a vast array of epigenetic mechanisms regulate atherogenesis. Table 1.5 summarises those studies that have shed light specifically on the epigenetic mediators of the effects of oxLDL.

1.6 Aims and objectives

Atherosclerosis-related diseases are the leading causes of death worldwide. An established aetiological factor in atherosclerosis is oxLDL, but as is clear from the discussion above, the molecular events that mediate its effects are complex and our understanding of these events remains unclear. In this study we seek to define the

responses of the vascular and immune cells involved in atherosclerosis to oxLDL at the transcriptional and gene regulatory levels, and to integrate this data with that derived from genome-wide association studies.

In chapter 2, we provide detailed methods for the experiments conducted as part of this study. In chapter 3, we discuss the phenotypic response of immune and vascular cells to oxLDL as determined by a variety of methods including light and fluorescence microscopy, flow cytometry, cell adhesion and angiogenesis assays. Chapter 4 is a detailed discussion of the statistical and bioinformatic principles underlying the analysis of the high-throughput sequencing data, followed by a description of the bioinformatic pipeline I have developed for this analysis. In chapter 5 we examine the changes in gene expression in the cells following treatment with oxLDL by RNA-seq. In chapter 6 epigenetic regulation of these gene expression changes is investigated using a combination of chromatin immunoprecipitation (ChIP) for acetylation of lysine 27 on histone 3 followed by high-throughput sequencing (H3K27ac ChIP-seq) and an assay for transposase-accessible chromatin (ATAC) followed by high throughput sequencing (ATAC-seq)¹³⁰. Finally, in chapter 7 the findings presented in this thesis will be summarised and discussed.

Table 1.5: Studies investigating the epigenetic effects of atherogenic lipids.

Author	Epigenetic mechanisms	Gene targets	Cell types
Mitra 2011	DNA methylation	LOX1, ANXA-5, BAX, CASP3, Bcl-2, cIAP-1	HUVEC
Kim 2012	DNA methylation	p66shc	HUVEC
Kumar 2013	DNA methylation, histone acetylation	KFL2	HUVEC
N'Guessan 2009	Histone acetylation and phosphorylation	IL8, MCP1	HUVEC
Choi 2005	Histone acetylation	CD36	RAW264.7
Li 2010	miRNA (miR-125a-5p)	ET-1	HUVEC
Qin 2011	miRNA (miR-365)	Bcl2	HUVEC
Bekkering 2014	Histone methylation	TNF α , IL-6, IL-18, MCP1, MMP2, MMP9, CD36, SRA	monocytes
Guay 2013	DNA methylation	LPL	leucocytes
Chen 2009	miRNA (miR-125a-5p)	ORP9	monocytes
Yang 2015	miRNA (miR-155)	SOCS1	Thp1 monocytes/ macrophages
Tian 2014	miRNA (miR-155), histone acetylation	HBP1	RAW264.7
Huang 2010	miRNA (miR-155)	LDLR, CD36, CD68, IL-6, IL-8, TNF α , MyD88	Thp1 monocytes/ macrophages
Chen 2011	miRNA (miR-155)	SCG2	dendritic cells, macrophages
Chen 2011	miRNA (miR-146a)	CD40L	dendritic cells
Chen 2011	DNA methylation	MMP2, MMP9	HASMC
Chen 2012	Histone acetylation	miR-29b	HASMC

Chapter 2

Materials and Methods

Contents

2.1	Basic cell culture techniques and cell types used	27
2.1.1	Primary cell culture	27
2.1.2	Cell lines	28
2.2	Isolation and purification of CD14 ⁺ and CD4 ⁺ cells	29
2.3	Cell treatments	31
2.4	Angiogenesis assays	31
2.5	LDL isolation	32
2.6	LDL oxidation	35
2.7	DiI-labelling of LDL	36
2.8	Flow cytometry	36
2.8.1	Cell surface staining	36
2.8.2	Intracellular staining	37
2.8.3	Flow cytometry data acquisition and analysis	37
2.9	RNA-seq sample preparation	38
2.10	ChIP-seq sample preparation	40
2.11	ATAC-seq sample preparation	42

2.1 Basic cell culture techniques and cell types used

All cells were cultured in a humidified tissue culture incubator at 37°C with 5% CO₂. Adherent cells were passaged by trypsinisation [Trypsin (0.05% Trypsin, 0.53M ethylenediaminetetraacetic acid (EDTA)), Invitrogen, Paisley, UK]. All cell cultures were carried out in a sterile class II cabinet.

2.1.1 Primary cell culture

Primary cells used were HAECs, VSMCs (derived from human coronary arteries), and CD14⁺ and CD4⁺ PBMCs. HAECs and VSMCs were used between passage 4-7;

PBMCs were used immediately after isolation.

HAECs were cultured in Medium 200 (Invitrogen) and supplemented with Low Growth Supplement (Invitrogen) and 50 units/ml penicillin + 50 µg/ml streptomycin (Sigma-Aldrich, Dorset, UK). VSMCs were propagated in Medium 231 (Invitrogen) supplemented with Smooth Muscle Growth Supplement (Invitrogen) and 50 units/ml penicillin + 50 µg/ml streptomycin (Sigma-Aldrich, Dorset, UK). For terminal differentiation of VSMCs, the culture medium supplement was changed to Smooth Muscle Differentiation Supplement (Invitrogen) along with 50 units 1/ml penicillin + 50 µg/ml streptomycin (Sigma-Aldrich). HAEC and pre-terminally differentiated VSMCs were passaged every 2-4 days at approximately 80-90% confluence.

PBMCs were cultured in R10 medium, which was prepared from standard RPMI purchased from Sigma Aldrich and foetal calf serum (FCS) purchased from Sigma-Aldrich or First Link UK. To prepare R10 medium, the FCS was first inactivated at 55°C for 15 minutes, and 50 ml (10%) of heat-inactivated FCS was added to each 500 ml of culture medium. The medium was supplemented with final concentrations of 2 mmol/l glutamine, 50 1/ml penicillin, and 50 µg/ml streptomycin (Sigma-Aldrich). For CD14⁺ cells, the R10 medium was also supplemented with 50 ng/ml of MCSF (eBioscience). For CD4⁺ cells, the R10 medium was supplemented instead with 50 units/ml of interleukin-2 (IL-2).

2.1.2 Cell lines

Human embryonic kidney 293T (HEK293T) cells were obtained from co-workers within the department and cultured in D10 medium, which was prepared from standard DMEM purchased from Sigma Aldrich and supplemented as described for R10 medium above. Cells were passaged every 3-4 days at approximately 80-90% confluence.

2.2 Isolation and purification of CD14⁺ and CD4⁺ cells

Cells were isolated from 50 ml of peripheral blood from healthy human volunteers. Ethical approval was obtained for this was obtained from NHS Research Ethics Committee (South Central-Hampshire B, reference 13/SC/0392) and all participants provided informed written consent. The blood was centrifuged over Ficoll Paque PLUS (GE Healthcare, LifeSciences) in a 2:1 ratio and centrifuged at 1200 revolutions per minute (rpm) for 30 minutes with no brake. The mononuclear cell layer (buffy coat layer) was removed with a pipette and transferred to a clean tube. Cells were resuspended in 50 ml of R10 and centrifuged at 1500 rpm at room temperature for 10 minutes. The supernatant was discarded and the remaining cell pellet resuspended in 25 ml R10 and centrifuged at 1500 rpm at room temperature for 5 minutes. The supernatant was again discarded and the the remaining cell pellet resuspended in R10 – samples distributed across several tubes were combined together to make a final cell suspension of 50 ml. The cell suspension was then re-centrifuged at 1500 rpm at room temperature for 5 minutes. The cells were then resuspended in a pre-chilled buffer composed of phosphate-buffered saline (PBS) containing 0.5% bovine serum albumin (BSA) and 2 mmol/l EDTA (magenetic cell separation (MACS) buffer) and centrifuged again for 5 minutes at 1500 rpm at room temperature. The supernatant was discarded and the cell pellet resuspended in 180 µl of MACS buffer. To this was added 20 µl of CD14 microbeads (Miltenyi Biotech, Surrey, UK) to make a final volume of 200 µl, and the cell/bead suspension incubated at 4°C for 15 minutes. The cells were once again washed after the incubation period with MACS buffer and then resuspended in 1 µl of cold MACS buffer to pass through an MS separation column

(Miltenyi Biotech) which was placed in a magnetic field. MS columns were first washed and equilibrated with 3 ml of MACS buffer. The labelled cell suspension was then applied to the column and the unbound cells allowed to pass through. The column and the bound cells were then washed a further three times with 3 ml of MACS buffer. The column was removed from the magnetic field and the bound cells were eluted in 5 ml of MACS buffer by adding the buffer to the column and then flushing the column with the plunger provided. The cells were pelleted by centrifugation for 5 minutes at 1500 rpm at room temperature and resuspended in 5 ml of R10 medium. 50 μ l of cells were taken for counting with a haemocytometer, and then the volume of the remainder was adjusted to make a cell suspension at a concentration of 5×10^5 cells 1/ml of medium. To this was added MCSF (eBioscience) to a final concentration of 50 ng/ml and 2 ml of the cell suspension added to each well of 6-well plate.

For isolation of CD4⁺ cells, the unbound cells from above (i.e., the CD14⁻ cell population) were centrifuged at 1500 rpm at room temperature for 5 minutes, the supernatant discarded, and the cell pellet resuspended in 180 μ l of pre-chilled MACS buffer. To this was added 20 μ l of CD4 microbeads (Miltenyi Biotech, Surrey, UK) to make a final volume of 200 μ l. The cell/bead suspension was incubated at 4°C for 15 minutes before resuspension in R10 medium at a concentration of 1×10^6 cells 1/ml. To this was added IL-2 to a final concentration of 50 units/ml and 1 ml of the the cell suspension added to each well of 12-well plate. The cells were activated by incubation with human T-activator CD3/CD28 dynabeads (ThermoFisher) overnight.

2.3 Cell treatments

At 70-80% confluence, HAECs were treated with 50 µg/ml of oxLDL or the equivalent volume of oxLDL buffer for 48 hours. For VSMCs, at the same level of confluence, the culture medium was changed from Medium 231 with Smooth Muscle Growth Supplement to Medium 231 with Smooth Muscle Differentiation Supplement. After five days, cells were treated with 50 µg/ml of oxLDL or the equivalent volume of oxLDL buffer for 48 hours.

Foam cells were generated from CD14⁺ macrophages by treatment with 50 µg/ml of oxLDL. Control cells were treated with the equivalent volume of oxLDL buffer for 48 hours.

CD4⁺ cells were activated by incubation with human T-activator CD3/CD28 dynabeads (ThermoFisher) overnight before treatment with 50 µg/ml of oxLDL or the equivalent volume of oxLDL buffer for 48 hours.

For DiI-labelling, cells were treated as described above, but substituting oxLDL for DiI-labelled oxLDL.

2.4 Angiogenesis assays

Geltrex (Invitrogen) was thawed overnight at 4°C. Once thawed, 100 µl of Geltrex was added to wells of a 96-well plate and allowed to solidify at 37°C for 60 minutes. Calcein AM (Santa Cruz) was added to HAECs in a 75 cm² flask at a final concentration of 2 µg/ml and allowed to incubate at 37°C for 30 minutes in 5% CO₂ in the dark. Following incubation with dye, cells were harvested by incubation with Trypsin/EDTA solution (0.05% Trypsin, 0.53 M EDTA) for 3 minutes at room temperature, followed

by neutralisation with medium and then centrifugation at 180 *g* for 7 minutes. The supernatant was discarded, and cells were then resuspended in DMEM, supplemented with 2% FCS and 2 mmol/l glutamine, 50 units/ml penicillin, and 50 µg/ml streptomycin. Cells were counted with a haemocytometer and density adjusted to 2×10^6 cells 1/ml of medium; for cells that had been cultured with oxLDL, this was added to the medium to a final concentration of 50 µg/ml. 200 µl of cell suspension was applied to each Geltrex-coated well. The cells were incubated overnight at 37°C, 5% CO₂, following which they were visualised using phase contrast and fluorescent microscopy. Calcein AM has a peak excitation at 495 nm and peak emission at 515 nm.

2.5 LDL isolation

50 ml of peripheral blood was drawn from healthy human volunteers into sterile EDTA (150 mmol/l) tubes. The blood was centrifuged for 30 minutes at 800 *g* at room temperature, and the upper plasma layer was transferred to a fresh tube. The plasma was centrifuged again for 30 minutes at 2500 *g* at room temperature to exclude platelets and remaining cells. The plasma was removed and samples from multiple donors was then pooled to minimise inter-individual variation. If not used immediately, plasma was stored in aliquots at -80°C for a maximum of three months. Otherwise, the density of plasma was adjusted to 1.210 g/ml by addition of potassium bromide (KBr; 0.325 g of KBr per ml of plasma).

A sterile stock salt solution of density 1.346 g/ml was prepared by addition of 17.5 g KBr to 50 ml of sterile, tissue-culture grade PBS. From this stock solution, solutions of density 1.019 g/ml (solution I) and 1.063 g/ml (solution II) were prepared by addition of PBS. If V_S is the starting volume of the stock solution, and d_F the final

density desired, then the volume of PBS to add (V_A) to the stock solution is given by the formula below (1.346 and 1.01 are the densities of the stock solution and PBS, respectively).

$$V_A = \frac{V_S \times (1.346 - d_F)}{d_F - 1.01} - V_S$$

7.5 ml of sterile PBS was transferred to a Beckman OptiSeal tube using a 10 ml syringe attached to a spinal needle. The same needle was used to gently underlay this with 7.5 ml of solution I, followed by 7.5 ml of solution II, and finally by 5.5 ml of the density-adjusted plasma, to set up a discontinuous density gradient. Typically, 50 ml of blood yields enough plasma to set up four discontinuous density gradients. Immediately upon completion, the gradients were centrifuged in a Beckman L5-50 ultracentrifuge with a SW28 rotor at 28,000 rpm for 24 hours at 4°C with no braking. Following ultracentrifugation, the LDL layer forms a discrete band as shown in Figure 2.1; this layer was removed by puncturing the bottle with a 21G needle and transferred to a sterile dialysis cartridge (Slide-A-Lyzer G2 Dialysis Cassettes, gamma-irradiated, 10K MWCO, 15mL) that had been pre-equilibrated with PBS for 5 minutes prior to use. The samples were extensively dialysed against a minimum of 200x the sample volume of PBS. Following dialysis, the samples were passed through a 0.22 μ m filter to sterilise and remove aggregates. The sample concentration was determined after filtration by bicinchoninic acid (BCA) assay, and adjusted to 0.5 μ g/ml with sterile PBS, after which the LDL was stored in aliquots in the dark at 4°C.

For Western blot, LDL samples were dissolved in a buffer containing 63 mmol/l Tris-HCl (pH 6.8), 2% sodium dodecyl sulphate (SDS), 10% glycerol, 1 mmol/l EDTA, and 0.001% bromophenol blue, heated for 5 min to 95°C, and then run for 50-60 minutes, at 200 V on a NuPage Novex 4-12% Bis-Tris gel (ThermoFisher) in 3-(N-morpholino)propanesulfonic acid (MOPS) SDS running buffer (50 mmol/l MOPS,

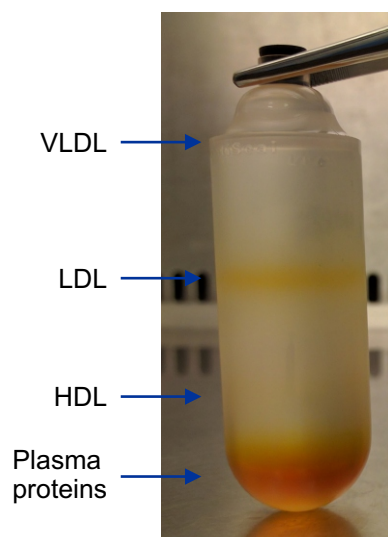


Figure 2.1: Lipoprotein isolation. After isopycnic ultracentrifugation, lipoprotein fractions separate out into different bands. The topmost layer is the very low density lipoprotein layer (VLDL). If the patient is unfasted, there may be an additional creamy white chylomicron layer here. The next layer is the low density lipoprotein layer (LDL), which is sandwiched between the solution I and solution II layers from the discontinuous gradient preparation. Below this is the high density lipoprotein (HDL) layer, and finally other plasma proteins.

50 mmol/l Tris base, 0.1% SDS, 1 mmol/l EDTA at pH 7.7). After electrophoresis, the proteins were transferred to polyvinylidene fluoride (PVDF) membranes in a 25 mmol/l Tris, 192 mmol/l glycine, 0.01% SDS, and 20% (v/v) methanol transfer buffer at 100V at 4°C for 3.5 hours. Membranes were blocked with 5% BSA for 1 hour at room temperature, and then incubated overnight with a 1:2000 dilution of goat IgG polyclonal anti-human apolipoprotein B (Abcam ab7616) in PBS/0.05%-Tween/2%-BSA solution. After incubation, the membranes were washed five times with PBS-Tween and then incubated for 1 hour with a 1:20000 dilution of bovine anti-goat IgG horse radish peroxidase (HRP)-conjugated antibody in the same buffer. Afterwards the membranes were washed five times with PBS-Tween and the immunoreactive proteins were visualized by treating the membranes with ECL detection reagents (Amersham, Arlington Heights, IL) followed by imaging in a BioRad ChemiDoc MP

imaging system.

2.6 LDL oxidation

A stock solution of sterile CuCl_2 (25 mmol/l in distilled water, dH_2O) was added to the LDL to a final concentration of 25 $\mu\text{mol/l}$. The sample was incubated at 37°C for 18 hours, after which the oxidation reaction was terminated by addition of EDTA to a final concentration of 0.5 mmol/l. The oxLDL was then dialysed extensively against at least 200x the volume of PBS and the resulting dialysate used for treatment of control samples.

The degree of LDL denaturation associated with copper oxidation was assessed by two criteria: agarose gel electrophoresis and Thiobarbituric acid reacting substances (TBARS) assay.

Agarose gel electrophoresis of native and modified LDL was performed in 0.5% (w/v) gel, prepared in 0.05 mol/l sodium barbital buffer, pH 8.6 and subjected to 25 mA constant current for 2.5 hours at 4°C . The gels were fixed for 1 hour in 5% glacial acetic acid in 75% ethanol (v/v) and air dried. LDL bands were visualized after staining the gel in a saturated solution of oil red O (0.04% in 60% ethanol, stored at 4°C). The relative electrophoretic mobility of oxLDL to native LDL gives a measure of degree of oxidation.

Malondialdehyde (MDA) is a byproduct of lipid peroxidation. Its production can be quantified colorimetrically at 530-540 nm by its reaction with TBARS at 90°C - 100°C . The assay procedure described in the TBARS Assay Kit (Cayman Chemical, 10009055) was used, and results were corrected for protein levels and reported in μmol MDA per g protein.

2.7 DiI-labelling of LDL

A stock solution of the fluorescent lipophilic carbocyanine dye, 1,1'-dioctadecyl-3,3,3',3'-tetramethylindocarbocyanine perchlorate (DiI), was prepared by dissolving 30 mg of DiI in 1 ml of dimethyl sulphoxide (DMSO). The stock solution was added to sterile purified LDL to a final ratio of 300 µg of DiI to 1 mg of LDL protein and incubated at 37°C for 18 hours. The DiI-LDL was purified by overlaying with a KBr solution of density 1.063 g/ml and ultracentrifuged in a Beckman L5-50 ultracentrifuge with a SW28 rotor at 28,000 rpm overnight at 4°C with no braking. The top-most layer, consisting of purified DiI-labelled LDL was removed with a sterile pipette and extensively dialysed against 200x the sample volume of PBS. DiI-labelled oxLDL was prepared by oxidation of the DiI-labelled LDL.

2.8 Flow cytometry

2.8.1 Cell surface staining

Cells were removed from incubator to class II tissue culture hood. For adherent cells, culture medium was removed and discarded, and cells washed once with sterile PBS. 5 mmol/l EDTA in PBS with 0.5% azide (PBSA) solution at 37°C was used to detach cells from plate. Cells were counted and the appropriate volume transferred to wells of a 96-well plate. For suspension cells, the cells were counted and the volume required centrifuged down, following which the supernatant was discarded and the cells resuspended in sterile PBSA before transfer to wells of a 96-well plate. The cells were centrifuged at 1500 rpm for 5 minutes, and the supernatant was discarded. Primary staining was carried out with 0.2 µg of primary antibody or isotype control

in 100 μ l of cold PBSA with 1% BSA per well. The cells were incubated with the primary antibody for 20-30 minutes at 4°C and then centrifuged at 1500 rpm for 5 minutes, following which, cells were washed with 150 μ l of cold PBSA twice. Secondary staining was carried out with 0.2 - 1 μ g of conjugated secondary antibody in 100 μ l cold PBSA/1%-BSA per well at 4°C for 20-30 minutes. The cells were then washed as described after the primary stain, before resuspension in 100 μ l cold PBSA for analysis by flow cytometry.

2.8.2 Intracellular staining

Cells were prepared as described above, and if cell surface staining was also to be performed this was done first. Cells were then fixed by resuspending in 4% paraformaldehyde for 10-15 minutes at room temperature, and then centrifuged at 1500 rpm for 5 minutes, following which, cells were washed with 150 μ l of cold PBSA twice. Permeabilisation of cells was performed by incubating cells with 100 μ l of Triton (0.5% in PBSA) for 15 minutes at room temperature. The cells were then centrifuged at 1500 rpm for 5 minutes, and the supernatant was discarded. Primary and secondary staining were performed as described for the cell surface staining protocol. All antibodies were prepared in the permeabilisation buffer to ensure the cells remained permeable throughout.

2.8.3 Flow cytometry data acquisition and analysis

Flow cytometry was carried out on a BDFACSCanto or BDFACSCanto 2, using BD FACSDiva software for data collection. Unstained cells were acquired to adjust forward scatter (FSc), side scatter (SSc) and fluorescence voltages to the appropriate

settings. In all cases, the height of the FSc voltage trace, and the area of the SSc voltage trace were used. After adjustment of FSc and SSc voltages to fit events into a detectable window, live cells were gated based on FSc/SSc characteristics. Where these characteristics were sufficient to distinguish live and dead cells, staining with propidium iodide was not routinely carried out so as to minimise fluorochrome interference and limit the need for compensation.

For single fluorochrome analysis, fluorescence voltages were adjusted to set the baseline voltage from unstained cells as zero. When two fluorochromes were used simultaneously, where the possibility of spectral overlap existed, compensation controls were used. For standard experiments, 10,000 gated events were recorded for each sample. All samples were measured in biological triplicates, unless otherwise stated.

Following data acquisition, the complete dataset was exported to FlowJo for further analysis. Typically, initial gating was carried out as described above. Having set the initial gate to include a maximum number of FSc/SSc healthy cells, a frequency distribution histogram of specific fluorescence within this population was created. The mean fluorescence intensity was recorded for each sample. Biological triplicate values were used to calculate the 95% confidence interval for the estimate of mean fluorescence intensity.

2.9 RNA-seq sample preparation

RNA extraction was carried out using TRIzol and the PureLink RNA mini kit (Invitrogen), largely following the manufacturer's instructions. For adherent cells, the culture medium was lifted from the plate, and the cells lifted off after incubating with Trypsin/EDTA solution (0.05% Trypsin, 0.53 mol/l EDTA) for 3-4 minutes

at room temperature. The Trypsin/EDTA was neutralised by addition of media and the cells centrifuged at 1500 rpm for 5 minutes. Suspension cells were simply centrifuged at 1500 rpm for 5 minutes. The supernatant was discarded and the cells resuspended in 1 ml of TRIzol and transferred to RNase-free 1.7 ml microcentrifuge tubes. After standing at room temperature for 5 minutes, 200 μ l of chloroform was added to each sample and the tubes shaken by hand for 15 seconds, before incubation at room temperature for 3 minutes. The samples were then centrifuged at 12,000 g at 4°C for 15 minutes. Following centrifugation, the clear aqueous upper phase was removed to a clean RNase-free 1.7 ml microcentrifuge tube. An equal volume of 70% ethanol made in diethylpyrocarbonate (DEPC)-treated water was added to each sample, for a final ethanol concentration of 35%. The samples were then added to the spin columns provided in the kit, and centrifuged at 12,000 g for 15 seconds, after which the flow-through was discarded. DNaseI digestion of the RNA samples was carried out using the Purelink DNase kit (Invitrogen), according to the manufacturer's instructions. The samples were washed on the spin-column with 350 μ l buffer I, and a working dilution of DNaseI was added to the column. The DNaseI reaction was allowed to proceed for 20 minutes at room temperature. The DNaseI was removed by washing the column with 350 μ l wash buffer I. The column was then washed twice with 500 μ l wash buffer II, before spin drying at 12,000 g . Samples were eluted in 20 μ l of RNase-free water.

Integrity of RNA samples was determined using the RNA 6000 pico assay on the Agilent Bioanalyzer (Agilent) and sample quantification was performed using Nanodrop. 2 μ g of total RNA was used for standard TruSeq library preparation with polyA selection (performed in-house by the Wellcome Trust Centre for Human Genetics Core Genomics Service). Libraries were subjected to 50 base-pair (bp) paired-end Illumina Hi-Seq

2500 sequencing, with the exception of the macrophage samples which were sequenced on the Illumina Hi-Seq 4000 platform with 75 bp paired-end reads.

2.10 ChIP-seq sample preparation

ChIP-seq was performed using the MAGnify Chromatin Immunoprecipitation System (ThermoFisher Scientific), largely following the manufacturer's instructions. For adherent cells, the culture medium was lifted from the plate, and the cells lifted off after incubating with Trypsin/EDTA solution (0.05% Trypsin, 0.53M EDTA) for 3-4 minutes at room temperature. The Trypsin/EDTA was neutralised by addition of media and the cells centrifuged at 1500 rpm for 5 minutes, before resuspension in 500 μ l of medium. For these harvested cells and suspension cells, the numbers were counted using a haemocytometer, aiming for $3\text{-}5 \times 10^5$ cells per ChIP. The required number of cells were centrifuged at 1500 rpm at room temperature for 5 minutes and resuspended in 500 μ l of medium. Cells were cross-linked by addition of 13.5 μ l of 37% formaldehyde (to a final concentration of 1%) in medium for 10 minutes at room temperature before neutralisation by incubating with 57 μ l of 1.25 mol/l glycine at room temperature for 5 minutes. The cross-linked samples were centrifuged at 200 *g* for 10 minutes at 4°C, the supernatant discarded and then the cell pellet resuspended in cold PBS to be centrifuged again as described. This was repeated once more, after which the cell pellets were resuspended in 50 μ l of lysis buffer with protease inhibitor (both provided in the kit) by mild pulses on a vortex mixer. The samples were incubated on ice for 5 minutes, before proceeding to sonication, which was performed with a Bioruptor (Diagenode) set to 'HI' with 16 cycles of 30 seconds on and 30 seconds off to generate sheared chromatin with an average size range of 100-300 bp as visualised by agarose gel electrophoresis. Sonicated supernatants were diluted 10-fold in dilution buffer with

protease inhibitor before immunoprecipitation. 2 μ l of antibody [total rabbit IgG, (ThermoFisher Scientific) or anti-H3K27ac (Abcam ab472)] was first incubated with 100 μ l of Dynabeads Protein G (ThermoFisher Scientific) beads for 4 hours at 4°C, then the antibody-bead conjugate was added to 100 μ l of the diluted sonicated lysate. The remainder of the sonicated lysates was kept aside as input control. The samples were incubated overnight at 4°C on a rotating mixer. Following incubation, samples were washed three times with IP Buffer 1, using a chilled DynaMag-PCR Magnet to precipitate the beads each time. For each wash, the tubes were placed on the magnet, until a tight pellet of beads was formed. The supernatant was then removed and discarded, the tubes taken off the magnet, and the beads resuspended in IP buffer 1 and placed in a rotating mixer at 4°C for 5 minutes. The washing steps were repeated twice more using IP Buffer 2. After the last wash, the beads were resuspended in 54 μ l of Reverse Crosslinking Buffer (containing Proteinase K); at the same time, 44 μ l of Reverse Crosslinking buffer was also added to 10 μ l of input control. The samples were vortexed to mix, and then incubated in a thermal cycler at 55°C for 15 minutes. The samples were then placed on the DynaMag-PCR magnet, and the supernatant, containing the DNA, was transferred to new tubes and incubated at 65°C for 15 minutes before cooling on ice for 5 minutes. Finally, the samples were purified using a Qiagen MinElute kit exactly as described by the manufacturer's, and eluted in 10 μ l of nuclease-free water. Standard TruSeq adaptors were ligated for library preparation (performed in-house by the Wellcome Trust Centre for Human Genetics Core Genomics Service), and samples were subjected to 50 bp paired-end Illumina Hi-Seq 2500 sequencing.

2.11 ATAC-seq sample preparation

ATAC-seq was performed as previously described¹³⁰. For adherent cells, the culture medium was lifted from the plate, and the cells lifted off after incubating with Trypsin/EDTA solution (0.05% Trypsin, 0.53 mol/l EDTA) for 3-4 minutes at room temperature. The Trypsin/EDTA was neutralised by addition of media and the cells centrifuged at 1500 rpm for 5 minutes, before resuspension in 1 ml of cold PBS. Cells were then counted using a haemocytometer. A volume containing 50-75,000 cells was centrifuged at 500 *g* at 4°C for 5 minutes, the supernatant discarded and the cells resuspended in 1 ml cold PBS, before centrifuging again at 500 *g* at 4°C for 5 minutes. The supernatant was once again discarded and the cell pellets were resuspended in 50 µl of cold lysis buffer (10 mmol/l Tris-HCl, pH 7.4, 10 mmol/l NaCl, 3 mmol/l MgCl₂, and 0.1% IGEPAL CA-630), before immediately centrifuging again at 500 *g* at 4°C for 15 minutes. Once lysis was complete, the supernatant was discarded and the nuclear pellets resuspended in transposition reaction mix (25 µl 2x TD buffer (from Illumina DNA sample preparation kit), 2.5 µl hyperactive Tn5 transposase (from Illumina DNA sample preparation kit), and 22.5 µl nuclease-free water) and incubated for 30 minutes at 37°C. Immediately following transposition, the samples were purified using a Qiagen MinElute kit, as per the manufacturer's protocol; the transposed DNA was eluted in 20 µl of elution buffer.

To amplify the transposed DNA fragments, 10 µl of this purified DNA was used, and the remainder was stored at -80°C. PCR was carried out using the reagents shown in Table 2.1 and the following cycling conditions: 72°C for 5 minutes, 98°C for 30 seconds, and then thermocycling at 98°C for 10 seconds, 63°C for 30 seconds, and 72°C for 1 minute.

Table 2.1: ATAC-seq PCR recipe. The forward primer used for all PCR reactions is the same (AD1); the reverse primer was barcoded to allow pooling of libraries (AD2.x).

Component	Volume
transposed DNA	10.0
nuclease-free water	10.0
forward primer (25 $\mu\text{mol/l}$)	2.5
reverse primer (25 $\mu\text{mol/l}$)	2.5
NEBNext High-Fidelity 2x PCR mastermix	25.0

For each library, the same forward primer (AD1) was used, but a different barcoded reverse primer was used (Table 2.2). To reduce GC and size bias in the PCR, the reaction was monitored using qPCR to stop amplification before saturation of libraries. To do this, the full libraries were amplified for five cycles, after which a 5 μl aliquot of the PCR reaction was taken and to which was added 4.41 μl of nuclease-free water, 0.25 μl of 25 $\mu\text{mol/l}$ solutions of the forward and reverse primers, 0.09 μl of 100x SYBR Green I and 5 μl of NEBNext High-Fidelity 2x PCR Master Mix. This reaction was run for 20 cycles; to determine the additional number of cycles needed the relative fluorescence was plotted against cycle, and the cycle number corresponding to one-third of the maximum fluorescent intensity was chosen (Figure 2.2A). The remaining 45 μl of original PCR reaction was then run using the same thermocycling conditions for the determined number of cycles. The quality of the generated libraries was assessed using the Agilent DNA High Sensitivity kit with the Agilent 2200 TapeStation, looking for the characteristic nuclear phasing pattern (Figure 2.2B). Libraries were then pooled and subjected to 50 bp paired-end Illumina Hi-Seq 2500 sequencing, with the exception of the macrophage samples which were sequenced on the Illumina Hi-Seq 4000 platform with 75 bp paired-end reads.

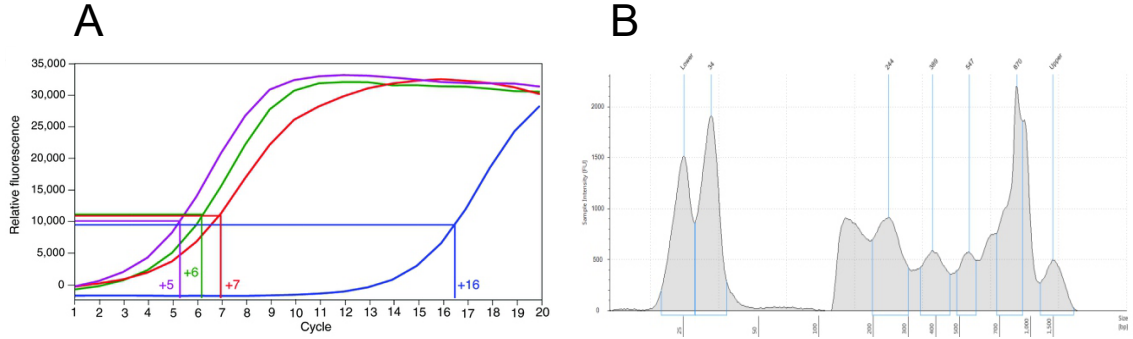


Figure 2.2: A) Representative amplification plot demonstrating the correct number of additional cycles to perform for four ATAC-seq libraries. B) Representative TapeStation run of an ATAC-seq library showing fragments of appropriate size with a characteristic nuclear phasing pattern. The peak at 34 bp comes from primers.

Table 2.2: ATAC-seq oligonucleotides used for PCR. Ad1 is the forward primer used for all reactions. Depending on the libraries to be pooled, a combination of Ad2.x oligonucleotides were used as reverse primers for individual libraries.

Name	Barcode	Sequence
Ad1		AATGATACGGCGACCACCGAGATCTACACTCGTCGGCAGCGTCAGATGTG
Ad2.1	TAAGGCGA	CAAGCAGAAGACGGCATAACGAGATTCGCCTTAGTCTCGTGGGCTCGGAGATGT
Ad2.2	CGTACTAG	CAAGCAGAAGACGGCATAACGAGATCTAGTACGGTCTCGTGGGCTCGGAGATGT
Ad2.3	AGGCAGAA	CAAGCAGAAGACGGCATAACGAGATTTCTGCCTGTCTCGTGGGCTCGGAGATGT
Ad2.4	TCCTGAGC	CAAGCAGAAGACGGCATAACGAGATGCTCAGGAGTCTCGTGGGCTCGGAGATGT
Ad2.5	GGACTCCT	CAAGCAGAAGACGGCATAACGAGATAGGAGTCCGTCTCGTGGGCTCGGAGATGT
Ad2.6	TAGGCATG	CAAGCAGAAGACGGCATAACGAGATCATGCCTAGTCTCGTGGGCTCGGAGATGT
Ad2.7	CTCTCTAC	CAAGCAGAAGACGGCATAACGAGATGTAGAGAGGTCTCGTGGGCTCGGAGATGT
Ad2.8	CAGAGAGG	CAAGCAGAAGACGGCATAACGAGATCCTCTCTGGTCTCGTGGGCTCGGAGATGT
Ad2.9	GCTACGCT	CAAGCAGAAGACGGCATAACGAGATAGCGTAGCGTCTCGTGGGCTCGGAGATGT
Ad2.10	CGAGGCTG	CAAGCAGAAGACGGCATAACGAGATCAGCCTCGGTCTCGTGGGCTCGGAGATGT
Ad2.11	AAGAGGCA	CAAGCAGAAGACGGCATAACGAGATTGCCTCTTGTCTCGTGGGCTCGGAGATGT
Ad2.12	GTAGAGGA	CAAGCAGAAGACGGCATAACGAGATTCTCTACGTCTCGTGGGCTCGGAGATGT
Ad2.13	GTCGTGAT	CAAGCAGAAGACGGCATAACGAGATATCACGACGTCTCGTGGGCTCGGAGATGT
Ad2.14	ACCACTGT	CAAGCAGAAGACGGCATAACGAGATACAGTGGTGTCTCGTGGGCTCGGAGATGT
Ad2.15	TGGATCTG	CAAGCAGAAGACGGCATAACGAGATCAGATCCAGTCTCGTGGGCTCGGAGATGT
Ad2.16	CCGTTTGT	CAAGCAGAAGACGGCATAACGAGATACAAACGGGTCTCGTGGGCTCGGAGATGT
Ad2.17	TGCTGGGT	CAAGCAGAAGACGGCATAACGAGATACCCAGCAGTCTCGTGGGCTCGGAGATGT
Ad2.18	GAGGGGTT	CAAGCAGAAGACGGCATAACGAGATAACCCCTCGTCTCGTGGGCTCGGAGATGT
Ad2.19	AGGTTGGG	CAAGCAGAAGACGGCATAACGAGATCCCAACCTGTCTCGTGGGCTCGGAGATGT
Ad2.20	GTGTGGTG	CAAGCAGAAGACGGCATAACGAGATCACCACACGTCTCGTGGGCTCGGAGATGT
Ad2.21	TGGGTTTC	CAAGCAGAAGACGGCATAACGAGATGAAACCCAGTCTCGTGGGCTCGGAGATGT
Ad2.22	TGGTCACA	CAAGCAGAAGACGGCATAACGAGATTGTGACCAGTCTCGTGGGCTCGGAGATGT
Ad2.23	TTGACCCT	CAAGCAGAAGACGGCATAACGAGATAGGGTCAAGTCTCGTGGGCTCGGAGATGT
Ad2.24	CCACTCCT	CAAGCAGAAGACGGCATAACGAGATAGGAGTGGGTCTCGTGGGCTCGGAGATGT

Chapter 3

Phenotypic and functional responses of vascular and immune cells to oxLDL

Contents

3.1	Introduction	46
3.2	LDL isolation and measurement of its oxidation	49
3.3	Oxidised LDL induces changes in cellular appearance	52
3.4	Oxidised LDL is taken up by cells in a concentration and time dependent manner	54
3.5	Oxidised LDL induces activation of vascular endothelial and smooth muscle cells.	59
3.6	Oxidised LDL induces cell death at high concentrations	62
3.7	Discussion and Summary	67
3.7.1	Characterisation of oxidised LDL	69
3.7.2	Cytotoxicity of oxidised LDL	70
3.7.3	Effects of oxidised LDL on cell morphology	71
3.7.4	Changes in expression of cell adhesion molecules in response to oxidised LDL	72
3.7.5	Summary	73

3.1 Introduction

Lipoprotein particles are responsible for carrying lipophilic substances around the aqueous environment of the blood. There are several classes of lipoproteins¹⁹, but in terms of atherogenesis, the class associated with the highest risk is LDL¹³¹, as its oxidation to form oxLDL is a key event in the progression and development of atherosclerosis. LDL particles are spherical particles, with a density between 1.019 and 1.063 g/ml and a diameter of approximately 22nm¹⁹. They are formed by hydrolysis of the triglyceride-rich core of VLDL particles by lipoprotein lipase on the surface of the vascular endothelium¹³². LDL has a complex structure but essentially comprises a core of cholesterol esters and triglyceride covered by a surface monolayer of phospholipids, free cholesterol and apoprotein B-100 (Apo-B100)¹³². ApoB-100 contains 4536 amino

acids and is one of the largest monomeric proteins known⁴³; it is an essential component of the LDL molecule and responsible for receptor-mediated endocytosis¹³³.

Oxidation of LDL results in irreversible modifications to both its protein and lipid components. These modifications can range from minimal to extensive, depending on the method of oxidation, and result in changes in the structure and function of the molecule. Oxidatively modified LDL has an increased buoyant density, an increased negative charge, a loss of enzyme activity associated with LDL, and is not recognised by the LDLR³⁶.

Oxidative modification occurs primarily within the intima of the arterial wall²⁹. Significant amounts of oxidation are not thought to occur in plasma due to the presence of antioxidants such as vitamin C and E and removal by SRs present on hepatic sinusoidal Kupffer cells²⁹.

The critical factor determining the susceptibility of LDL to oxidation is the abundance of PUFAs in the particle. Other factors influencing the susceptibility of LDL to oxidation (at least *in vitro*) include particle size, composition of other lipids and presence of lipoprotein(a)²¹. The oxidation of LDL is thought to occur in three phases²¹: an initial lag phase where endogenous LDL antioxidants (such as α -tocopherol) are consumed, a propagation phase with rapid oxidation of PUFAs to lipid hydroperoxides, and a decomposition phase when hydroperoxides are converted to reactive aldehydes. The means by which cell mediated oxidative modification of LDL occurs include non-enzymatic (mediated by transition metal ions such as iron and copper, or ROS) and enzymatic (mediated by lipoxygenases and MPO).

OxLDL induces a broad range of proatherogenic effects in the cell types involved in atheroma formation^{39–42,51,52,77,89}. These have been discussed in detail in Chapter One, and are summarised in Table 3.1.

Table 3.1: Important atherogenic effects of oxLDL.

Potential mechanisms by which oxLDL may influence atherogenesis	
1	OxLDL uptake is taken up 3-10x more rapidly than native LDL by macrophages and VSMCs, leading to foam cell formation.
2	Products of oxLDL are chemotactic for monocytes and T cells, and inhibit the motility of tissue macrophages.
3	OxLDL and its products are mitogenic for VSMCs and macrophages.
4	OxLDL can alter gene expression in vascular cells, for example, induction of MCP-1, colony-stimulating factors, and IL-1 and expression of adhesion molecules.
5	OxLDL increases expression of scavenger receptors, thus enhancing its own uptake.
6	OxLDL can induce expression of and activate PPAR γ .
7	OxLDL is immunogenic and elicits autoantibody formation and activates T cells.
8	Oxidation renders LDL more susceptible to aggregation, which independently leads to increased uptake.
9	OxLDL may enhance procoagulant pathways, for example, by induction of tissue factor and platelet aggregation.
10	OxLDL adversely affects arterial vasomotor properties.
11	OxLDL is cytotoxic to cells, promoting both cell necrosis and apoptosis.

In order to more fully understand how oxLDL results in these atherogenic effects, it is important to first characterise its effects on cells known to be components of the atherosclerotic plaque. There are inconsistent reports in the literature of the biological effects of oxLDL^{51,134}. This is partly because the effect of oxLDL is related not only to its concentration, but also due to the extent to which it has been modified⁴³. Bearing these factors in mind, the primary goal of the data presented in this chapter is to characterise the effects of oxLDL at the cellular level against the vascular and immune cells involved in atherosclerosis.

The first part of this chapter begins by reporting the key biochemical measurements of native and oxidised LDL. To ensure a reproducible level of oxidation of LDL for all experiments, maximally oxidised LDL using copper ions was produced. The prepared oxLDL was analysed for apo-B100 oxidation by measurement of its relative electrophoretic mobility (REM) compared to native LDL, and for lipid oxidation by measurement of MDA using the TBARS assay. The second part of the chapter characterises the cellular effects of the oxLDL prepared in this way against HAECs, VSMCs, macrophages and CD4⁺ T cells. In particular, this chapter presents the morphological and cytotoxic changes induced by oxLDL at different concentrations and over time and establishes the concentration and duration of exposure to oxLDL used for further studies in subsequent chapters.

3.2 LDL isolation and measurement of its oxidation

LDL was isolated from plasma as described in the methods section using an isopycnic ultracentrifugation technique¹³⁵ followed by dialysis overnight at 4°C against PBS to remove excess salts and EDTA. The LDL was sterile-filtered and its concentration

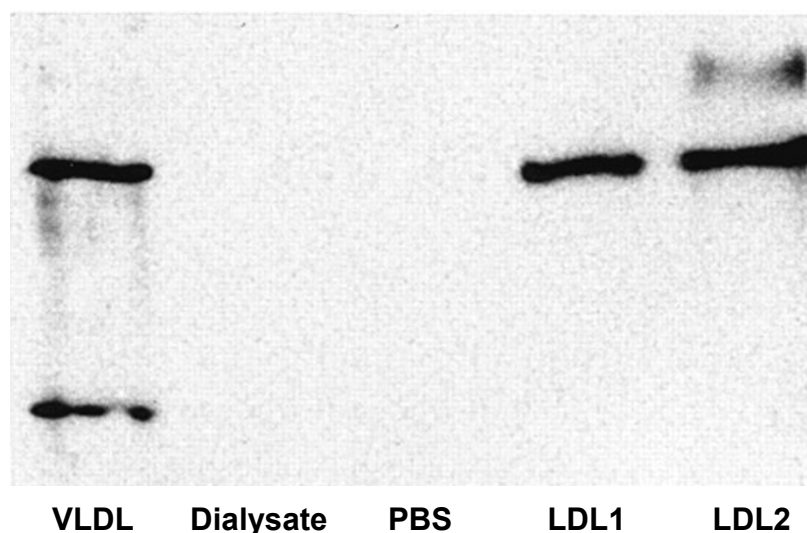


Figure 3.1: Western blot analysis of lipoprotein samples. 10 μ g samples of lipoprotein were resolved by gel electrophoresis on a 4-12% SDS-polyacrylamide gel and used for Western blots and probed with a human apo-B100-specific antibody. A VLDL sample and commercial LDL sample (LDL1) were used as positive controls for the sample of LDL isolated by isopycnic ultracentrifugation from healthy human volunteers (LDL2). The dialysate from the LDL preparation and phosphate-buffered saline (PBS) were included as negative controls.

measured using a BCA assay, after which the concentration was adjusted to 0.5 μ g/ml with sterile PBS. The resultant LDL was stored in aliquots in the dark at 4°C. To ensure that the lipoprotein samples isolated using this technique were indeed LDL, SDS-polyacrylamide gel electrophoresis followed by Western Blotting using antibodies directed against the apo-B100 component of LDL was performed. As positive controls, VLDL isolated using the same technique as for LDL and commercially bought LDL (LDL1 in Figure 3.1) were used. The dialysate buffer and PBS were used as negative controls. As expected, the antibody detected both apo-B100 and apo-B48 (cleavage product of apo-B100) in VLDL. Only apo-B100 was detected in the LDL preparations. A Western Blot band of higher molecular weight than expected was also detected in the samples isolated from volunteers (LDL2 in Figure 3.1); this is likely to represent the heterogeneous nature of the pooled samples.

Oxidation of LDL was carried out by incubation with CuCl_2 at 37°C . The oxidation reaction was terminated by addition of EDTA to a final concentration of 0.5 mmol/l . The oxLDL was then dialysed extensively against PBS and the resulting samples stored in aliquots in the dark at 4°C . The oxidation of LDL results in an increased overall negative charge. Increasing concentrations of native LDL and oxLDL were separated by agarose gel electrophoresis and the resultant gels fixed and stained with oil red O to determine the appropriate quantity of protein required to accurately enable detection of LDL and ascertain the level of oxidation. As shown in Figure 3.2A, oxLDL migrated much further towards the positively charged anode than native LDL and was confidently visualised at $20\text{ }\mu\text{g}$. This amount was therefore used for subsequent gel runs.

To determine a duration of incubation of LDL with CuCl_2 that would result in a reproducible oxidation of LDL, native LDL was oxidised for different durations and the REM of the resultant oxLDL preparations was measured. The REM of oxLDL is calculated as the distance travelled by oxLDL divided by the distance travelled by the native LDL it is derived from. As shown in Figure 3.2B, after 12 hours of oxidation, the REM of the resultant oxLDL remains consistent. Oxidation of LDL also results in the formation of highly reactive lipid hydroperoxides. Decomposition of these unstable peroxides results in the formation of MDA, which can be quantified by its colorimetric reaction with thiobarbituric acid in the TBARS assay. The level of MDA within the native LDL and the oxLDL preparations derived from increasing durations of oxidation were examined. As seen in Figure 3.2B, the MDA levels in the preparations oxidised for 16 and 18 hours were similar, and much higher than that of native LDL. On the basis of these results, the duration of oxidation used to prepare oxLDL in subsequent experiments was 18 hours. Batch-to-batch variation in oxidation

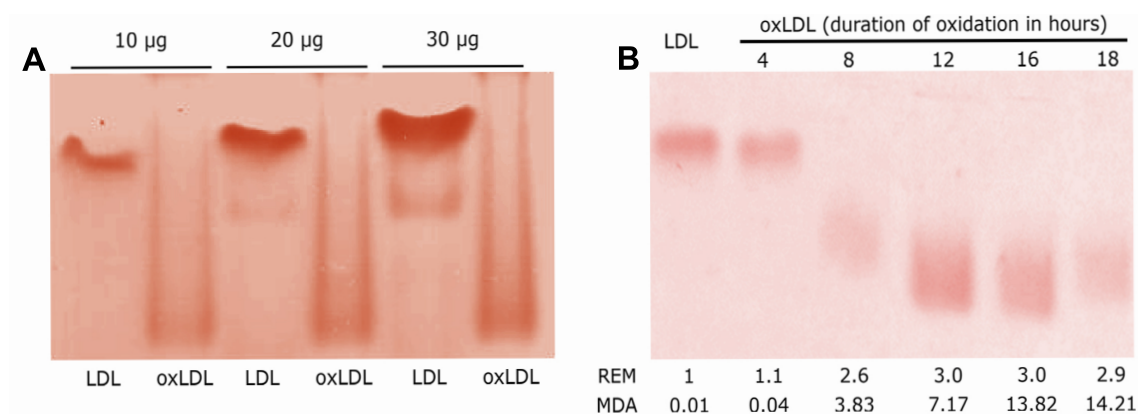


Figure 3.2: Measurement of the oxidation of low density lipoprotein (LDL). A) Increasing doses of native LDL and oxidised LDL (oxLDL) were separated by agarose gel electrophoresis and the resultant gel stained with oil red O. Both native and oxidised LDL can be visualised confidently at 20 μ g. B) The relative electrophoretic mobility (REM) and malondialdehyde (MDA) content of oxLDL prepared by incubating native LDL with copper (II) chloride for increasing durations was determined. After 12 hours, the varying oxLDL preparations had similar REMs and after 16 hours, similar MDA content.

of LDL was determined by measuring the REM and MDA content of multiple batches (Figure 3.3). The mean REM of the different oxLDL batches was 3.26 (± 0.46) and the levels of MDA in the same preparations was 14.90 (± 3.4) nmol/mg of protein. Both the REM and the MDA levels were significantly greater than for native LDL ($p < 0.001$).

3.3 Oxidised LDL induces changes in cellular appearance

Work in our group has previously shown that in macrophages, 50 μ g/ml of oxLDL is sufficient to induce foam cell formation with minimal cell death. Therefore, the primary cell types were incubated with this concentration of oxLDL for 24 hours and lipid content visualised by staining with Oil Red O and examining under phase

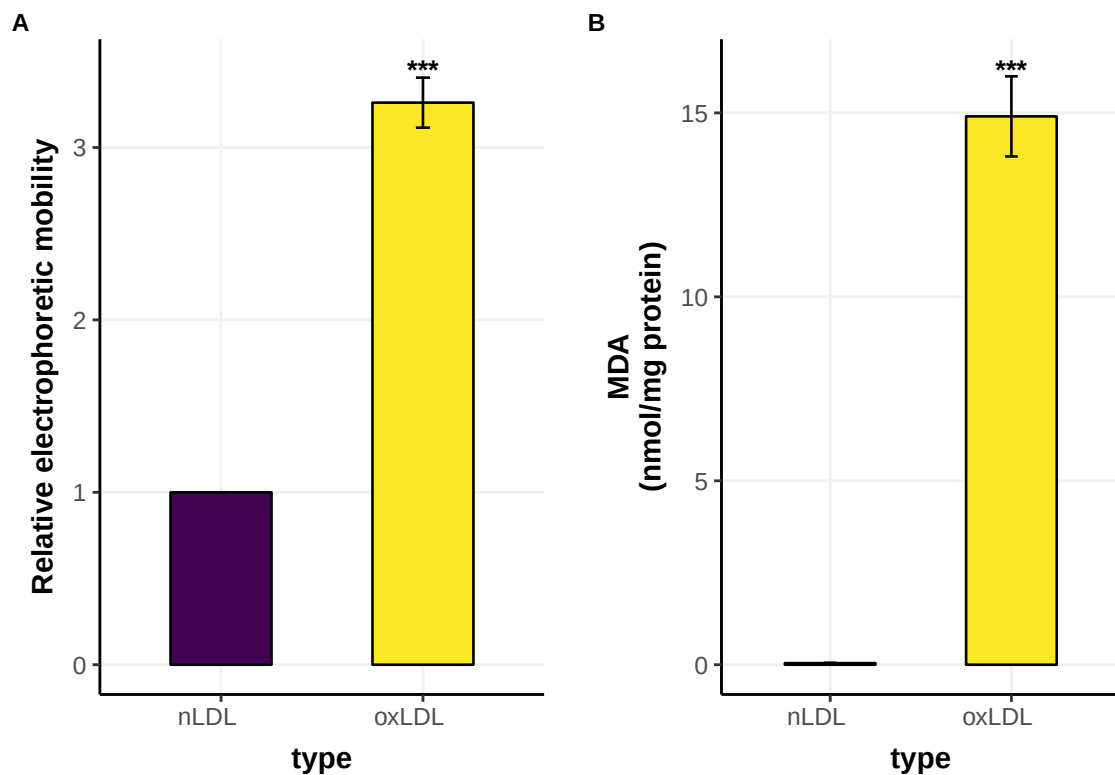


Figure 3.3: Measurement of batch-to-batch variation in oxLDL. Relative electrophoretic mobility (A) and MDA content (B) were measured in oxLDL from 10 separate preparations. nLDL = native low density lipoprotein, oxLDL = oxidised low density lipoprotein, MDA = malondialdehyde.

contrast microscopy (Figure 3.4).

In VSMCs, the most striking change was accumulation of lipid droplets. However, the overall morphology of the VSMCs remained similar. The HAECs also showed lipid droplet accumulation following oxLDL exposure, but in addition, they acquired a more elongated phenotype. There was a clear change in morphology of the macrophages, with them becoming larger and more rounded, in keeping with foam cell formation. There was no increased Oil Red O staining in CD4⁺ lymphocytes incubated with oxLDL. However, these cells became larger and more clumped, suggesting activation.

3.4 Oxidised LDL is taken up by cells in a concentration and time dependent manner

Having established that the cells were morphologically affected by oxLDL, the functional effects of oxLDL were next examined. Primary HAECs were incubated with DiI-labelled oxLDL at concentrations of 0, 10, 25, 50, and 75 µg/ml for 12 hours and then examined by flow cytometry. Live and dead cell gates were set up based on forward and side scatter flow cytometry plots, which characterise cells based on size and granularity (Figure 3.5A). The uptake of DiI-labelled oxLDL was measured by increase in emission in the phycoerythrin (PE) channel (Figure 3.5B). The specificity of uptake was demonstrated by pre-treating cells with an excess of unlabelled oxLDL at a concentration of 100 µg/ml for 6 hours before incubation with 25 µg/ml of DiI-labelled oxLDL. This resulted in saturation of the cells with unlabelled oxLDL and prevented uptake of the DiI-labelled oxLDL (Figure 3.5B). The experiment was repeated in triplicate for all cell lines under investigation, using incubation times of 12, 24 and 48 hours.

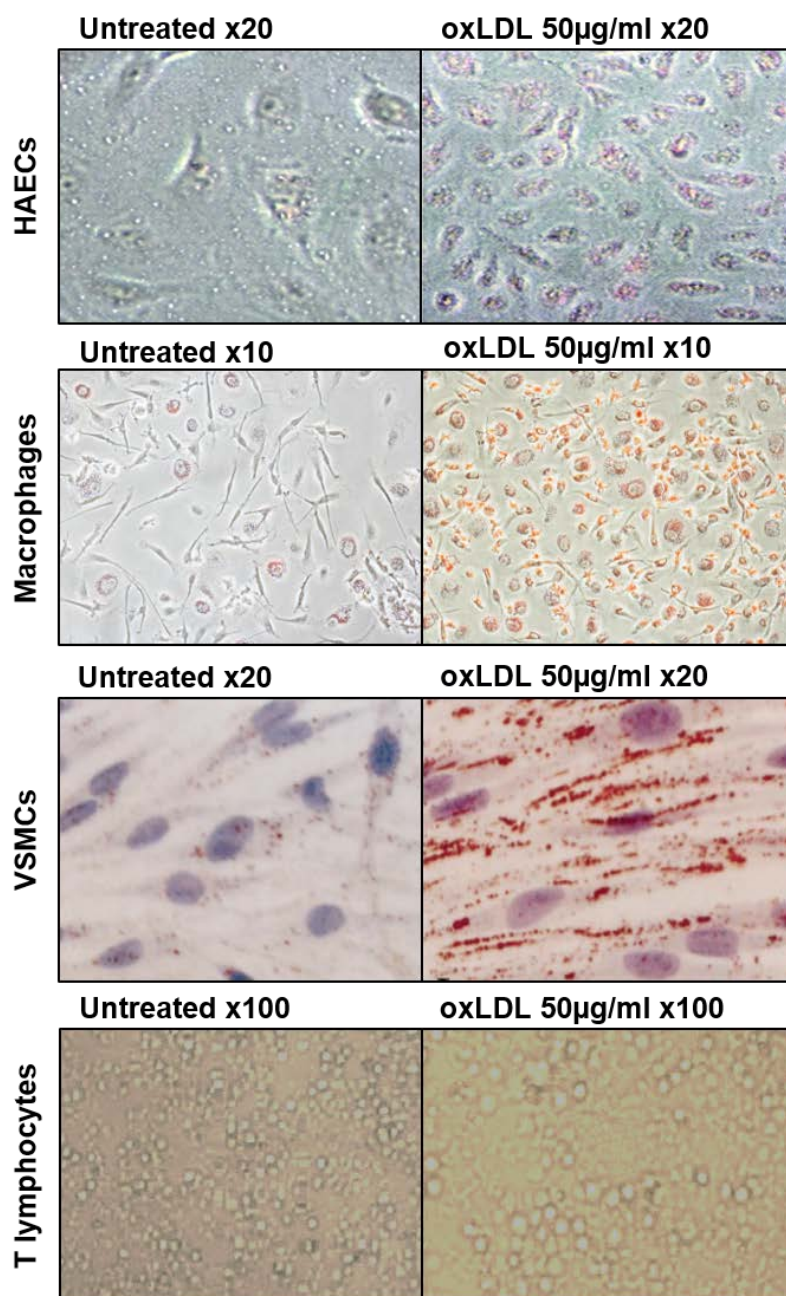


Figure 3.4: Oil Red O staining and phase contrast microscopy of cells treated with oxidised low density lipoprotein (oxLDL). Phase contrast microscopy demonstrates extensive staining with Oil Red O in human aortic endothelial cells (HAECs), vascular smooth muscle cells (VSMCs), and macrophages following incubation with 50 µg/ml of oxidised LDL for 24 hours. There is little change in Oil Red O staining in the T lymphocytes, but the cells appear larger and more clustered together, suggesting activation. Images for macrophages produced by M Reschen.

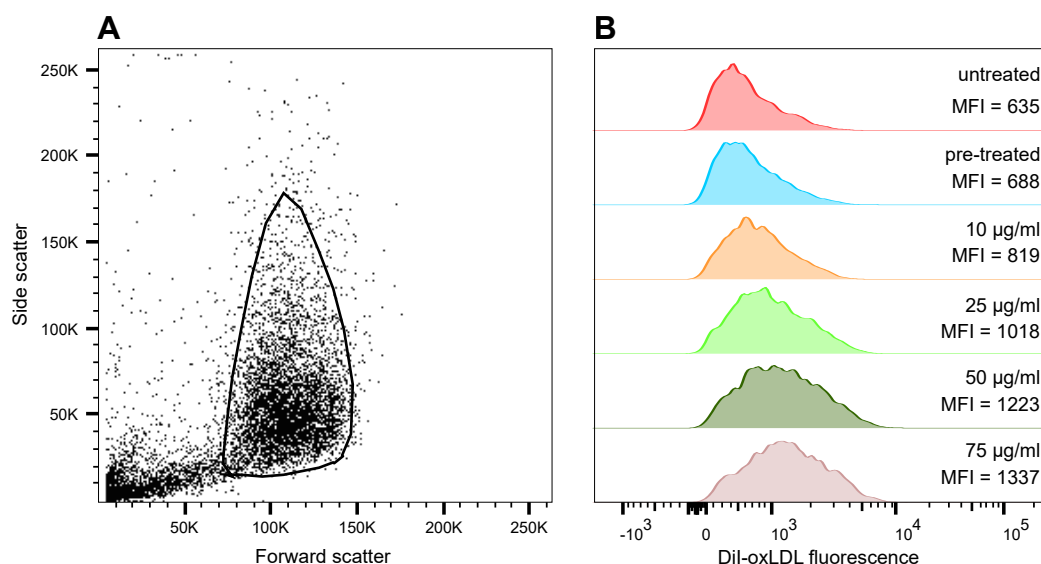


Figure 3.5: Flow cytometric analysis of human aortic endothelial cells (HAECs) treated with DiI-labelled oxLDL. A) Gating strategy used to identify live and dead cells on the basis of forward and side scatter, which measure cell size and granularity respectively. Cells treated with oxLDL have a wide range of granularity depending on the degree of oxLDL uptake, reflecting the greater width of the gate on the side scatter axis. B) Representative flow cytometric analysis of HAECs treated with DiI-labelled oxLDL. Shown is the fluorescence for cells treated without oxLDL or with DiI-labelled oxLDL at increasing concentrations for 12 hours. The specificity of uptake was tested by pre-treating cells with an excess of unlabelled oxLDL at a concentration of 100 $\mu\text{g/ml}$ for 6 hours before incubation with 25 $\mu\text{g/ml}$ of DiI-labelled oxLDL.

Figure 3.6 shows there was an increase in uptake of oxLDL, as indicated by an increase in mean fluorescence intensity (MFI), in a concentration-dependent manner for HAECs, VSMCs, and macrophages. In HAECs following a 12 hour incubation period, the uptake of oxLDL was significant at 50 and 75 $\mu\text{g/ml}$, at which concentrations there were 1.56 and 1.67-fold increases in MFI. This trend carried through to the longer incubations where the equivalent fold changes were 1.77 and 1.86 at 24 hours, and 2.16 and 2.41 at 48 hours. In VSMCs, a similar pattern was seen. At 12 hours, the increase in MFI was significant at 50 $\mu\text{g/ml}$ (1.53-fold), and 75 $\mu\text{g/ml}$ (1.71-fold). This pattern was replicated in the 24 and 48 hour incubations, where the fold increases for 50 and 75 $\mu\text{g/ml}$ were 1.79 and 2.07, and 2.10 and 2.30, respectively. The macrophages took up oxLDL more avidly than the vascular cell types. At 12 hours, the significant increases in MFI were once again seen at 50 and 75 $\mu\text{g/ml}$ (1.44 and 1.64-fold, respectively), but with longer incubation, the change in uptake became significant at even lower concentrations. At 24 hours, the increase in MFI was significant at 25, 50 and 75 $\mu\text{g/ml}$ (1.57, 1.84, and 2.21-fold, respectively) and at 48 hours the uptake was significant even at 10 $\mu\text{g/ml}$. The fold-change increases seen at 10, 25, 50 and 75 $\mu\text{g/ml}$ following 48 hours incubation were 1.57, 1.77, 1.99 and 2.15 respectively. There was much less uptake of oxLDL by CD4⁺ cells. At 48 hours there was a trend of increasing MFI with concentration, but this was not statistically significant.

To determine whether incubation time had a significant effect on oxLDL uptake, we reanalysed the same flow cytometry data to compare the change in MFI of DiI at each concentration over time, using the uptake at 12 hours time as reference. As Figure 3.7 shows, the general trend is for an increase in oxLDL uptake over time for HAECs, VSMCs, and macrophages. This was seen at all concentrations in macrophages, but for the other two cell types, this was seen best in cells treated with 50 $\mu\text{g/ml}$ of oxLDL

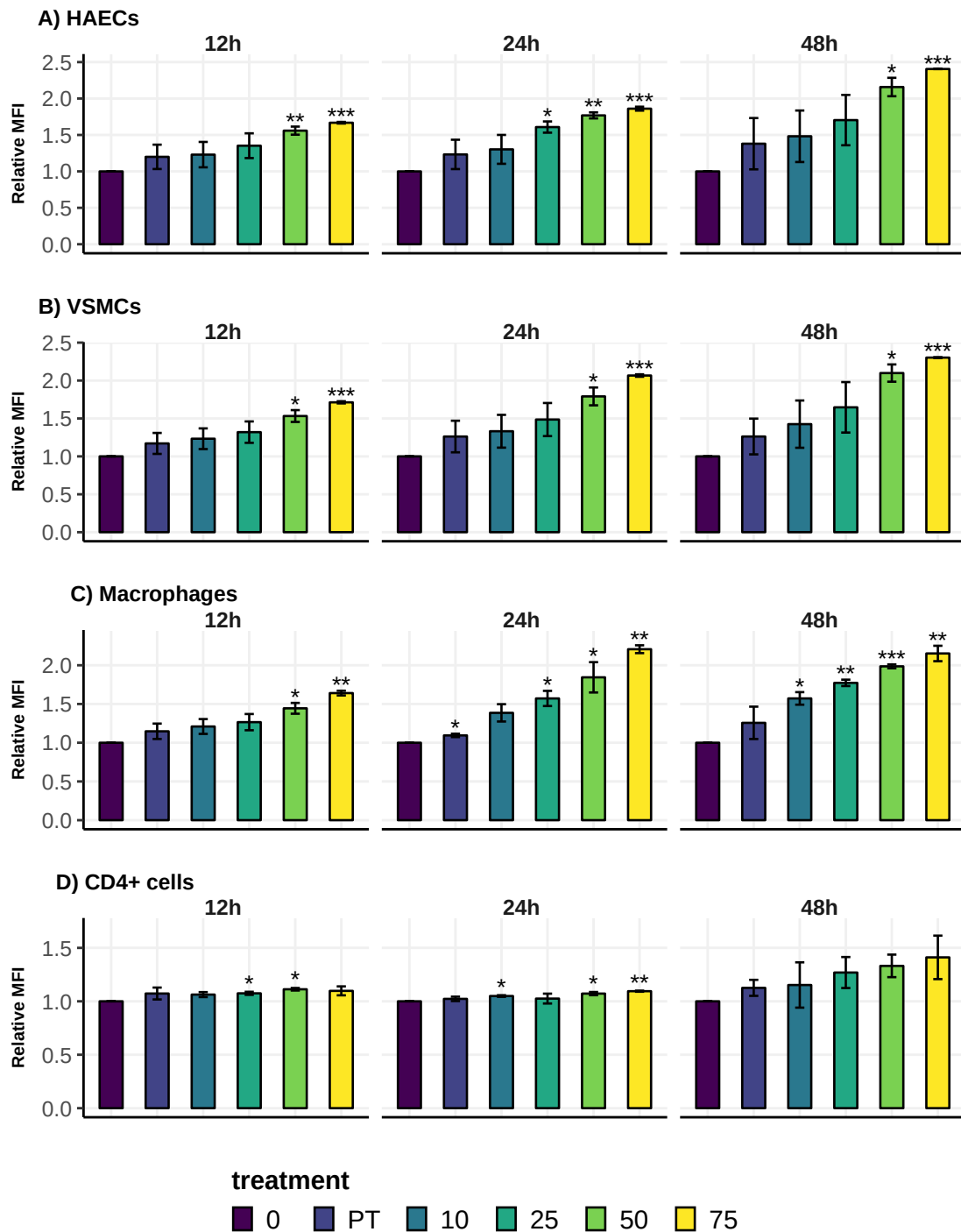


Figure 3.6: Oxidised LDL is taken up by cells in a concentration-dependent manner. Cells were incubated with no DiI-labelled oxLDL (0) or DiI-labelled oxLDL at concentrations of 10, 25, 50 or 75 $\mu\text{g/ml}$ for 12, 24 or 48 hours and DiI fluorescence measured by flow cytometry. The specificity of uptake was tested by pre-treating cells with an excess of unlabelled oxLDL at a concentration of 100 $\mu\text{g/ml}$ for 6 hours before incubation with 25 $\mu\text{g/ml}$ of DiI-labelled oxLDL (PT). Results are shown for HAECs (A), VSMCs (B), macrophages (C) and CD4+ T cells (D). All incubations were performed in triplicate. Significant results are indicated by asterisks (* < 0.05 , ** < 0.01 , *** < 0.001) and indicate a significant change in MFI compared to untreated cells. Error bars indicate standard error of the mean.

for 48 hours. There was no significant time-dependent increase in oxLDL in CD4⁺ T cells.

3.5 Oxidised LDL induces activation of vascular endothelial and smooth muscle cells.

Atherosclerotic lesions contain activated vascular endothelial and smooth muscle cells. These activated cells are characterised by increased cell surface expression of cellular adhesion molecules such as ICAM1. To determine whether oxLDL produced *in vitro* by copper oxidation could induce this phenotype, cells were incubated with oxLDL for 12 hours and flow cytometry used to measure the cell surface expression of ICAM1 after staining with anti-human ICAM1 antibody conjugated with allophycocyanin (APC). Figure 3.8 shows a representative analysis using HAECs as an example. Similar results were obtained for VSMCs. Live and dead cell gates were set up based on forward and side scatter flow cytometry plots, and APC-fluorescence measured for live cells. Figure 3.8A illustrates the live cell gating strategy and Figure 3.8B shows the specificity of the anti-human ICAM1 antibody. Staining of the cells with APC-conjugated isotype control antibody showed only marginal background staining above the unstained cells. Having established that ICAM1 is expressed on the cell surface of HAECs and VSMCs, we sought to establish the effect of oxLDL on its expression in more detail. As a preliminary experiment, HAECs were incubated with increasing concentrations of oxLDL for twelve hours and the change in cell surface expression of ICAM1 was measured. As Figure 3.8C shows, with increasing concentrations of oxLDL there was an increase in ICAM1 expression as determined by an increase in MFI of APC.

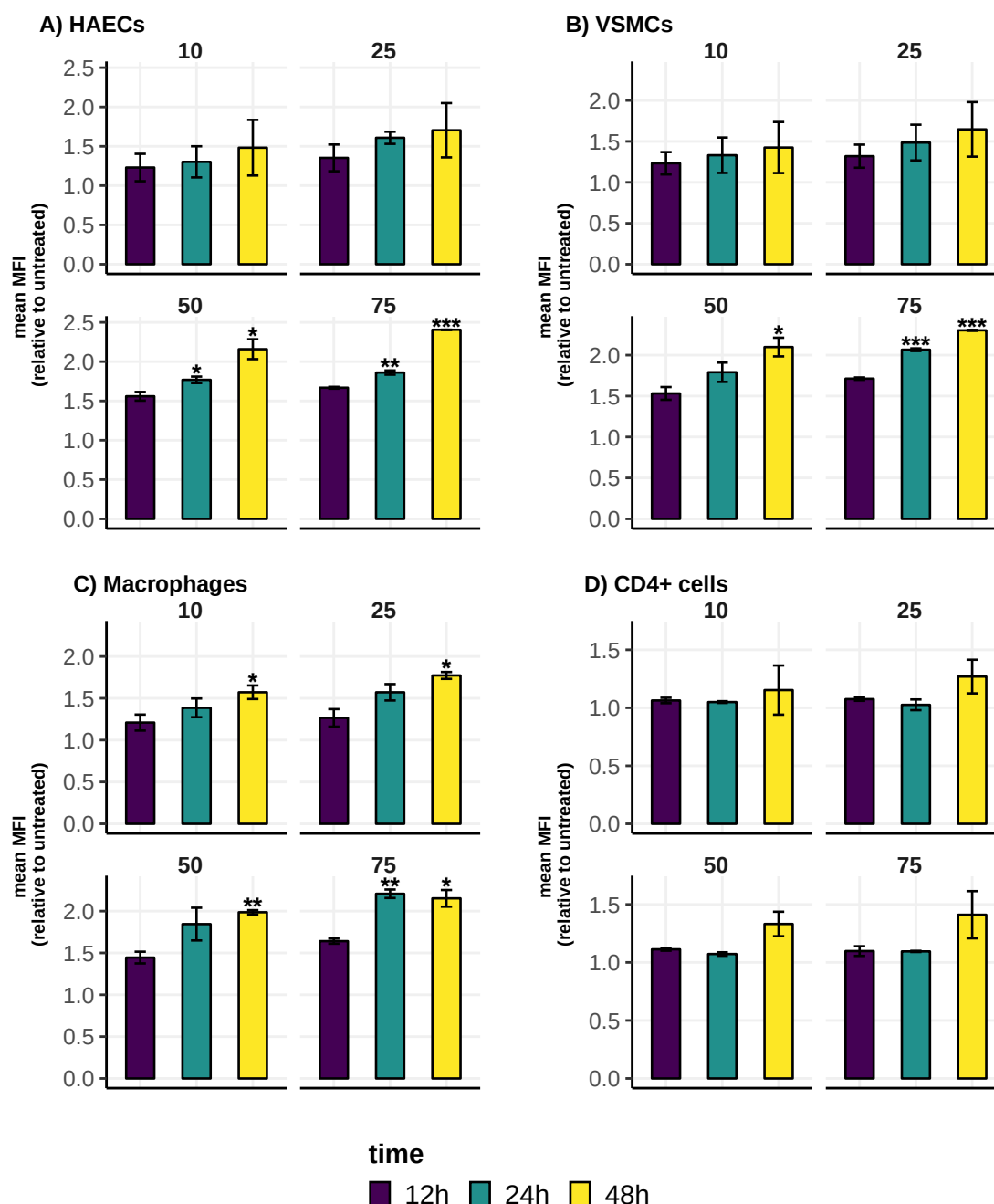


Figure 3.7: Oxidised LDL is taken up by cells in a time-dependent manner. Cells were incubated with DiI-labelled oxLDL at concentrations of 10, 25, 50 or 75 µg/ml for 12, 24 or 48 hours and flow cytometry performed to measure DiI fluorescence. Results are shown for HAECs (A), VSMCs (B), macrophages (C) and CD4+ T cells (D). All incubations were performed in triplicate. Significant results are indicated by asterisks (* < 0.05, ** < 0.01, *** < 0.001) and indicate a significant change in MFI compared to cells treated at the same concentration for 12 hours. Error bars indicate standard error of the mean.

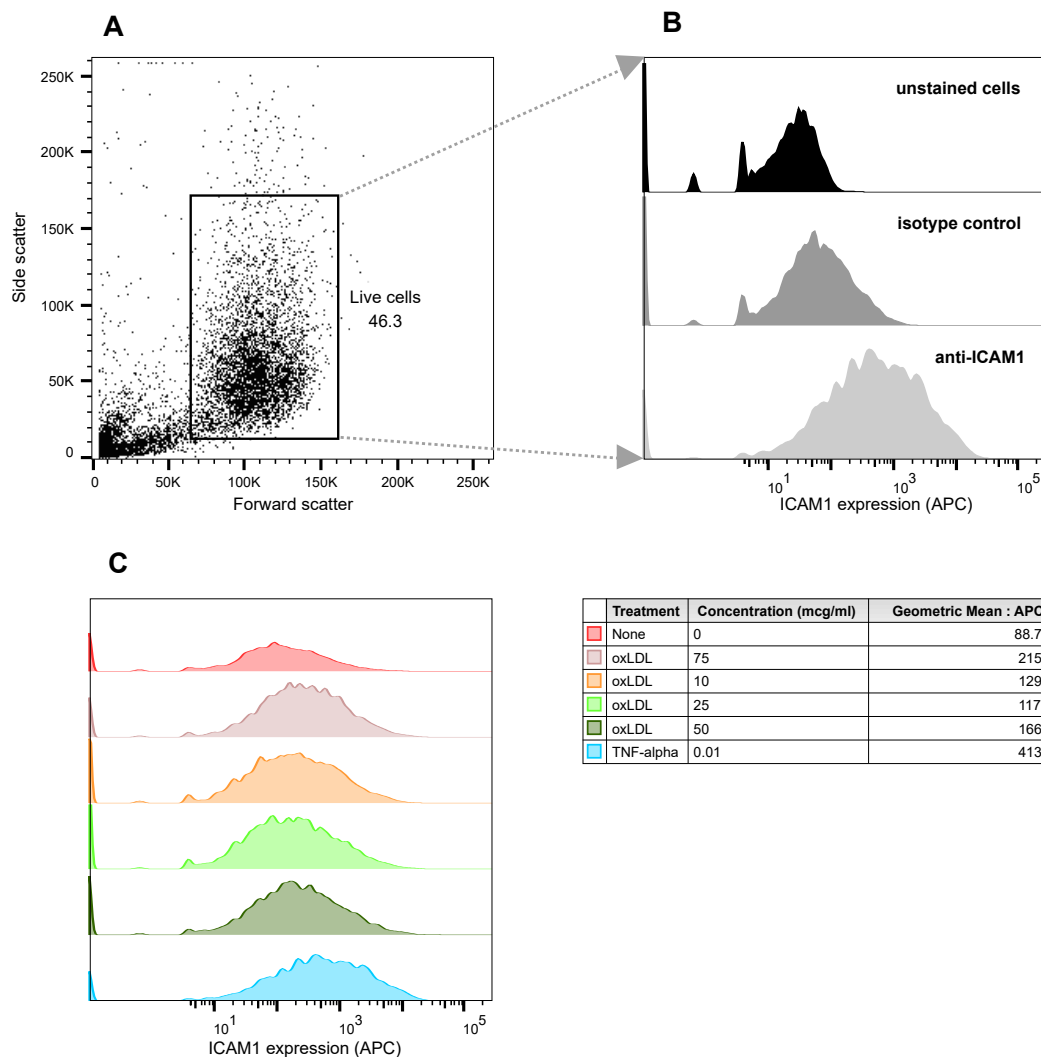


Figure 3.8: Flow cytometric analysis of intercellular cell adhesion molecule 1 (ICAM1) expression in human aortic endothelial cells (HAECs) treated with oxLDL. Cells were incubated with oxLDL at the concentrations shown for 12 hours and cell surface expression of ICAM1 measured by flow cytometric analysis after staining with APC-conjugated anti-human ICAM1 antibody. Untreated cells are used as negative control and TNF-alpha-treated cells used as positive control. The results shown are representative of all replicate HAEC experiments, and similar to those for vascular smooth muscle cells (VSMCs). A) Gating strategy used to identify live and dead cells on the basis of forward and side scatter, which measure cell size and granularity respectively. B) Illustration of the specificity of the APC-conjugated anti-human ICAM1 antibody. Shown is the APC-fluorescence of untreated HAECs that were unstained, stained with APC-conjugated isotype control, or stained with APC-conjugated anti-ICAM1 antibody. C) Measurement of ICAM1 expression on the surface of HAECs treated with increasing concentrations of oxLDL or TNF-alpha (positive control).

To assess the effect of oxLDL on ICAM1 expression quantitatively, the incubation of HAECs and VSMCs with oxLDL at concentrations of 0, 10, 25, 50, and 75 $\mu\text{g}/\text{ml}$ was performed in triplicate, using incubation durations of 12, 24, and 48 hours. Following incubation, cells were stained with APC-conjugated anti-human ICAM1 antibody as described in the methods, and then analysed by flow cytometry. As Figure 3.9 and Figure 3.10 illustrate there was a concentration and time-dependent increase in ICAM1 expression in live HAECs and VSMCs following incubation with oxLDL. At all time points examined, there was a significant increase in ICAM1 expression in cells treated with 75 $\mu\text{g}/\text{ml}$ compared to untreated cells. The fold-changes in expression (as determined by change in MFI of APC) in HAECs treated at this concentration for 12, 24 and 48 hours were 1.53, 1.73, and 2.11, respectively. The equivalent fold-changes in VSMCs were 1.39, 2.01, and 1.96. When incubated with oxLDL at a concentration of 50 $\mu\text{g}/\text{ml}$ the increase in expression of ICAM1 became significant in HAECs at 24 and 48 hours only (1.56, and 1.88-fold increase). In VSMCs incubation with 50 $\mu\text{g}/\text{ml}$ of oxLDL resulted in significant increase in ICAM1 expression at 48 hours (1.79-fold increase), but not at 24 hours.

3.6 Oxidised LDL induces cell death at high concentrations

We have shown that oxLDL affects cell phenotype and function in a concentration and time-dependent manner, but that when incubated with oxLDL at concentrations above 50 $\mu\text{g}/\text{ml}$ there was increased cell debris in all cell types (Figure 3.4). Previous work by this group [unpublished] and elsewhere^{39–42} has shown that high concentrations of oxLDL can induce cell death. To investigate whether the increase in cell debris at

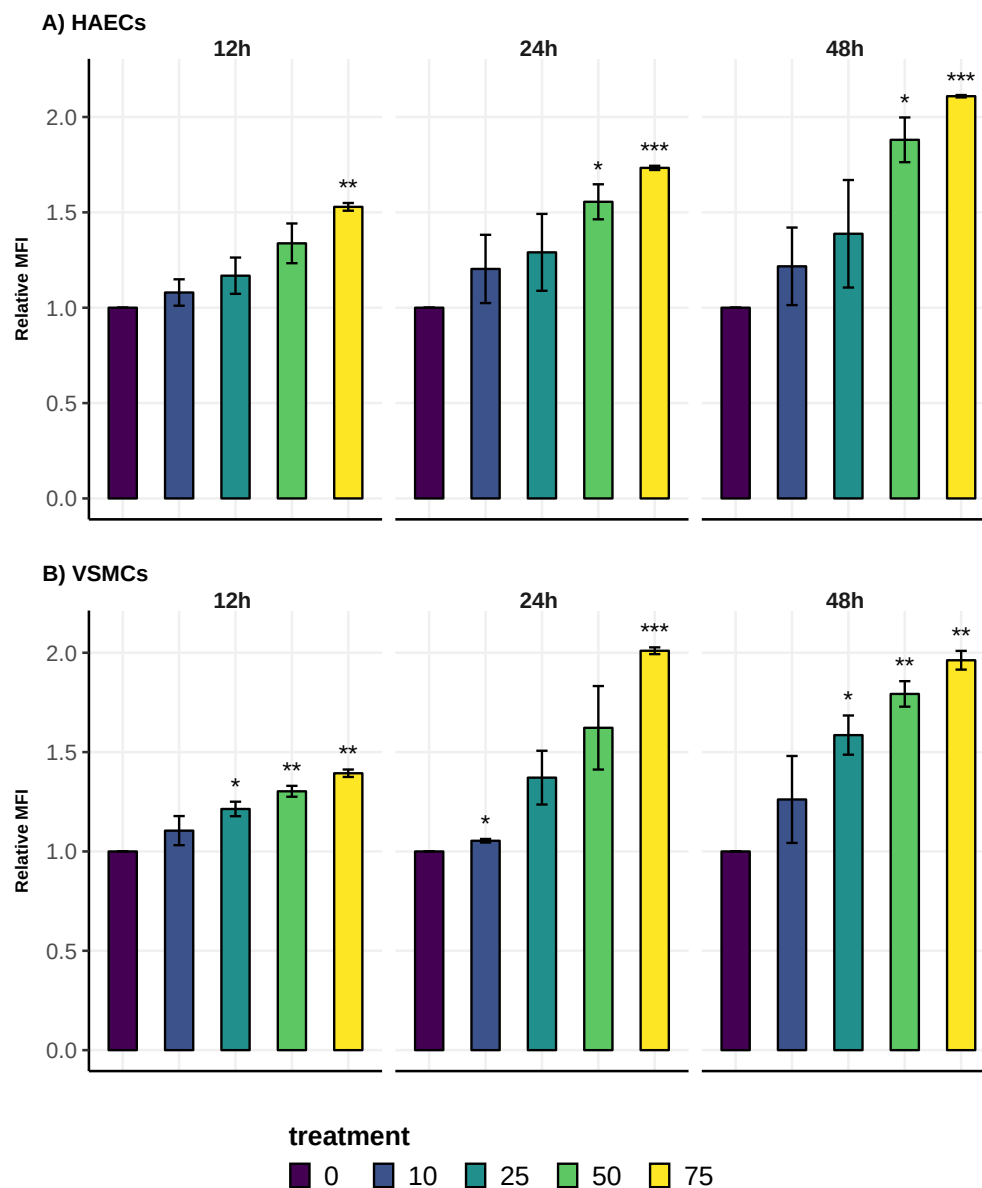


Figure 3.9: Oxidised LDL induces cell surface expression of ICAM1 in human aortic endothelial and coronary artery smooth muscle cells in a concentration-dependent manner. Cells were incubated with oxLDL at concentrations of 10, 25, 50 or 75 µg/ml for 12, 24 or 48 hours and ICAM1 expression was measured by flow cytometry after labelling the cells with APC-conjugated anti-ICAM1 antibody. Results are shown for HAECs (A), and VSMCs (B). All incubations were performed in triplicate. Significant results are indicated by asterisks (* < 0.05, ** < 0.01, *** < 0.001) and indicate a significant change in MFI compared to untreated cells. Error bars indicate standard error of the mean.

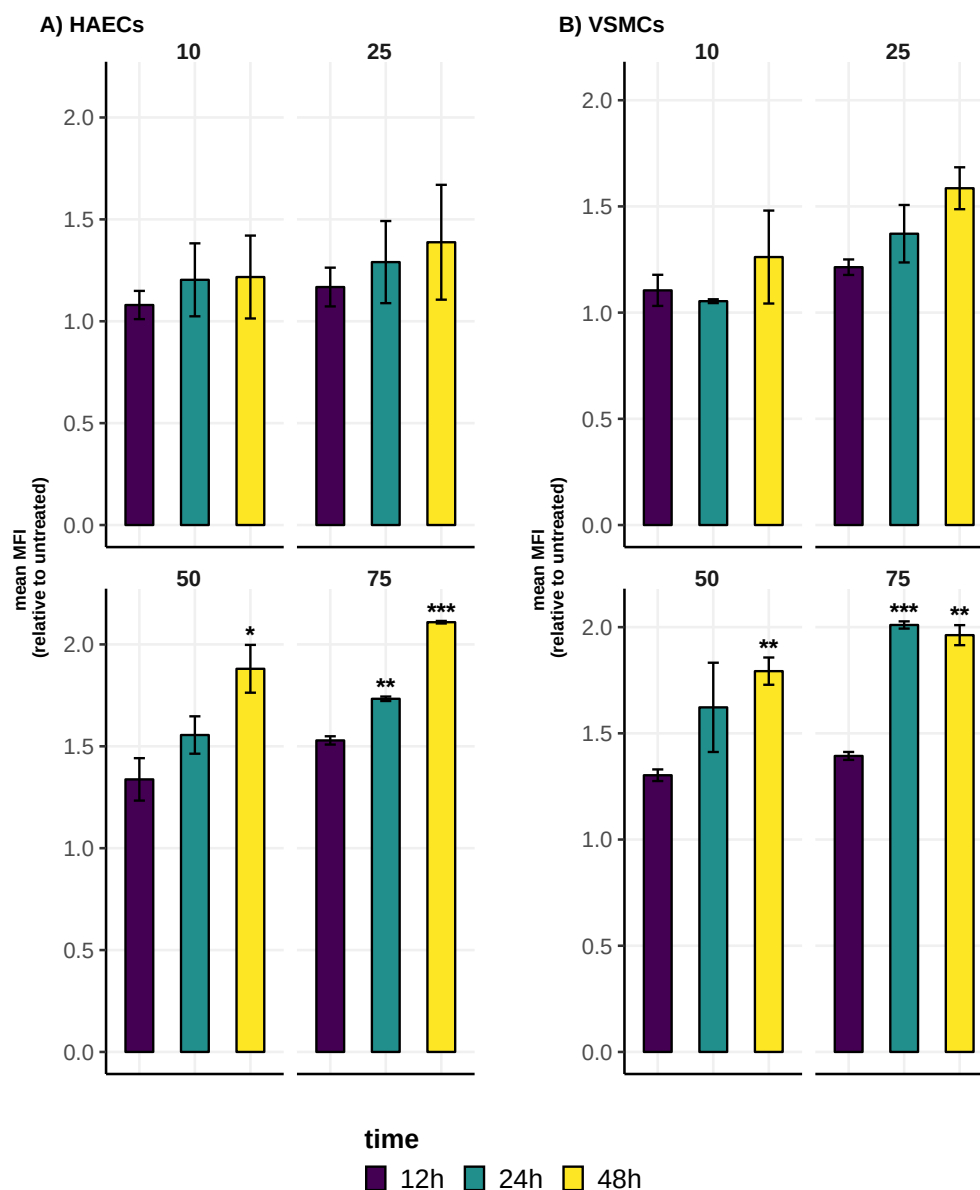


Figure 3.10: Oxidised LDL induces cell surface expression of ICAM1 in human aortic endothelial and coronary artery smooth muscle cells in a time-dependent manner. Cells were incubated with oxLDL at concentrations of 10, 25, 50 or 75 µg/ml for 12, 24 or 48 hours and ICAM1 expression was measured by flow cytometry after labelling the cells with APC-conjugated anti-ICAM1 antibody. Results are shown for HAECs (A), and VSMCs (B). All incubations were performed in triplicate. Significant results are indicated by asterisks (* < 0.05 , ** < 0.01 , *** < 0.001) and indicate a significant change in MFI compared to cells treated at the same concentration for 12 hours. Error bars indicate standard error of the mean.

higher concentrations of oxLDL might be attributable to cell death, the cells were incubated for 48 hours with oxLDL at concentrations of 10, 25, 50, 75, or 100 µg/ml or with the equivalent volume of buffer (i.e. the vehicle medium for the oxLDL). As a negative control, the cells were incubated in normal growth medium alone, and as a positive control the cells were incubated with camptothecin, which is known to chemically induce apoptosis. After 48 hours, the cells were stained with APC-labelled Annexin V and 7-amino-actinomycin D (7-AAD), and then analysed by flow cytometry.

Figure 3.11 illustrates a typical analysis and set of results, using HAECs as an example. At 48 hours, 81% of cells treated with camptothecin were in early or late apoptosis, compared to 34.3% of untreated cells. In this example, which is typical of the experimental results, buffer-treated cells had a similar number of viable cells regardless of the volume of buffer used. In contrast, with increasing concentration of oxLDL, the number of viable cells decreased. At a concentration of 10, 25, 50, 75 and 100 µg/ml, 66.1%, 66.2%, 65.8%, 53.4%, and 26.3% of cells respectively, were viable at 48 hours.

To confirm this result across the different cell types, we performed the experiment in triplicate on HAECs, VSMCs, macrophages and CD4⁺ T cells. Figure 3.12 shows the results for each cell type. The significance values are for the proportion of live cells following incubation with oxLDL or camptothecin compared to untreated cells. The overall survival of HAECs was static up to 50 µg/ml with survival of more than 60% of cells, although amongst the injured cells, the proportion of apoptotic cells decreased whilst the proportion of necrotic cells increased. At concentrations above 50 µg/ml, there was a marked decrease in cell survival. At 75 µg/ml cell survival was 54.6% and at 100 µg/ml, the survival was 26.6%, which was very similar to that seen with camptothecin-treated cells (18.8%). The pattern of survival in VSMCs was

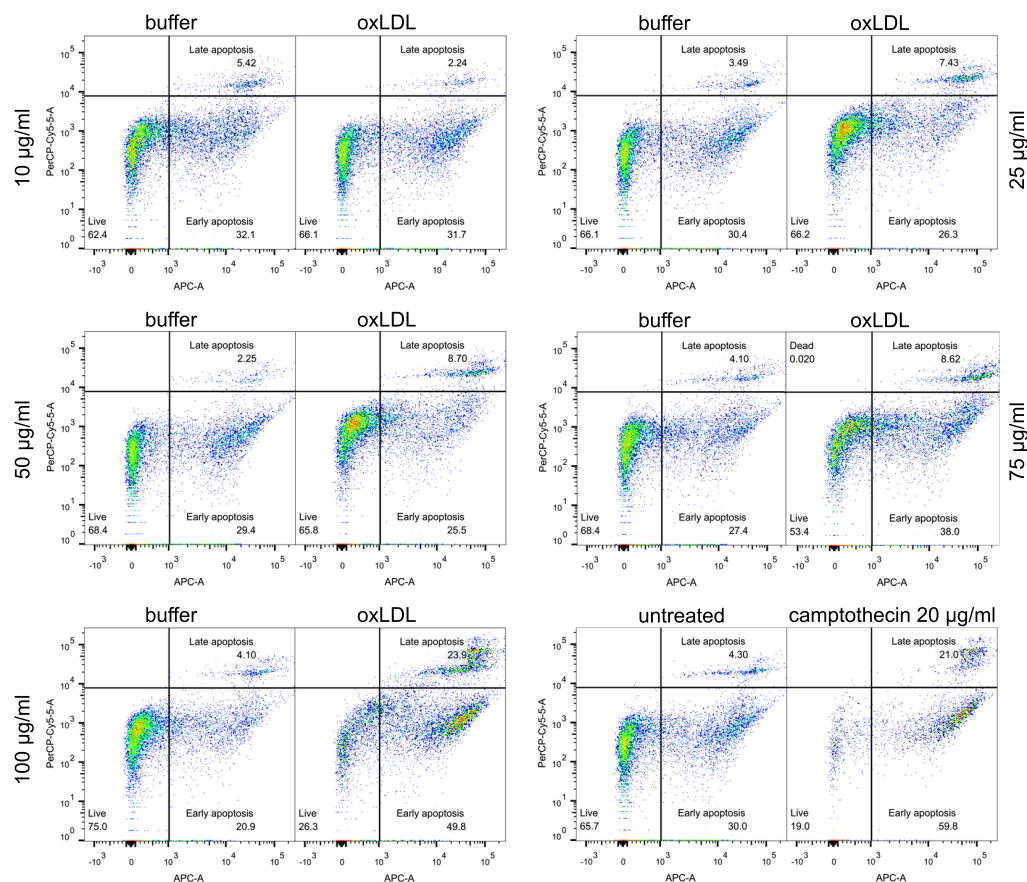


Figure 3.11: Oxidised LDL induces cell death in cells at high concentrations. HAECs were incubated with culture medium supplemented with oxidised LDL at concentrations of 0, 10, 25, 50, 75 and 100 µg/ml. As negative controls, the equivalent volume of buffer solution was used, and camptothecin was used as a positive control. After incubation for 48 hours, cells were stained with APC-labelled Annexin V and 7-AAD and analysed by flow cytometry. At concentrations above 50 µg/ml there was significant cell death. LDL = low density lipoprotein, HAEC = human aortic endothelial cells.

very similar. Again, with incubation with oxLDL up to concentrations of 50 $\mu\text{g/ml}$, cell survival was in excess of 60%. However, at 75 $\mu\text{g/ml}$, this dropped to 51.7% and at 100 $\mu\text{g/ml}$ to 19.4% (similar to camptothecin treatment, with which cell survival was 22.3%). In macrophages, the deleterious effect of oxLDL on cell survival showed statistical significance even at 50 $\mu\text{g/ml}$, where the cell survival was 57.7%. This was even more pronounced at 75 and 100 $\mu\text{g/ml}$, where the respective survivals were 51.4% and 19.5%, respectively. The macrophages were very sensitive to camptothecin exposure with a cell survival of only 12.5%. CD4^+ T cells had a very static survival response to incubation with oxLDLs up to and including a concentration of 75 $\mu\text{g/ml}$ with no apparent change in the proportion of live cells ($>60\%$ throughout). However, at 100 $\mu\text{g/ml}$ there was a sharp and significant decrease in the number of live cells down to 36.7% and following camptothecin treatment, cell survival dropped to 13.7%.

3.7 Discussion and Summary

The main aim of this chapter was to establish a reproducible method of oxLDL preparation and to assess its effect on vascular and immune cells to identify the optimal concentration and duration of exposure to use for subsequent experiments. The main findings were that oxLDL can indeed be reproducibly made and that it has a time and concentration dependent effect on cell morphology and function, but that above 50 $\mu\text{g/ml}$ it induces significant cell death, with fewer than 50% viable cells in most cell types. At this concentration, the maximal effects of oxLDL were seen at 48 hours.

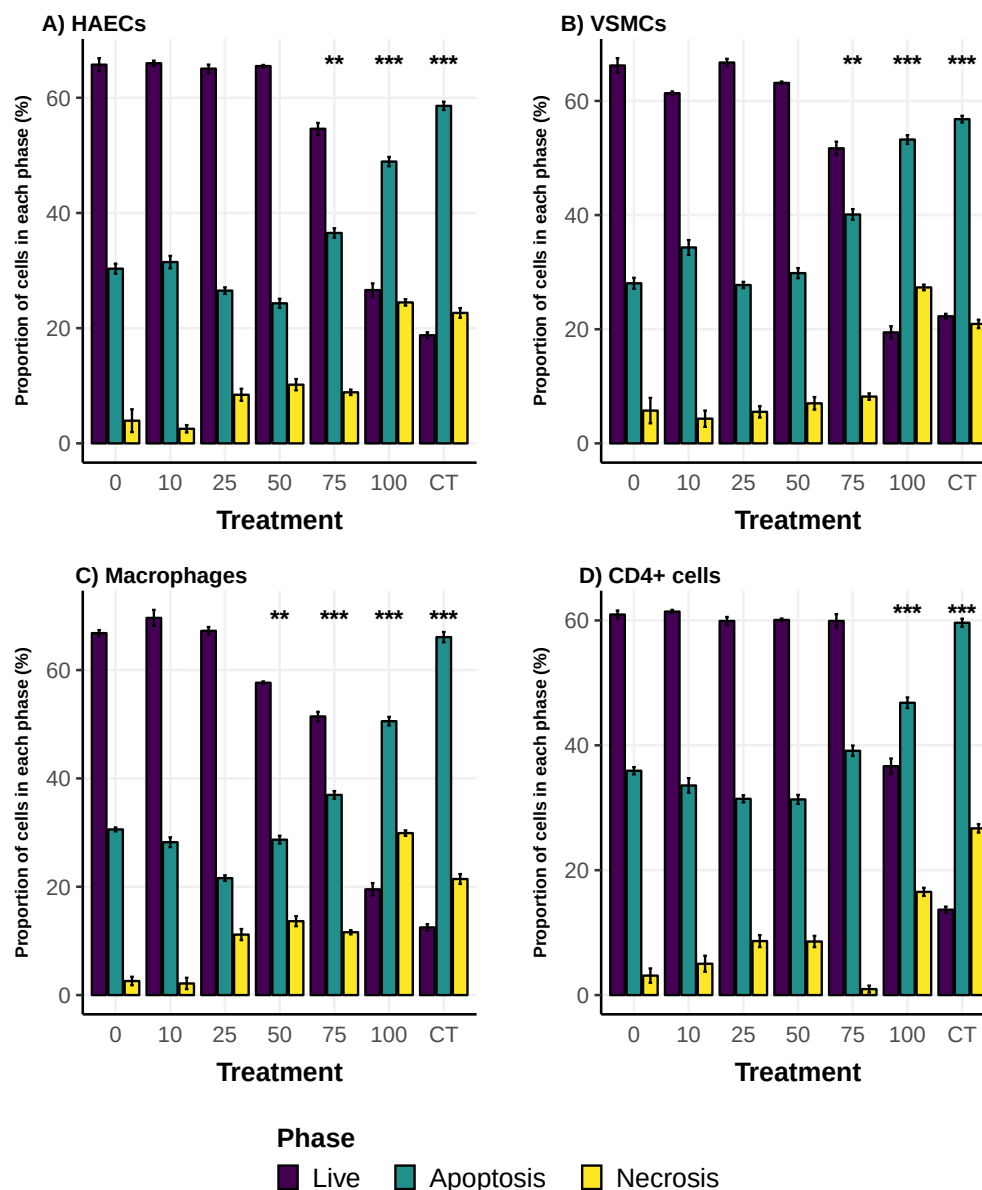


Figure 3.12: Oxidised LDL at high concentrations induces cell death. Cells were incubated for 48 hours with oxLDL at concentrations of 10, 25, 50, 75, or 100 µg/ml. As a negative control, the cells were incubated in normal growth medium alone, and as a positive control the cells were incubated with camptothecin, which is known to chemically induce apoptosis. After 48 hours, the cells were stained with APC-labelled Annexin V and 7-AAD, and then analysed by flow cytometry. Results are shown for HAECs (A), VSMCs (B), macrophages (C) and CD4+ T cells (D). All incubations were performed in triplicate. Significant results are indicated by asterisks (* < 0.05, ** < 0.01, *** < 0.001) and indicate a significant change in the proportion of live cells compared to cells that were incubated without oxLDL. Error bars indicate standard error of the mean. CT = camptothecin.

3.7.1 Characterisation of oxidised LDL

The precise mechanisms of oxLDL-induced changes in cellular physiology are difficult to determine as the oxidation of LDL is a complex and heterogeneous process resulting in the formation of a variety of modifications to either the free fatty acid, lipid or protein components, the properties of which vary depending on the method and extent of oxidation. Surprisingly few authors have characterised the oxLDL used in their studies except to state that it is oxidised; this may well be the explanation for the discrepancies in the reported biological effects of oxLDL as discussed in chapter 1. To overcome this variability we developed a robust and reproducible protocol that results in oxLDL preparations with a similarly high level of oxidative modification, which we confirmed using two assays.

We first used agarose gel electrophoresis to measure the REM of oxLDL compared to native LDL. Oxidation of the apoB-100 component of LDL results in an overall net negative charge due to reactions between positively charged lysine and arginine residues and aldehydes formed from modification of lipid components, and due to conversion of histidine and proline residues to aspartate and glutamine³⁶. The results here show increased mobility of oxLDL confirming its oxidation. The mean REM of the oxidised LDL was 3.1 compared to native LDL. These values are similar to those reported in the literature^{45,48} where very highly oxidised LDL was used.

We also used the measurement of TBARS to determine the oxidation status of the modified LDL. The TBARS assay measures the amount of MDA, formed during oxidation, by colorimetric reaction with thiobarbituric acid, which can be spectrophotometrically quantified by absorbance at 532nm. The oxLDL prepared in this study had a very high TBARS level of 372-fold greater than the native LDL. This is in contrast to some other studies which reported between 9-fold⁴⁵ and 52-fold⁴⁸ greater TBARS than the

native LDL from which it was derived. The differences in the disparity in the TBARS values result partly from the naturally heterogeneous nature of LDL, but probably more importantly from the differences in preparation method. The present study used 25 $\mu\text{mol/l}$ CuCl_2 to oxidise 0.5 mg/l of oxLDL whereas previous studies have used 300 $\mu\text{mol/l}$ to oxidise 3 mg/ml⁴⁸ and 500 $\mu\text{mol/l}$ to oxidise 10 mg/ml⁴⁵. These studies have also used 12 hours of incubation with oxLDL compared to the 18 hours used in the present study.

The high value of TBARS and the change in REM seen with the modified LDL preparations indicate that the oxLDL prepared in this study has been heavily oxidised.

It is important to mention that the TBARS assay is not specific for oxidation, and other substances such as amino acids and sugars can also react with thiobarbituric acid. Other assays that have been used to overcome this lack of specificity include measurement of conjugated diene formation and measurement of oxysterols by high performance liquid chromatography. These assays are cumbersome to use routinely, but nonetheless, they would be of interest in future work to further assess the oxidation status of LDL.

3.7.2 Cytotoxicity of oxidised LDL

Toxicity of oxLDL to the cells was measured using an Annexin V and 7-AAD-based flow cytometry assay. Other studies have used the trypan blue assay to measure failure of the cell membrane¹³⁶, but this is a late sign of cell injury, and is less sensitive to early cell death. Annexin V is an intracellular protein that binds to phosphatidylserine (PS), which is normally located on the intracellular side of the plasma membrane. In early apoptosis, membrane asymmetry is lost and PS translocates to the extracellular side, allowing labelled Annexin V to target and identify apoptotic cells. Annexin V

binding alone cannot differentiate between apoptotic and necrotic cells. To distinguish between these, we used 7-AAD, a DNA-binding molecule that is efficiently excluded by cells with intact plasma membranes. Early apoptotic cells will exclude 7-AAD, while late stage apoptotic or necrotic cells will stain positively, due to the passage of these dyes into the nucleus where they bind to DNA. When excited by 488 laser light, 7-AAD fluorescence may be detected in the far red range of the spectrum. The loss of cell viability was dependent on the oxLDL concentration. In this study, oxLDL concentrations of above 50 µg/ml were toxic to HAECs and VSMCs. Below this concentration there were no significant changes in cell viability at 48 hours compared to untreated cells, with more than 60% cell viability. Above this concentration, there was a significant decrease in cell viability. In contrast, in macrophages, there was a significant decrease in macrophage viability even at 50 µg/ml, although the number of viable cells remained above 50%. CD4⁺ cells were the least sensitive to oxLDL; a significant tail off in cell viability was seen only when they were incubated with oxLDL at a concentration of 100 µg/ml. The serum concentration of oxLDL *in vivo* is unclear; the ranges reported in the literature vary depending on the assay used and population studied, but under normal circumstances appears to be no more than 50 µg/ml²²⁻²⁸. This is consistent with the results from this study, where concentrations above 50 µg/ml resulted in significant cell death.

3.7.3 Effects of oxidised LDL on cell morphology

The cytotoxic effects of oxLDL are in part related to its cellular uptake. Phase contrast microscopy of oxLDL-treated cells stained with Oil Red O show that the cells most vulnerable to cytotoxicity (i.e., HAECs, VSMCs, and macrophages) are those which take up oxLDL. At a concentration of 50 µg/ml there was evidence of lipid droplet

accumulation in these cells; in contrast, this was not apparent in the CD4⁺ cells, which were also the most resistant to oxLDL-induced cytotoxicity. OxLDL uptake was also associated with changes in cell shape in macrophages, which became larger and more rounded, in keeping with foam cell formation. Foam cell formation has also been reported in VSMCs, but this was not apparent in this study. Previous studies which have reported this have only observed this phenotype when using minimally oxidised LDL and probably reflects a balance between lipid accumulation and cell toxicity. It is likely that minimally oxidised LDL is able to accumulate to higher levels than the highly oxidised LDL used in this study, thus allowing foam cell formation in VSMCs. The lack of uptake of oxLDL by CD4⁺ cells is in keeping with the known distribution of scavenger receptors - SR-A and CD36 are both extensively expressed in macrophages and VSMCs and SR-E is highly expressed in ECs. No specific oxLDL scavenger receptors are expressed in CD4⁺ T cells.

3.7.4 Changes in expression of cell adhesion molecules in response to oxidised LDL

In this study, vascular cells treated with oxLDL had significantly increased expression of the pro-inflammatory molecule ICAM1, which is recognised as a marker of activation. ICAM1 plays a pivotal role in adhesion of circulating immune cells to the vascular wall, and facilitates the transition from cell rolling to crawling and finally to firm adhesion. Following arrest of circulating immune cells on the vascular wall, transmigration of these cells into the arterial intima is mediated by chemokines secreted by vascular wall cells and intimal macrophages. Reports of the degree of induction of cell surface expression of ICAM1 by oxLDL is variable. Some studies have indicated that less heavily modified oxLDL is more effective in inducing ICAM1 expression than more

heavily modified preparations^{137,138}. This may in part be due to oxLDL-induced cytotoxicity, but also due to the formation of oxLDL aggregates by highly oxidised oxLDL¹³⁹ which are less effective in stimulating ICAM1 expression. To ameliorate this the oxLDL preparations were sterile-filtered before storage, but it is possible that aggregates formed subsequently. OxLDL has also been shown to alter the expression of other cell adhesion molecules such as E-selectin, P-selectin and VCAM1⁶³, and also CD40 and CD40L⁶⁴. In future studies it would be of interest to investigate these additional cell surface markers; combining expression data of multiple cell surface molecules would be useful in establishing a more robust method of identifying activated vascular cells.

3.7.5 Summary

To summarise, this chapter has provided evidence that the oxLDL used in this study is of a highly oxidised nature, and when cells are incubated with this oxLDL at a concentration of 50 µg/ml for 48 hours, there is more than 50% cell viability across all cell types, and that there is uptake of oxLDL by HAECs, VSMCs, and macrophages, and that macrophages acquire a foam cell phenotype. In addition, vascular cells become activated and show increased expression of ICAM1. Under the incubation conditions used here, CD4⁺ cells became larger and more clumped, suggestive of cell activation. An important point to mention is that whilst these cells were CD4⁺, as the selection method was a positive selection method, it is possible that these cells represented a mixed population of CD8⁺ and CD8⁻ cells. This may have had some influence on the responses seen, and in the future it would be interesting to see if there was a more pronounced response to oxLDL if the CD8⁺ population were removed. In subsequent chapters we use genome wide approaches to investigate the potential

mechanisms for these and other effects of oxLDL.

Chapter 4

Development of custom bioinformatic and statistical analysis tools

Contents

4.1	Introduction	76
4.2	RNA-seq analysis: the <i>rseqR</i> package	81
4.2.1	Methods	81
4.3	ChiP- and ATAC-seq analysis: the <i>chompR</i> package	87
4.3.1	Methods	87
4.4	Trait-associated SNP enrichment: the <i>snpR</i> package	94
4.4.1	Methods	95
4.5	Discussion	97
4.5.1	Comparisons with other tools	98
4.6	Conclusions	102

4.1 Introduction

Over the past 15 years, there have been tremendous advances in genome sequencing technologies¹⁴⁰. Highly parallelised methods, termed next generation sequencing (NGS) or high throughput sequencing (HTS), allow rapid sequencing of millions of DNA molecules in a single sequencing run¹⁴¹, and as the technology has become more available, the cost of sequencing has greatly reduced. Various applications of HTS technologies have provided important insights into the mechanisms and regulation of genome architecture and gene expression^{130,142}. Despite the diversity of techniques available, the work flow of these methods are conceptually similar. In general, the first step is enrichment of the genomic sequences of interest and library preparation from small fragments of DNA (Figure 4.1). In RNA-seq, total RNA is first extracted, but must then be purified for messenger ribonucleic acid (mRNA) which comprises only a small percentage of all RNA species. Enrichment of eukaryotic mRNAs is done by virtue of their poly(A) tail, which is hybridised to oligo-dT molecules connected

to a carrier. Non-binding nucleic acids are ‘washed off’ and the remaining poly(A)-mRNA is then eluted from the oligo-dT/carrier, sheared into small fragments and then reverse transcribed into complementary DNA (cDNA), to which library adaptors are ligated. To perform a histone ChIP assay, histones are first covalently crosslinked to genomic DNA with formaldehyde, and chromatin is then isolated and fragmented by sonication. Target protein-DNA complexes are captured using antibodies specific to the histone of interest. After reversal of crosslinks, the DNA obtained by chromatin immunoprecipitation is then purified for sequencing and library adaptors ligated. ATAC-seq depends on a modified version of the hyperactive Tn5 transposase to simultaneously cut and ligate adaptors at regions of chromatin accessibility, but is prevented from doing so in genomic regions tightly packed into nucleosomal arrays. The resulting DNA fragments are thus enriched for open/accessible chromatin, and are ready to be sequenced.

There are now several HTS technologies available, and all share the same general steps but perform them in different ways. In general, following library preparation, the DNA fragments are clonally amplified following their physical separation by immobilisation to a solid surface or beads. Sequencing is then applied in parallel to all the amplified fragments. A variety of sequencing techniques exist. Illumina technologies depend on cyclic reversible termination, Applied Biosystems SOLID on sequencing by ligation, and Roche 454 on pyrosequencing. Whichever technology is used, the output of the process is a sequence of images consisting of bright spots indicating the base incorporated. In the case of Illumina and Applied Biosystems SOLID, the image is produced using fluorescently labelled nucleotides, and for Roche 454 using a bioluminescence method. Base calling tools are then used to convert these images into text files containing the nucleotide sequence for each cluster. These files

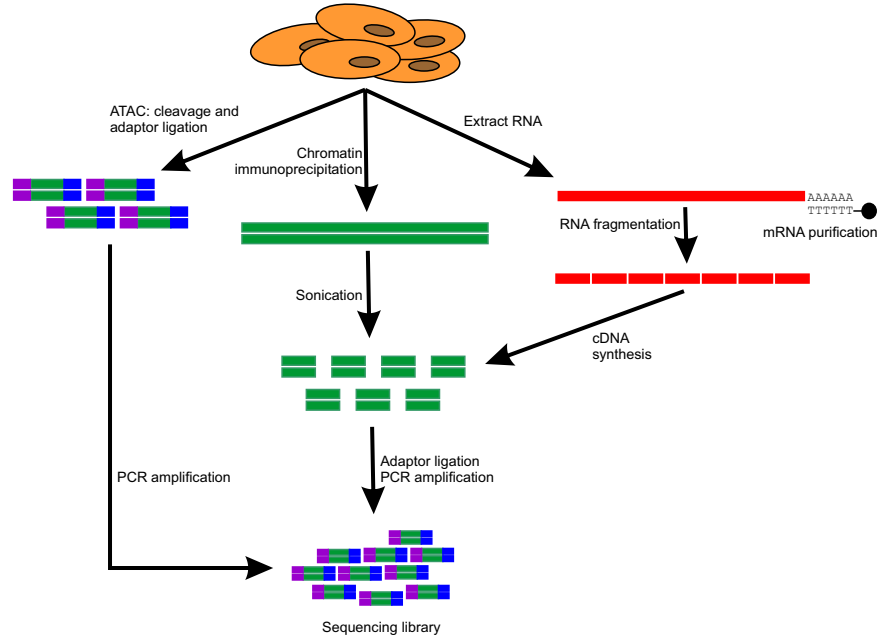


Figure 4.1: Library preparation for high-throughput sequencing. Fragmented DNA is ligated to adaptors and subjected to amplification. For ATAC-seq, fragmentation and ligation occur as a single step. In ChIP-seq, chromatin is first immunoprecipitated before sonication to shear into smaller fragments. In RNA-seq, mRNA is first sheared and then reverse transcribed to cDNA.

are typically in the FASTQ format. In this format, each read (cluster) is represented by a description line, containing a unique ID, the sequence for that read¹⁴³, and the quality score per base¹⁴⁴. The quality score, also known as the Phred quality score Q is determined by the base calling tool, and is an integer value representing the estimated probability that a base has been called incorrectly. If P is the probability that a base has been called incorrectly, then the Phred score is defined as $Q = -10\log_{10}P$.

As HTS technologies have become more available, the emphasis in sequencing experiments has shifted from generation of data to its storage and analysis. A single sequencing run can generate data in the range of tens to hundreds of gigabytes depending on sequencing depth and genome coverage, and the demand for software that allows researchers to handle these large data sets and extract biologically meaningful

information in a straight-forward manner has escalated. Analysis of HTS data involves a number of steps: quality checking of raw data and filtering of poor quality or contaminant sequences, mapping of reads to a reference genome, feature detection and estimation, and statistical analysis of these estimates to extract meaningful data Figure 4.2.

Several tools have been developed to support the individual aspects of this workflow. FastQC¹⁴⁵ is arguably the most widely used software for assessing quality of raw data, and FASTX¹⁴⁶ and CutAdapt¹⁴⁷ provide a collection of tools to process and filter this data. A large number of aligners, many of which are splice-aware are now available to map reads to a reference genome^{148–152}. Feature detection and estimation is performed by a range of tools. Popular software for gene/transcript estimation in RNA-seq reads include HTseq¹⁵³, featureCounts¹⁵⁴, bedtools¹⁵⁵, and Salmon¹⁵⁶. The most popular tool for peak detection in ChIP-seq and ATAC-seq reads is arguably MACS2¹⁵⁷. Strategies to identify differentially expressed genes or peaks have been implemented in a number of tools^{158–161}, and for more downstream analyses such as gene set enrichment, gene association, motif enrichment and SNP enrichment, a plethora of tools exist.

Although tools for these individual steps exist, combining them into a single reliable analysis pipeline poses significant challenges. The output of one tool is often not compatible with the input of another, and the configuration options chosen for one tool may affect those downstream, making it necessary to have a thorough knowledge of the theory behind and workings of each bit of software used. Hence, running a complete HTS analysis requires considerable bioinformatics skills, and this is an obstacle for many non-specialist researchers.

To this end, I have developed a set of tools that allow for an integrated, platform-

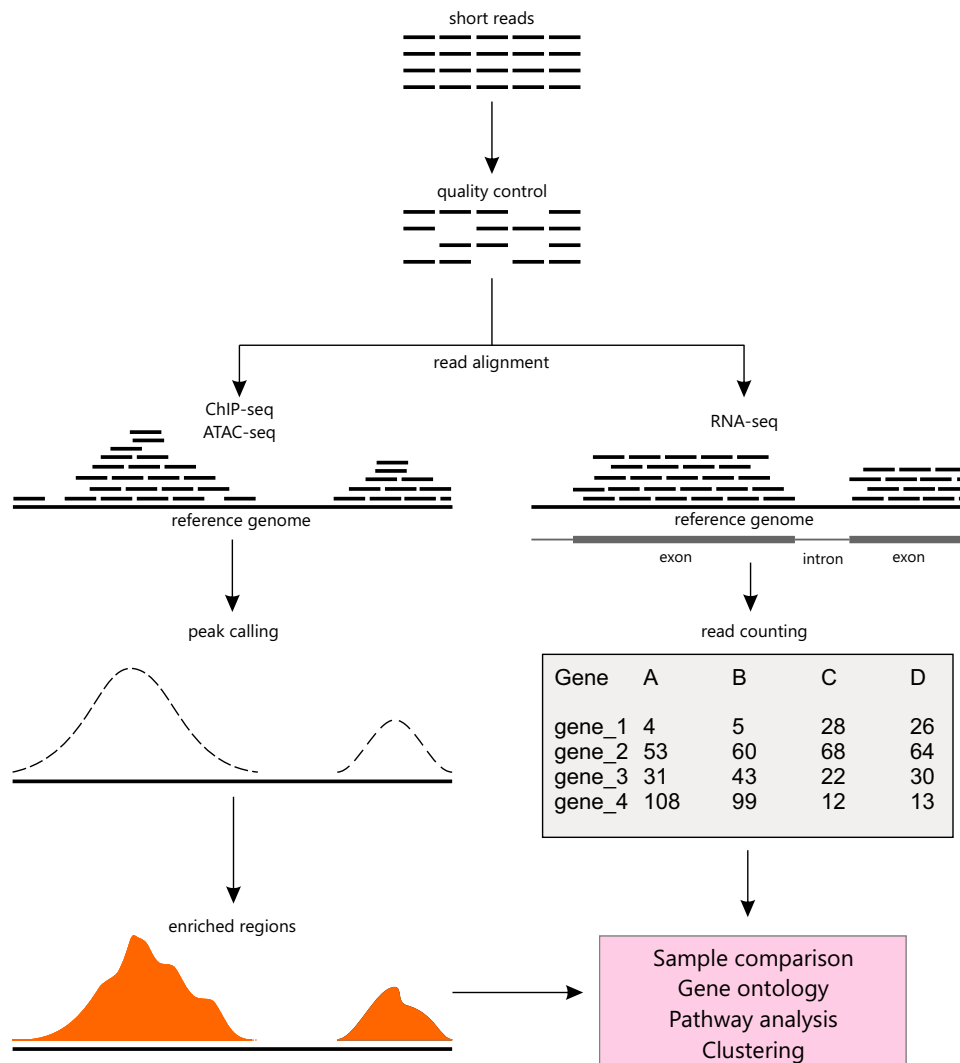


Figure 4.2: Generalised workflow for HTS data analysis. Raw sequencing data contains millions of sequences of short length, termed short reads. The reads with good quality scores are aligned to a reference genome. For RNA-seq, reads align to exonic regions and are counted to determine expression of the genes they are part of. For ChIP-seq and ATAC-seq regions of high density are inferred to be binding sites or accessible chromatin, respectively, and are identified using peak calling methods. These enriched regions can be visualised in coverage tracks. Downstream analysis of the data includes comparison of samples, pathway analysis, motif enrichment analysis, etc.

agnostic analysis of HTS data. The set consists of three packages that run within the R environment. The *rseqR* package performs gene expression quantification using the read estimation tool Salmon, followed by differentially expressed gene (DEG) analysis using the R/Bioconductor¹⁶² packages edgeR¹⁵⁸, DESeq2¹⁵⁹, and limma¹⁶⁰, with the added option to integrate the analyses from these packages to provide more robust results. The *chompR* package provides a wrapper around the MACS2 software to perform peak-calling for both ChIP-seq and ATAC-seq data. Both *rseqR* and *chompR* provide functions for motif enrichment and pathway analysis. Finally the *snpr* package provides tools to identify trait-associated SNP enrichment in genomic loci, such as differential peaks identified by the *chompR* package.

4.2 RNA-seq analysis: the *rseqR* package

The *rseqR* package is an automated workflow that integrates all the steps for analysis of RNA-seq data. Detailed instructions on its use may be found at this link: <https://anilchalissey.github.io/rseqR/>. The package is able, from raw RNA-seq reads, to determine the main differences in gene expression and biological processes between two (or more) biological conditions. This pipeline provides the basis for the RNA-seq analysis described in subsequent chapters.

4.2.1 Methods

A diagram illustrating the successive steps of the *rseqR* workflow is shown in Figure 4.3. The package incorporates both publicly available tools and custom scripts developed by me. The entire workflow from start to finish may be run by a single call to the function `run_dea()` whose main argument is a configuration file, in the form of a tab-delimited

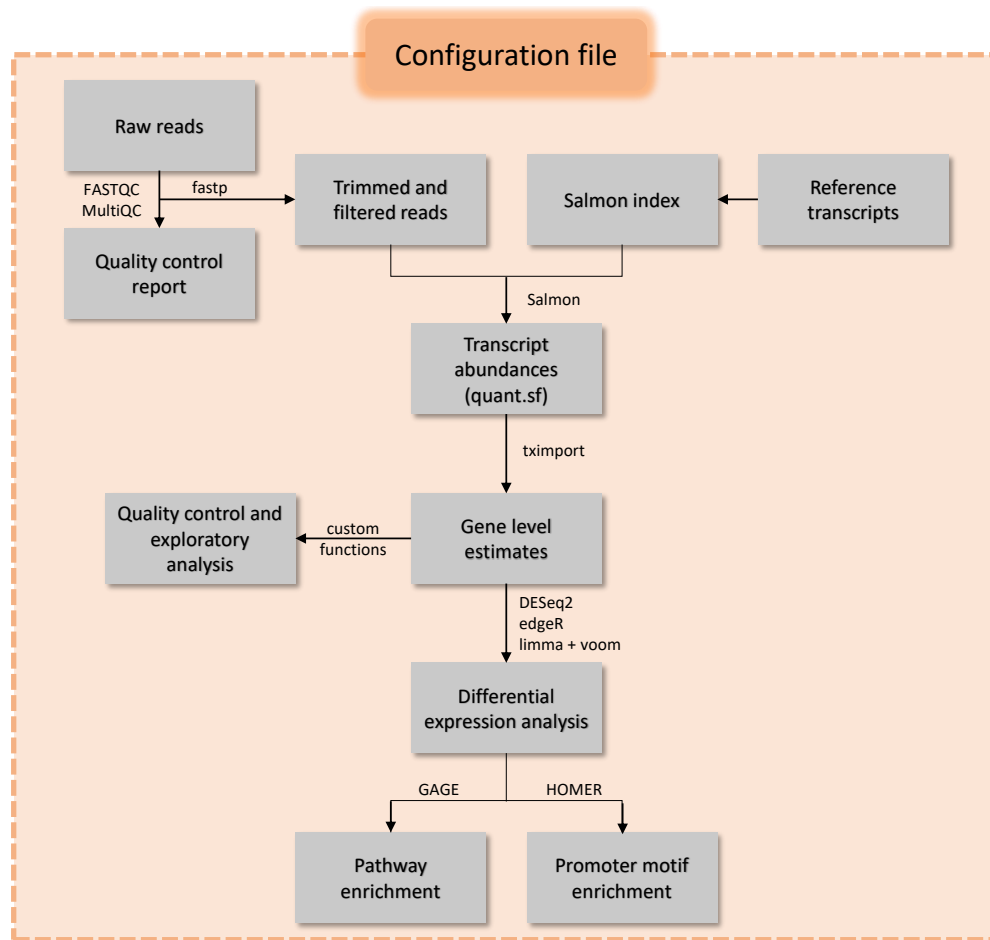


Figure 4.3: Workflow of the rseqR package. Successive steps of the analysis with the tools used to perform each step.

table, containing information on the location of data files, sample names, experimental conditions, and batch effects. This configuration file is used as a means to track all analysis performed by the package and the metadata associated to the project. Most importantly, it allows future reproducibility of the results. Several other parameters are also available, and these are detailed either on the website or manual for the package, and may be adjusted by more advanced users. However, for the majority, the default parameters may be left unchanged. The source code of the package and instructions to install it are available on GitHub (<https://github.com/anilchalisey/rseqR>).

Data analysis begins with the analysis of the raw read files, which may be either single-end or paired-end reads. There are multiple potential sources of noise in this data, some of which can be removed before downstream analysis. Aside from the reduction of noise in the data, this also reduces the size of the read file, thereby saving computational resources. Read cleaning approaches are of three main types: quality-based clipping or filtering¹⁶³, sequencing-error correction¹⁶⁴, and contaminant-removal¹⁶⁵. In the first approach, bases with low Phred quality scores are clipped from the ends of the reads. If the overall quality of a read is low, the whole read may also be discarded. Error-correction methods aim to correct bases likely to be sequencing errors based on information gathered from other reads. Contaminant removal may be necessary if there is a high proportion of adaptor sequences, or contamination from other sources (e.g. ribosomal RNA in mRNA-sequencing). All three of these approaches may be used in the *rseqR* pipeline, but the default settings are to omit this step so as to reduce the risk of losing valuable information. If it is decided to perform read cleaning, this may be done using `fastp`¹⁶⁶, for which a wrapper function `run_fastp()` exists.

The actual read quality assessment is done using `FASTQC`¹⁴⁵, a JAVA-based program designed to perform quality control of HTS data. The results of `FASTQC` are provided as individual reports for each sample; these are integrated into a single HTML report, allowing comparison of samples, using `MultiQC`¹⁶⁷.

Had this package been developed even a year ago, the next step to discuss would be read alignment. However, it is now increasingly recognised that alignment-independent quantification methods are often more accurate (and considerably faster) than alignment-dependent methods¹⁶⁸. These methods rely on a known transcriptome, and use a ‘pseudo-alignment’ or ‘light-weight mapping’ approach that is capable of rapidly de-

termining the set of transcripts compatible with a given sequenced fragment, and then determining which of those transcripts is the best fit. In the *rseqR* we use the filtered reads as input to the transcript quantification tool Salmon¹⁵⁶, which uses a quasi-mapping procedure coupled with a two-phase inference process, to rapidly estimate transcript abundances. With the default settings of the pipeline, bootstrap sampling is turned off, sequence bias and GC bias are both corrected for, but not positional bias, and orphan reads are not allowed. The latter increases accuracy but has the potential to decrease sensitivity, which we feel is an acceptable pay-off. Prior to this, an index of the reference transcriptome must be provided; if this is not specified in the parameters, then the *rseqR* package will automatically create one based on the species under evaluation. The transcript-level results are aggregated into gene-level abundances using the Bioconductor package tximport¹⁶⁸, which takes into account potential changes in gene length across samples arising from differential isoform usage and creates an offset matrix to be used by downstream steps for normalisation of the counts.

Prior to further analysis of these derived abundances, they need be modelled onto a statistical framework. Both edgeR and DESeq2, which are used in the *rseqR* pipeline model the counts on a negative binomial (NB) distribution. The limma+voom approach, also used in this package, first transforms the data using the variance-mean relationship of the gene counts, and then uses a linear-modelling strategy similar to that used for microarray analyses.

Once the statistical model has been defined, exploratory analysis of the data is performed. This includes measurement of dispersion and biological variability, and the assessment of sample similarity (using principal component analysis). This is done with a custom script which utilises functions within base R. The next step is to

perform a test for differential expression between different conditions. As part of this step, the various algorithms also perform count normalisation and outlier correction. The program is sophisticated enough to perform paired analyses that take account of batch effect provided this information is included in the metadata. Depending on the statistical model used, the test methods include the exact test, the likelihood ratio test, the Wald test, or the T-test. Since thousands of tests are performed (one for each gene considered), multiple-testing correction must be performed to avoid uncontrolled inflation of type I errors, typically by applying the Benjamini-Hochberg false discovery rate (FDR) procedure.

For all three differential gene expression methods, the output is a table containing information about the mean expression of each gene across all samples, the fold-change in expression between the conditions, the p -value for statistical significance and the p -value adjusted for multiple-testing using the Benjamini-Hochberg procedure. To select the most meaningful results, the pipeline considers an adjusted p -value threshold of 0.05 as significant for a change in gene expression.

In order to infer meaningful functional changes in biological processes between the conditions being investigated, the gene expression changes are associated with gene ontology (GO) terms using the bioconductor package `gage`¹⁶⁹, which utilises a gene set analysis (GSA) method called Generally Applicable Gene-Set Enrichment (GAGE). This is a functional class scoring method, whereby co-ordinated changes in gene expression over gene sets are identified. This approach derives a score for all genes belonging to a given gene set, based on change in expression, regardless on whether or not it is a significant change in expression. The idea is that the overall effect of co-ordinated changes in expression of genes within a gene set can result in a significant functional or mechanistic effect, regardless of whether or not the individual genes

are significantly differentially expressed between conditions. Here we use gene sets derived from GO terms and Kyoto Encyclopaedia of genes and genomes (KEGG) pathways. The GO database provides a dictionary of terms for describing gene product characteristics and annotation data. In particular it describes gene products in terms of biological processes (BP), cellular components (CC), and molecular functions (MF). The KEGG pathway database is a collection of manually drawn pathway maps derived from knowledge of the molecular interactions, reactions, and relation networks for a number of metabolic, genetic, sensory, and organismal processes. It also includes pathways for human diseases and drug development.

Promoter motif enrichment analysis is performed using HOMER¹⁷⁰. It uses a probabilistic approach, relying on representation of motifs by position weight matrices, to identify statistically over-represented motifs in the promoter regions of co-expressed genes.

The running time for the pipeline depends on the starting point of the analysis and the number and size of samples. If starting from raw reads, then the main bottleneck is the pseudo-alignment phase to generate transcript abundances. The package has been written to leverage parallel processing where possible to speed this step up. Once transcript abundances have been estimated, the rest of the analysis takes less than 10 minutes, unless motif enrichment analysis is performed, which can take between 30-60 minutes on a desktop computer with 8Gb of RAM. It is possible to start the analysis with pre-computed transcript abundances. In this case, as detailed in the manual, the configuration file should point towards the quant.sf files generated by Salmon rather than the raw reads.

4.3 ChiP- and ATAC-seq analysis: the *chompR* package

As the *rseqR* package is to RNA-seq, the *chompR* package provides an automated workflow interface that integrates all the steps for analysis of HTS data derived from ChIP-seq and ATAC-seq experiments. Detailed instructions on its use may be found at this link: <https://anilchalisey.github.io/chompR/>. Starting from raw reads, the *chompR* package identifies regions of differential chromatin modification and determines the key biological processes affected by these changes. This pipeline provides the basis for the ChIP-seq and ATAC-seq analyses described in subsequent chapters.

4.3.1 Methods

A diagram illustrating the successive steps of the *chompR* workflow is shown in Figure 4.4. As with the *rseqR* package, both publicly available tools and custom scripts developed by me are incorporated within the package. The entire workflow from start to finish may be run by a single call to the functions `run_chip()` or `run_atac`, and as for the `run_dea()` function in *rseqR*, the main argument for these functions is a configuration file containing information on the location of data files, sample names, experimental conditions, and batch effects. In the case of ChIP-seq experiments, this file should also include details of input controls. The source code of the package and instructions to install it are available on GitHub (<https://github.com/anilchalisey/chompR>).

As with the *rseqR* package, read quality assessment and processing is performed using FASTQC, MULTIQC and fastp. The same principles discussed earlier apply here.

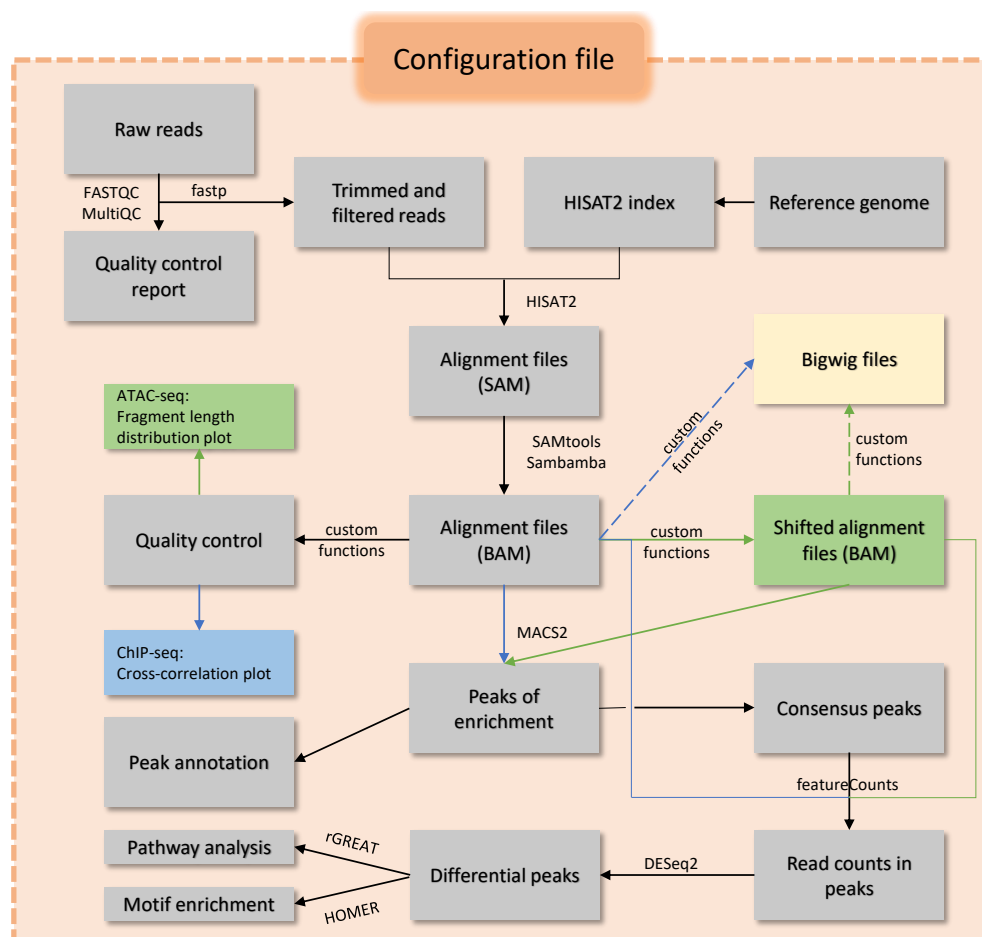


Figure 4.4: Workflow of the chompR package. Successive steps of the analysis with the tools used to perform each step. The boxes and arrows in green show output or processes only performed for ATAC-seq analysis, and those in blue for ChIP-seq analysis. Dashed arrows indicate optional functionality.

Following read processing, the next step is read mapping, the purpose of which is to identify, for each read, its originating position in the genome. In this pipeline, we use the HISAT2 aligner¹⁵⁰, which is a fast and memory-efficient alignment tool. Prior to mapping, the reference genome must be indexed; if this is not available, then the pipeline will automatically create the relevant index based on the species specified. The output of the alignment process is a sequence alignment map (SAM) file which stores the information about the alignment of the reads against the reference genome. SAM files are large human-readable files, and so the next steps in the pipeline convert it to the more memory efficient binary alignment map (BAM) format. At the same time, duplicate reads are marked, and mitochondrial reads (which can form up to 40% of reads in ATAC-seq experiments) are removed. Following these post-mapping processing steps, quality control is performed. For both ChIP-seq and ATAC-seq reads, library complexity is measured using the Non-Redundant Fraction (NRF), PCR bottlenecking coefficient 1 (PBC1), and PCR bottlenecking coefficient 2 (PBC2). The NRF is defined as the fraction of reads that are uniquely mapping. PBC1 and PBC2 provide a measure of how skewed the distribution of read counts per location is towards one read per location. For ChIP-seq, the preferred values are $\text{NRF} > 0.9$, $\text{PBC1} > 0.9$, and $\text{PBC2} > 10$. The respective values for ATAC-seq are 0.9, 0.9, and 3. For ChIP-seq experiments, the pipeline also performs cross-correlation analysis, which is an important measure of ChIP success and enrichment (Figure 4.5A). Coverage signal tracks for forward and reverse strands are computed separately, and these tracks are shifted towards and away from each other by incremental distances. For each shift, the Pearson correlation coefficient is determined, resulting in a cross-correlation profile showing the correlation between forward and reverse strand coverage at different amounts of shift. The highest cross-correlation should be obtained at a shift equal to

the predominant fragment length as a result of clustering of the fragments around the binding site of the ChIP target. Alongside this peak at the fragment length, a ‘phantom peak’ is also observed at the read length, which is an effect of variability in mapping at different positions in the genome. For a successful ChIP-seq experiment, the fragment-length cross-correlation peak should be larger than the phantom-peak. A more objective measure of enrichment can be determined by calculation of the normalised strand cross-correlation coefficient (NSC), which is the ratio of the maximum cross-correlation and the minimum cross-correlation; higher values indicate more enrichment, and 1 indicates no enrichment. Another measure is the relative strand cross-correlation coefficient (RSC) which is defined as $(CC_{max} - CC_{pp}) / (CC_{pp} - CC_{min})$, where CC_{max} is the maximum cross-correlation, CC_{min} the minimum cross-correlation and CC_{pp} the phantom-peak cross-correlation. Significantly enriched experiments have values > 1 . Both these measures of quality are calculated by the *chompR* package and a cross-correlation plot is also produced.

For ATAC-seq experiments, a fragment length distribution plot is produced as a measure of a successful experiment. Unlike in ChIP-seq, where fragments are of a relatively fixed length, in ATAC-seq fragments span a range of lengths. Smaller fragments correspond to truly open regions, but longer fragments represent regions of accessible open chromatin, which are separated by nucleosome-bound regions. A plot of the distribution of fragment lengths therefore reveals a characteristic nucleosomal phasing, showing free, and mono-, di-, and tri-nucleosome protected regions (Figure 4.5B). If this is absent, then the ATAC-seq experiment has been unsuccessful.

Once reads have been aligned, sorted, filtered and quality control has been performed, the next step in the pipeline is to call peaks of enrichment. However, for ATAC-seq, prior to doing this, read-shifting is performed. The Tn5 transposon used to perform

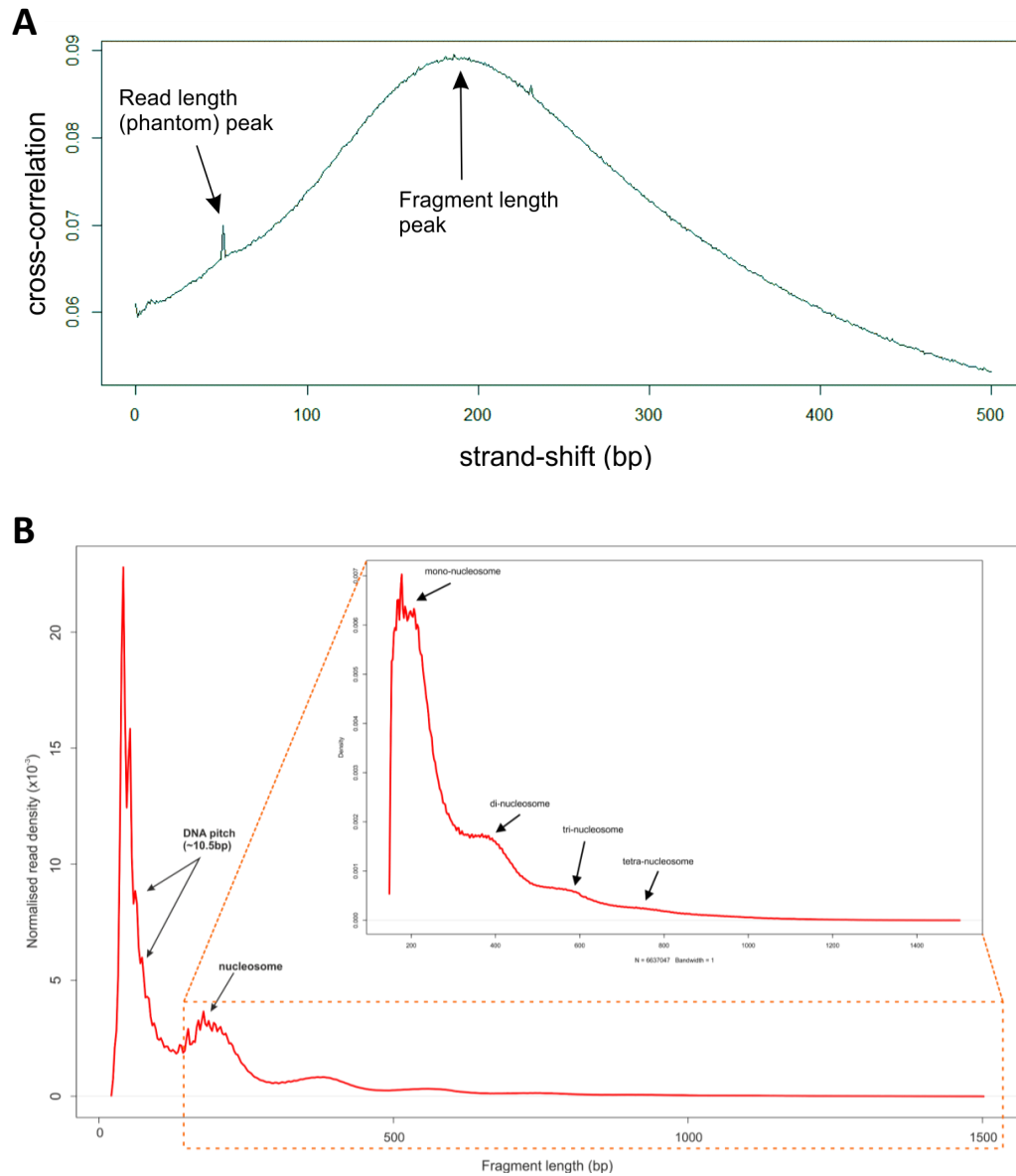


Figure 4.5: Quality control plots of ChIP-seq and ATAC-seq experiments. (A) Cross-correlation plot for ChIP-seq. Generated by shifting coverage plots for forward and reverse strands and measuring the Pearson's correlation at each shift. The highest cross-correlation should be at the fragment length, and an additional (but smaller) phantom peak is often seen at the read length. (B) Fragment length distribution of a successful ATAC-seq experiment. Nucleosome-free fragments are <150 bp in length. Longer fragments represent mono-, di-, and tri-nucleosome fragments.

ATAC-seq actually binds as a dimer and inserts two adaptors separated by 9 bp. The cutting site is therefore somewhere between these sites and so all reads aligning to the positive strand are offset by +4 bp, and all reads aligning to the negative strand are offset by -5 bp. A custom function is included within the *chompR* pipeline to perform this shift automatically for ATAC-seq reads, and all downstream analyses are based on these shifted BAM files.

Peak-calling in the *chompR* package is performed using MACS2¹⁵⁷, which is also the main tool used by the ENCODE consortium. MACS2 uses a sliding window-based computational algorithm to identify peaks of enrichment. In its default setting, which is designed for ChIP-seq, MACS2 first performs read shifting and extension so that peaks are centred in the middle of fragments.

The *chompR* pipeline keeps this setting for ChIP-seq experiments but turns it off for ATAC-seq where peaks should be centred on the free ends, which are the regions of accessible chromatin. For both ChIP-seq and ATAC-seq, if paired end sequencing has been performed, the default setting is to use actual fragment lengths rather than an estimate, which MACS2 would otherwise do. MACS2 next identifies regions where reads are enriched relative to either local background (default) or the global background. For sharp binding events, such as transcription factor ChIP-seq, the default local background is appropriate, but for broad peaks as might be seen in some histone modifications, global background is often used.

There is a lack of consensus about which is appropriate for ATAC-seq, but the Buenrostro group who first developed ATAC-seq use the narrow peaks setting for MACS2 but with global background normalisation. We adopt that strategy in this pipeline. Finally, if an input control sample is present, as in ChIP-seq, this can be specified in the metadata, and this information will be fed to MACS2 to refine the

peak calling. This is used to estimate an empirical FDR.

Annotation of peaks is performed using a combination of custom functions and the *ChIPSeeker* package. For each peak, the position and strand of the nearest genes, and the distance to the nearest TSS is reported. Peaks are also classified on the basis of their genomic location - either promoter, 5'UTR, 3'UTR, exon, intron, or intergenic. At this stage a further quality control step is performed. The *featureCounts* tool¹⁵⁴, which is a highly efficient read summarisation program, is used to count the number of reads in each sample falling in peaks. This is used to calculate the fraction of reads in peaks (FRiP); current ENCODE standards for ChIP-seq are a FRiP $> 1\%$ and for ATAC-seq FRiP $> 30\%$.

Following annotation, for each experimental condition, a set of consensus peaks is derived. In the *chompR* pipeline we define a consensus peak as one that is present in at least half of the samples for each condition, or, if only two samples are present, then that peak must be present in both. The consensus peaks from the different experimental conditions are then merged to generate a master set of peaks. The *featureCounts* tool is used to count the number of reads aligning to the peak regions across all samples. The output of this is a table in which the entry in the i th row and j th column is the number of reads that have been mapped to peak i for sample j . This count table is used as input for DESeq2, using similar methodologies to that described for RNA-seq analysis earlier. The final of this analysis is a table containing information about the mean enrichment in each peak across all samples, the fold-change in enrichment between the conditions, the p -value for statistical significance and the p -value adjusted for multiple-testing using the Benjamini-Hochberg procedure. To select the most meaningful results, the pipeline considers an adjusted p -value threshold of 0.05 as significant for a change in enrichment.

To analyse the functional significance of changes in the genomic regions identified by the above analysis, we use the Genomic Regions Enrichment of Annotation Tool (GREAT) tool¹⁷¹. This is a web-based tool that provides functional annotation of *cis*-regulatory regions identified by ChIP and ATAC-seq experiments. It takes in to account regions proximal to genes as well as more distal sites, and incorporates annotation from 20 different ontologies. In order to access this tool in an automated fashion we use a R-based application programming interface¹⁷² to the GREAT website that constructs a query according to user input and then retrieves the results from the GREAT server.

Motif enrichment analysis is again performed using HOMER¹⁷⁰.

As with *rseqR*, the running time for the *chompR* pipeline depends on the starting point of the analysis and the number and size of samples. If starting from raw reads, then the main bottlenecks are the alignment phase and the optional phase of BigWig track creation. The package has been written to leverage parallel processing where possible to speed these steps up. If for any reason the pipeline needs to be re-run, provided the directory structure is maintained, *chompR* is able to recognise which steps have been completed and skip over these.

4.4 Trait-associated SNP enrichment: the *snpR* package

The advent of the genome wide association study (GWAS) era has led to the identification of many sequence variants associated with diseases and traits. Tens of thousands of these trait-associated SNPs (taSNPs) have been catalogued, but there are few tools available to easily mine this resource. To this end, I have developed the *snpR* package

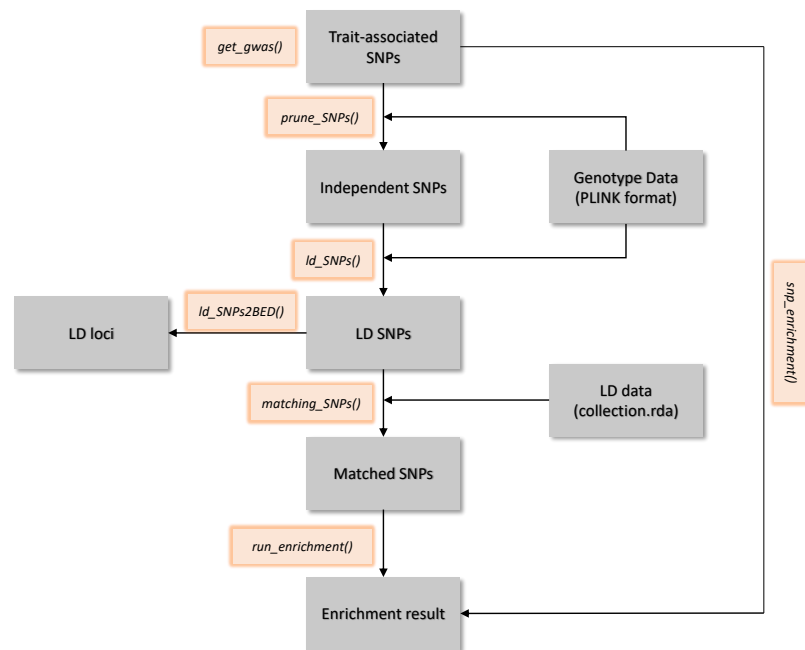


Figure 4.6: Workflow of the *snpR* package. Successive steps of the analysis with the functions used to perform each step.

to identify whether a set of genomic regions, such as differential regions identified from a ChIP-seq or ATAC-seq experiment, is likely to be functionally associated with a set of taSNPs.

4.4.1 Methods

A diagram illustrating the successive steps of the *snpR* workflow is shown in Figure 4.6. The package incorporates custom scripts only and depends on PLINK¹⁷³, an open source C/C++ tool designed for the analysis of GWAS. The entire workflow from start to finish may be run by a single call to the functions `snp_enrichment()` or individual functions may be called as detailed in Figure 4.6. The source code of the package and instructions to install it are available on GitHub (<https://github.com/anilchalisey/snpR>).

The first step in the analysis is to generate a list of taSNPs. This is performed by the `get_GWAS()` function, which downloads the latest version of the GWAS catalog¹⁷⁴ and filters it according to criteria specified by the user (e.g., `get_GWAS(filter = "coronary|cardiovascular")`). This list is used as input to the `prune_SNPs()` function, which generates a list of independent SNPs by filtering the taSNPs to keep only those with a linkage disequilibrium (LD) $r^2 < 0.2$ and more than 500kb apart. Having generated the list of independent SNPs, we recognise that this variant may not be a causal variant. We therefore, select all other SNPs from the 1000 Genomes Project Phase 3 European data (i.e. GBR, FIN, IBS, TSI, and CEU) which are in strong LD ($r^2 > 0.8$) with these index taSNPs. The reference data for this is freely available at this link: http://www.broadinstitute.org/mpg/depict/depict_download/1kg/1000_genomes_project_phase3_CEU.tar.gz. Once downloaded and uncompressed, the path to this directory should be specified in the parameters to the function `ld_SNPs()`, which performs this analysis. A BED file of the locus for each index taSNPs is generated by the `ld_SNPs2BED()` function; the locus is defined as the genomic region spanning the distance between the most-5' and most-3' variants in LD to the index SNP. The program next calculates the total number of overlaps of the index taSNPs or their LD proxies with genomic regions of interest; for example, differential ATAC-seq or ChIP-seq regions detected by the *chompR* package. To determine the significance of this observed overlap, a set of matched variants is used to estimate the probability of the observed overlap compared to expectation. To derive these matched variants, for each index taSNPs, we identify a set of 1000 control SNPs randomly selected from across the entire genome that match the index SNP for mean allelic frequency, number of genes in the SNP locus, and number of LD 'buddies' (i.e. number of other SNPs in LD). To generate the matching random SNPs, I provide precomputed LD r^2 boundaries for **every SNP** in the 1000 Genomes Project

Phase 3 database, binned by mean allelic frequency, and saved in Rdata format. This data was obtained by parsing the SNPsnap database¹⁷⁵. The data is downloadable from <https://www.dropbox.com/s/s9roomzkq98zvhb/collection.rda?dl=0>. With this data to hand, the enrichment test can be performed with a call to the function `run_enrichment()`, which runs a Monte Carlo bootstrapping simulation to determine a p-value to estimate the significance of the overlap. In essence, SNPs from the provided database are sampled with replacement to generate a random set the same size as the original set, and the degree of overlap of the random set with the loci is compared with the original set. If this is repeated many times (default being 1000), then a p-value can be calculated based on the observed overlap of the original set compared with the expected overlap of the multiple random sets.

The entire pipeline has been optimised to be able to run on a home computer very speedily.

The entire analysis can be performed in under half an hour depending on the number of permutations tested - there is a linear relationship between the run time and the number of permutations.

4.5 Discussion

HTS is now a relatively commonplace technique being used in many different settings. The field of bioinformatics is moving rapidly to cope with the increased data and applications HTS is being used for and new algorithms are published very frequently. As a result, although some aspects of analysis have become standard, the number of tools to choose from can be overwhelming. Numerous pipelines have been developed to address this, but there is still a strong need for new pipelines that are easy

to use by people of all levels of computational experience. The pipelines presented here go some way in addressing this. All three pipelines can be run using a single command and require minimal user input. The metadata files that constitute the main argument to *rseqR* and *chompR* can be made in any text or table editor. As will be shown in subsequent chapters, all three pipelines output informative figures or tables that clearly illustrate the results of the analyses. The R backbone provides the major advantage of allowing most dependencies to be installed alongside the packages, although does not provide a mechanism for automated installation of the command line tools. However, these are straight-forward to install, and detailed instructions are included in the online documentation.

4.5.1 Comparisons with other tools

The *rseqR* and *chompR* packages are not the only non-commercial analysis tools available for analysis of HTS data; several other pipelines have been developed. However, many of these pipelines are web or cloud-based with the inherent limitations of upload speed and cost and file size limitation^{176–179}, or do not allow execution of a complete pipeline from read mapping to advanced processing tasks^{180,181}. In addition, almost all tools that allow execution of a complete pipeline require an underlying Unix-based system to work. With the advent of the windows subsystem for linux in Windows 10, it is now possible to run a bioinformatics pipeline in a platform-agnostic manner, as we have done here.

In terms of RNA-seq analysis, other recently published or widely used pipelines include VIPER¹⁸², HppRNA¹⁸³, TRAPLINE¹⁸⁴, QuickRNASeq¹⁸⁵, systemPipeR¹⁸⁶, and SARTools¹⁸⁰. Whilst many of the features and software packages used in these pipelines are shared by *rseqR*, there is also considerable variation (Table 4.1).

Table 4.1: Comparison of features of *rseqR* with other RNA-seq pipelines.

Features	<i>rseqR</i>	VIPER	HppRNA	TRAPLINE	QuickRNASeq	<i>systemPipeR</i>	SARTools
Raw Read Quality Control	X	X	X	X	X	X	
Read Alignment	Pseudo	X	X	X	X	X	
Differential Expression	X	X	X	X		X	X
Pathway Analysis	X	X			X	X	
Consolidated Report		X		X	X	X	X
Dependencies Packaged	Partly	X	X	X		Partly	X
Batch Correction	X	X				X	X
Framework	R	Snakemake	Snakemake	Galaxy	Bash, Perl, Java	R	R
Windows-Compatible	X					Partly	X

The VIPER RNA-seq pipeline is feature rich and in addition to having all the features seen in *rseqR*, also produces a consolidated report, which I aim to provide in future versions of this package. However, the VIPER has been developed to be used only on cluster or supercomputer environments and requires a recommended 128GB of RAM to run. This is because it performs the analysis using a traditional count-based method, which depends on read alignment with the very fast but very memory-consuming STAR aligner. In fact, *rseqR* is the only pipeline which incorporates the new alignment-independent method of gene abundance estimation. One advantage that VIPER and the less feature-rich HppRNA have over *rseqR* is the fact that they run on Snakemake framework and so all their dependencies come packaged with the program. In a similar vein, the TRAPLINE pipeline manages its dependencies through the Galaxy server¹⁷⁹, but it requires a Galaxy installation and comes with the limitations of transferring files to an online server. In contrast, as *rseqR* is R-based, only the R package dependencies can be automatically installed; command-line tools must be installed separately. To make this easy, I have provided detailed instructions on the online documentation. Of the two other R based pipelines in Table 4.1, *systemPipeR* is also as feature-rich as *rseqR* and suffers the same limitations in terms of package dependencies. However, the tool is designed not just for RNA-seq analysis, but also for other HTS analysis, and it offers an impressive number of algorithms. However, this flexibility can be

confusing as its configuration and set up is much more complicated than *rseqR*, and this could be a barrier to those new to computational biology and bioinformatics. In addition, it has not been written to use the Windows Subsystem for Linux, so many of its features are unavailable in the Windows environment. The *SARTools* package is a pipeline developed solely for differential analysis and only works on count-based data - it performs none of the earlier steps necessary to get that data, and also does not perform downstream analysis.

With regard to sequencing data derived from chromatin modification experiments, there are many excellent tools available for ChIP-seq analysis, but a relative paucity for ATAC-seq analysis. Most tools used for ATAC-seq analysis are repackaged versions of ChIP-seq tools and a careful review of methods sections in published literature demonstrates that often settings and parameters have not been adjusted to optimise the analysis for ATAC-seq. With this in mind, there is a definite need for an automated pipeline, such as *chompR*, which addresses this issue. Commonly used pipelines for ChIP-seq (and ATAC-seq) analysis are shown in Table 4.2^{170,186–191}. It is immediately apparent that very few of these pipelines cover the whole of the analysis, and instead have been developed to focus on specific aspects of the analysis. They therefore provide a specific and less comprehensive evaluation based on a smaller number of metrics. The *chompR* pipeline is one of only two with clearly optimised peak calling settings for ATAC-seq, and for read shifting of aligned files. It is also the most feature-rich pipeline and performs all the major steps in both ChIP-seq analysis and ATAC-seq analysis. In addition, because it has been written with the Windows Subsystem for Linux in mind, all of its functions can operate in both a Unix and Windows environment. As with the *rseqR* package, it lacks the option for a consolidated report, which will be included in future versions, and it also has the same problems with command-line tool

dependencies.

Table 4.2: Comparison of features of chompR with other ChIP-seq and ATAC-seq pipelines.

Features	chompR	ChiLin	CisGenome	HOMER	CHIPQC	ATACseqQC	systemPipeR	ChIPseeker
Raw Read Quality Control	X	X					X	
Read Alignment	X	X					X	
Library complexity	X	X				X		
FRiP	X	X			X			
Fragment Length Distribution	X					X		
Read shifting	X					X		
Generate Tracks to View	X	X					X	
Peak Calling	X	X	X				X	
ATAC-seq Optimised Peak Calling	X			X				
Peak Annotation	X	X	X					X
Differential analysis	X			X				
Pathway Analysis	X	X	X	X				X
Motif Enrichment	X	X	X	X				
Consolidated Report		X			X		X	X
Dependencies Packaged	Partly	X	X		X		Partly	X
Framework	R	Python	C/C++	PERL, C++	R	R	R	R
Windows-Compatible	X		X		X	X	X	X

The final package discussed was the *snpR* package, which has been designed to detect enrichment of taSNPs in a set of genomic regions. To my knowledge, only two published tools exist to evaluate enrichment of trait-associated variants in genomic regions. The first of these is *GREGOR*¹⁹², which is a PERL-based program that uses a very similar methodology to that used by *snpR*. However, I have been unable to compare its output to *snpR* due to installation problems on a number of platforms. The other tool is *traseR*¹⁹³, which is an R-based package that takes all taSNPs and then performs a user-specified enrichment test (one of a contingency table based test, binomial test or non-parametric testing procedure) of all these variants combined, as well as for each set of taSNPs associated with the phenotypes listed in the GWAS catalog. However, the *traseR* package does not by default look at variants in LD with these taSNPs and so does not consider that these may not be causal. In addition, the background control for the enrichment tests does not include matched controls, instead non-trait associated variants are used.

4.6 Conclusions

In this chapter I have presented three new tools for (A) analysis of RNA-seq data, (B) analysis of ChIP and ATAC-seq data, and (C) evaluating enrichment of trait-associated variants in genomic features. Both *rseqR* and *chompR* are comprehensive pipelines for analysis of HTS analysis conforming to the latest standards and are fast, efficient, and are able to be used in a manner easy enough for those without a significant computational background. Both pipelines are as, if not more, feature-rich than existing tools, and can operate in a platform-agnostic manner. To my knowledge these are the first tools specifically designed to be run both directly on a UNIX-based operating system and on a Windows 10 based operating system with leverage of the Windows Subsystem for Linux. The *snpR* tool is an easy to use R package that performs enrichment analysis of trait-associated SNPs, allowing identification of regulatory elements and potential functional variants that affect the regulation of transcription. The method used in the *snpR* package takes into account all potential causal variants by taking into account those in LD with the index trait-associated variant and determines enrichment using matched control variants. Detailed instructions on the use of these programs and their source code are freely available on my GitHub repository at <https://github.com/anilchalisey>.

In the next chapters we use these tools to evaluate the effects of oxLDL on transcriptional regulation in vascular and immune cells.

Chapter 5

Transcriptomic Changes in Vascular and Immune Cells in Response to oxLDL

Contents

5.1	Introduction	104
5.2	Total RNA isolation and quality	106
5.3	Quality control of raw reads	107
5.4	Abundance estimation and quality control of count data	114
5.5	Differential gene expression and pathway analysis	117
5.5.1	Oxidised LDL alters expression of lipid-related and hypoxia-modulated genes in CD4 ⁺ cells	120
5.5.2	Oxidised LDL significantly alters expression of genes involved in angiogenesis and lipid metabolism in HAECs	121
5.5.3	In macrophages oxidised LDL causes significant changes in expression of genes involved in lipid and carbohydrate metabolism, and immune function and inflammation	121
5.5.4	Oxidised LDL alters expression of lipid-handling and oxidative stress related genes in VSMCs	122
5.5.5	Genes differentially expressed across all cell types play roles in lipid metabolism, angiogenesis and oxidative stress regulation	123
5.5.6	VCAM1 is downregulated in response to oxLDL	124
5.6	Genes differentially regulated by oxLDL are associated with cardiovascular risk variants	126
5.7	Discussion	126

5.1 Introduction

Studies in the last two decades have clearly identified oxLDL as a primary factor leading to atherosclerosis. Internalised oxLDL triggers complex signalling cascades leading to production of ROS, cellular dysfunction, recruitment and trans-endothelial migration of monocytes with subsequent transformation into foam cells, recruitment of CD4⁺ T cells, and proliferation and migration of VSMCs^{5,7,10,13,17}. The transcriptional changes underlying these effects remain poorly understood, as oxLDL has been shown

simultaneously to increase the expression of both cell cycle activating and inhibiting proteins, to both augment and inhibit cell signalling pathways, and to have potential to cause both cell proliferation and death^{51,54}.

Assays that allow interrogation of the entire cell transcriptome have been available for some time now, but there is a remarkable paucity in the literature of their use in the investigation of the molecular mechanisms underlying atherosclerosis. Microarray studies of gene expression in peripheral blood cells of patients with clinically manifest atherosclerotic disease or disease states that predispose to atherosclerosis have been performed to try and identify biomarkers of the disease with limited success^{194,195}. There have been no RNA-seq based transcriptomic studies of primary human peripheral blood cells from atherosclerotic individuals, although there have been attempts at RNA-seq based analyses of atherosclerotic plaque tissue from patients with stroke^{196,197}. However, these studies were limited by the relative scarcity of available tissue, and the heterogeneous nature and cellular composition of the plaques, which resulted in insufficient power to detect differentially expressed genes. A thorough and detailed review of the literature identified only fourteen studies in which vascular or immune cells had specifically been exposed to oxLDL for the purpose of gene expression analysis. All of these studies used cDNA microarray probe based transcriptional assays (Table 5.1)^{63,198–209} and whilst they have provided useful insights into some of the molecular mechanisms underlying the effects of oxLDL, they have all employed different methods of oxLDL treatment and statistical analysis and have resulted in identification of different numbers of DEGs, ranging from 149 to 1500.

In an attempt to overcome the inconsistencies and limitations inherent to these studies, we have for the first time performed RNA-seq on multiple primary human cell types known to be involved in the atherogenic process before and after exposure to a set

Table 5.1: Microarray studies of oxLDL-treated vascular and immune cells. Where the oxLDL concentration or duration of exposure is not defined, this was either not specified or multiple conditions were used. HCAECs = human coronary artery smooth muscle cells, HUAECs = human umbilical artery endothelial cells, HAECs = human aortic endothelial cells, HCASMCs = human coronary artery smooth muscle cells, HUVECs = human umbilical vein endothelial cells, HVSMCs = human vascular smooth muscle cells, HECs = human endothelial cells, SMCs = smooth muscle cells.

Author	Cell Type	oxLDL conc (ug/ml)	duration (hr)
Sukhanov 2003	HASMCs	-	-
Virgili 2004	HECs	-	-
Mattaliano 2009	HAECs	-	-
Chen 2009	Primary human peripheral blood monocytes	30	6-12
Hammad 2009	U937 cells	150	4
Minta 2010	HCASMCs	2	3-21
Khaidakov 2012	HCAECs	5	2-12
Hu 2014	THP-1 cells	50	48
Laguna-Fernandez 2015	HUAECs	100	-
Reschen 2015	Primary human peripheral blood monocytes	50	48
Damian-Zamacona 2016	HVSMCs	7.5	1-24

concentration of oxLDL prepared in a reproducible manner. The same treatment has been used in studies of chromatin accessibility, which we will discuss in subsequent chapters, thus allowing investigation of gene regulation of these cells in a consistent manner. Throughout this study, cells have been treated for 48 hours with 50 $\mu\text{g}/\text{ml}$ of human plasma-derived LDL oxidised by 18 hours of treatment with 25 μM Cu_2SO_4 , or with the equivalent volume of buffer vehicle (or no treatment).

5.2 Total RNA isolation and quality

Total RNA was extracted as described in the methods and prior to submission for sequencing underwent QC to determine purity and RNA integrity. RNA purity was evaluated using the NanoDrop Spectrophotometer (Thermofisher) to measure the absorbance of the sample: all samples had a 260/280 and 260/230 ratio > 2 . A representative analysis is shown in Figure 5.1A. RNA integrity was determined using

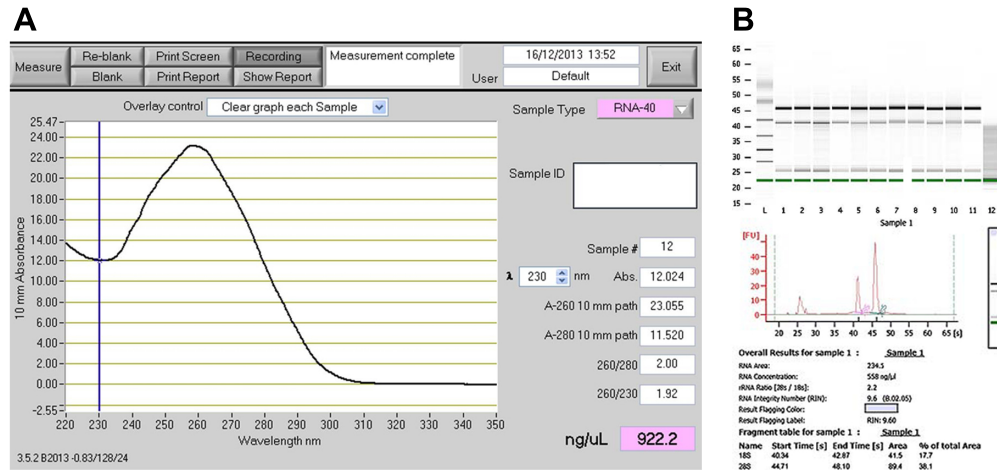


Figure 5.1: RNA quality. A) NanoDrop analysis of total RNA samples extracted using Qiagen's RNeasy kit and eluted in TE showed absorbance peaks at 260nm suggesting no contamination of the RNA. B) Electrophoresis of total RNA on the Agilent 2100 Bioanalyzer using the Eukaryote Total RNA nano assay. An electropherogram of one of the samples with relevant statistics is shown in the bottom half of the figure, and above this is a reconstructed gel image of all the samples on the run, allowing a qualitative comparison.

the Agilent High Sensitivity RNA high sensitivity kit on the Agilent 2200 TapeStation. All samples submitted had RNA integrity numbers (RIN) > 9. A representative trace is shown in Figure 5.1B.

5.3 Quality control of raw reads

RNA samples underwent TruSeq library preparation with polyA selection and libraries were subjected to 50-100 base-pair (bp) paired-end Illumina Hi-Seq 2500 or 4000 sequencing. Following sequencing, the raw reads were processed according to the *rseqR* pipeline discussed in Chapter 4.

Quality control of the raw read files was performed with *FastQC* and aggregate reports created using *multiQC*. A representative snapshot of the plots and elements of the report produced by *rseqR* pipeline are shown in Figure 5.2.

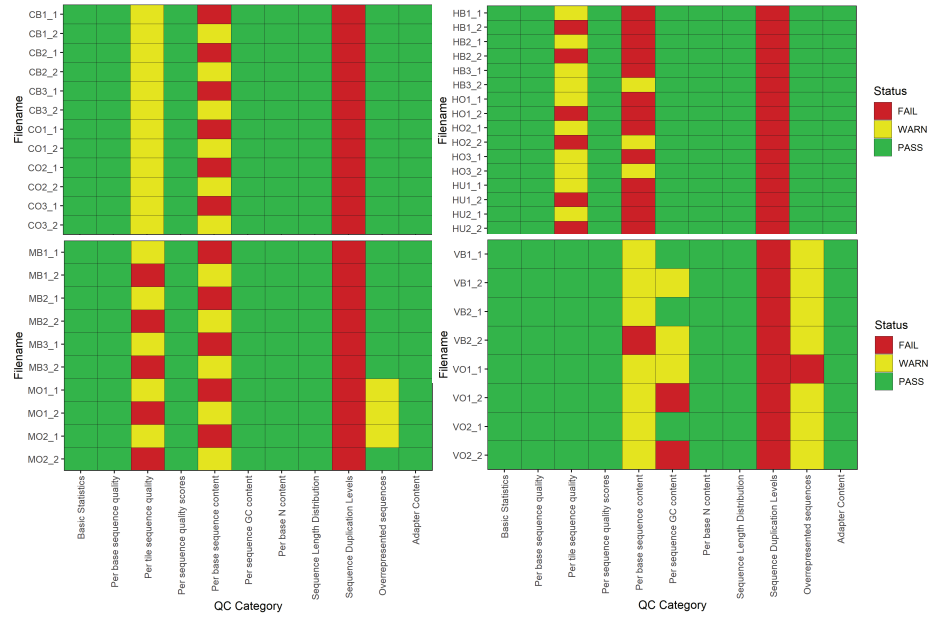


Figure 5.3: Summary of PASS/WARN/FAIL results of FASTQC modules.

An overall summary of the QC statistics is shown in Table 5.2 and a plot summarising the outcome of each QC module in *FastQC* is shown in Figure 5.3. The average number of read pairs across all samples was 33 464 231. Overall, the quality of reads was very high with all samples having a mean Phred score > 30 (Table 5.2).

There was a tendency to a slightly lower Phred score in the reverse (i.e. mate 2) reads than in the forward (i.e. mate 1 reads), which is not surprising as the accuracy/quality of read assignment decreases with longer reads. In Figure 5.3 we can see that all samples raised a warning or failure in the per base sequence content module. This is universal for RNA-seq libraries produced by TruSeq kit, and results from the annealing of ‘random’ hexamer primers at the start of the fragments during sequencing. These primers are not truly random, leading to a sequence bias at the start of the read. This is not usually a problem which can be fixed by processing and has not been shown to adversely affect downstream analysis. All libraries also failed the sequence duplication levels module; again this is common in RNA-seq data, where many of the duplicates

Table 5.2: Basic summary statistics of raw RNA-seq reads.

Cell	Treatment	Replicate	Reads	Mate 1		Mate 2	
				%GC	Phred	%GC	Phred
CD4	Buffer	1	36203863	47	39.02	47	38.22
CD4	Buffer	2	43329488	48	39.02	48	38.04
CD4	Buffer	3	37234583	47	39.02	47	38.04
CD4	oxLDL	1	38953724	48	39.05	48	38.30
CD4	oxLDL	2	37448242	48	39.03	48	38.04
CD4	oxLDL	3	40286215	48	39.03	48	38.14
HAEC	Buffer	1	37095366	48	34.55	49	33.40
HAEC	Buffer	2	32572103	48	34.57	49	33.47
HAEC	Buffer	3	47425304	49	39.00	49	38.09
HAEC	oxLDL	1	37027469	48	34.55	49	33.43
HAEC	oxLDL	2	33499491	48	34.55	49	33.43
HAEC	oxLDL	3	41295208	48	39.02	48	38.08
HAEC	Untreated	1	24720477	47	34.65	48	33.54
HAEC	Untreated	2	28157440	47	34.68	48	33.60
MP	Buffer	1	23889394	49	38.93	49	38.27
MP	Buffer	2	23931335	48	38.93	48	38.36
MP	Buffer	3	27407799	49	38.93	49	38.20
MP	oxLDL	1	29383615	49	38.96	49	38.40
MP	oxLDL	2	28212549	49	38.96	49	38.37
VSMC	Buffer	1	30282650	53	35.27	52	34.57
VSMC	Buffer	2	29122246	52	35.33	52	34.70
VSMC	oxLDL	1	29909176	54	34.89	54	34.10
VSMC	oxLDL	2	32289567	54	35.03	53	34.33

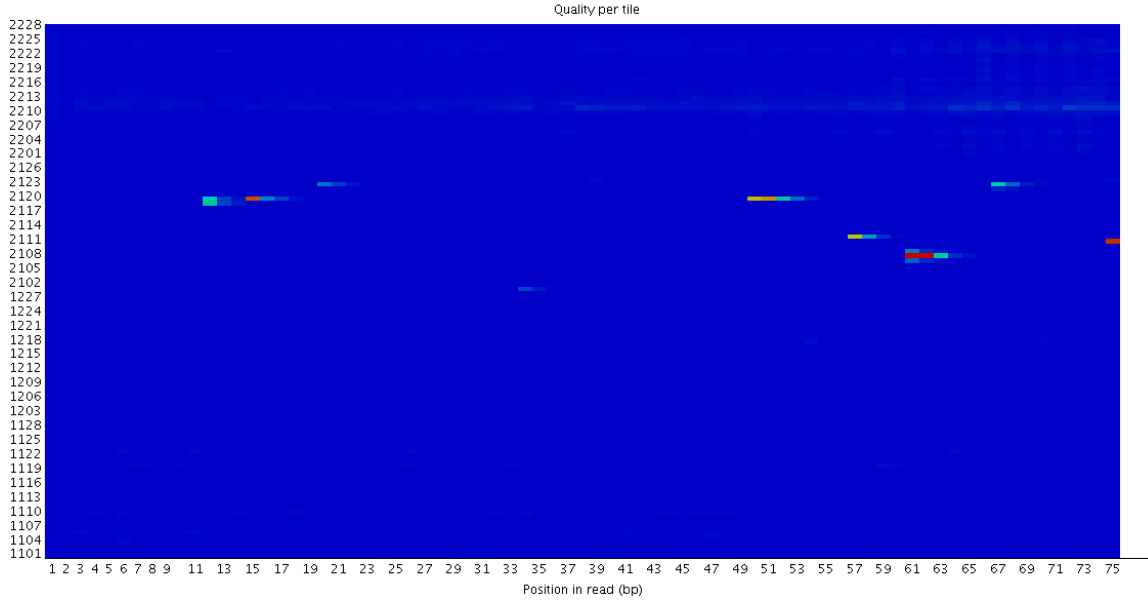


Figure 5.4: Heatmap plot of per tile sequence quality for a sample that failed this FASTQC module (MB1_2). This plot is representative of all samples that raised a warning or failure for this module. The errors are minor and restricted to small areas in the same place for all samples, suggesting a flow cell error.

represent high levels of gene expression, and so we have not removed these. Potentially more concerning are the flags raised in the per tile sequence quality modules in all cell types except the VSMCs. A representative plot of a sample that failed this module is shown below; in all samples that raised a warning or failure, the errors are at the same locations, suggesting a problem with the flow cell/sequencing process rather than the sample. However, despite raising a failure message, these errors are very minor and restricted, and overall these would not be expected to have any impact on downstream analysis and can be safely ignored.

The overall GC percentage was as expected for the CD4⁺ cells, HAECs, and macrophages but higher for the VSMCs and analysis of the GC distribution of these cells revealed a significant deviation from the theoretical (Figure 5.5). Alongside the deviation in the GC content distribution, the VSMCs samples also raised warnings in

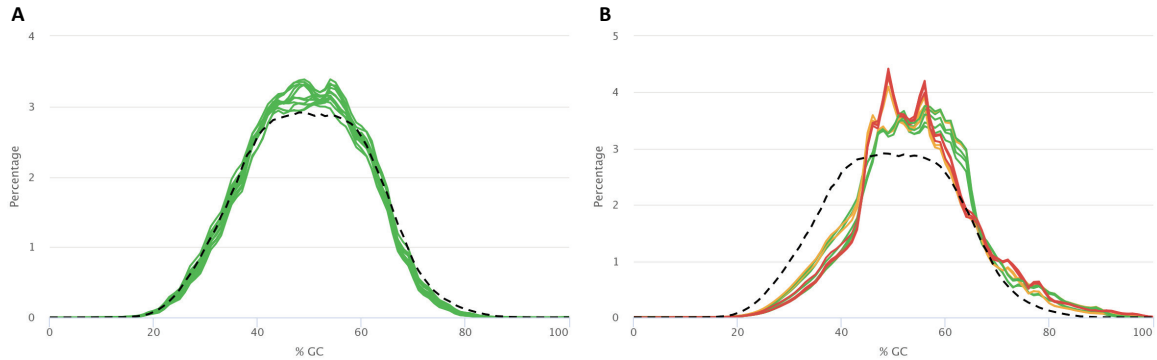


Figure 5.5: Per sequence GC content of RNA-seq samples. Plots displaying the GC content distribution in (A) macrophage libraries and (B) vascular smooth muscle cell-derived libraries. The x-axis shows the GC percentage, and the y-axis the proportion of reads in each library that contained that percentage of GC content. Also shown is the theoretical GC content for the hg38 transcriptome. There is a significant deviation in the GC content of the vascular smooth muscle cell-derived libraries compared to the macrophage-derived libraries. The human aortic endothelial cell and CD4+ T-cell derived libraries had similar GC content distributions to (A).

the over-represented sequences module. Taken together, this is very suggestive of a contaminating sequence causing a skew in the GC content distribution.

To investigate this further we first identified the over-represented sequences in the libraries and determined whether they matched known adapter sequences. Whilst a large number did indeed match the TruSeq Adapter sequences, no hit was found for most of them. Thus a BLAST search of a random subset of reads from the suspect libraries was performed. This revealed significant overlap with sequences from ribosomal RNA (rRNA). Whilst rRNA contamination can cause problems with *de novo* assembly or for detection of novel isoforms, it does not usually affect expression analysis as only reads aligning to features in the annotation file are counted. Nonetheless, it was decided to perform a computational rRNA depletion using SortMeRNA²¹¹. The RNA-seq files were blasted against the Rfam 5s and 5.8s databases²¹², and the Silva 18s and 28s eukaryote databases²¹³ using an E-value of 1, which is equivalent to a NCBI BLAST E-value of 1×10^{-5} . Reads that did not align to these databases were

Table 5.3: Alignment of VSMC RNA-seq reads to ribosomal RNA database.

Sample	Reads filtered (%)	Silva-euk-18s (%)	Silva-euk-28s (%)	Rfam-5s (%)	Rfam-5.8s (%)
VB1	36.79	25.08	11.55	0.12	0.04
VB2	33.40	20.93	12.29	0.13	0.04
VO1	50.86	31.84	18.86	0.10	0.05
VO2	45.13	27.81	17.17	0.11	0.03

retained. As SortMeRNA only accepts a single file at a time, the paired reads were interleaved into a single merged file for analysis and analysed in such a way that if a read successfully aligned to the rRNA database, then its matching pair was also filtered out with it. Table 5.3 shows the percentage of reads in each file that were filtered out as they mapped to the rRNA databases - the vast majority of rRNA reads were of the 18s variety, followed by 28s; only a very small percentage mapped to the 5s and 5.8s databases. In total, 42.8% of the reads were removed. Following filtering, the reads were unmerged and QC performed on the individual forward and reverse read files showed complete resolution of the spikes in the GC profile plot; however, downstream analyses using these filtered files showed reduced power to detect differential gene expression. A detailed analysis of the reads that were filtered out suggested that the filtering criteria was resulting in many false positive alignments to rRNA. We therefore opted to use the unfiltered reads for mapping and downstream analyses. As expected, these read libraries had a lower alignment rate than the other, uncontaminated libraries, but the identified differential genes were the same as those found using the filtered reads, but with higher significance values.

5.4 Abundance estimation and quality control of count data

Gene level abundance was determined by using the quasi-mapping approach of *Salmon* through the *rseqR* pipeline. The reference transcriptome used was for GRCh38/hg38, and included both protein-coding and non-coding RNA species. Visualisation of the pseudo-bam files generated by *Salmon* confirmed alignment at and across exons. Overall, the mean rate of alignment was 83.76% with a mean sequencing depth of 33 464 231 reads (Table 5.4). However, assessment of individual sample alignments showed that, as expected, this metric was much lower for the VSMCs samples which had rRNA contamination. Despite this, however, similar number of genes were detected across all samples (at least one mapped read per million reads), and the distribution of normalised gene counts between samples showed no significant differences.

The QC metrics assessed to this point provide a global overview of the count data and show similar trends across all samples. To assess how similar samples are to one another, we performed inter-correlation (principal component) analysis and hierarchical clustering based on the euclidean distances between rlog-transformed gene counts in each sample. When comparing the samples in their entirety, they segregate and cluster on the basis of their cell type (Figure 5.7A and B). The VSMCs and macrophages are widely separated on the first principal component which accounts for 50.5% of the variance. Both of these cell types are separated from the CD4⁺ cells and HAECs on the second principal component, which accounts for 26.2% of the variance. Not shown here is the third principal component, which accounted for 14.9% of the variance, and on which there was separation of the CD4⁺ cells and HAECs. The separation of the cell types is shown clearly on the heatmap of hierarchical clustering. This shows that

Table 5.4: Summary statistics of read mapping.

Cell	Treatment	Replicate	Total	Mapped	Mapped (%)
CD4	Buffer	1	36203863	31675426	87.492
CD4	Buffer	2	43329488	37554266	86.671
CD4	Buffer	3	37234583	32053610	86.086
CD4	oxLDL	1	38953724	35015936	89.891
CD4	oxLDL	2	37448242	32368611	86.436
CD4	oxLDL	3	40286215	35656692	88.508
HAEC	Buffer	1	37095366	34027101	91.729
HAEC	Buffer	2	32572103	30046352	92.246
HAEC	Buffer	3	47425304	43622747	91.982
HAEC	oxLDL	1	37027469	34332172	92.721
HAEC	oxLDL	2	33499491	30900239	92.241
HAEC	oxLDL	3	41295208	37784760	91.499
HAEC	Untreated	1	24720477	22520728	91.102
HAEC	Untreated	2	28157440	25505719	90.583
MP	Buffer	1	23889394	21585395	90.356
MP	Buffer	2	23931335	21307409	89.036
MP	Buffer	3	27407799	25181333	91.877
MP	oxLDL	1	29383615	26373526	89.756
MP	oxLDL	2	28212549	25198639	89.317
VSMC	Buffer	1	30282650	17826439	58.867
VSMC	Buffer	2	29122246	17990694	61.776
VSMC	oxLDL	1	29909176	13700071	45.806
VSMC	oxLDL	2	32289567	16314429	50.525

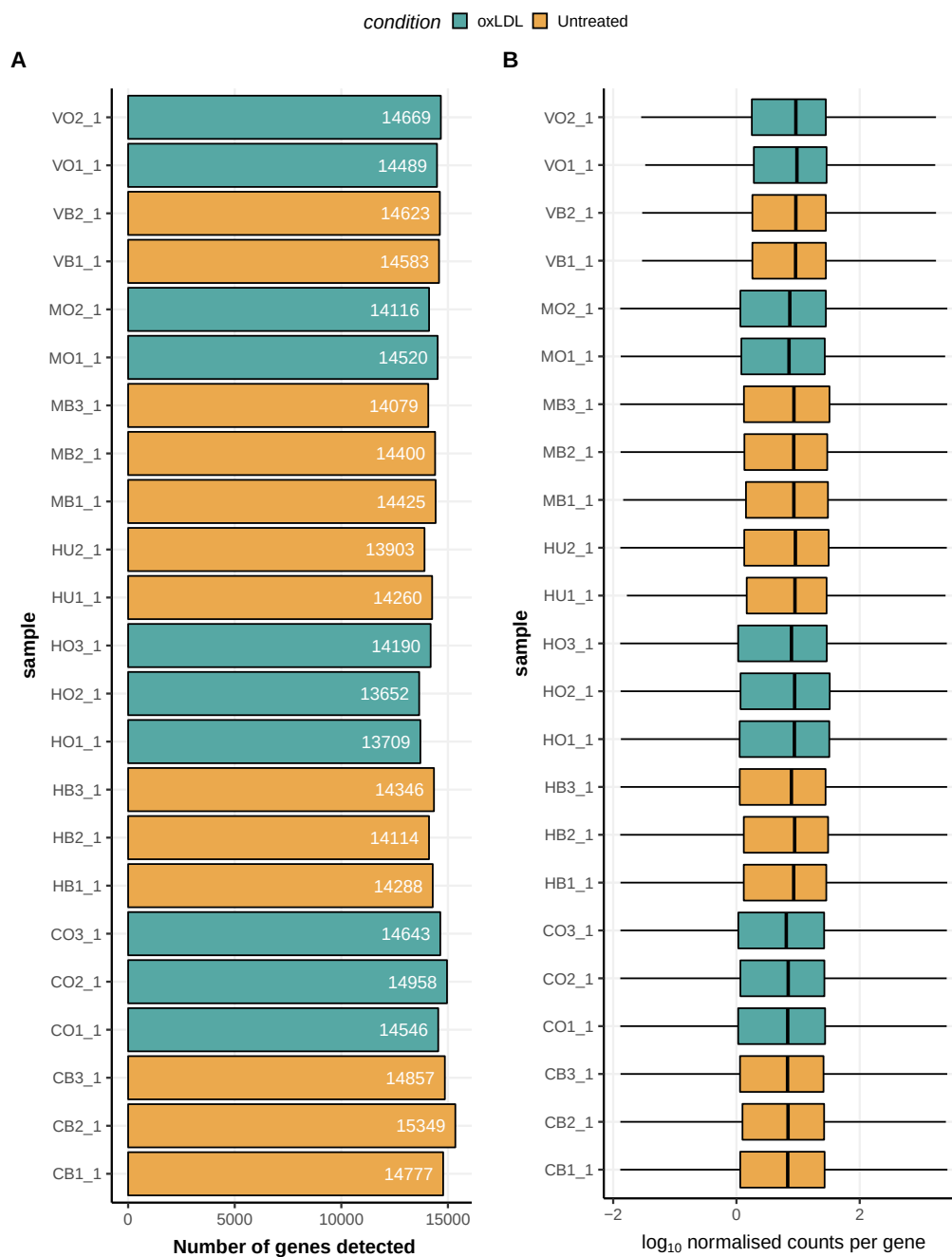


Figure 5.6: Gene counts statistics. (A) Number of genes detected per sample. A gene is defined as detected if it had at least one mapped read per million reads. (B) Distribution of normalised counts (trimmed mean of M-values (TMM) normalisation) for detected genes per sample.

not only do the individual cell types cluster, but there is a clear segregation between the vascular cell types (HAECs and VSMCs) and immune cell types (macrophages and CD4⁺ cells). Within each cell type subset, there is further segregation of samples on the basis of treatment (Figure 5.7C-D), indicating that oxLDL does indeed have an effect on gene expression. For the HAECs there is a clear batch effect for the samples HB3 and HO3 which has been incorporated into the downstream analysis.

5.5 Differential gene expression and pathway analysis

Differential expression analysis was performed by the *rseqR* package, using all three of *edgeR*, *DESeq2*, and *limma+voom* (Figure 5.8). The greatest number of differential genes were identified by *limma+voom*, whilst *edgeR* and *DESeq2* were much more similar in their outputs with a large overlap between genes identified as differentially expressed. Overall, *edgeR* was the most conservative, and on the basis of this, further downstream analyses were conducted using the *edgeR* output.

In total, *edgeR* identified 3907, 5414, 2832, and 2471 differentially expressed genes following oxLDL treatment of CD4⁺ cells, HAECs, macrophages, and VSMCs, respectively (Figure 5.9). To prioritise genes we used a ranking system that took into account both the p-value and fold change. Each significantly differentially expressed gene was assigned a score between 0 and 1 for p-value (with the lowest p-value resulting in 1 and the highest p-value a value of 0) and for fold change (with the highest absolute fold change resulting in 1 and the lowest 0). The mean of these two scores was then used to rank the genes.

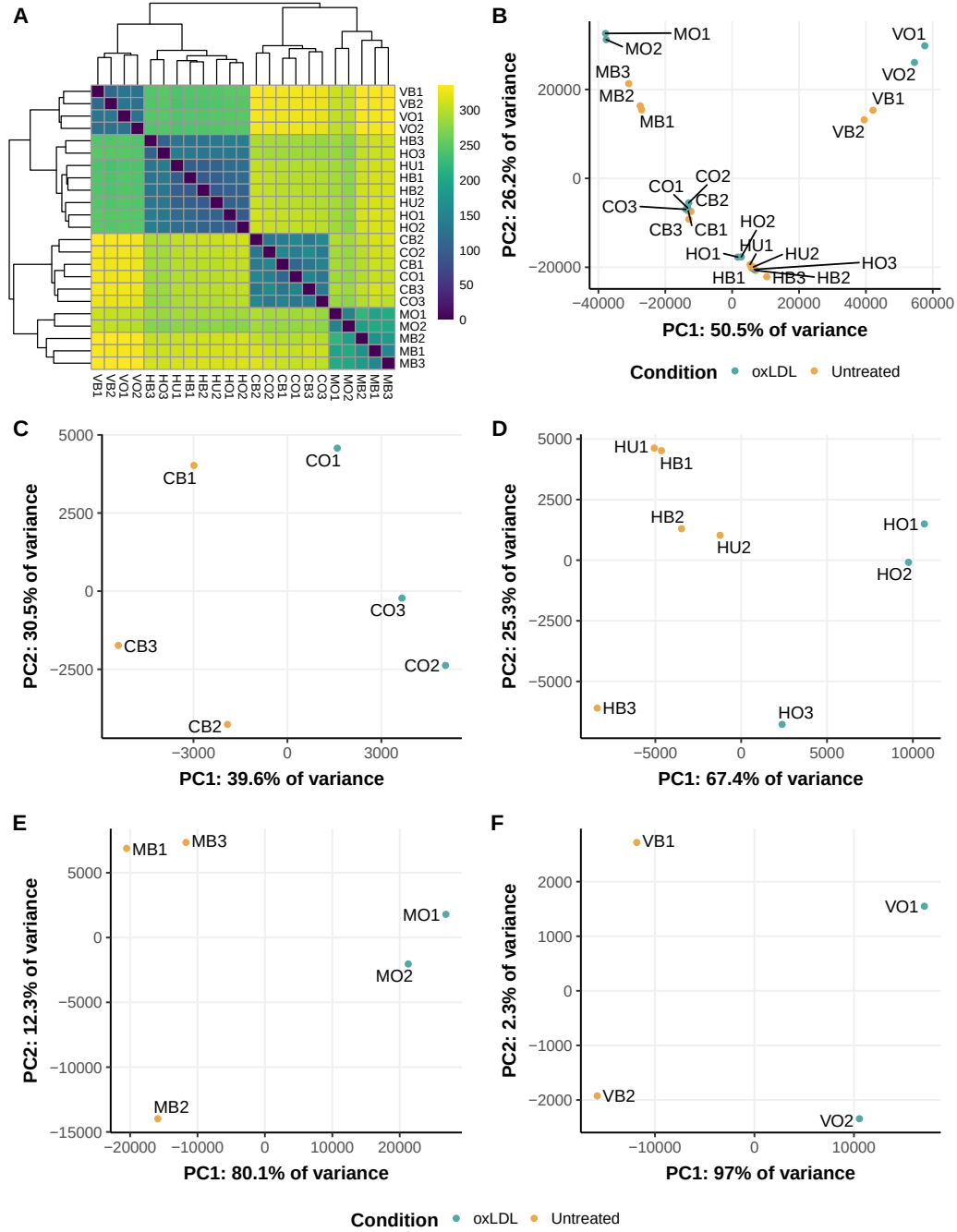


Figure 5.7: Sample similarity. A) Heatmap of the euclidean distances (based on rlog normalised gene counts) between samples, illustrating clustering of samples from the same cell type. This is also shown in the PCA plot of the entire sample set (B). Plots C-F illustrate PCA plots for individual cell types. Sample labelling strategy: C = CD4+ cells, H = human aortic endothelial cells, M = macrophages, V = vascular smooth muscle cells, O = oxLDL-treated, B = buffer-treated, U = untreated, 1-3 indicates replicate number. For example, MO1 = macrophage, oxLDL-treated, replicate 1.

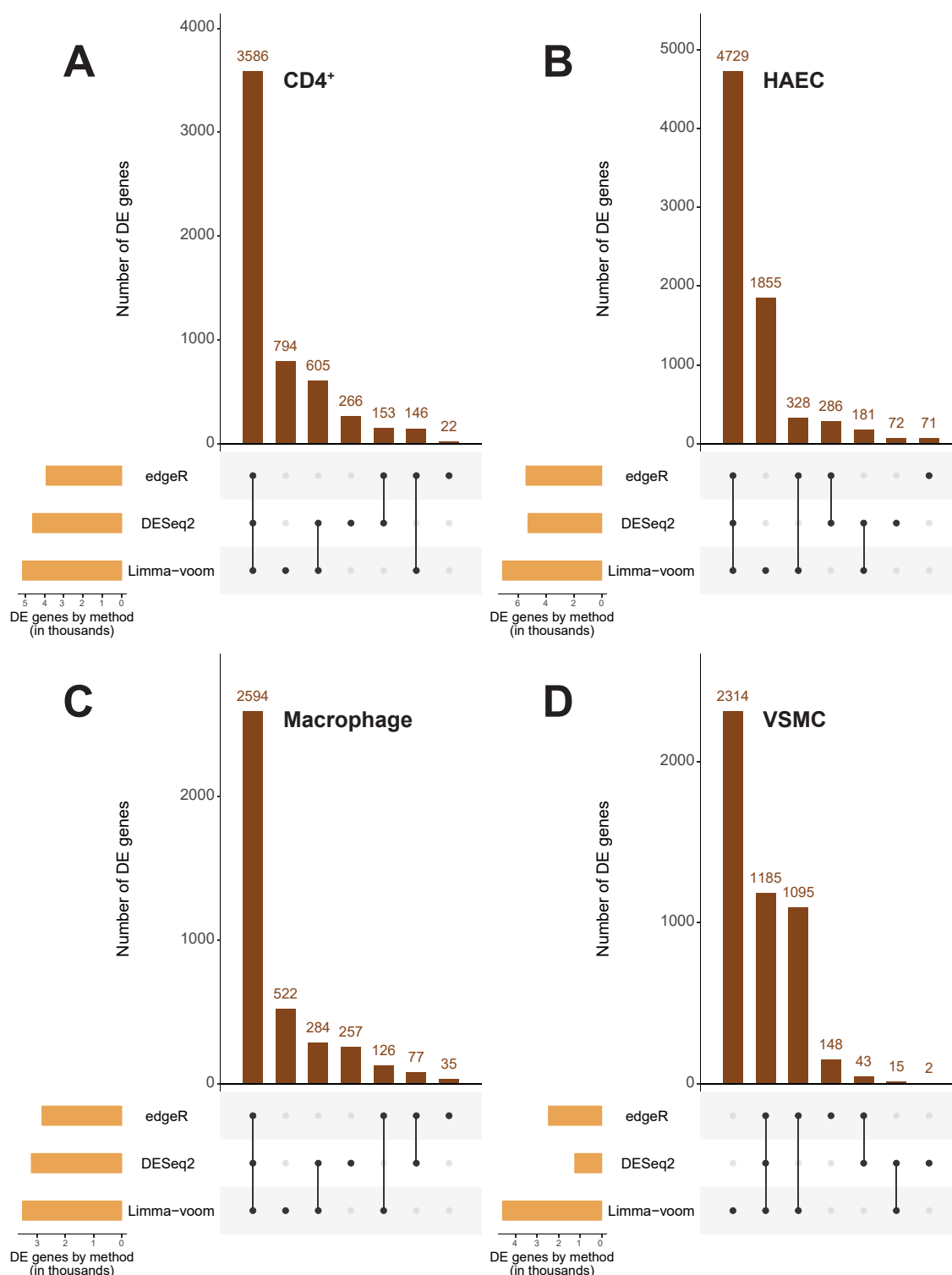


Figure 5.8: Upset plots displaying the number of differentially expressed genes (DEGs) detected by edgeR, DESeq2, and limma+voom between oxLDL-treated and untreated cells. Plots A-D display the results for CD4⁺ cells, HAECs, macrophages, and VSMCs, respectively. The x-axis displays the number of DEGs identified by each method, and the y-axis, the overlap between the methods.

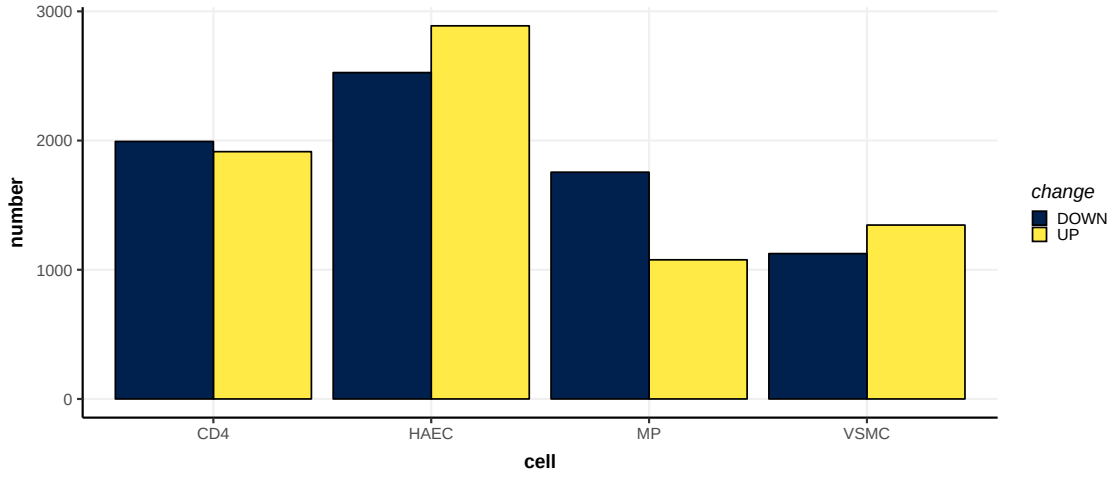


Figure 5.9: Differentially regulated genes in response to oxLDL. Barplot illustrating the number of up and down-regulated genes in each cell type in response to oxLDL.

5.5.1 Oxidised LDL alters expression of lipid-related and hypoxia-modulated genes in CD4⁺ cells

Of the 3907 genes differentially expressed in CD4⁺ cells, 1914 were up-regulated and 1993 were down-regulated. Amongst the most differentially regulated genes, a significant number are directly involved in lipid handling and metabolism (*RXRA*, *LRP2*, *PRKAA2*, *SMARCD3*, *ABCA4*, *PRELID2*, *GPER1*, *EEF1A2*). In keeping with this, there was statistical enrichment of lipid-related GO, KEGG, and reactome terms. In addition, many of the top genes were also related to immune function (*TNFRSF6B*, *MUCL1*, *MUC4*), cell-signalling (*IGFB2*, *GPER1*, *EPHA3*, *KLF4*, *BARX1*, *GPRC5C*), cell structure and vesicle-mediated transport (*PTPN3*, *ACTRT3*, *MISP*, *SCG3*, *STC1*), and apoptosis (*IGFBP2*, *PNMA2*), and the identified gene set pathways were consistent with this. Several of the differentially regulated genes are also known to be mediators of angiogenesis (*GPER1*, *ADM*, *MT3*, *MT1G*, *ADORA2B*), and whose expression is modulated by hypoxia. Gene set enrichment analysis identified

enrichment of pathways related to hypoxia signalling, lipid metabolism, inflammation, immune response, angiogenesis and transforming growth factor β (TGF β) signalling (Table 5.9).

5.5.2 Oxidised LDL significantly alters expression of genes involved in angiogenesis and lipid metabolism in HAECs

In HAECs, 5414 genes were differentially expressed; of these 2888 were up-regulated and 2526 were down-regulated. Amongst these genes, the most differentially regulated include some that are very important in lipid handling and metabolism (*NGFR*, *AKR1C1*, *NR1D1*, *CYP26B1*), angiogenesis and ECM remodelling (*GEM*, *EDNRB*, *GJA5*, *ECSCR*, *RASD1*, *TNXB*), and inflammatory responses (*HMOX1*, *GJA4*, *SERPIND1*). Pathway analysis identified enrichment of terms related to angiogenesis, cholesterol metabolism, regulation of macrophage and smooth muscle cell responses, and hypoxia and TGF β signalling. The most differentially regulated genes and important pathways are shown in Table 5.6 and Table 5.9.

5.5.3 In macrophages oxidised LDL causes significant changes in expression of genes involved in lipid and carbohydrate metabolism, and immune function and inflammation

In macrophages we identified 2832 DEGs, of which 1077 were upregulated and 1755 were downregulated. An overview of the top ten up and down-regulated genes (Table 5.7) identified diverse roles including mediation of lipid and carbohydrate metabolism (*NGFR*, *CES1*, *MBOAT7*, *MBOAT4*, *OLAH*, *PSMB8*, *ALOX5*, *NEU1*, *CD1C*), immune cell function and inflammation (*LURAP1L*, *CRYAB*, *PSMB8*, *HLA-*

DMA, *HLA-DOA*), and thrombosis (*F13A1*). Once again, pathway enrichment analysis (Table 5.9) identifies terms related to lipid handling, regulation of angiogenesis, and hypoxia response. Inflammation, immune response, and oxidative stress response terms are also enriched.

5.5.4 Oxidised LDL alters expression of lipid-handling and oxidative stress related genes in VSMCs

The VSMC dataset had the lowest number of differential genes; this was predominantly due to loss of sensitivity arising from ribosomal RNA contamination of mRNA. We identified 1346 upregulated genes and 1125 downregulated genes for a total of 2471 DEGs. Many of the top differentially regulated genes were related to lipid handling (*AKR1B10*, *AKR1C2*, *AKR1C1*, *RARRES1*, *SAA1*, *LDLR*). Others were related to oxidative stress (*HMOX1*, *NQO1*, *OSGIN1*), smooth muscle cell proliferation and migration (*HSPA6*, *ITGAB*, *MYOC*, *PPP1R18*, *TUBB*), and vasorelaxation (*DDAH2*, *NPPB*). Interestingly two different isoforms of the apoptosis and cell differentiation regulating gene *TCF19* were differentially regulated in completely different directions (Table 5.8), suggesting not only changes in gene expression in response to oxLDL but also changes in isoform usage. Of all the cell types, the VSMCs showed the greatest enrichment for terms related to lipid handling (Table 5.9). Other enriched terms related to ECM organisation, blood vessel tone, and TGF β signalling.

5.5.5 Genes differentially expressed across all cell types play roles in lipid metabolism, angiogenesis and oxidative stress regulation

We next compared the genes differentially expressed across all cell types. To do this we identified overlaps between the differentially expressed gene lists for each cell type (Figure 5.10). The greatest overlap was between HAECs and CD4⁺, but this was a reflection of the number of differential genes identified in each of those cell types. In fact, a formal assessment of significance of overlap based on odds ratio shows that the most significant overlap was between differential genes identified in HAECs and VSMCs, in keeping with their roles as vascular cells. Across all cell types, we identified 89 genes as being commonly differentially expressed, although not all in a concordant fashion. A striking number of these genes are involved in lipid metabolism. *TXNRD1*, *PLIN2*, and *NPAS2* all regulate lipid metabolism by modulating the activity of PPAR α . *PTGRN* and *PTGR1* are mediators of prostaglandin activity; prostaglandins are lipid-derived autocooids that have also been shown to have potent anti-atherogenic effects. *SQLE* and *GPX1* are both catalysts of lipid redox reactions, and *FADS2* is a key component of the polyunsaturated fatty acid biosynthesis pathway. *SLC44A1* is one of the main choline transporter proteins, whose downregulation has been shown to impair lipid metabolism due to defective phosphatidylcholine synthesis. As previously discussed, phosphatidylcholine is the major phospholipid moiety of LDL.

Several of the commonly differentially regulated genes also play important roles in vasculogenesis. *HSPG2* encodes the large ECM protein perlecan which maintains the endothelial barrier function, inhibits smooth muscle proliferation and promotes FGF2 activity, thereby stimulating endothelial cell regeneration. *ITGA6* is an integrin that

binds to ECM proteins and promotes angiogenesis. *ATF4* is a positive regulator of VEGF production and is thus pro-angiogenic.

Finally, many of the genes identified here have potent redox activity including *TXNDR1*, *SOD2*, *GPX1*, *PTGR1*, and *SQLE*. The differential regulation of these genes is in keeping with the inflammatory nature of atherosclerosis.

We also compared the top most significant pathways identified by enrichment analysis in each cell type following treatment with oxLDL. As Figure 5.12 shows, pathways related to lipid metabolism were the most strongly enriched and the greatest similarity in the pattern of enrichment was between *glspl{haec}*, macrophages, and *glspl{vsmc}*.

5.5.6 VCAM1 is downregulated in response to oxLDL

In Chapter 3, we showed that cells treated with oxLDL had increased expression of ICAM1. We did not find that to be the case here - the overall expression of ICAM1 remained static on exposure to oxLDL, and it may well be that the flow cytometric data is reflecting an increase in cell surface localisation of ICAM1 rather than an increase in transcription and translation. Of more interest are the changes in VCAM1 expression. VCAM1 was not found to be significantly expressed in CD4⁺ cells at all. However, in the remaining three cell types, it was significantly downregulated in response to oxLDL with log fold changes of -2.1, -8.4, and -1.0 in macrophages, HAECs, and VSMCs respectively. Other studies have found that oxLDL causes upregulation of VCAM1, and we discuss these findings further below.

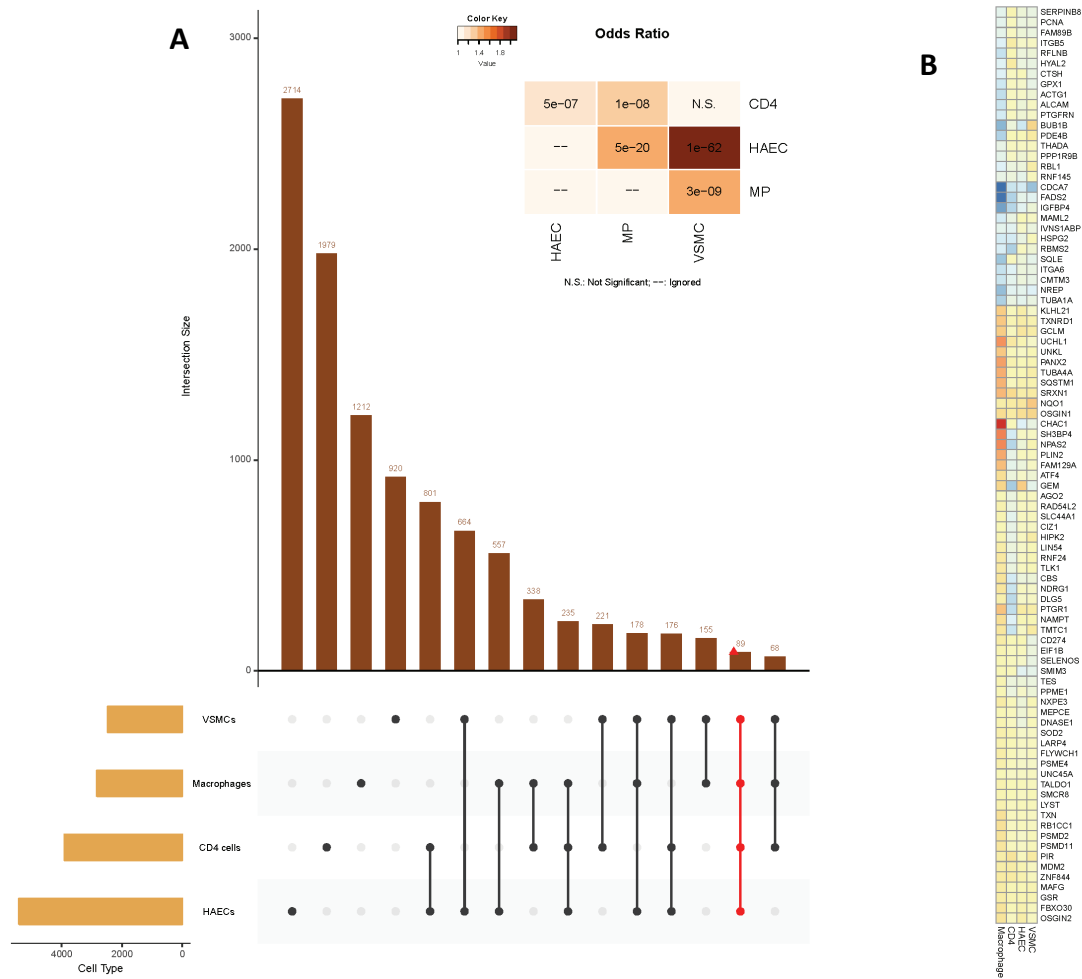


Figure 5.10: Overlap of differential genes. A) UpSet plot showing the overlap of the differential genes from the differential analysis of each cell type. Marked by a red line are the genes differentially expressed across all 4 cell types. The inset shows the statistical significance of the overlap based on the p-value for the odds ratio of the overlap. B) Heatmap showing the log2 fold change in expression of the genes differentially expressed across all 4 cell types.

5.6 Genes differentially regulated by oxLDL are associated with cardiovascular risk variants

We next explored the relationship between gene expression changes in response to oxLDL and CAD genetic risk variants that have been identified by GWAS. Using the *snpR* package, all variants related to CAD were extracted from the GWASCatalog and pruned to generate independent SNPs. We then identified all SNPs in high linkage disequilibrium ($r^2 > 0.8$) with these independent SNPs to generate CAD locus intervals. CAD variant associated genes were then defined as any gene within a 50 kbp distance of one of these risk loci. In total, we identified 563 such genes and pathway analysis identified these genes as being enriched for processes related to atherosclerosis. Of these genes, 74, 134, 79, and 61 were differentially regulated by oxLDL in CD4⁺ cells, HAECs, macrophages, and VSMCs, respectively (Figure 5.11). We regarded these genes as especially interesting as their action may be influenced by the variants identified in GWAS. In particular, we identified 4 of these CAD-associated genes that were differentially expressed across all 4 cell types: *HIPK2*, *CTSH*, *CBS*, and *GPX1* (Figure 5.11).

5.7 Discussion

Oxidised LDL plays a key role in the development and progression of atherosclerosis. In an attempt to understand the molecular mechanisms underlying its effects, it is necessary to study the transcriptional changes it exerts on the cells involved in atherosclerosis. Over the last one and a half decades, the development of genome wide approaches to studies of transcription have led to a number of studies comparing

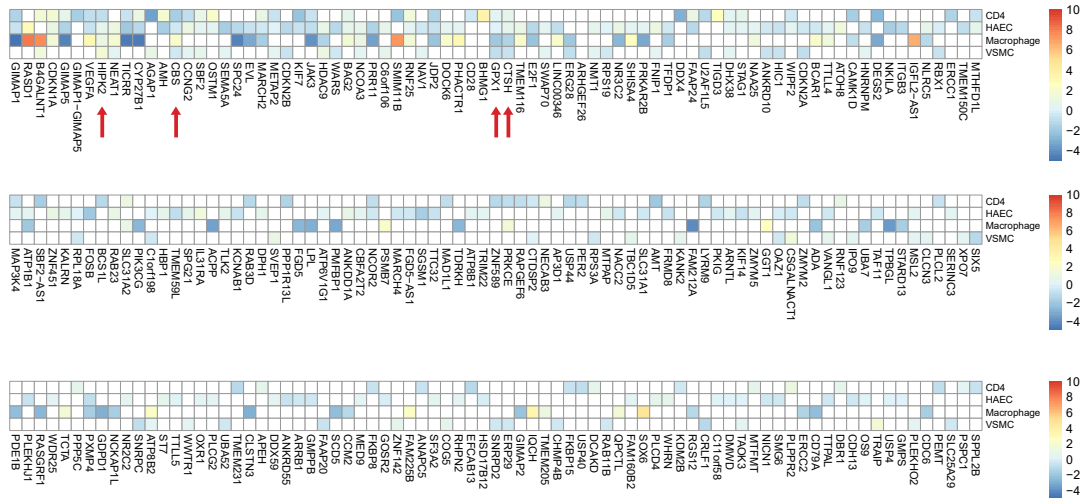


Figure 5.11: Genes differentially expressed in each cell type that are also associated with coronary artery disease (CAD) loci. CAD variants were obtained from the GWAS catalog and used to generate loci consisting of an index SNP plus SNPs in high linkage disequilibrium (r^2 0.8). All genes within 50kb of these loci were identified. Those that were differentially expressed in at least one cell type following exposure to oxLDL are shown with the colour gradient indicating the fold change (in a base 2 logarithmic scale). The red arrows highlight those genes that were differentially expressed in all 4 cell types.

gene expression changes in untreated and oxLDL-treated cells. Whilst these studies have provided insights into the mechanisms of the actions of oxLDL, they have several limitations which has led to the analyses described here.

Firstly, all the studies looking directly at the effects of oxLDL have been microarray based. RNA-seq is now recognised as being superior to DNA microarray in transcriptome profiling for several reasons²¹⁴: RNA-seq has a broader dynamic range and is better at detecting low abundance transcripts, thus allowing detection of more differential genes; it avoids the technical issues inherent to microarrays such as crosshybridisation, non-specific hybridisation and limited detection range of individual probes; and, because it does not rely on a pre-designed set of probes, it reduces transcript detection bias, whilst at the same time avoids problems related to probe redundancy and annotation, which greatly simplifies analysis. Secondly, several of these stud-

ies have investigated cell lines (e.g., THP-1 or U937) or cell types not involved in atherosclerosis in vivo (e.g., human umbilical vein or artery endothelial cells), and generalising those results to other cells may not be appropriate. Thirdly, several studies have used what would now be considered inadequate statistical measures of differential gene expression, and where robust methodologies have been used they have differed between studies. Finally, in all cases, there have been differences in cell type, methods of cell isolation, culture, oxLDL production and treatment. As discussed earlier, oxLDL has pleiotropic and sometimes opposing effects depending on the degree of oxidation and which moieties have been most oxidised. Thus, comparisons of studies in which the preparation method of, and duration of treatment with, oxLDL have been different is difficult. Our study overcomes these limitations by using RNA-seq to perform an unbiased genome-wide expression analysis across a variety of cell types all exposed to oxLDL oxidised to the same amount, for the same duration and at the same concentration. The data has been analysed using the most sophisticated quality control and statistical methodologies.

Comparison of gene expression in untreated and oxLDL-treated cells identified a large number of differentially expressed genes in each cell type. Gene set enrichment analysis identified the key influences of oxLDL on these cells were to alter the pathways involved in cellular lipid handling, oxidative stress, and inflammation. These pathways allow organisms to quickly adapt to changing energy requirements and to oxidative status.

Interestingly, a very recent study (currently in pre-print) has used RNA-seq to investigate HDL-induced changes in gene expression in human monocyte-derived macrophages. This study identified three key genes as upregulated - fatty acid desaturase(*FADS1*), insulin induced gene 1 (*INSIG1*), and low density lipoprotein receptor (*LDLR*). All

three of these genes are involved in lipid handling and oxidative stress regulation, and in our study were differentially regulated in the opposite direction in macrophages. This directional difference in fold change is in keeping with the biology of HDL and LDL particles as so-called ‘good’ and ‘bad’ cholesterol, with the former being involved in cholesterol efflux and the latter in cholesterol influx.

In macrophages, HAECs, and VSMCs we showed that expression of the cell adhesion molecule VCAM1 was significantly down-regulated following exposure to oxLDL. In some ways, this is counter-intuitive, and previous studies have shown the opposite effect [63; 198; 199; Li_2002]. In tumour angiogenesis, vascular cells have been shown to downregulate VCAM1 expression in response to basic fibroblast growth factor and VEGF²¹⁵. This has been shown only to occur after treatment with VEGF for > 24h; in studies where cells were exposed to VEGF for less than this duration, upregulation was seen²¹⁶. The previous studies which have assessed the expression of VCAM1 in endothelial cells in response to oxLDL have also used durations of exposure < 24h^{62,63,198,199}. It is tempting to speculate, therefore, that whilst a short incubation with oxLDL leads to increased VCAM1 expression, a longer incubation impairs its expression. Biologically, this makes sense, as decreased expression of cellular adhesion molecules is critical for vascular cell migration and neovascularisation.

Amongst all cell types, terms related to cytoskeletal structure and organisation were also heavily represented, in keeping with the proliferative and migratory phenotypes seen in these cells following exposure to oxLDL. In the vascular cell types, angiogenesis-related terms also featured, and indicates a pivotal role for oxLDL in the neovascularisation seen in atherosclerotic plaques. The TGF β signalling pathway has been shown to be involved in angiogenesis, particularly in solid organ cancers. In both the HAECs and VSMCs, this pathway emerges as a strongly enriched pathway, and

many genes related to the pathway are also differentially regulated in the macrophages and CD4⁺ cells, where the pathway is also enriched to a lesser extent.

Members of the TGF β super family are involved in a diverse range of processes including cell proliferation and differentiation, cell migration, and ECM production. Signalling via this pathway has both anti- and pro-atherogenic effects. There are two main TGF β signalling pathways, the TGF β /activin/nodal pathway and the BMP/GDF pathway. Intracellular signalling from the former is mediated by SMAD2 and SMAD3, and from the latter by SMAD1, SMAD5, and SMAD9. The pathways are negatively controlled by inhibitory SMADs (SMAD6 and SMAD7) and ubiquitination. Across all cell types, there was differential regulation of a number of E3 ubiquitin-ligases (e.g. *SMURF1*) which promote the degradation of SMADs, thus attenuating TGF β signalling. In all cell types, at least one of the SMAD proteins was differentially regulated as were TGF β -beta receptors and bone morphogenic protein (BMP) receptors, and in HAECs and VSMC several of these had changes in expression.

There has been one RNA-seq study which have used smooth muscle cells derived from carotid arterial tissue of patients undergoing endarterectomy for significant arterial stenosis, with patients divided into asymptomatic and symptomatic groups²¹⁷. The former had more than 80% stenosis of their vessels detected incidentally or as part of another investigation, but were yet to have a neurological event. The major limitation of this study was that it was not a comparison of healthy vs diseased state and, as such, only small numbers of differential genes were identified (67 genes in total). However, the analysis from this study also identified TGF β signalling as the crucial pathway mediating symptomatic carotid arterial disease, which is in keeping with the results from the pathway enrichment analyses performed here. In addition, they noted that the carotid arterial VSMCs from symptomatic patients had lower

levels of BMP2 expression than asymptomatic patients, and suggested that higher levels might promote a protective osteogenic phenotype that is less likely to result in plaque rupture. Whether this suggestion is true is unclear, as previously, in a population based study, high levels of BMP2 have been found to be associated with plaque progression in atherosclerosis²¹⁸. In our study *BMP* was downregulated in macrophages; this may be a reflection of the fact that our study looks at a much earlier stage of atherosclerosis when plaque calcification would not be expected.

The HIF1 α signalling pathway also emerged as an important signalling pathway across all the cell types. The internal milieu of advanced atherosclerotic plaques is characterised by a hypoxic environment. Cell responses to hypoxia are mediated by HIF1 α which mediates inflammatory responses, angiogenesis, and metabolic switches key to cell survival. Although hypoxia is the major stimulant to HIF1 α activation, this was not the case here, as all cells were cultured under normoxic conditions. This suggests that the activation of the HIF1 pathway in response to oxLDL is independent of hypoxia, and occurs even in early atherosclerosis. A potential mechanism for this activation may be downregulation of *EGLN3*. This protein encodes a prolyl hydroxylase which hydroxylates a specific prolyl in HIF1 α and targets it for degradation. *EGLN3* is downregulated in both CD4⁺ cells and HAECs (by more than 8-fold in the latter). To our knowledge a direct effect of oxLDL on levels of prolyl hydroxylases has not previously been reported, and this may be a novel mechanism of action.

We considered whether any of the differentially regulated genes might be associated with known CAD-associated variants and identified this to be the case for 563 of them. Four of these genes were differentially regulated across all 4 cell types: *HIPK2*, *CBS*, *CTSH*, and *GPX1*.

HIPK2 encodes the serine/threonine kinase homeodomain-interacting protein kinase

2, which plays roles in diverse processes. Of particular relevance, it has been shown to regulate TGF β -dependent angiogenesis during embryonic development and to inhibit signalling via SMAD1 and SMAD3-dependent pathways²¹⁹. In addition, HIPK2 is part of a multi-protein corepressor complex with histone deacetylase 1 that targets the HIF1 α promoter, reducing its expression²²⁰. Thus, *HIPK2* may represent a common mediator of oxLDL-induced changes in TGF β and HIF1 α signalling across multiple cell types.

CTSH encodes the cysteine petidase, cathepsin H. Cathepsin H has both endopeptidase and exopeptidase activity, and is the only lysosomal aminopeptidase. It has been shown have increased expression in a variety of inflammatory conditions and is co-localised with modified LDL in atherosclerotic plaques²²¹, particularly in macrophages and VSMCs. Incubation of macrophages with recombinant cathepsin H, LDL particles and cholesterylesterase causes a significant increase in intracellular lipid concentration and foam cell formation²²¹. Variants in this gene may well promote atherosclerosis through an impaired ability in lysosomal degradation of ingested cholesterol.

The product of *CBS* is cystathionine β -synthase, which converts homocysteine to cysteine. Hyperhomocysteinaemia is a key risk factor for atherosclerosis and coronary artery disease²²². The mechanism for this is unclear, but it is thought that hyperhomocysteinaemia may alter the methylation status of cells resulting in changes in global DNA methylation and transcription. Whilst it has been shown that hyperhomocysteinaemia potentiates the effects of oxLDL on cells, to our knowledge it has not been shown that oxLDL directly affects homocysteine levels. Thus, it is possible that oxLDL mediates its affects through changes in DNA methylation.

Finally *GPX1* encodes Glutathione Peroxidase 1, which is part of the principal anti-oxidant system of mammalian cells. As such it plays an important protective role

against cell-mediated oxidation of LDL and also oxLDL-mediated damage²²³, and alteration in its expression or activity will crucially alter the intracellular redox status.

To summarise, in this chapter we have investigated the transcriptional changes induced by oxLDL in CD4⁺ cells, HAECs, macrophages, and VSMCs. We have shown that the key effect of oxLDL is on genes involved in lipid handling and metabolism, oxidative stress and inflammation, and we have identified differential genes that may be affected by CAD variants. Oxidised LDL has potent effects on ECM modelling, cell-to-cell interactions and angiogenesis. We have identified the TGF β and HIF1 α signalling pathways as key mediators of these effects. The *HIPK2* gene is differentially expressed in response to oxLDL across all four cell types and may be a common mediator of its effects on these pathways. In addition, oxLDL may cause changes in transcription by altering DNA methylation globally through its effects on *CBS*.

Overall, we believe that our gene expression analysis provides a robust transcriptomic signature of the response to oxLDL in the key cells involved in the atherosclerotic process. To our knowledge, it is the first RNA-seq based study of the effects of oxLDL, and certainly the first to investigate the response of CD4⁺ cells to oxLDL. We have identified a number of candidate genes as being important in understanding the response of these cells to oxLDL and, in particular, have identified novel candidate genes as being potential mediators of the effects of oxLDL across multiple cell types.

Some limitations of this study should be acknowledged. Firstly, due to cost constraints, we had limited number of replicates for each condition and cell type. It would be worthwhile repeating the study with larger numbers of samples, particularly for the VSMCs, where the sensitivity was limited by rRNA contamination. Secondly, RNA-seq is inherently limited to the measurement of transcripts and so can not identify other mechanisms potentially regulating the effects of oxLDL such as by epigenomic

Table 5.5: Top ten most up- and down-regulated genes in CD4+ cells following treatment with oxLDL.

Gene	log2FC	FDR	Symbol	name
Up-regulated				
ENSG00000104112	6.751	0.003	SCG3	secretogranin III
ENSG00000243509	6.969	0.011	TNFRSF6B	TNF receptor superfamily member 6b
ENSG00000260903	5.320	0.009	XKR7	XK related 7
ENSG00000109832	4.910	0.003	DDX25	DEAD-box helicase 25
ENSG00000214553	4.374	0.006	LRRC37A11P	leucine rich repeat containing 37 member A11, pseudogene
ENSG00000006128	4.406	0.007	TAC1	tachykinin precursor 1
ENSG00000099812	4.256	0.007	MISP	mitotic spindle positioning
ENSG00000172551	4.474	0.011	MUCL1	mucin like 1
ENSG00000237452	3.536	0.007	BHMG1	basic helix-loop-helix and HMG-box containing 1
ENSG00000150773	3.360	0.006	PIH1D2	PIH1 domain containing 2
Down-regulated				
ENSG00000204072	-7.991	0.007	ARMCX7P	armadillo repeat containing X-linked 7, pseudogene
ENSG00000044524	-6.164	0.005	EPHA3	EPH receptor A3
ENSG00000159167	-5.660	0.004	STC1	stanniocalcin 1
ENSG00000115457	-5.348	0.004	IGFBP2	insulin like growth factor binding protein 2
ENSG00000164638	-5.335	0.003	SLC29A4	solute carrier family 29 member 4
ENSG00000131668	-5.256	0.005	BARX1	BARX homeobox 1
ENSG00000141622	-5.229	0.006	RNF165	ring finger protein 165
ENSG00000149596	-4.903	0.004	JPH2	junctophilin 2
ENSG00000176383	-4.804	0.007	B3GNT4	UDP-GlcNAc:betaGal beta-1,3-N-acetylglucosaminyltransferase 4
ENSG00000164850	-4.725	0.008	GPER1	G protein-coupled estrogen receptor 1

mechanisms, post-translational modifications, and protein modifications. We hope to address the former in subsequent chapters.

A next step forward would be to compare how the transcriptomic signature in response to oxLDL compares to that of the same cells derived from an atherosclerotic plaque. Limited studies employing laser dissection of macrophages from murine atherosclerotic plaques to measure gene expression exist, but not for other cell types; no comparable human studies have been reported. However, with the advent of single-cell RNA-seq techniques, gene expression profiling of individual cell types in an atherosclerotic plaque may become more feasible.

Table 5.6: Top ten most up- and down-regulated genes in HAECs following treatment with oxLDL.

Gene	log2FC	FDR	Symbol	name
Up-regulated				
ENSG00000064300	5.557	0.000	NGFR	nerve growth factor receptor
ENSG00000100292	4.489	0.006	HMOX1	heme oxygenase 1
ENSG00000187134	3.294	0.001	AKR1C1	aldo-keto reductase family 1 member C1
ENSG00000164949	2.870	0.001	GEM	GTP binding protein overexpressed in skeletal muscle
ENSG00000108602	2.841	0.002	ALDH3A1	aldehyde dehydrogenase 3 family member A1
ENSG00000108551	2.815	0.002	RASD1	ras related dexamethasone induced 1
ENSG00000126368	2.792	0.003	NR1D1	nuclear receptor subfamily 1 group D member 1
ENSG00000025039	2.606	0.003	RRAGD	Ras related GTP binding D
ENSG00000171658	2.450	0.000	NMRAL2P	NmrA like redox sensor 2, pseudogene
ENSG00000196611	2.405	0.001	MMP1	matrix metalloproteinase 1
Down-regulated				
ENSG00000168477	-7.118	0.018	TNXB	tenascin XB
ENSG00000228883	-5.086	0.009	LY6G5C	lymphocyte antigen 6 family member G5C
ENSG00000003137	-3.822	0.001	CYP26B1	cytochrome P450 family 26 subfamily B member 1
ENSG00000171864	-3.566	0.002	PRND	prion like protein doppel
ENSG00000187513	-3.840	0.006	GJA4	gap junction protein alpha 4
ENSG00000265107	-3.506	0.001	GJA5	gap junction protein alpha 5
ENSG00000169715	-3.063	0.000	MT1E	metallothionein 1E
ENSG00000240583	-2.884	0.001	AQP1	aquaporin 1 (Colton blood group)
ENSG00000147113	-2.966	0.004	CXorf36	chromosome X open reading frame 36
ENSG00000116133	-2.631	0.001	DHCR24	24-dehydrocholesterol reductase

Table 5.7: Top ten most up- and down-regulated genes in macrophages following treatment with oxLDL.

Gene	log2FC	FDR	Symbol	name
Up-regulated				
ENSG00000205609	11.476	0.005	EIF3CL	eukaryotic translation initiation factor 3 subunit C like
ENSG00000137337	10.811	0.008	MDC1	mediator of DNA damage checkpoint 1
ENSG00000064300	10.122	0.011	NGFR	nerve growth factor receptor
ENSG00000198848	9.725	0.010	CES1	carboxylesterase 1
ENSG00000277025	9.720	0.011	MBOAT7	membrane bound O-acyltransferase domain containing 7
ENSG00000153714	9.656	0.011	LURAP1L	leucine rich adaptor protein 1 like
ENSG00000109846	8.907	0.003	CRYAB	crystallin alpha B
ENSG00000228582	8.874	0.007	TAP2	transporter 2, ATP binding cassette subfamily B member
ENSG00000148677	8.291	0.003	ANKRD1	ankyrin repeat domain 1
ENSG00000177669	9.133	0.014	MBOAT4	membrane bound O-acyltransferase domain containing 4
Down-regulated				
ENSG00000235715	-12.542	0.003	PSMB8	proteasome subunit beta 8
ENSG00000281887	-11.383	0.005	GIMAP1-GIMAP5	GIMAP1-GIMAP5 readthrough
ENSG00000229300	-10.674	0.006	C6orf48	chromosome 6 open reading frame 48
ENSG00000214357	-10.602	0.007	NEURL1B	neuralized E3 ubiquitin protein ligase 1B
ENSG00000242361	-10.051	0.006	HLA-DMA	major histocompatibility complex, class II, DM alpha
ENSG00000124491	-9.189	0.006	F13A1	coagulation factor XIII A chain
ENSG00000235744	-9.350	0.010	HLA-DOA	major histocompatibility complex, class II, DO alpha
ENSG00000125810	-8.722	0.003	CD93	CD93 molecule
ENSG00000012779	-9.306	0.011	ALOX5	arachidonate 5-lipoxygenase
ENSG00000158481	-9.325	0.011	CD1C	CD1c molecule

Table 5.8: Top ten most up- and down-regulated genes in VSMCs following treatment with oxLDL.

Gene	log2FC	FDR	Symbol	name
Up-regulated				
ENSG00000096155	9.903	0.001	BAG6	BCL2 associated athanogene 6
ENSG00000224472	8.370	0.005	TCF19	transcription factor 19
ENSG00000228589	8.330	0.005	SPCS2P4	signal peptidase complex subunit 2 pseudogene 4
ENSG00000281500	7.927	0.007	SLC37A4	solute carrier family 37 member 4
ENSG00000275882	7.887	0.008	IKBKGP1	inhibitor of nuclear factor kappa B kinase subunit gamma pseudogene 1
ENSG00000137491	6.603	0.003	SLCO2B1	solute carrier organic anion transporter family member 2B1
ENSG00000198074	5.628	0.000	AKR1B10	aldo-keto reductase family 1 member B10
ENSG00000173110	4.777	0.002	HSPA6	heat shock protein family A (Hsp70) member 6
ENSG00000204351	4.697	0.012	SKIV2L	Ski2 like RNA helicase
ENSG00000100292	3.476	0.000	HMOX1	heme oxygenase 1
Down-regulated				
ENSG00000236428	-9.631	0.001	PPP1R18	protein phosphatase 1 regulatory subunit 18
ENSG00000137273	-8.625	0.004	FOXF2	forkhead box F2
ENSG00000233890	-8.272	0.005	TCF19	transcription factor 19
ENSG00000227739	-8.110	0.007	TUBB	tubulin beta class I
ENSG00000277101	-7.818	0.009	ARHGEF26	Rho guanine nucleotide exchange factor 26
ENSG00000099956	-8.405	0.016	SMARCB1	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily b, member 1
ENSG00000205572	-7.325	0.015	SERF1B	small EDRK-rich factor 1B
ENSG00000118849	-3.272	0.002	RARRES1	retinoic acid receptor responder 1
ENSG00000206395	-3.279	0.005	DDAH2	dimethylarginine dimethylaminohydrolase 2
ENSG00000120937	-2.833	0.000	NPPB	natriuretic peptide B

Table 5.9: Selected enriched 'Gene Ontology: Biological Process' terms for genes differentially expressed in CD4+ cells, HAECs, macrophages and VSMCs.

GO Term	adj pvalue
CD4+ cells	
cellular response to hypoxia	0.00002
cell-cell junction organization	0.00029
fatty acid elongation, saturated fatty acid	0.00467
cellular response to oxidative stress	0.00846
positive regulation of acute inflammatory response	0.00877
phosphatidylcholine biosynthetic process	0.01249
adipose tissue development	0.02172
negative regulation of blood vessel endothelial cell migration	0.02212
cellular response to transforming growth factor beta stimulus	0.03062
T cell chemotaxis	0.03065
Human aortic endothelial cells	
positive regulation of angiogenesis	0.00008
negative regulation of cholesterol storage	0.00131
negative regulation of macrophage derived foam cell differentiation	0.00157
regulation of lipid metabolic process	0.00174
cellular response to hypoxia	0.00187
positive regulation of cell proliferation	0.00254
positive regulation of vascular endothelial growth factor production	0.00270
cholesterol homeostasis	0.00441
negative regulation of smooth muscle cell proliferation	0.00459
positive regulation of transforming growth factor beta production	0.02319
Macrophages	
T cell receptor signaling pathway	0.00000
regulation of transcription from RNA polymerase II promoter in response to hypoxia	0.00000
inflammatory response	0.00000
cholesterol biosynthetic process	0.00000
cellular response to oxidative stress	0.00002
positive regulation of cell migration	0.00003
T cell costimulation	0.00022
positive regulation of vascular smooth muscle cell proliferation	0.00061
positive regulation of angiogenesis	0.00846
fatty acid beta-oxidation	0.00896
Vascular smooth muscle cells	
cholesterol biosynthetic process	0.00000
cellular response to oxidative stress	0.00003
oxidation-reduction process	0.00060
regulation of cholesterol biosynthetic process	0.00172
negative regulation of cell death	0.00265
sterol biosynthetic process	0.00356
regulation of blood vessel diameter	0.00356
pathway-restricted SMAD protein phosphorylation	0.01720
extracellular matrix organization	0.02512
positive regulation of activated T cell proliferation	0.03545

Transcriptomic Changes in Vascular and Immune Cells in Response to oxLDL

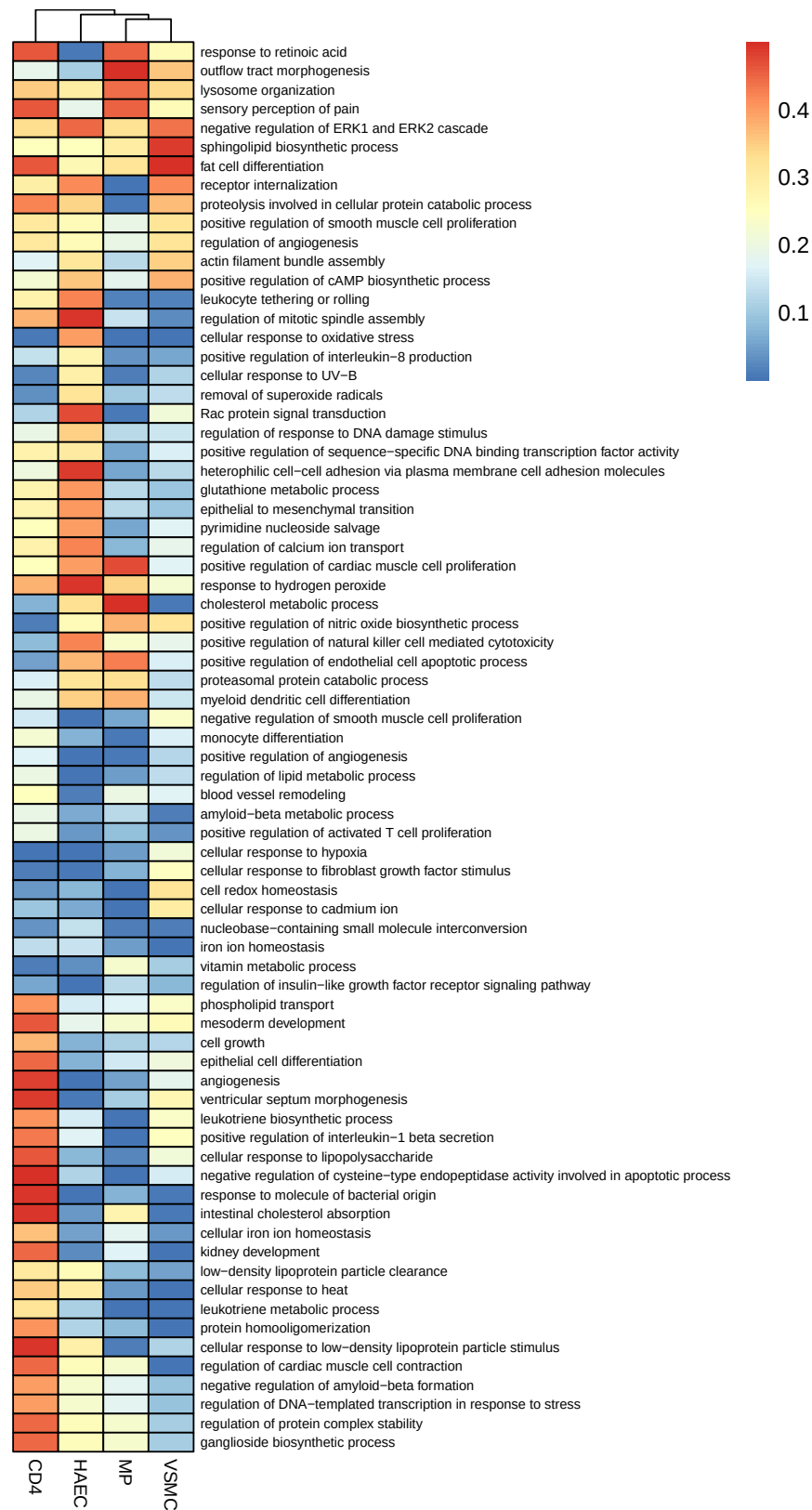


Figure 5.12: Heatmap of enriched 'Gene Ontology: Biological Process' terms for genes differentially expressed in CD4+ cells, HAECs, macrophages and VSMCs. This heatmap takes the top most significant terms in each cell type and compares the p-values across all the cell types.

Chapter 6

Epigenetic Changes in Vascular and Immune Cells in Response to oxLDL

Contents

6.1	Introduction	140
6.2	Chromatin accessibility: ATAC-seq	144
6.2.1	DNA isolation and assessment of quality	144
6.2.2	Quality control of ATAC-seq reads	145
6.2.3	Read alignment and quality control	148
6.2.4	Oxidised LDL induces chromatin remodelling at gene promoters	151
6.2.5	Oxidised LDL induces changes in chromatin accessibility at sites distal to promoters	154
6.3	Oxidised LDL-induced changes in enhancer activity: H3K27ac CHIP-seq analysis	157
6.3.1	DNA isolation and quality assessment	157
6.3.2	Oxidised LDL induced changes in chromatin accessibility are located at enhancers	159
6.4	Dynamic enhancer sites show strong enrichment for ETS family of transcription factors	164
6.5	Coronary artery disease associated variants overlap with dynamic accessible and enhancer elements	165
6.6	Discussion	167

6.1 Introduction

The development and maintenance of cell phenotypes depends on *cis*-regulatory elements. These are genomic sites which allow binding of transcription factors and consequently affect the rate of transcription of a target gene, often in a tissue-specific manner^{224,225}. *Cis*-regulatory elements are often highly evolutionarily conserved, and have also been connected to human disease. In fact, most of the variants associated with human diseases and phenotypes by genome-wide association studies (GWAS) do not lie in protein-coding genes, and lie in these regulatory regions²²⁶. A large

scale study of causal autoimmune disease variants identified that approximately 90% were located in non-protein coding regions of the genome and up to 60% of these locations had been identified as a *cis*-regulatory region in one cell type or another. A typical *cis*-regulatory element is in the order of 100 to 1,000 DNA base pairs²²⁵ and contains binding sites for multiple transcription factor proteins, thereby mediating combinatorial control of target gene expression²²⁷.

The two major classes of *cis*-regulatory elements in the genome are promoters and enhancers. They have different functional roles and are correlated with different characteristic epigenetic modifications (as discussed in Chapter One)²²⁸. Promoters lie proximal to a target gene and interact with distal regulatory elements to control gene expression. Enhancers are distal elements that activate gene expression by interacting with gene promoters. Enhancers are also characterized by sensitivity to the enzyme DNase I, and are often bound by the transcriptional-coactivator p300²²⁹. Enhancers typically exhibit cell-type specific behaviour, and the enhancer repertoires of different cell types are often very different²²⁸.

A pre-requisite for activity of a *cis*-regulatory element is nucleosome depletion and accessibility to transcription factors. A key method of regulating chromatin accessibility is the modification of histones in the surrounding nucleosomes¹⁰³. Histone modifications exert their activity by one of two mechanisms. They may cause structural changes in local chromatin as a result of changes in electrostatic charge (e.g. acetylation and phosphorylation reduce the positive charge of a histone), or by direct steric hindrance (e.g. ubiquitylation results in the addition of a very large molecule to a histone). Alternatively, histone modifications may act as recognition sites for interaction with chromatin-binding proteins thus recruiting them to that region. For example, the ETS1 transcription factor can be recruited to the *VEGFR* promoter by acetylation of

histone H3, a key mechanism for the action of TGF β in renal mesangial cells²³⁰.

Cis-regulatory elements may, in part, be recognised by combinations of these histone modifications^{103,104} (Table 1.4). The presence of high levels of tri-methylation of histone 3 at lysine 4 (H3K4me3), acetylation of histone 3 at lysine 27 (H3K27ac), and the absence of mono-methylation of histone 3 at lysine 4 (H3K4me1) are characteristic of promoters, whilst the presence of high levels of H3K4me1 and H3K27ac are characteristic of active enhancer regions. The result of these modifications is a change in chromatin structure so that the local chromatin becomes more open and accessible to transcription factor binding.

Several HTS techniques have been developed to allow assessment of histone modifications and chromatin accessibility on a genome-wide scale, and the principles underlying these have been discussed in Chapter Four and detailed methods given in Chapter Two. For histone modifications, chromatin immunoprecipitation with an antibody directed against the modification of interest is performed before high throughput sequencing. To measure chromatin accessibility, three main sequencing assays have been developed (Table 6.1): DNase-seq²³¹, FAIRE-seq²³², and ATAC-seq¹³⁰. In FAIRE-seq chromatin is crosslinked with formaldehyde to capture *in vivo* protein-DNA interactions, and subsequently the chromatin is sheared and sonicated. Following phenol-chloroform extraction, accessible regions of chromatin are released to the aqueous phase due to cross-linking of histones to the non-accessible DNA. DNase and ATAC-seq rely on enzymatic digestion or insertion of the Tn5 transposon, respectively, to detect open regions of chromatin. DNA fragments are then isolated, and the start and end sites of the fragments define the accessible regions. Both DNase-seq and ATAC-seq have a much higher signal-to-noise ratio than FAIRE-seq, and ATAC-seq in particular is becoming increasingly popular due to the low number of cells required and its ease

Table 6.1: Genome-wide high throughput chromatin accessibility assays.

Technique	Cell Numbers	Method	Genomic Target	Considerations
DNase-seq	10-50 million	DNase I cuts within unprotected DNA	Open chromatin	Requires many cells Time consuming and complicated sample preparation Laborious enzyme titrations
FAIRE-seq	1-10 million	Based on the phenol-chloroform separation of nucleosome-bound and free sonicated areas of a genome, into the interphase and aqueous phase respectively	Open chromatin	Low signal-to-noise ratio Results very dependent on fixation efficiency
ATAC-seq	500-50,000	Unfixed nuclei are tagged in vitro with sequencing adapters by purified Tn5 transposase. Adapters are integrated into regions of accessible chromatin	Open chromatin TF and nucleosome occupancy	High contamination with mitochondrial DNA Immature data analysis tools

compared to DNase-seq.

We have shown in previous chapters that epigenetic mechanisms play a role in the pathogenesis of atherosclerosis. We have also shown that oxLDL, a key risk factor for atherosclerosis, induces considerable phenotypic and genotypic changes on cells. We thus hypothesise that oxLDL changes are, in part, epigenetically mediated. In this chapter we have employed ATAC-seq to generate epigenomic profiles in primary CD4⁺ cells, HAECs, macrophages, and VSMCs that were untreated or exposed for 48 hours to 50 μ g/ml of human plasma-derived LDL oxidised by 18 hours of treatment with 25 μ M Cu₂SO₄. We integrate this data with ChIP-seq profiles for the active enhancer modification H3K27ac to identify *cis*-regulatory elements and with RNA-seq data to determine the induced changes in gene expression. We also incorporate findings from publicly available epigenomic and GWAS data. Taken together, we aim to untangle some of the mechanisms underlying the regulation of atherosclerotic vascular disease by oxLDL.

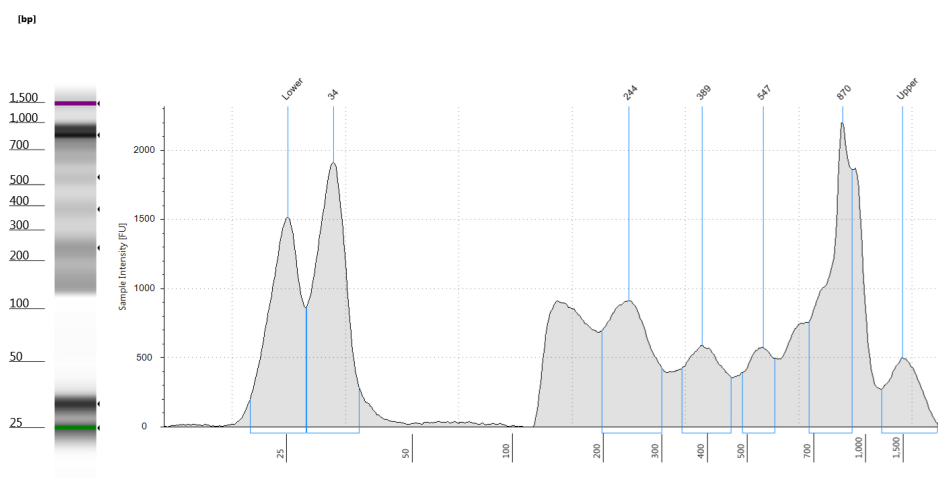


Figure 6.1: Representative TapeStation run of an ATAC-seq library showing fragments of appropriate size with a characteristic nuclear phasing pattern. On the right is an electropherogram and to the left a reconstructed gel image of the sample. The peak at 34 bp comes from primers.

6.2 Chromatin accessibility: ATAC-seq

To understand the transcriptional regulation of the oxLDL-mediated changes on gene expression we began by first identifying genome-wide changes in open chromatin in cells before and after exposure to oxLDL using ATAC-seq.

6.2.1 DNA isolation and assessment of quality

Genomic DNA for ATAC-seq was prepared as described in the methods section. Prior to submission for sequencing, the samples were analysed using the Agilent DNA High Sensitivity kit with the Agilent 2200 TapeStation, looking for the characteristic nuclear phasing pattern (Figure 6.1). Only samples that exhibited this profile were sequenced using the Illumina HiSeq 2500 or 4000 platforms. Following sequencing, reads were processed following the *chompR* pipeline.



Figure 6.2: Summary of PASS/WARN/FAIL results of FastQC modules for raw untrimmed reads.

6.2.2 Quality control of ATAC-seq reads

Quality control of the raw reads generated by sequencing was performed using *FastQC*. An overall summary of the quality control statistics is shown in Table 6.2 and a plot summarising the outcome of each QC module in *FastQC* is shown in Figure 6.2. The average number of read pairs across all samples was 50 506 700, but there was considerable intersample variability. Overall, the quality of reads was very high with all samples having a mean Phred score > 30 (Table 6.2).

The *FastQC* module summary shows widespread warnings or failures in the ‘Per base sequence content’, ‘Per sequence GC content’ and ‘Sequence duplication levels’. A significant cause for these is adapter contamination, the module for which also generates warnings across the board.

To correct for this reads were subjected to adapter trimming using *fastp* within the

Table 6.2: Basic summary statistics of raw ATAC-seq reads.

Cell	Treatment	Replicate	Reads	Mate 1		Mate 2	
				%GC	Phred	%GC	Phred
CD4	Buffer	1	41691265	46	37.94	46	37.74
CD4	Buffer	2	42085275	46	38.01	46	37.62
CD4	Buffer	3	79496028	46	39.16	46	38.10
CD4	oxLDL	1	43670091	47	37.94	47	37.61
CD4	oxLDL	2	132814355	48	37.92	48	37.57
CD4	oxLDL	3	78457617	47	39.16	47	37.94
HAEC	Buffer	1	17315729	46	37.91	46	37.21
HAEC	Buffer	2	142884244	47	38.28	46	37.79
HAEC	Buffer	3	13533259	47	39.03	47	37.75
HAEC	oxLDL	1	29151779	47	37.88	47	37.47
HAEC	oxLDL	2	38169034	47	38.33	47	37.74
HAEC	oxLDL	3	9912968	47	38.97	47	37.78
HAEC	Untreated	1	27076716	46	37.92	46	37.43
MP	Buffer	1	24287481	47	38.87	47	37.74
MP	Buffer	2	39341967	49	38.94	48	37.91
MP	oxLDL	1	69105448	48	38.92	48	38.33
MP	oxLDL	2	129559938	47	38.93	47	38.11
VSMC	Buffer	1	23441012	48	38.34	48	37.69
VSMC	Buffer	2	23562040	48	38.36	48	37.66
VSMC	oxLDL	1	27445011	47	38.33	47	37.76
VSMC	oxLDL	2	27639446	47	38.35	47	37.73

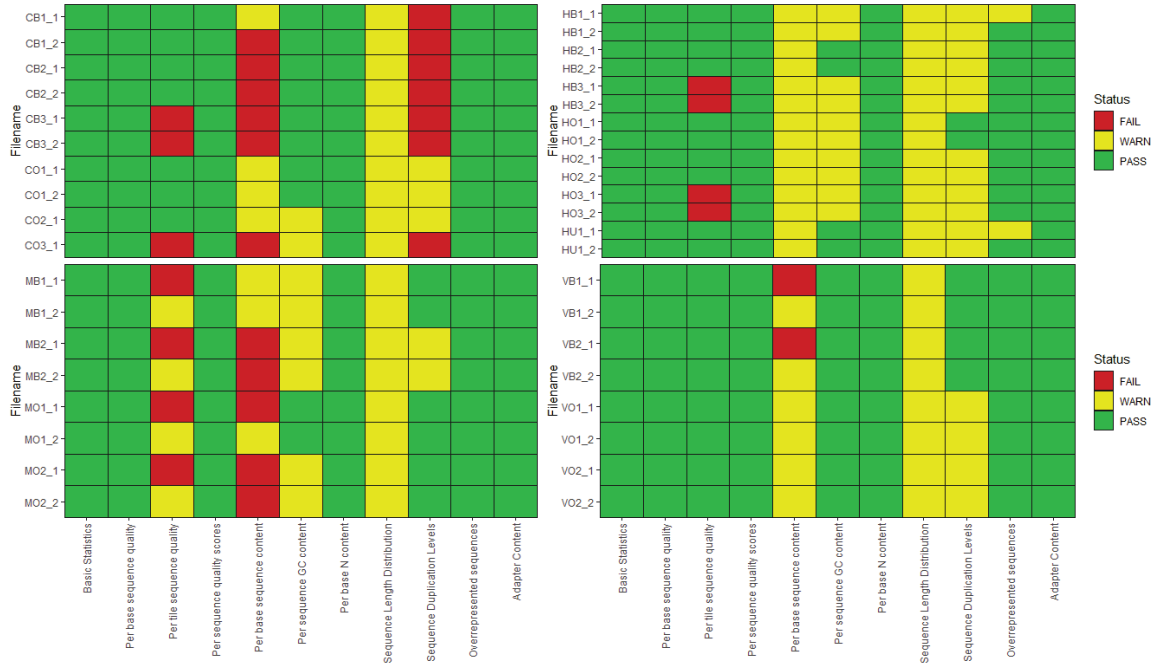


Figure 6.3: Summary of PASS/WARN/FAIL results of FastQC modules for trimmed reads.

framework of the *chompR* pipeline. The repeat module summary plot is shown in Figure 6.3 and shows significant improvement. However, not all of the warnings/failures have been resolved. This is because ATAC-seq generates large numbers of mitochondrial reads which are not filtered out by *FastQC*. We remove these reads in the post-alignment phase. In addition, there is now a warning for the ‘sequence length distribution’; this arises from the fact that read trimming has resulted in reads of different lengths, and is not of concern.

As with the RNA-seq data, flags are raised in the ‘Per tile sequence quality’ module in some samples, but as before, a look at the actual data for this module confirms a minor and restricted problem, that would not be expected to have any impact on downstream analysis and can be safely ignored.

6.2.3 Read alignment and quality control

Reads were aligned to the human genome (GRCh38/hg38) using *HISAT2* and converted from SAM format to BAM using *SAMtools*. Duplicate reads were marked using the *Sambamba* tool and the BAM files then filtered to remove reads marked as duplicate or aligning to the mitochondrial genome. Table 6.3 summarises the read alignment results and important quality control metrics.

Table 6.3: General and quality metrics of ATAC-seq alignment.

Library	Total Reads	Proper Reads	Mitochondrial Reads	Total Duplicated	Autosomal Duplicated	Mitochondrial Duplicated	FMNR	TSSE	Mean	MAPQ
CB1.bam	94739895	79896912	40523340	43715612	4441993	33633589	1.60	5.59	45.63	
CB2.bam	95173708	84079105	49235015	53372533	6825780	41927961	1.17	5.03	46.15	
CB3.bam	178195536	132843912	89887537	117573956	10825365	80374696	0.68	7.02	45.60	
CO1.bam	99769314	82620564	30250317	33890666	4304235	24614172	1.32	5.95	45.40	
CO2.bam	303612909	255846927	70012500	101996793	27606299	60907043	0.99	6.36	45.96	
CO3.bam	176172133	133672431	72974186	97649202	12218378	64964998	0.65	7.28	45.76	
HB1.bam	38962568	30774394	7667231	9171199	2505388	5155583	1.97	6.24	45.68	
HB2.bam	316429640	248594971	64175629	113348059	41987364	54953719	2.06	6.50	44.61	
HB3.bam	29829836	21293001	6781958	9761448	2459587	4876385	1.68	8.28	46.31	
HO1.bam	66295557	51276999	8616378	11740745	3851492	5869570	1.64	5.45	45.36	
HO2.bam	83837715	63785037	9472248	21280806	11175650	6862223	2.42	6.39	44.07	
HO3.bam	21934128	15694483	5286700	6510975	1227744	3646490	1.56	8.87	46.40	
HU1.bam	60835625	48592017	13217845	15806379	3797272	9546851	1.51	5.65	45.89	
MB1.bam	52833433	38731088	2736138	6490528	3518274	1425968	1.39	7.25	47.58	
MB2.bam	86930005	63785998	15411086	22296994	5149827	12156080	1.85	7.26	47.12	
MO1.bam	158816988	106164005	3566934	15099985	7709241	2309291	0.89	4.13	44.33	
MO2.bam	289675649	209130381	19555528	53099355	22636155	16213306	1.94	3.14	46.22	
VB1.bam	51629347	42388216	7165744	8967268	2706925	5129695	2.47	6.75	46.22	
VB2.bam	51854081	42561500	7192255	9022415	2736246	5149868	2.47	6.73	46.24	
VO1.bam	60386377	48213952	11154451	15285019	4908865	8405213	2.15	6.45	44.77	
VO2.bam	60753679	48497252	11199811	15394322	4968802	8442390	2.15	6.41	44.80	

We obtained an average of 113 million Tn5-integrated mapped paired reads per sample. Of these, an average of 26 million reads mapped to the mitochondrial genome, although this proportion varied considerably between samples. Overall, the average number of properly mapped paired reads was 88 million. At first glance, the overall duplication rate appeared to be very high (as picked up by *FastQC*), but a closer analysis shows that 58% of these duplicates were actually in the mitochondrial portion of the fragments, although this proportion did show a sample-to-sample variation. The overall quality of the read alignment was excellent with mean mapping quality (MAPQ) scores greater than 44 throughout. All samples showed the characteristic nucleosomal phasing expected of a successful ATAC-seq experiment (Figure 6.4A). The free-to-mononucleosomal fragment ratio (FMNR, the ratio of fragments 50-100bp long to 150-200bp long) was generally around 1. Taken together with the fragment length distribution plots, this value gives an indication of whether there has been over- or under-transposition. A significant deviation from 1 does not necessarily indicate a problem, and may be a reflection of underlying biology. For example, in this data set, the drop in FMNR in the CD4⁺ cells following exposure to oxLDL may be an indication of transcriptional repression. Across all samples, the signal-to-noise ratio was excellent as indicated by the transcriptional start site enrichment (TSSE) scores (Figure 6.4B); this score is used by ENCODE to assess ATAC-seq quality. To determine this score, each transcriptional start site (TSS) in the genome was placed at the centre of a 2000bp window and the aggregate coverage (i.e. across all TSS) at each base pair in that window calculated. The distribution was normalised by taking the average aggregate read depth in the 100bp at the two extremes of the 2000bp window and using that value to determine a fold-change at each base pair in the window. As TSS are expected to be open regions of chromatin, in a successful ATAC-seq experiment there should be an increase in fold change up to the centre of

Table 6.4: Genomic distribution of open chromatin sites. The first row shows the total number of sites for each sample group and subsequent rows express the distribution as a proportion.

	CD4+		HAECs		Macrophages		VSMCs	
	Untreated	oxLDL	Untreated	oxLDL	Untreated	oxLDL	Untreated	oxLDL
All	88067	188018	164965	122352	135827	176481	137720	125245
3' UTR	0	0	0	0	0	0	0	0
5' UTR	0	0	0	0	0	0	0	0
Downstream (1-5kb)	15	19	17	16	17	21	17	16
Exon	2	3	2	2	2	4	2	2
Intergenic	23	25	26	25	25	25	24	25
Intron	33	31	35	34	33	27	33	34
Promoter	24	17	17	20	19	19	20	20

the window. The value at the peak is taken as the TSSE score.

6.2.4 Oxidised LDL induces chromatin remodelling at gene promoters

We identified open chromatin peaks using *MACS2*, and merged those from samples belonging to the same cell type and exposed to the same treatment condition. Overall, across all cell types and conditions, we identified 142334 open chromatin peaks (Table 6.4). The fraction of reads in peaks (FRiP) was determined using feature counts and met ENCODE standards for ATAC-seq, with FRiP scores > 0.3 . Peaks were annotated using the *chompR* pipeline. Sites at promoters, defined as within 1kb upstream of the TSS of all UCSC known genes, accounted for 19.8% of open chromatin sites overall, and there was no significant change in proportions of peaks between untreated and treated cells.

Although there was no change in the proportion of open chromatin peaks at promoters, there was a change in the read count that varied concordantly with oxLDL-induced changes in gene expression. Across all cell types, genes that were upregulated by

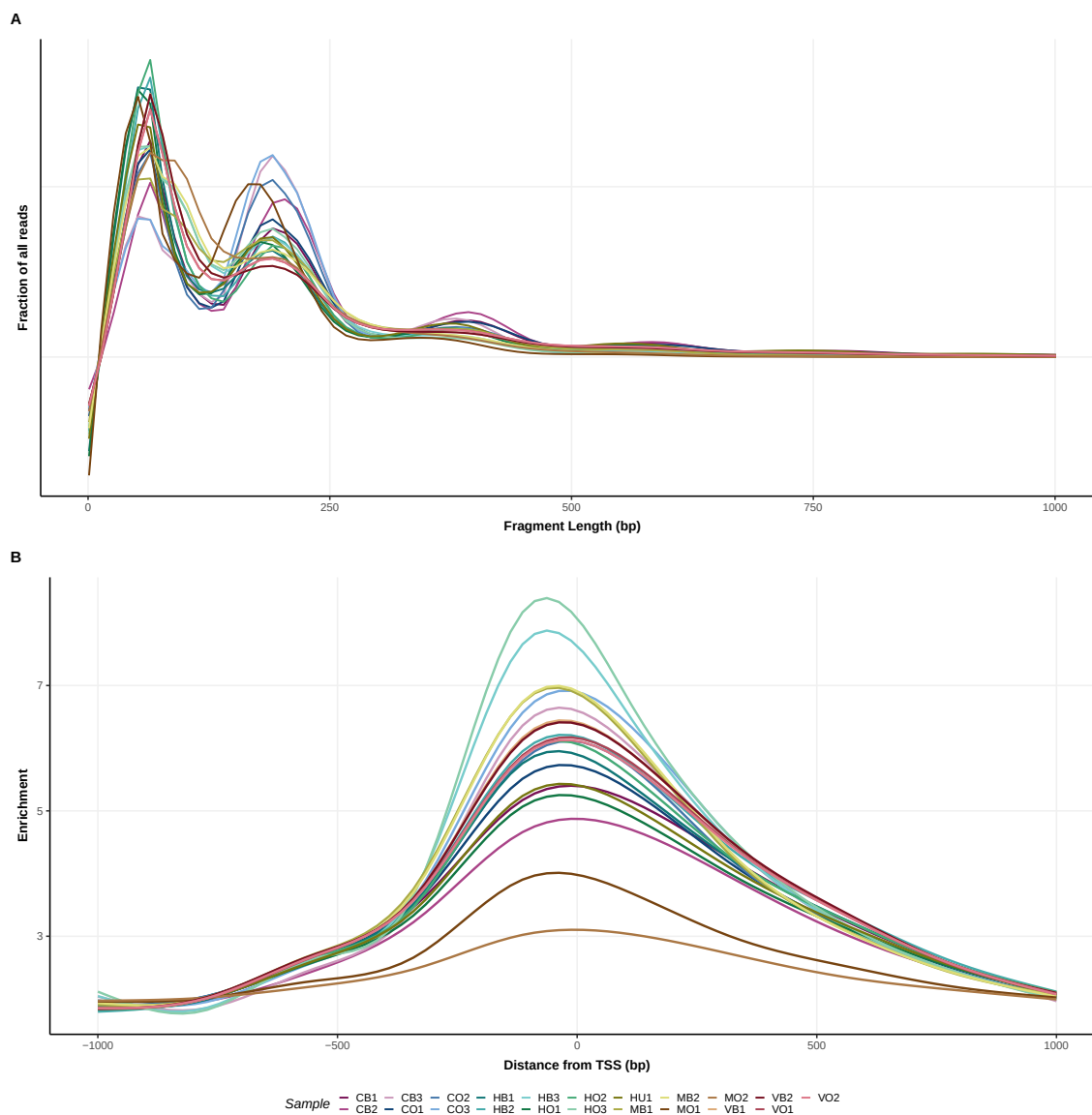


Figure 6.4: QC plots for ATAC-seq alignment. (A) Fragment length distribution plot with characteristic nucleosomal phasing of successful ATAC-seq experiments. (B) Transcriptional start site enrichment plot showing signal-to-noise ratio, calculated as distribution of ATAC-seq reads centred on TSS in a 2,000-bp window, normalized to reads mapped to the 100bp flanking ends of the window.

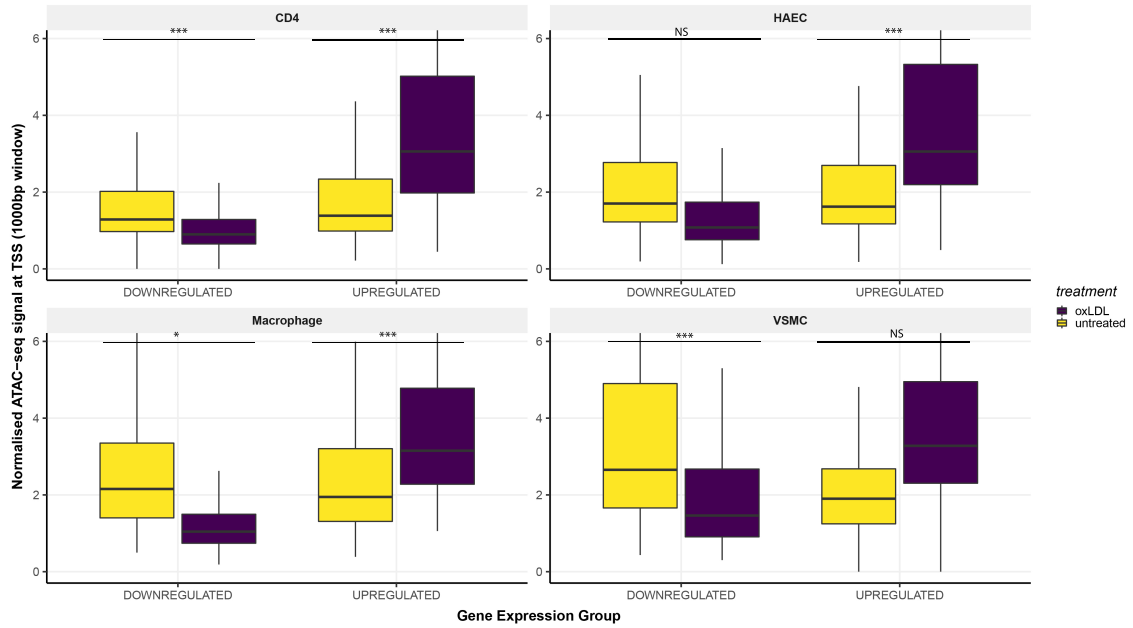


Figure 6.5: Changes in chromatin structure at promoters in response to oxLDL. Normalised ATAC-seq signal in 2000bp windows centred on the TSS for down- and up-regulated genes. Changes in chromatin structure were associated with concordant changes in gene expression (* < 0.05, ** < 0.01, *** < 0.001, NS = non-significant)

oxLDL had greater signal at their promoters in oxLDL-treated cells and genes that were down-regulated had lower signals in oxLDL-treated cells (Figure 6.5). This was highly significant when the results from all cell types was aggregated ($p = 1.81 \times 10^{-21}$ for upregulated genes, and $p = 4.16 \times 10^{-20}$ for down-regulated genes). When assessed for individual cell types, however, although the trend was in the expected direction, it did not reach significance for genes downregulated by HAECs or upregulated in VSMCs.

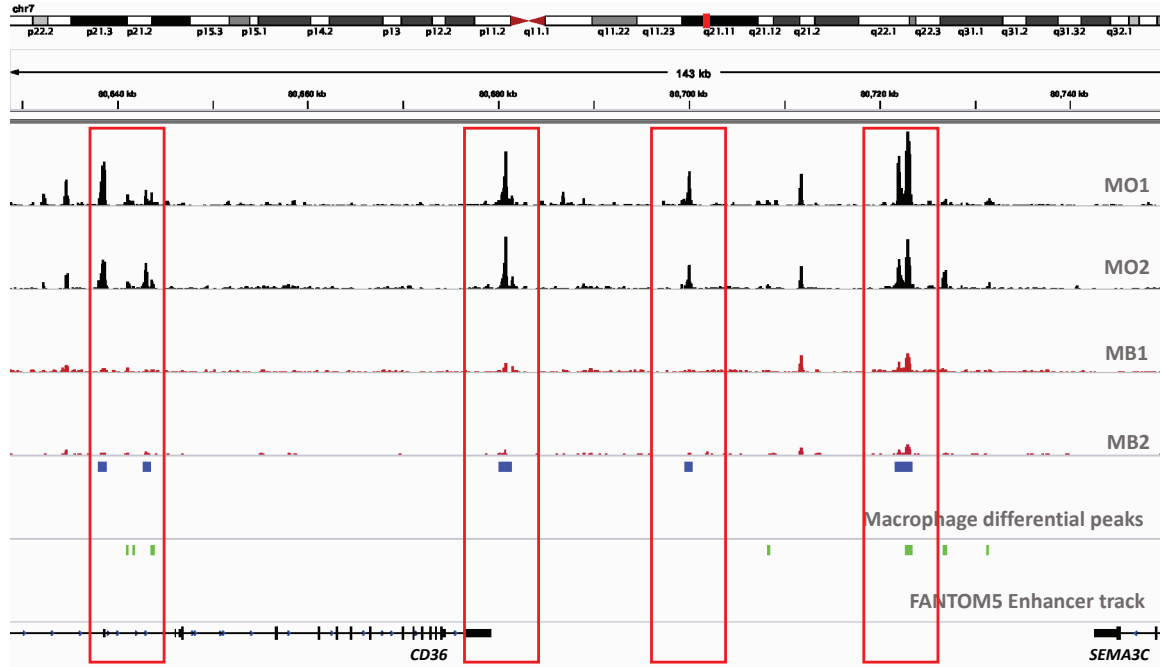


Figure 6.6: Representative chromatin accessibility changes at the CD36 gene locus for macrophages. The dynamic accessible sites here comprise five sites, each with greater accessibility in the oxLDL-treated cells (MO1 and MO2, in black) compared to the untreated cells (MB1 and MB2, in red). Also shown is the track for differential peaks identified in macrophages (in blue) and underneath this the track for enhancers identified as part of the FANTOM5 project (in green; co-ordinates have been lifted over from hg19 to hg38). Dynamic sites are boxed in red; one site directly overlaps with an enhancer site and another set of enhancers are in close proximity to another two DAS.

6.2.5 Oxidised LDL induces changes in chromatin accessibility at sites distal to promoters

All genomic sites with significant changes in chromatin accessibility in response to oxLDL were identified; we termed these sites dynamic accessible sites (DAS) (Figure 6.6).

Oxidised LDL changed the accessibility of chromatin at 6.8, 3.9, 9.4, and 7.1% of all open chromatin sites (Table 6.5) in CD4⁺ cells, HAECs, macrophages, and VSMCs, respectively. Of these, the majority were non-promoter sites, located either

Table 6.5: Changes in chromatin accessibility following oxLDL treatment. The table shows the number and proportion of sites that changed their accessibility following treatment with oxLDL and the genomic locations of these sites as a proportion of all DAS. DAS = dynamic accessible site.

	CD4	HAEC	Macrophage	VSMC
Total Accessible Sites	276085	287317	312308	262965
Dynamic Accessible Sites	18904	11330	29250	18612
Proportion DAS (%)	7	4	9	7
Genomic Location of DAS (%)				
Promoter	9	11	12	13
Downstream (1-5kb)	8	10	8	8
5' UTR	0	1	0	1
3' UTR	0	0	0	0
Exon	1	2	1	1
Intron	24	22	22	21
Intergenic	57	55	57	56

in intergenic or intronic regions.

To validate the distal DAS and show that they represent true *cis*-regulatory elements we determined the overlap of these sites with chromatin states from ChromHMM for 9 different cell lines²³³ and FANTOM5 permissive enhancers determined from bi-directional CAGE transcription²³⁴. The former were downloaded via the UCSC Table Browser and the latter from the FANTOM5 resource. The annotations were lifted over from hg19 to hg38 using the hg19 to hg38 chain file available from UCSC. To evaluate the significance of the intersection between known regulatory sites and the DAS we compared it to the overlap of control sites chosen by circularising each chromosome and then shifting the regions along to maintain their order and distance. We also applied a mask to avoid blacklisted and centromeric regions. The overlap analysis was significant at all regulatory sites with Fisher's exact test p-values of < 0.05. The greatest number of DAS overlapped with enhancers.

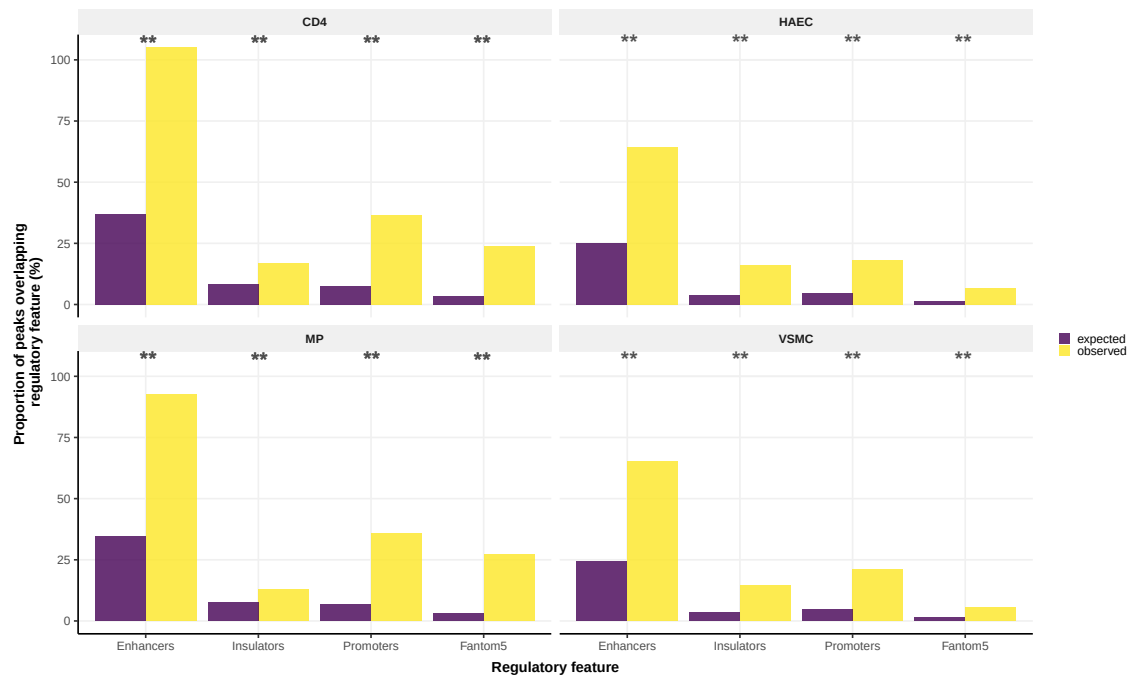


Figure 6.7: Dynamic accessibility sites are enriched for non-coding regulatory elements. Dynamic accessibility sites were overlapped with ChromHMM states for HESC, HMEC, HUVEK, NHEK, GM12878, HEPG2, HSMM, K562, and NHLF, and with enhancer sites from the FANTOM5 project. Overlap was compared to the expected overlap of 500 control regions. The overlap is shown as a percentage of the total number of regions. The value is greater than 100 per cent as some regions overlap multiple sites.

6.3 Oxidised LDL-induced changes in enhancer activity: H3K27ac CHIP-seq analysis

The greatest proportion of the the DAS were distal to promoters and located in intergenic or intronic regions that overlapped with sites previously identified as regulatory elements. We, therefore, hypothesised that many of these sites must be enhancers. Given that accessible chromatin alone may not identify an enhancer, we also performed ChIP-seq for the H3K27ac modification on histones, as this has been shown to mark active enhancers.

6.3.1 DNA isolation and quality assessment

Genomic DNA for ChIP-seq was prepared as described in the methods section. Prior to submission for sequencing, the samples were analysed using the Agilent DNA High Sensitivity kit with the Agilent 2200 TapeStation and were found to be of high quality. The raw reads underwent the same assessment using *FastQC* as described earlier for the ATAC-seq data (data not shown). Overall the read quality was high with Phred scores > 30 throughout, and the only *FastQC* modules to raise warnings were those for sequence duplication. The reads were aligned using HISAT2 and Table 6.6 summarises important alignment statistics and quality metrics.

Table 6.6: General and quality metrics of ChIP-seq alignment. Samples containing 'I' in their name are input control samples.

Library	Total Reads	Proper Reads	Mitochondrial Reads	Total Duplicated	Autosomal Duplicated	Mitochondrial Duplicated	Mean MAPQ	PBC	NSC	RSC
CB1.bam	80516195	71898822	0	4339091	3988767	0	51.22	0.87	1.30	1.84
CB2.bam	90315368	80121871	0	13991651	12793132	0	53.99	0.75	1.18	1.87
CI1.bam	95760834	82823548	0	1577790	1010080	0	47.44	0.75	1.01	0.35
CO1.bam	64306817	58278927	0	4233329	3928462	0	52.16	0.90	1.32	1.34
CO2.bam	95815023	80221270	0	55558129	48654923	0	53.51	0.61	1.20	1.67
HB1.bam	44915758	37823196	670	1404394	1231998	16	52.49	0.78	1.36	1.28
HB2.bam	44596840	37535540	630	1383037	1212314	24	52.48	0.80	1.09	1.35
HI1.bam	63683756	53021332	380080	969944	646326	22704	52.42	0.71	1.05	0.34
HI2.bam	62775648	52237942	374908	946354	630036	22228	52.41	0.79	1.10	0.35
HO1.bam	41461434	34985742	400	1024120	882358	2	52.71	0.76	1.18	1.05
HO2.bam	41032148	34608794	384	1004843	865296	0	52.70	0.87	1.20	1.00
HU1.bam	32030580	27055422	248	446936	369840	4	52.75	0.92	1.29	1.24
HU2.bam	31517700	26610282	212	435014	359618	0	52.74	0.81	1.35	1.53
MB1.bam	78182708	57858426	2236	4183739	3373848	94	45.94	0.75	1.34	1.31
MB2.bam	79322874	61349912	1404	6390462	5535508	100	48.20	0.75	1.27	1.82
MB3.bam	80677372	71062048	788	13441870	12409996	74	54.38	0.90	1.21	1.57
MB4.bam	137318120	117807458	902	60575058	55615166	224	54.44	0.86	1.25	1.29
MI1.bam	78751320	58296550	65590	2722156	1912774	2400	47.27	0.80	1.03	0.63
MI3.bam	62186542	52856290	62442	14970017	13415968	16742	53.40	0.88	1.06	0.05
MI4.bam	80954040	68856200	144624	17931306	15826624	36486	53.61	0.83	1.06	0.48
MO1.bam	79030390	60430188	1544	3082793	2363918	44	47.54	0.77	1.37	1.32
MO2.bam	69568364	48404236	1968	4458139	3649126	130	43.63	0.85	1.24	1.84
MO3.bam	74462346	65636804	778	11291975	10402088	102	54.46	0.73	1.33	1.04
MO4.bam	144969320	118362170	1456	96734330	86264592	652	53.08	0.82	1.15	1.61
VB1.bam	148221360	126299848	280	13655355	12196186	12	53.19	0.75	1.37	1.02
VB2.bam	109691068	93282760	232	6776226	5960166	4	53.04	0.76	1.32	1.67
VI1.bam	117222442	97077480	237886	4322443	3324612	13032	52.45	0.79	1.02	0.63
VO1.bam	124187956	106169906	590	29080803	26849316	40	53.23	0.82	1.19	0.97
VO2.bam	148192590	127390604	426	16054640	14319828	12	53.48	0.88	1.40	1.48

We obtained an average of 83 million mapped paired reads per sample. Overall, the average number of properly mapped paired reads was 69 million. The duplication rate was low, aside from one sample (CO2), but this was also the sample with the highest number of reads and so with the greatest tolerance for deduplication. The quality of the read alignment was excellent with mean mapping quality (MAPQ) scores greater than 40 throughout. Almost all samples met ENCODE quality standards; PBC1 was generally mild (although sample CO2 showed more moderate bottlenecking), and NSC and RSC were as expected, showing good cross-correlation for the ChIP samples, and absent cross-correlation for input samples. The mean mapping quality was greater than 40 throughout.

6.3.2 Oxidised LDL induced changes in chromatin accessibility are located at enhancers

We identified enrichment of H3K27ac using *MACS2*, and generated consensus peaks in the same way as described earlier. Overall, across all cell types and conditions, we identified an average of 72983 H3K27ac peaks (Table 6.7) per sample. Peaks were annotated using the *chompR* pipeline. Sites at promoters, defined, as before, as within 1kb upstream of the TSS of all UCSC known genes, accounted for 20.7% of open chromatin sites overall. However, interestingly, oxLDL induced a significant change in localisation of H3K27ac signal in CD4+ cells from distal to proximal locations - the proportion of H3K27ac peaks at promoter rose from 37 to 50% and dropped at distal regions from 43 to 19%. In the other cell types, the remainder of peaks were located predominantly at intergenic (25.4%) and intronic regions (32.6%).

We determined co-localisation of H3K27ac modification and open chromatin by identifying those genomic regions in which peaks were detected in both (Table 6.8). There

Table 6.7: Genomic distribution of H3K27ac sites. The first row shows the total number of sites for each sample group.

	Genome	CD4+		HAECs		Macrophages		VSMCs	
		Untreated	oxLDL	Untreated	oxLDL	Untreated	oxLDL	Untreated	oxLDL
All		39703	20209	48680	50609	62669	60616	174394	126986
3' UTR	1	0	0	0	0	0	0	0	0
5' UTR	0	0	0	0	0	0	0	0	0
Downstream (1-5kb)	10	15	9	17	17	18	18	17	17
Exon	3	2	1	3	3	3	3	3	3
Intergenic	46	13	12	29	29	29	29	30	30
Intron	50	30	17	36	35	32	32	37	37
Promoter	3	37	58	11	11	13	13	9	9

Table 6.8: Co-localisation of open chromatin and H3K27ac marked sites.

	CD4+	HAEC	Macrophage	VSMC
Accessible chromatin sites	276085	287317	312308	262965
H3K27ac-marked sites	59912	99289	123285	301380
Dynamic accessible sites (DAS)	18904	11330	29250	18612
Dynamic H3K27ac	674	569	711	767
Accessible + H3K27ac	57505	18914	36166	42377
DAS + H3K27ac	6846	1241	6265	4290
DAS + dynamic H3K27ac	0	15	3	56

was significant co-localisation between H3K27ac-enriched regions and open chromatin regions. For each cell type, the overlap between these regions showed Fisher's exact test p-values of <0.0005 .

Although we detected large numbers of enhancer sites, very few of these had a significant change in H3K27ac signal (less than a 1000 for each cell type) outside of promoters. However, a significant proportion of DAS did have static H3K27ac signal (36.2, 11, 21.4, and 23 in CD4+ cells, HAECs, macrophages, and VSMCs respectively). We considered these sites, which had both change in chromatin accessibility and were marked by H3K27ac as being true enhancer sites that were most likely to be the mediators of oxLDL-induced responses; we termed these sites dynamic enhancer sites (DES) (Figure 6.8).

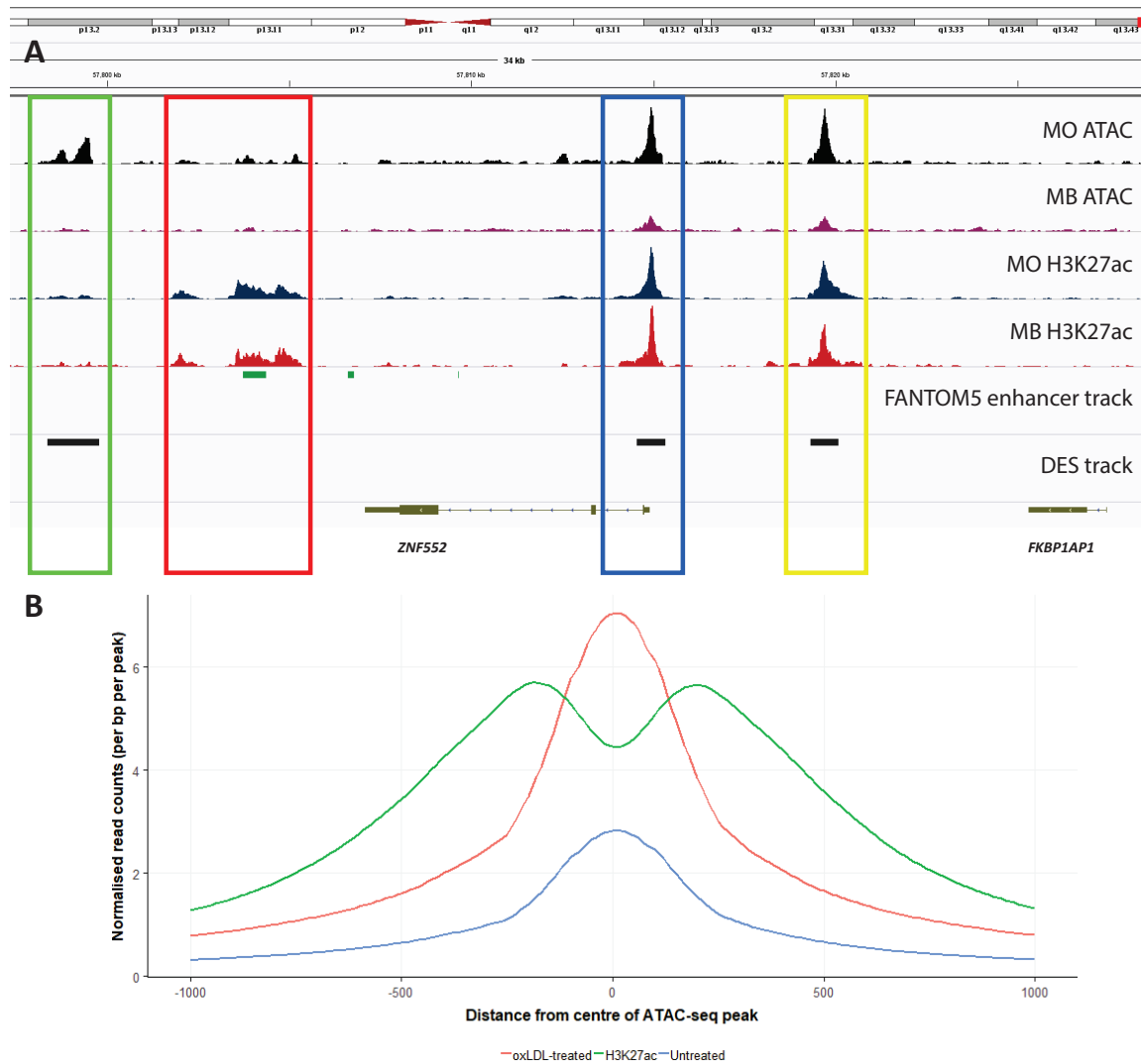


Figure 6.8: Characterisation of dynamic enhancer sites (DES). A) Chromatin changes at the ZNF552 region. Upstream of the promoter (in the yellow box) is a dynamic enhancer site: it is located distal to the promoter by more than 1 Kb, it has dynamic chromatin accessibility and is also marked by H3K27ac. The dynamic accessible region downstream of the gene (in the green box) is not marked by H3K27ac, and as such is not classified as a DES. There is a further region which is marked only by H3K27ac and does not show chromatin accessibility (in the red box); this site was identified as an enhancer site in the FANTOM5 project. The promoter of the ZNF552 gene is shown in the blue box. B) A multi-profile plot showing the ATAC-seq and H3K27ac signal at sites in CD4⁺ cells identified as having increased chromatin accessibility following oxLDL exposure. The profile encompasses a 2000bp window centred at the ATAC-seq peak summit. The red line shows the ATAC-seq signal for cells treated with oxLDL and the blue for untreated cells. The green profile shows the biphasic H3K27ac signal with the peaks at the histones flanking the trough at the nucleosome-free region. MO = oxLDL-treated macrophages; MB = untreated macrophages.

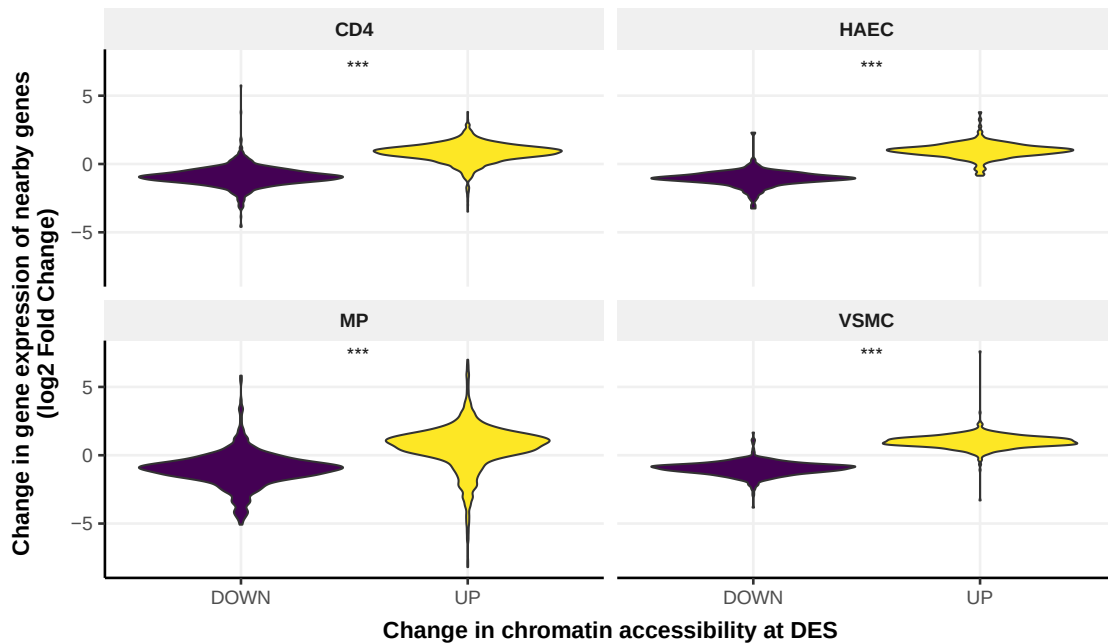


Figure 6.9: OxLDL-induced changes at dynamic enhancer sites is associated with a concordant fold-change in nearby genes (***) ($p < 0.05$).

These DES were concordantly associated with changes in expression of nearby genes (Figure 6.9). The difference in mean log fold change between genes associated with DES that became more open and those that became more closed in CD4+ cells, HAECs, macrophages, and VSMCs was 1.7, 2.1, 1.6, and 2, respectively, with t-test values $< 2.2 \times 10^{-16}$ for all. We also observed enrichment of these genes for pathways relevant to immune function, cell migration and ECM interaction, and atherosclerosis. As with the RNA-seq results, pathways related to $TGF\beta$ and HIF-1 α signalling were also enriched. Interestingly, there was also strong enrichment for pathways related to thromboxane signalling and platelet function and dysfunction (??).

Table 6.9: 'MSigDb' terms for dynamic enhancer sites in the different cell types show enrichment for pathways related to lipid handling, TGF-beta and HIF1-alpha signalling.

Term	adj pvalue
CD4+ cells	
Genes involved in Immune System	1.695e-28
HIF-1-alpha transcription factor network	1.218e-06
Cytokines can induce activation of matrix metalloproteinases, which degrade extracellular matrix.	1.315e-06
Genes involved in TGF-beta receptor signaling activates SMADs	8.463e-06
Thromboxane A2 receptor signaling	5.454e-05
Human aortic endothelial cells	
Genes involved in Thromboxane signalling through TP receptor	1.343e-10
LPA receptor mediated events	1.632e-08
Genes involved in Downregulation of SMAD2/3:SMAD4 transcriptional activity	2.671e-08
Genes involved in Phospholipid metabolism	2.972e-08
Genes involved in Fatty acid, triacylglycerol, and ketone body metabolism	3.938e-08
Macrophages	
Genes involved in Immune System	2.882e-12
HIF-1-alpha transcription factor network	9.447e-08
Thromboxane A2 receptor signaling	2.951e-07
Cytokines can induce activation of matrix metalloproteinases, which degrade extracellular matrix.	3.768e-05
Genes involved in Signaling by TGF-beta Receptor Complex	4.536e-05
Vascular smooth muscle cells	
Genes involved in Platelet activation, signaling and aggregation	3.707e-13
HIF-1-alpha transcription factor network	3.035e-10
PPAR signaling pathway	3.277e-08
Hypoxia-Inducible Factor in the Cardiovascular System	6.521e-08
Genes involved in Signaling by TGF-beta Receptor Complex	0.0002112

6.4 Dynamic enhancer sites show strong enrichment for ETS family of transcription factors

We used the DES to identify transcription factors that might be regulating the responses of oxLDL. We performed motif enrichment analysis in each cell type using the genomic sequences of the DES. There was significant enrichment for AP1-related factors, in particular, for members of the Fos and Jun families. The AP1 transcription factor mediates a wide range of cellular processes, including cell growth, response to stress, and apoptosis. Members of the CREB family of transcription factors, such as ATF3 were also enriched; ATF3 has been shown to be involved in immune modulation and atherogenesis and metabolic regulation²³⁵. We also found enrichment of several members of the ETS family of transcription factors (ELF3, ELF4, ELF5, ETV1, ETV2, ERG, SPIB, FLI1, GABP α). These transcription factors are involved in a wide variety of function including cell proliferation, cell migration and angiogenesis. A key mechanism for this is by modulation of the TGF β signalling pathway. ETS transcription factors also regulate platelet-derived growth factor signalling and may explain the strong enrichment for platelet-related functions seen in the gene enrichment analysis.

A: CD4 ⁺ cells			
Rank	Motif	Best match	P-value
1		BATF(bZIP)/Th17-BATF-ChIP-Seq	1×10^{-113}
2		ERG(ETS)/VCaP-ERG-ChIP-Seq	1×10^{-122}
3		IRF3(IRF)/BMDM-Irf3-ChIP-Seq	1×10^{-56}
4		YY1(Zf)/Promoter	1×10^{-45}
5		RUNX1(Runt)/Jurkat-RUNX1-ChIP-Seq	1×10^{-43}
B: HAECs			
Rank	Motif	Best match	P-value
1		CTCF(Zf)/CD4+CTCF-ChIP-Seq	1×10^{-35}
2		AP-1(bZIP)/ThioMac-PU.1-ChIP-Seq	1×10^{-32}
3		NFATC1/MA0624.1	1×10^{-17}
4		POL010.1_DCE_S_III	1×10^{-14}
5		PB0192.1_Tcfap2e_2	1×10^{-14}
C: Macrophages			
Rank	Motif	Best match	P-value
1		PB0058.1_Sfp1_1	1×10^{-447}
2		BATF(bZIP)/Th17-BATF-ChIP-Seq	1×10^{-98}
3		MF0006.1_bZIP_cEBP-like_subclass	1×10^{-60}
4		Atf1/MA0604.1	1×10^{-37}
5		RUNX1(Runt)/Jurkat-RUNX1-ChIP-Seq	1×10^{-34}
D: VSMCs			
Rank	Motif	Best match	P-value
1		Jun-AP1(bZIP)/K562-Jun-ChIP-Seq	1×10^{-395}
2		BORIS(Zf)/K562-CTCF-ChIP-Seq	1×10^{-40}
3		EWS:ERG-fusion(ETS)/CADO_ES1-EWS:ERG-ChIP-Seq	1×10^{-37}
4		NFY(CCAAT)/Promoter	1×10^{-31}
5		TEAD3(TEA)/HepG2-TEAD3-ChIP-Seq	1×10^{-28}

Figure 6.10: Enrichment of transcription factor motifs in dynamic enhancer sites. Dynamic enhancer sites were analysed for transcription factor motif using HOMER. The top 5 most-enriched motifs and their p-values are shown

6.5 Coronary artery disease associated variants overlap with dynamic accessible and enhancer elements

Trait-associated variants have the potential to affect transcriptional regulation. We therefore explored the relationship between changes in chromatin structure in response to oxLDL and CAD genetic risk variants that have been identified by GWAS using the *snpR* package as described in Chapter 4. We used the filter term `coronary artery disease|myocardial ischaemia|angina|myocardial infarction|ischaemic stroke|ischaemic cva|atherosclerosis` to identify the relevant SNPs. We identified multiple CAD-associated variants overlapping with enriched peaks for open chromatin or H3K27ac, although the degree of overlap did not meet statistical significance (Table 6.10). We also considered overlap with cell-specific DAS and DES. Although the enrichment of the CAD variants did not

Table 6.10: Numbers of CAD-associated variants overlapping regulatory elements in untreated and oxLDL-treated vascular and immune cells.

Region	CD4+		HAECs		Macrophages		VSMCs	
	Untreated	oxLDL	Untreated	oxLDL	Untreated	oxLDL	Untreated	oxLDL
ATAC	45	94	126	80	90	88	80	78
H3K27ac	71	55	65	68	128	126	201	178
DAS		1		3		7		2
DES		0		0		2		0

meet statistical significance in these regions either, the variants that did overlap merit further discussion.

We identified 13 overlaps with the DAS, comprising 11 unique variants (Table 6.11). Most of these variants were located in intronic regions and have potential to change the binding motif for a number of transcriptions factors. Many of the transcription factors identified as being differentially enriched in the analysis in the previous section were included in this list, in particular, AP-1 and ETS-related transcription factors. The genes associated with these variants were *YTHDF3*, *ICA1L*, *SMAD3*, *LIPA*, *SWAP70*, *DOT1L*, and *LDLR*, but a search of the GTEx portal did not suggest these variants acted as eQTLs for these genes in these cell types.

The product of *YTHDF3* is a mediator of mRNA translation and stability; it binds to N6-methyladenosine (m6A)-containing mRNAs, and can either promote translation (in conjunction with YTHDF1) or can accelerate mRNA decay (in conjunction with YTHDF2)²³⁶. The gene was upregulated across all cell types following exposure to oxLDL, and significantly so in HAECs. In addition, the *YTHDF1* gene was also upregulated by oxLDL across all cell types.

DOT1L is a histone methyltransferase which methylates lysine-79 of histone H3, a modification which has been linked to transcriptional activation²³⁷. There was no significant change in its expression following exposure of the cells to oxLDL.

Little is known about the function of *ICA1L*. However, it has recently been shown to be a major binding partner for the product of the *PICK1* gene in the brain and spermatids²³⁸ and is crucial for its function in these cells. *PICK1* in turn interacts with *PRKCA* and regulates its subcellular localisation and function; *PRKCA* is a lipid-regulated serine-threonine kinase that regulates cell migration, adhesion, and proliferation, and also angiogenesis, platelet function and inflammation. *ICA1L* was down-regulated across all cell types.

SWAP70 regulates actin cytoskeletal structure and is necessary for efficient B cell migration to lymphoid tissue; it has not been shown to have a role in atherosclerosis previously. *LPA* and *LDLR* are directly involved in lipid handling and metabolism and have established roles in atherosclerosis.

The variant associated with *SMAD3* is particularly interesting. It has not previously been directly linked to CAD in any GWAS, but has been associated with inflammatory bowel disease. In a recent study comparing untreated and TGF β and PDGF-BB-treated human coronary artery smooth muscle cells (HCASMCs), this variant was also found to lie in a site marked by dynamic accessible chromatin and by H3K27ac²³⁹. They found the C allele (major risk for inflammatory bowel disease) was more associated with open chromatin and resides in a canonical AP-1 motif TGACT[C>T]A. The minor risk T allele disrupted AP-1 binding.

6.6 Discussion

In previous chapters we have discussed the phenotypic and transcriptional effects of oxLDL on immune and vascular cells. In this chapter, we have used epigenetic techniques to understand how these effects might be mediated by oxLDL induced

Table 6.11: Annotation of CAD-associated variants overlapping with dynamic accessible sites. INT = intron.

rsID	location	Motifs changed	Nearest gene	Distance	Annotation
rs9643384	8:63147099	ATF3	YTHDF3	21453	.
rs145741252	2:202871479	E2F, Egr-1, HDAC2, HEY1, Nrf1, Pax-5, SETDB1, YY1, Zfp1613	ICA1L	0	INT
rs149846585	2:202871528	AP-10, AP-2, BCL, BHLHE40, CHD2, CTCF, ELF1, ERalpha-a, Ets, Irf, SP1, YY1, Znf143	ICA1L	0	INT
rs17293632	15:67150258	AP-1, AP-1, AP-2, Arid3a2, BAF155, BATF, BCL, Bach1, Bach2, GATA, GR, HMGN3, Hoxb7, Irf, KAP1, Mef2, Myc, PRDM1, Phox2a, Pou3f4, Pou4f3, RXRA, STAT, Sox5, TCF4, p300	SMAD3	0	INT
rs9643384	8:63147099	ATF3	YTHDF3	21453	.
rs1412445	10:89243047	Ets, Maf, Mef2	LIPA	0	INT
rs1332327	10:89251924	AP-41, LBP-13, SP1, TATA	LIPA	0	.
rs173396	11:9738371	AP-2rep, Esr2, RBP-Jkappa1	SWAP70	0	INT
rs2019090	11:103798234	Foxo3, HDAC2, HMG-IY2, Lhx32	RP11-563P16.1	0	.
rs7256735	19:2169122	TCF11::MafG	DOT1L	0	INT
rs57217136	19:11090448	SP1	LDLR	0	INT
rs9643384	8:63147099	ATF3	YTHDF3	21453	.
rs3824754	10:102854593	Egr-1, Ets, HMG-IY2, SETDB1, Smad1	C10orf32	0	INT

chromatin remodelling. We mapped the effects of oxLDL on chromatin accessibility to provide a map of the changes arising from exposure of these cells to atherogenic lipid. As open chromatin alone may not delineate functional regulatory elements, we combined the analysis of changes in chromatin accessibility with ChIP-seq for H3K27ac, which marks active enhancers^{103,104}.

We found that roughly a fifth of all accessible sites were located at promoters, and this proportion did not change on exposure to oxLDL. However, the degree of accessibility at these sites did change in response to oxLDL - genes that were found to be upregulated by oxLDL displayed increased accessibility and genes which were downregulated had less accessible promoters. This is as expected to allow for transcription factor binding at gene promoters.

Most of the changes in chromatin accessibility induced by oxLDL were at distal sites, and there was a significant overlap between these sites and those previously found to have regulatory potential from other studies^{233,234}. We found that the number of DAS induced by oxLDL were far larger than the number of genes differentially expressed. This is likely to reflect the combinatorial control of gene expression by multiple regulatory sites.

In this study, oxLDL did not appear to have a significant effect on H3K27ac signal. However, DAS that were also marked by H3K27ac demonstrated a significant correlation between changes in chromatin accessibility at these sites and gene expression. This suggests that the effects of oxLDL on transcription are regulated by changes in accessibility at enhancers and not by modification of the status of these regulatory elements as enhancers; we termed these regions dynamic enhancer sites. The importance of this set of sites was evident from the pathway enrichment analysis which identified terms related to the cellular mechanisms underlying atherogenesis.

Given the correlation of these sites with changes in gene expression, we identified the transcription factors that bind at these regions. A recent epigenomic study of PDGF-stimulated human coronary artery smooth muscle cells and *ex vivo* coronary arteries also identified AP1-related transcription factors as potential mediators of atherogenesis²³⁹. We also identified the motifs of the ETS family of transcription factors as strongly enriched in DES. Previous studies have shown that AP1 and ETS factors bind the majority of active enhancers in HAECs²⁴⁰; our study confirms this, but also shows that they are important in the other cell types involved in atherosclerosis. Of note, functional interactions between these two transcription factors have previously been described in T cells²⁴¹, and would be a plausible general mechanism for the effects of oxLDL in atherosclerosis.

There were a number of CAD-related variants overlapping the identified DAS, and many of these variants were located in transcription factor binding motifs, and thus had potential to affect transcription factor binding. Not all of the genes associated with these variants have previously been directly linked to atherosclerosis, but plausible mechanisms for their effects exist. Oxidised LDL may act via *YTHDF3* and *DOT1L* to directly upregulate the translation and transcription of mRNA of target genes. *ICA1L* may mediate the effects of PRKCA through its binding partner PICK1; PRKCA has the capacity to modulate a number of processes essential to atherosclerosis.

It is surprising that the overlap of the CAD-associated variants did not reach statistical significance, particularly given this has been shown in related studies^{210,239,240}. The reason for this is unclear, but may be related to a lack of sufficient replicate samples leading to reduced sensitivity to detect regulatory sites.

In summary, this study provides an epigenomic characterisation of immune and vascular cells at baseline and following exposure to atherogenic lipid. There have recently

been other studies which have utilised similar approaches in individual cell types, but this is the first time that such techniques have been applied across the spectrum of cells involved in atherosclerosis exposed to atherogenic lipids. This work provides a valuable foundation upon which to expand and a means of prioritising potential regulatory elements for further investigation.

Chapter 7

General Discussion

In this study we investigated the fundamental role of atherogenic lipids in atherosclerosis, and in particular investigated the transcriptional and epigenomic changes induced by oxLDL on the vascular and immune cells involved in this process. A key strategy in the prevention of atherosclerotic vascular disease is the use of lipid-lowering agents such as statins, but beyond this, there have been limited advances in the development of therapeutic strategies. As such, the ultimate purpose of this study was to identify potential targets for further investigation to help move this area forward. We postulated that the effect of oxLDL on CD4+ cells, HAECs, macrophages, and VSMCs would include phenotypic, transcriptional, and epigenetic changes, and that integration of these changes would provide fundamental information about the mechanisms by which oxLDL acts. As part of this study, we also developed bioinformatic pipelines and tools for the analysis of HTS data, which are straightforward and easy to use, even for the non-bioinformatically inclined.

In the first part of this study, we developed a robust and reproducible protocol for the preparation of oxLDL from human plasma, and investigated methods for confirming its oxidation. We felt this was a vital aspect of the experimental set-up, as oxLDL has heterogeneous effects, and this is predominantly due to differences in the nature of minimally and extensively oxidised forms of the molecule. A review of the literature shows that very few studies have characterised the oxLDL used and this is likely to be a key explanation for the disparities in findings between investigators. Throughout this study we have used extensively modified oxLDL generated from human plasma-derived fresh LDL with pooling of samples to avoid inter-sample variability, and have taken pain-staking precautions to avoid endotoxin contamination.

We showed that the effects of oxLDL on cells is dependent on the concentration and duration of exposure to oxLDL; in particular, at concentrations above 50 µg/ml and

durations greater than 48 hours, there was a significant drop in cell viability. We showed that the cytotoxic effects of oxLDL are related to its uptake, and that cells that most avidly internalise oxLDL (macrophages and VSMCs) are most sensitive to its cytotoxic effects. We also saw that oxLDL modulates cell surface expression of cellular adhesion molecules and is able to induce angiogenesis in HAECs when they were exposed to 50 µg/ml for 48 hours.

Taken together it is clear that oxLDL does indeed induce relevant phenotypic and functional changes in vascular and immune cells and that these changes are relevant to the process of atherosclerosis, and that they warrant further investigation.

An integrative approach incorporating data from transcriptomic and epigenomic analyses, integrated with publicly available GWAS data has identified key mechanisms for the regulation of the effects of oxLDL. Our results showed striking transcriptional changes in cells exposed to oxLDL. We revealed that it induced profound changes in the expression of genes involved in lipid handling and metabolism, oxidative stress and inflammation, and gene set enrichment analysis identified additional effects on ECM modelling, cell-to-cell interactions and angiogenesis. A key mediator of these effects appears to be the $TGF\beta$ signalling pathway. Across all cell types, there was differential regulation of genes encoding proteins that mediate degradation of SMADs, of SMAD proteins themselves, and $TGF\beta$ -beta and BMP receptors. Terms related to this pathway were significantly enriched in both the transcriptomic and epigenomic analysis, and the CAD-variant enrichment analysis identified a SNP located in an intronic region of SMAD3 as lying within a DAS. $TGF\beta$ has been shown to exert diverse effects on all of the cell types involved in atherosclerosis, with the precise response being modulated by the local environment. In endothelial cells, $TGF\beta$ has a concentration and VEGF and FGF-modulated effect²⁴². It appears to stimulate

proliferation and migration at low concentrations and has the opposite effect at higher concentrations. Biologically this would make sense - it attracts cells to areas of injury from a distance but then acts to retain them at the site of injury. TGF β has similar effects in smooth muscle cells, but also plays a critical role in VSMC phenotype alteration by its effects on cytoskeletal structure²⁴². The overall effects of TGF β in macrophages is unclear - on the one hand, it has been shown to inhibit differentiation of monocytes to macrophages²⁴³ (anti-atherogenic role) and on the other hand knock-outs of components of the TGF β 2 pathway have resulted in reduced macrophage chemotaxis in mice, suggesting that the intact pathway is pro-atherogenic²⁴⁴. In lymphocytes, TGF β has a potent anti-proliferative effect under most conditions²⁴⁵, and knocking out the TGF β 2R in mice results in increased vessel wall infiltration by lymphocytes²⁴⁶.

The HIF1 α signalling pathway also emerged as an important signalling pathway across all the cell types. It is recognised that HIF1 α pathway plays a role in atherosclerosis, but the predominant theory for its activation is the ‘anoxaemia theory’ where vascular disease results in local tissue hypoxia, with this leading to its activation. However, in this study, cells were cultured in a monolayer in normoxic conditions, suggesting that the activation of the HIF1 α pathway in response to oxLDL is independent of hypoxia, and occurs even in early atherosclerosis. One possibility is that oxLDL-induced HIF1 α activation occurs via changes in cellular redox status rather than by hypoxia^{???}. A further potential mechanism of pathway activation is downregulation of *EGLN3*, the prolyl hydroxylase which targets HIF1 α it for degradation. In response to oxLDL, *EGLN3* is very significantly downregulated in both CD4⁺ cells and HAECs. To our knowledge a direct effect of oxLDL on levels of prolyl hydroxylases has not previously been reported, and this may represent a novel mechanism of action.

Of the differentially regulated genes we identified four that were concordantly differentially regulated across all the cell types. Amongst these was *HIPK2* which is known to regulate TGF β -dependent angiogenesis during embryonic development and to inhibit signalling via SMAD1 and SMAD3-dependent pathways²¹⁹. It is also part of a multi-protein corepressor complex with histone deacetylase 1 that targets the HIF1 α promoter, reducing its expression²²⁰. Thus, *HIPK2* may well be a common mediator of oxLDL-induced changes in TGF β and HIF1 α signalling across multiple cell types.

Epigenomic analysis identified accessible areas of chromatin whose accessibility appeared to regulate the expression of nearby genes. Many of these dynamic accessible sites were marked by the active enhancer marker H3K27ac and changes in accessibility at these dynamic enhancer sites occurred in a concordant direction to changes in expression of nearby genes. Transcription factor motif enrichment analysis identified enrichment of AP-1 motifs in these regions, suggesting that the ‘pioneer’ AP-1 factors are able to prime the chromatin, allowing subsequent recruitment of other transcription factors, and play an important role in mediating the effects of oxLDL. Amongst the other transcription factors recruited to these regions are likely to be members of the ETS family, whose motifs were also enriched. Functional interactions between these two transcription factor families have previously been described in T cells²⁴¹, and warrant further investigation. CAD variant analysis identified several SNPs overlapping with DAS and DES in these cells. Annotation of these variants links them to several potentially important genes including *YTHDF3* and *DOT1L*; the former was upregulated across all cells and in conjunction with *YTHDF1* promotes mRNA translation. *DOT1L* causes H3K79me3 of chromatin, which marks promoters for transcriptional activation. This suggests two distinct epigenetic mechanisms for

the effects of oxLDL, one at an RNA level and the other at the chromatin level.

We took several factors into consideration in the course of this study. The oxLDL we have used is heavily modified, and it is difficult to know to what extent this truly reflects the degree of oxidation *in vivo* as no assays exist for this. It is likely that circulating oxLDL exists along the spectrum of mildly modified to extensively modified, but for the purposes of reproducibility we felt that having a maximally modified form was likely to give the most consistent and interpretable results. The oxLDL we used was generated by copper oxidation, which is unlikely to be the main mechanism *in vivo*, where enzymatic processes are likely to predominate. Nonetheless, we used this method, because again, it ensured reproducibility. To control for the potential effects of copper ions in the oxLDL buffer we extensively dialysed the samples and then added the dialysate to the growth medium of the untreated cells. In the transcriptomic and epigenomic analyses, due to cost constraints, we had limited number of replicates for each condition and cell type; as such it would be useful to repeat the study with larger numbers of samples. In addition, number of samples aside, it would be useful to repeat the studies at a higher sequencing depth. ATAC-seq has the potential to allow for transcription factor foot-printing, but this requires larger biological samples and sequencing depth, which is substantially costlier, but would be worthwhile to further investigate the specific transcription factors that might be involved in atherogenesis.

Our results show that striking epigenetic changes are associated with the responses of cells to atherogenic lipids, and these were correlated with transcriptomic changes. ATAC-seq has proven to be an extremely useful tool to investigate these changes, and integration of chromatin accessibility signals with histone markers was highly useful in prioritising ATAC-seq signals. Analysis of the data generated by these HTS technologies can be difficult, and as part of this study we have developed a

bioinformatic pipeline to aid this. ATAC-seq is an emerging technology, and on a weekly basis new tools and adaptations are developed around it. We aim to continue to update the pipelines to reflect this and are in the process of submitting these pipelines to the Bioconductor framework. The study highlights the utility of combining multiple genomic data sets and publicly available GWAS to uncover the mechanisms underpinning complex disease processes.

Chapter 8

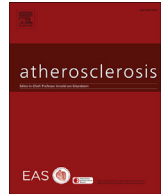
Appendix

8.1 Published work

Some of my work has been included in peer-reviewed publications, which is attached in the following pages.

Genetic and environmental risk factors for atherosclerosis regulate transcription of phosphatase and actin regulating gene PHACTR1. Reschen ME, Lin D, Chalissey A, Soilleux EJ, O'Callaghan CA Atherosclerosis. 2016 May 2;250:95-105

Genetic fine-mapping and genomic annotation defines causal mechanisms at type 2 diabetes susceptibility loci. DIAGRAM Consortium Nature Genetics. 2015 Dec;47(12):1415-25.



Genetic and environmental risk factors for atherosclerosis regulate transcription of phosphatase and actin regulating gene *PHACTR1*



Michael E. Reschen^a, Da Lin^a, Anil Chalisey^a, Elizabeth J. Soilleux^b,
Christopher A. O'Callaghan^{a,*}

^a Wellcome Trust Centre for Human Genetics, Nuffield Department of Medicine, University of Oxford, Roosevelt Drive, Oxford, OX3 7BN, United Kingdom

^b Nuffield Division of Clinical Laboratory Sciences, Radcliffe Department of Medicine, University of Oxford and Department of Cellular Pathology, John Radcliffe Hospital, Oxford, OX3 9DU, United Kingdom

ARTICLE INFO

Article history:

Received 20 October 2015

Received in revised form

20 March 2016

Accepted 26 April 2016

Available online 2 May 2016

Keywords:

Myocardial infarction

Atherosclerosis

Functional genomics

Genetic polymorphism

Genomics

Low-density lipoprotein (LDL)

Genetic disease

PHACTR1

Expression quantitative trait locus (eQTL)

Oxidized low density lipoprotein (oxLDL)

ABSTRACT

Background and aims: Coronary artery disease (CAD) risk is associated with non-coding genetic variants at the phosphatase and actin regulating protein 1 (*PHACTR1*) gene locus. The *PHACTR1* gene encodes an actin-binding protein with phosphatase regulating activity. The mechanism whereby *PHACTR1* influences CAD risk is unknown. We hypothesized that *PHACTR1* would be expressed in human cell types relevant to CAD and regulated by atherogenic or genetic factors.

Methods and results: Using immunohistochemistry, we demonstrate that *PHACTR1* protein is expressed strongly in human atherosclerotic plaque macrophages, lipid-laden foam cells, adventitial lymphocytes and endothelial cells. Using a combination of genomic analysis and molecular techniques, we demonstrate that *PHACTR1* is expressed as multiple previously uncharacterized transcripts in macrophages, foam cells, lymphocytes and endothelial cells. Immunoblotting confirmed a total absence of *PHACTR1* in vascular smooth muscle cells. Real-time quantitative PCR showed that *PHACTR1* is regulated by atherogenic and inflammatory stimuli. In aortic endothelial cells, oxLDL and TNF- α both upregulated an intermediate length transcript. A short transcript expressed only in immune cells was upregulated in macrophages by oxidized low-density lipoprotein, and oxidized phospholipids but suppressed by lipopolysaccharide or TNF- α . In primary human macrophages, we identified a novel expression quantitative trait locus (eQTL) specific for this short transcript, whereby the risk allele at CAD risk SNP rs9349379 is associated with reduced *PHACTR1* expression, similar to the effect of an inflammatory stimulus.

Conclusions: Our data demonstrate that *PHACTR1* is a key atherosclerosis candidate gene since it is regulated by atherogenic stimuli in macrophages and endothelial cells and we identify an effect of the genetic risk variant on *PHACTR1* expression in macrophages that is similar to that of an inflammatory stimulus.

© 2016 The Authors. Published by Elsevier Ireland Ltd. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Coronary artery disease (CAD) is caused by atherosclerosis, a

form of chronic inflammation. Deposition of pro-inflammatory lipoproteins in arterial walls is associated with the development of atherosclerotic plaques that can obstruct blood flow and trigger the formation of occlusive thrombi [1,2]. CAD is a complex disease with a strong heritable component [3]. Genome-wide association studies (GWAS) have identified 58 genomic regions where single nucleotide polymorphisms (SNPs) are associated with CAD risk [4,5]. Most of these disease-associated variants are in non-coding DNA and likely exert their effect predominantly by altering transcription factor binding to regulatory DNA elements [4,6–8]. A number of CAD risk loci contain genes that were not previously implicated in coronary artery disease and for which no mechanism

Abbreviations: CAD, coronary artery disease; GWAS, genome wide association studies; SNP, single nucleotide polymorphism; *PHACTR1*, phosphatase and actin regulator 1; oxLDL, oxidized low density lipoprotein; PP1, protein phosphatase 1; PBMC, peripheral blood mononuclear cells; TSS, transcription start site; RACE, rapid amplification of cDNA ends; HAEC, human aortic endothelial cells; LPS, lipopolysaccharide; NLS, nuclear localization signal; DAB, diaminobenzidine.

* Corresponding author.

E-mail address: chris.ocallaghan@ndm.ox.ac.uk (C.A. O'Callaghan).

has been identified to account for the risk associated with the reported genetic variant. Understanding the effect of these genetic risk variants on the expression and function of the genes that they regulate will advance our understanding of atherosclerosis and could lead to new therapeutic targets.

Several studies have independently replicated an association between CAD and SNPs at a genetic locus containing the phosphatase and actin regulator 1 (*PHACTR1*) gene [9–12]. SNPs at the *PHACTR1* locus are associated with the specific phenotypes of early onset myocardial infarction, coronary artery calcification [11,12] and with an intermediate phenotype of impaired central hemodynamic indices, indicating abnormal vascular stiffness [13]. The *PHACTR1* locus is pleiotropic since the protective alleles of the CAD risk SNPs are associated with an increased risk of ischemic stroke caused by cervical artery dissection, a form of non-atherosclerotic vascular disease [14]. The variants reported in these studies lie in an intronic region of *PHACTR1* over 250 kb away from any other gene. A genetic fine mapping study suggested that rs9349379 was the most likely causative variant at the locus and it was associated with expression of *PHACTR1* mRNA in composite right coronary artery tissue [15]. An expression quantitative trait locus (eQTL) study assaying gene expression in diverse human tissues showed that rs9349379 also affected *PHACTR1* mRNA expression in aortic artery and tibial artery tissue [16]. In a whole genome epigenetic and expression study, we recently demonstrated that *PHACTR1* is one of the most highly upregulated genes in macrophages exposed to oxLDL [17].

The *PHACTR* family consists of 4 genes encoding proteins that interact directly with both actin and protein phosphatase 1 (PP1) [18]. *PHACTR1* was originally cloned from a rat brain cDNA library using a yeast two-hybrid system with PP1 as bait [18]. The human *PHACTR1* gene is on chromosome 6 and a transcript with a 1743 bp open reading frame has been cloned from human brain [19]. The resulting 580 amino acid protein has 4 highly conserved actin-binding RPEL domains and both mouse and rat orthologues have been shown to bind actin [18,20]. Human *PHACTR1* protein contains nuclear localization signal (NLS) motifs at both ends of the protein that, in the mouse ortholog, have been shown to facilitate importin-dependent nuclear translocation in response to serum stimulation [20]. Binding of the RPEL domains to G-actin maintains *PHACTR1* in a cytoplasmic location [20]. Serum induces Rho-dependent remodeling of actin from G- to F-actin and translocation of *PHACTR1* into the nucleus [20].

PP1 is part of a family of serine/threonine phosphatases that are present in both the nucleus and cytoplasm [21]. The enzymatic specificity of PP1 enzymes is achieved in part by their association with accessory proteins [22]. *PHACTR1* may function as an accessory protein since *PHACTR1* binds to PP1 and inhibits its activity *in vitro* [18]. *PHACTR1* may also affect actin structure as siRNA-mediated *PHACTR1* knockdown in human umbilical vein endothelial cells (HUVEC) reduces F-actin filament numbers and repartitioning as well as increasing cell protrusion dynamics [23]. Although the precise cellular role of *PHACTR1* remains to be determined, overexpression and siRNA-mediated knockdown experiments suggest a role in cell motility and vascular morphogenesis [20,23,24].

Atherosclerosis involves the interplay of genetic factors with atherogenic stimuli, such as modified low-density lipoprotein, in cell types including macrophages, lymphocytes, endothelial cells and vascular smooth muscle cells [1,25–27]. The role of *PHACTR1* in these cells and the mechanism whereby it alters CAD risk are unknown. Understanding the role of *PHACTR1* is complicated by the fact that the human *PHACTR1* gene is predicted to encode multiple transcripts that have not been characterized. In this study, we hypothesized that *PHACTR1* is regulated by atherogenic or

inflammatory stimuli and that its expression is influenced by CAD-associated genetic variants. After establishing the expression of *PHACTR1* protein in atherosclerotic lesions and the profile of human *PHACTR1* transcripts in primary cell types involved in atherosclerosis, we determined the responses of these transcripts to inflammatory stimuli and to atherogenic lipid. The role of *PHACTR1* in CAD was highlighted by its *in vivo* abundance in macrophages and foam cells in human atherosclerotic plaque. The CAD risk allele at SNP rs9349379 was associated with significantly reduced expression of a short transcript in macrophages such that the risk genotype mirrors the effect of an inflammatory stimulus.

2. Materials and methods

2.1. Ethical approval

Ethical approval for the study was obtained from the NHS Research Ethics Committee (South Central-Hampshire B, reference 13/SC/0392) and all participants provided written informed consent.

2.2. Cell isolation, culture and reagents

Primary human coronary artery vascular smooth muscle cells (Invitrogen, Carlsbad, CA) were cultured in Medium 231 (Invitrogen) and Smooth Muscle Growth Supplement (Invitrogen). Primary human aortic endothelial cells (Invitrogen) were cultured in Medium 200 (Invitrogen) with Low Serum Growth Supplement. CD14⁺ monocytes, T cells (CD3⁺) and B cells (CD19⁺) were isolated from the peripheral blood of healthy human volunteers. Blood was centrifuged over Ficoll-Paque PLUS (GE Healthcare LifeSciences, Buckinghamshire, UK) to isolate peripheral blood mononuclear cells (PBMCs). Subsets of cells were isolated using positive selection with magnetic bead-conjugated antibodies (Miltenyi Biotec, Bergisch Gladbach, Germany). Individual populations were confirmed to be >95% pure using flow cytometry for the markers used for positive selection. The cells remaining after CD14⁺ cells were magnetically separated from PBMCs were termed CD14⁺-depleted PBMCs and constitute a mixed predominantly lymphocytic population. CD14⁺ monocytes were differentiated into macrophages by 7 days of culture in RPMI 1640 medium with 10% fetal calf serum, 4 mM L-glutamine, 50 units/ml penicillin and 50 mcg/ml streptomycin (Sigma, St Louis, MO), supplemented with 50 ng/ml macrophage colony stimulating factor (M-CSF, eBioscience, San Diego, CA). Foam cells were generated by treating macrophages with human oxLDL (50 mcg/ml) for 48 h. Foam cell formation was confirmed by oil red O staining for intracellular neutral lipid by light microscopy. Macrophage and foam cell viability was shown to be >98% using the Invitrogen LIVE/DEAD microscopy kit.

Ultracentrifugation of freshly isolated human plasma using a discontinuous potassium bromide gradient was used to purify LDL with density 1.019–1.063 g/ml [28]. LDL was extensively dialyzed against PBS, and incubated with 25 μ M CuCl₂ at 37° for 18 h to produce oxidized-LDL. To terminate oxidation 1 mM EDTA was added and oxLDL was stored at 4°. The TBARS assay was used to confirm that LDL was oxidized (Cayman Chemical, Ann Arbor, Michigan). Careful precautions were taken to ensure low endotoxin levels and LDL was tested for endotoxin using the gel clot method after heating to 75° for 15 min to remove the plasma inhibitor (Associates of Cape Cod, East Falmouth, MA). LDL was used at levels of <0.1 EU/ml.

2.3. Genotyping of rs9349379

Individuals providing blood samples were genotyped by the

Oxford Biobank and additional primary cells were genotyped using an inventoried Taqman genotyping assay according to the manufacturer's instructions (Applied Biosystems, Foster City, CA, ID 4351379).

2.4. RNA extraction and cDNA synthesis and 5' RACE

RNA was extracted using the Trizol RNA Plus extraction kit with on-column DNase treatment according to the manufacturer's instructions (Invitrogen). cDNA was produced by reverse transcription of 1 mcg of RNA using Bioscript reverse transcriptase (Bioline, London, UK) with random hexamers (Qiagen, Venlow, Limburg, Netherlands). For 5' RACE, 5 mcg of total RNA was reverse transcribed using 20 pmol of gene specific primer (CO5441 TTTGGCTGAAGGATCTTGGG) and then poly A tailed at the 3' cDNA end. First round PCR was performed using an inner gene specific primer (CO5439 CCATCCATGATGTCTGACGGTTGG) and oligo(dT) primer (with additional 5' sequence for round two PCR, CO1352) for 18 cycles using Bioscript (Bioline). Second round PCR was performed using gene specific primer (CO5400 CTTCTCTGCTTTGCCTCATAGATATTT) and a primer (CO3984 CTAGGAATTCTAGAGGTACCTCGAG) annealing to a 5' prime site created in the 1st round PCR. PCR products were gel extracted and ligated into Strataclone vector pSC-A-amp/kan using the strataclone PCR cloning kit (Agilent, Santa Clara, CA). Colonies were cultured, plasmid DNA extracted using a miniprep kit and sequenced using BigDye (Invitrogen) with both inner gene-specific primer (CO5513 CTGACGTGTGTTGAACCTTTTCG) and also (CO5400 CTTCTCTGCTTTGCCTCATAGATATTT). Sequences were aligned to the hg19 reference genome using the BLAT program in the UCSC genome browser.

2.5. PHACTR1 cloning

The long *PHACTR1* transcript was amplified from an Image clone (IRATp970E05120D) using primers CO4400 ACTGATGATCAATG-GATTATCCCAAAATGGATTATTTTC, and CO4807 TCAGTCTCGAGT-TAAGGTCGGTGAAACCTTGTTAAGTG and cloned into pCDNA3.1 with an N-terminal myc tag to produce plasmid pOC1259. The intermediate transcript was amplified from CD14⁺ macrophage cDNA using primers CO4976 ACTGATGATCAATGCGTTCTGACTCCCTCGTCC and CO4807 and cloned into pCDNA3.1 with an N-terminal myc tag to produce plasmid pOC1256. The short *PHACTR1* transcript (ENST00000379335) was amplified from macrophage cDNA using CO4806 AATGAGGATCCATGTATCTGCAAGGGCCGAGG and CO4807 and cloned into pCDNA3.1 with an N-terminal Myc tag to make plasmid pOC1257. Plasmids were sequenced using BigDye.

2.6. Tissue bank and primary cell expression analysis

An oligo(dT)-primed cDNA library set made from adult and fetal human tissue was screened for expression of *PHACTR1* transcripts. *PHACTR1* transcripts and beta-actin controls were amplified using BioTaq polymerase with 33 cycles and 25 cycles respectively. Products were run on a 1% agarose gel and stained using ethidium bromide. The primer pairs used were as follows: long transcript, CO3948 TGCTCACAGACTCTTGGATGTTG and CO3880 CTTCTCTGCTTTGCCTCATAGATATTT; intermediate transcript, CO3879 GGAAGAAGAAAAGCGAAAAGTTCA and CO3876 GGCAGTAGCAGGACTGGTTTG; short transcript, CO4987 ATGTATCTGCAAGGGCCGAG and CO3944 TGAATTCATTGAGCTCCTTTCG; control beta-actin, CO1851 TCCAGCCTTCCTTCCTGGGCAT and CO1852 GTCAAGAAAGGGTGTAACGCAACT.

2.7. Quantitative PCR and eQTL analysis

Expression levels of *PHACTR1* were determined by RT-qPCR with SYBR green reagents (Invitrogen) using either cDNA from macrophages or CD14⁺-depleted PBMCs from genotyped healthy individuals from the Oxford Biobank. RT-qPCR primers used were: CO5399 GGAAGAA-GAAAAGCGAAAAGTTCA and CO5400 CTTCTCTGCTTTGCCTCATAGATATTT (intermediate transcripts); CO4987 ATGTATCTGCAAGGGCCGAG and CO5122 CTTCTGATCTCTCTCTCTCCTCCTG (short transcript); CO3744 TTGCCATCAATGACCCCTTCA and CO3745 CGCCCCACTTGATTTTGGG (control *GAPDH*). Melt curve analysis demonstrated a single product for each primer pair. For each sample, triplicates were assayed. A Kruskal-Wallis test was applied across the three possible genotypes to identify a significant genotype-expression relationship with significance defined as a *p* value < 0.05. To assess *PHACTR1* expression in response to stimuli, primary human macrophages (rs9349379 AA) were treated with 50 mcg/ml oxLDL for 48 h, 10 ng/ml LPS (Sigma) for 24 h, 10 ng/ml TNF α for 24 h (eBioscience), 30 mcg/ml OxPAPC (Invivogen, San Diego, CA) for 48 h, or 10 μ g/ml cyclodextrin-cholesterol (Sigma) for 48 h and assayed in biological triplicates. HAEC (rs9349379 AG) were treated similarly with 50 μ g/ml oxLDL for 48 h, 10 ng/ml TNF α or 10 μ g/ml cyclodextrin-cholesterol for 48 h. Coronary artery smooth muscle cells (rs9349379 AG) were treated with 10 μ g/ml cyclodextrin-cholesterol for 48 h and assayed in triplicate. Student's *t*-test was applied with significance defined as a *p* value < 0.05.

2.8. Western blotting

Cells were lysed in RIPA buffer with protease inhibitors and subjected to SDS-PAGE analysis using 10 mcg of protein per lane and PageRuler Plus pre-stained marker lane. Endogenous *PHACTR1* was detected using a rabbit polyclonal anti-*PHACTR1* antibody (Novus Biologicals, Littleton, CO) and secondary HRP-conjugated goat anti-rabbit IgG antibody (Vector labs, Burlingame, CA). Chemiluminescent reactions were performed using ECL Advance (GE Healthcare LifeSciences) and imaged using either a BioRAD gel doc imaging station (BioRad, Hercules, CA) or Amersham hyperfilm ECL (GE Healthcare LifeSciences).

2.9. Immunohistochemistry

Formalin-fixed paraffin-embedded tissue samples of non-atherosclerotic aorta (*n* = 5), and atherosclerotic aorta (*n* = 5), obtained with full ethical approval from the National Research and Ethics Service (Oxfordshire Research and Ethics Committee A: reference 04/Q1604/21), were immunostained for *PHACTR1* with a rabbit polyclonal anti-*PHACTR1* antibody (Abcam, Cambridge, UK) detected with the NovolinkTM max polymer detection system (Leica Microsystems, Buffalo Grove, IL), as per the manufacturer's instructions, or with detection reagents only, as a negative control. Slides were mounted in Aquatex mounting medium (Merck, White House Station, NJ). Stained sections were photographed with a Nikon DS-F11 camera with a Nikon DS-L2 control unit (Nikon, Tokyo, Japan) and an Olympus BX40 microscope (Olympus, Tokyo, Japan). All immunohistochemical staining was analyzed by and performed under the supervision of an experienced board certified consultant pathologist (EJS).

3. Results

3.1. *PHACTR1* expression is increased in human atherosclerotic lesions

To investigate the role of *PHACTR1* in CAD, we used immunohistochemistry to evaluate human *PHACTR1* protein expression

in vivo in atherosclerotic lesions in arteries from patients with atherosclerosis ($n = 5$) and in normal arteries from healthy controls ($n = 5$). In healthy arteries, PHACTR1 was expressed in endothelial cells and in the occasional immune cells that were present, but not in vascular smooth muscle cells (Fig. 1A). In atherosclerotic plaque lesions, an abundance of infiltrating macrophages and foam cells were present that all strongly expressed PHACTR1, but no expression was seen in vascular smooth muscle cells (Fig. 1B, C). Plaque endothelial cells also expressed PHACTR1 (Fig. 1B). Lymphocytes in peri-plaque adventitial aggregates also expressed PHACTR1 and this was present in both the cytoplasm and nucleus (Fig. 1D). PHACTR1 was also expressed by lymphocytes in secondary lymphoid tissue (tonsillar), but here it was predominantly in the cytoplasm, rather than the nucleus (Fig. 1E). Overall, the immunohistochemistry results highlight the marked expression of PHACTR1 in immune cells in atherosclerosis, particularly in macrophages and lipid-laden foam cells.

3.2. PHACTR1 expresses multiple transcripts from two promoters in vascular and immune cells

We next determined the expression pattern of PHACTR1 in primary human atherosclerosis-related cell types involved in atherosclerosis. A combined genomic analysis of human expressed sequence tag (EST) data, chromatin state mapping based on ChIP-seq data and CAP analysis of gene expression with high throughput sequencing (CAGE-seq) [29,30] suggested the presence of a transcriptional start site (TSS) around 240 kb downstream from that of the reported RefSeq transcript (NM_030948.2) and a further TSS close to the 3' end of this transcript (Fig. 2A). To determine the origin of transcripts in cell types involved in atherosclerosis, we mapped the TSS using Rapid Amplification of cDNA Ends (RACE) in primary human macrophages, aortic endothelial cells (HAEC), monocytes and peripheral blood mononuclear cells (PBMC) depleted of CD14⁺ monocytes (Fig. 2A, Supplementary data 1). This demonstrated that the known TSS of the human RefSeq transcript (NM_030948.2) was not used, but instead the RACE clones confirmed a TSS 240 kb downstream, that was supported by the EST, chromatin state and CAGE-seq data (Fig. 2A). We used the RACE results to design primers to amplify and clone a 1674 bp open reading frame from primary human macrophages and arterial endothelial cells. This transcript encodes a 557 amino acid PHACTR1 isoform that matches the predicted NCBI Reference Sequence: XP_011512694.1 (Fig. 2A/B, Supplementary data 1). On the basis of its length, we refer to this transcript as the 'intermediate transcript'. The open reading frame of the intermediate transcript begins partway through exon 3 of NM_030948.2 and contains all subsequent downstream exons. Additionally, the intermediate transcript includes an extra exon (exon 6) that is not present in NM_030948.2 (Fig. 2B/C). In CD14⁺-depleted PBMCs, the transcriptional start site was around 600 bp upstream of that seen in both primary arterial endothelial cells and CD14⁺ cells, but this did not alter the open reading frame.

The genomic analysis supported the presence of a distal TSS that matched the beginning of ENSEMBL transcript ENST00000379335 [29,30] (Fig. 2 A/B). Additional analysis of chromatin mapping data from our previous study of primary human macrophages demonstrated an open chromatin site at the TSS that was flanked by H3K27ac-marked histones consistent with promoter activity [17]. Using this information, we cloned a 435 bp transcript from macrophages, which encoded a 144 amino acid protein. This 'short' transcript shares its last four exons with the longer transcripts, but begins with an extended version of exon 10 that starts 5' to the exon 10 start position seen in the longer transcripts (Fig. 2B/C).

3.3. PHACTR1 transcripts display tissue-specific differential expression

We designed transcript-specific primer pairs to examine the expression pattern of the long, intermediate and short transcripts in atherosclerosis-related cell types (Table 1). The long transcript was not expressed in primary human monocytes, macrophages, HAEC, VSMC, T cells, B cells or CD14⁺-depleted PBMC (Fig. 3A). The intermediate transcript was expressed in macrophages, HAEC and VSMC (Fig. 3A), but was barely detectable in monocytes and B cells and absent from T cells. The short transcript was expressed in primary human monocytes, macrophages, B cells, CD14⁺-depleted PBMC, and to a lesser extent in T cells, but was absent from HAEC and VSMC (Fig. 3A).

We screened a panel of adult and fetal tissues and the long transcript was only present in adult brain tissue (Fig. 3B). The intermediate transcript was expressed in pancreas and fetal heart (Fig. 3B). The short transcript was expressed in adult spleen, fetal spleen, PBMC, colon and at a low level in the thymus (Fig. 3B). Overall, the long transcript is restricted to neuronal tissue only and the intermediate transcript is more variable, being present in HAEC, VSMC and macrophages. The short transcript is restricted to immune cells.

3.4. PHACTR1 protein is expressed in macrophages and HAEC but not VSMC

To assess PHACTR1 protein expression in primary human cells, western blotting was undertaken using an antibody raised against a region of the protein shared by all PHACTR1 isoforms. Recombinant PHACTR1 isoforms were expressed and analyzed in parallel to facilitate molecular size comparison (Fig. 3C). The isoform encoded by the long transcript migrated with an apparent molecular weight of about 75 kDa (predicted 66 kDa, pI 6.51), similar to that of the rat brain isoform (Fig. 3C) [18]. The intermediate isoform is 23 amino acids shorter, but migrated with an apparently higher molecular weight, equivalent to about 80 kDa (predicted 64 kDa, pI 6.81) (Fig. 3C). This apparently anomalous migration pattern may arise from differences between the protein isoforms such as shape, rigidity, charge or post translational modification. Notably, exon 6 is rich in proline residues that may increase rigidity (14 of 69 residues in exon 6 are proline). Exon 6 in the intermediate isoform does not contain any additional predicted glycosylation motifs. PHACTR1 protein was completely absent from primary vascular smooth muscle cells despite moderate level mRNA expression in these cells (Fig. 3C). In primary human macrophages, foam cells, HAEC and oxLDL-treated HAEC there was strong expression of the intermediate isoform (Fig. 3C). We did not detect protein expression of the short transcript but detection of this protein was difficult even following overexpression in 293T cells (Fig. 3C and D). Although it was not possible to assess the endogenous short isoform reliably, these findings are consistent with our transcript studies and indicate that primary macrophages and arterial endothelial cells express the intermediate, but not the long protein isoform.

3.5. PHACTR1 expression is regulated by oxLDL and inflammatory stimuli in macrophages and endothelial cells

Given the association of the PHACTR1 locus with CAD and its expression in atherosclerotic lesions, we hypothesized that PHACTR1 would be regulated by atherogenic and inflammatory stimuli. We used transcript-specific RTqPCR to assess expression of PHACTR1 in response to a variety of stimuli. Lipopolysaccharide (LPS) treatment of primary human macrophages upregulated the intermediate transcript (fold = 4.68, $p < 0.0001$), but

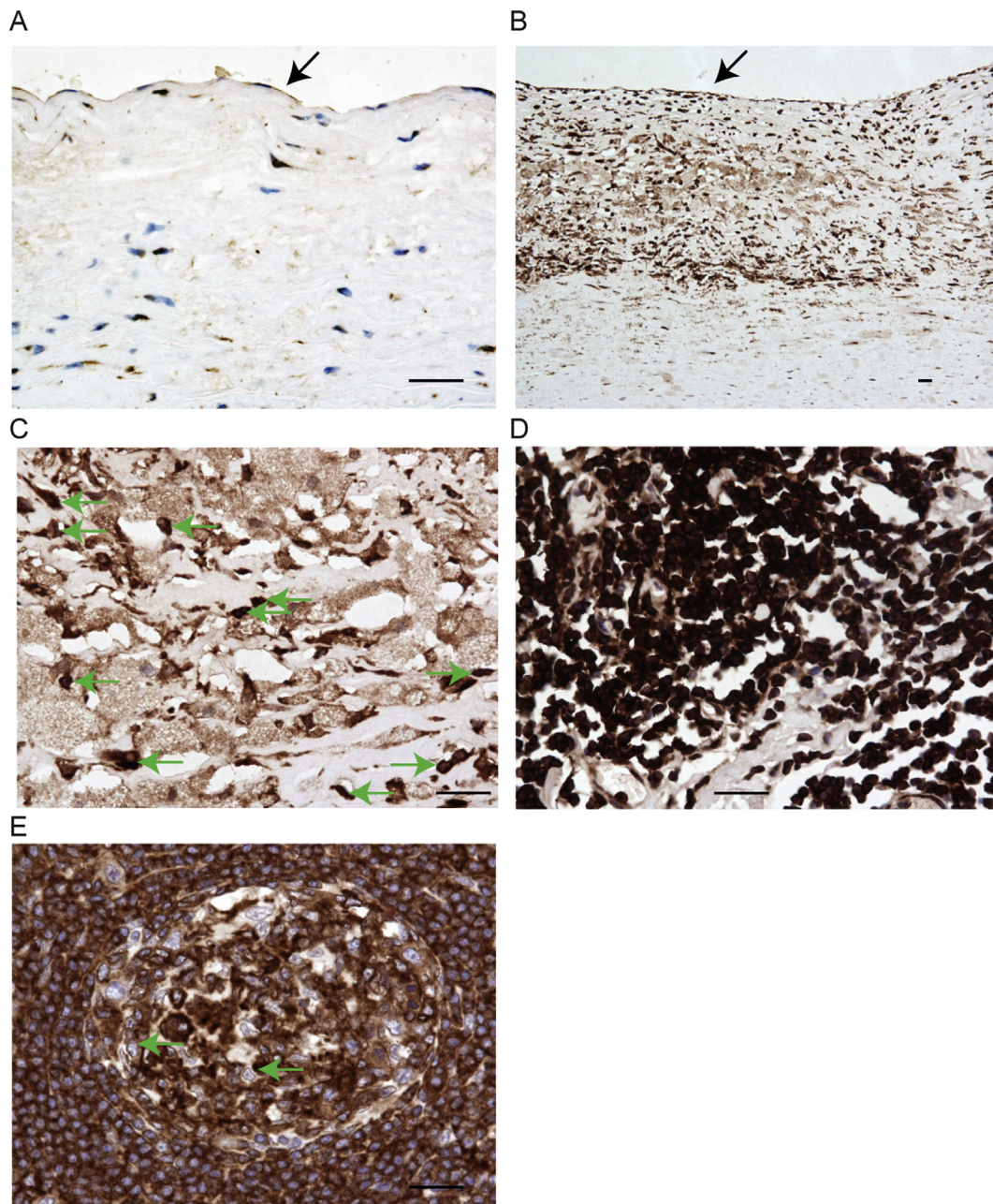


Fig. 1. Immunohistochemistry demonstrates *in vivo* PHACTR1 expression in atherosclerotic plaque. (A) Healthy aorta demonstrates PHACTR1-positive endothelium (arrowed), but vascular smooth muscle cells in the sub-intimal layer lack PHACTR1 staining. (B) Atherosclerotic plaque showing endothelial positivity (arrowed) and positivity in the sub-intimal space, comprising predominantly foam cells. (C) Higher power view of foam cells in atherosclerotic plaque showing cytoplasmic and nuclear staining for PHACTR1 (Green arrows indicate examples of strongly stained foam cell nuclei. Cytoplasmic foam cell staining is present in the lighter brown areas that asymmetrically encircle the nuclei). (D) Lymphocytic aggregate in aortic adventitia of atherosclerotic vessel showing strong positivity for PHACTR1. The PHACTR1 staining is present throughout the cells, indicating nuclear and cytoplasmic distribution. (E) Germinal centre in tonsillar lymphoid tissue; lymphocytes lack nuclear PHACTR1 as demonstrated by a thin rim of brown cytoplasmic staining with unstained, blue nuclei (example lymphocytes arrowed). Staining for human PHACTR1, using HRP/DAB, each section representative of data from 5 individuals per condition, scale bars indicate 50 μ m.

downregulated the short transcript more than 10-fold (fold = 0.06, $p < 0.0001$) (Fig. 4A). The inflammatory cytokine Tumor Necrosis Factor- α (TNF α) produced a similar, but weaker effect with upregulation of the intermediate transcript (fold = 1.41, $p = 0.04$) and downregulation of the short transcript (fold = 0.64, $p = 0.02$) (Fig. 4A). OxLDL exerted a converse effect on expression of PHACTR1 compared to LPS and TNF α ; the short transcript was upregulated 6 fold ($p = 0.004$) and the intermediate transcript was downregulated (fold = 0.74, $p = 0.01$) (Fig. 4B). Oxidized 1-palmitoyl-2-arachidonoyl-sn-glycero-3-phosphorylcholine

(oxPAPC) is a mixture of oxidized phospholipids that have a similar biological effect to minimally oxidized LDL [31]. OxPAPC contains components of the biologically active fractions of HPLC-separated minimally-modified LDL [31]. In macrophages OxPAPC produced a similar directional effect to oxLDL—upregulation of the short transcript (fold = 2.28, $p = 0.002$) and a trend towards downregulation of the intermediate transcript (fold = 0.83, $p = 0.12$) (Fig. 4B). To refine the lipid species responsible for the effect on PHACTR1 expression we loaded macrophages with cholesterol itself. Cells were treated with a cyclodextrin-cholesterol

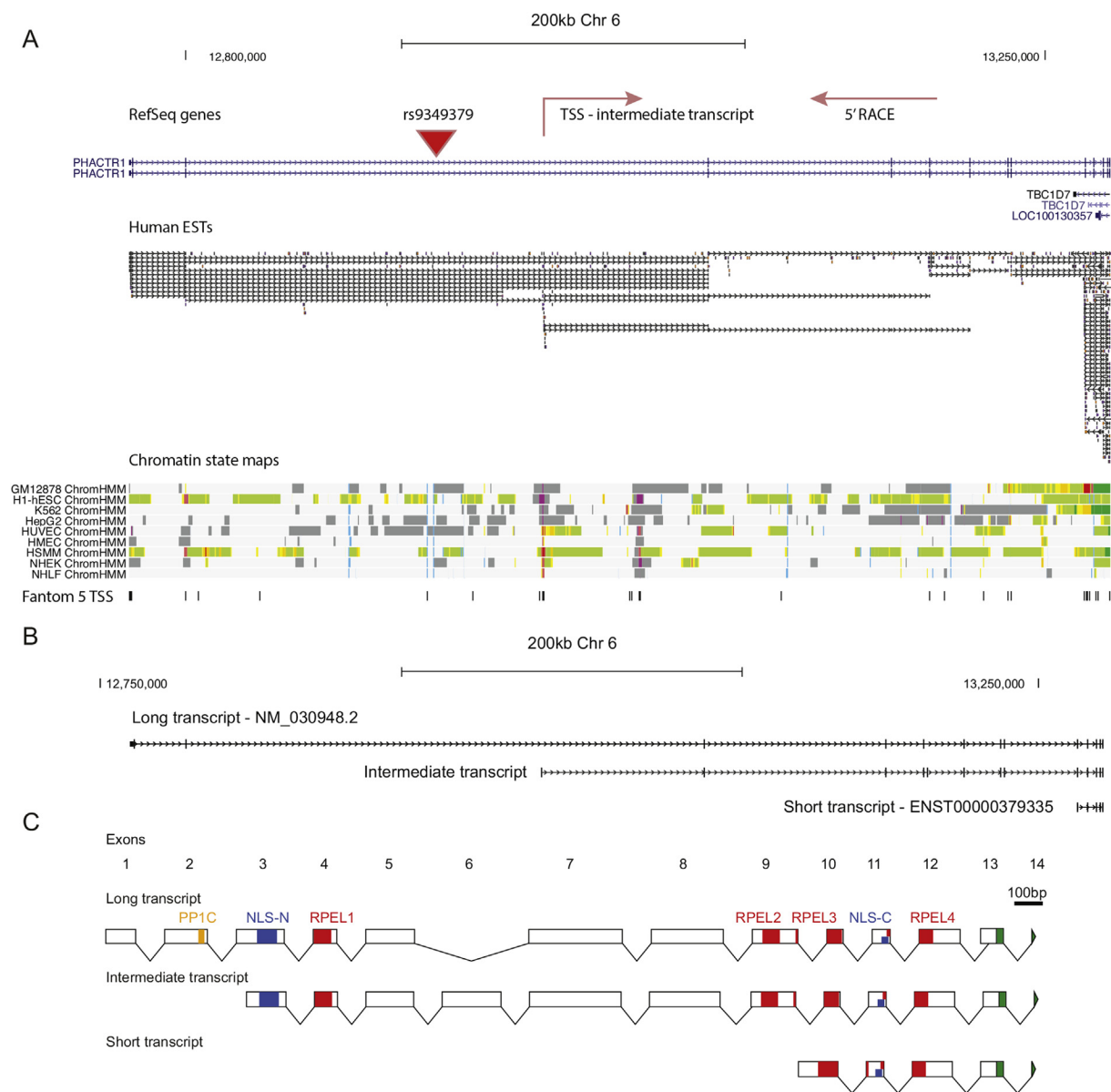


Fig. 2. The structure of human *PHACTR1* transcripts. (A) The transcript structure for *PHACTR1* is shown for RefSeq curated transcripts showing 2 transcripts in blue. Shown below is human expressed sequence tag data, which is suggestive of a transcription start site halfway along the gene and a further possible site at the 3' end of the gene. These data were compared to chromatin state data that uses multiple ChIP-seq markers to classify the genome into different types of regulatory element or transcription state. Chromatin state maps are shown for 9 cell types with the chromatin state colored (red - promoter, orange - enhancer, yellow - weak enhancer, blue - insulator, green - transcribed, dark grey - not transcribed). Red regions depicting active promoter elements are seen half way along the gene, coinciding with a TSS suggested by EST data. In one cell type, a promoter is also seen at the 3' end. These sites coincide with precise TSSs identified by CAGE-seq in a variety of human cell types. Taken together, these data suggest three main TSS in three main promoter regions. The position of coronary artery disease risk variant, rs9349379, is indicated by the red triangle. (B) The transcript structure for the RefSeq transcript is shown uppermost and is aligned to the intermediate transcript that was cloned from macrophages and endothelial cells. The short transcript cloned from macrophages is also shown below. Vertical lines indicate exons. (C) Coding structure of *PHACTR1* transcripts with exons drawn to scale and motifs annotated. Motif locations were predicted using the Eukaryotic Linear Motif resource. The region of experimentally proven PP1 binding is shown in green.

Table 1
Summary of expression pattern of 3 different *PHACTR1* transcripts (mRNA) in primary human cell types relevant to atherosclerosis.

	Long	Intermediate	Short
Macrophages	No	Yes	Yes
Arterial endothelial cells	No	Yes	No
Vascular smooth muscle cells	No	Yes	No
PBMC depleted of CD14 ⁺ cells	No	Yes	Yes
B cells (CD19 ⁺)	No	Barely detectable	Yes
T cells (CD3 ⁺)	No	No	Barely detectable

complex that promotes cholesterol loading of macrophages and vascular smooth muscle cells [32,33]. Unlike the oxidized stimuli, cholesterol loading had no effect on *PHACTR1* expression (Fig. 4B). Overall, these data indicate that in macrophages inflammatory pathways downregulate the short transcript and upregulate the intermediate transcript, whereas oxidized lipid-responsive pathways upregulate the short transcript. In human aortic endothelial cells—which only express the intermediate transcript—both TNF α and oxLDL upregulated *PHACTR1* (fold = 2.5, p = 0.007, and fold = 1.8, p = 0.009, respectively) (Fig. 4C). As with macrophages, cholesterol loading with cyclodextrin-cholesterol had no effect on

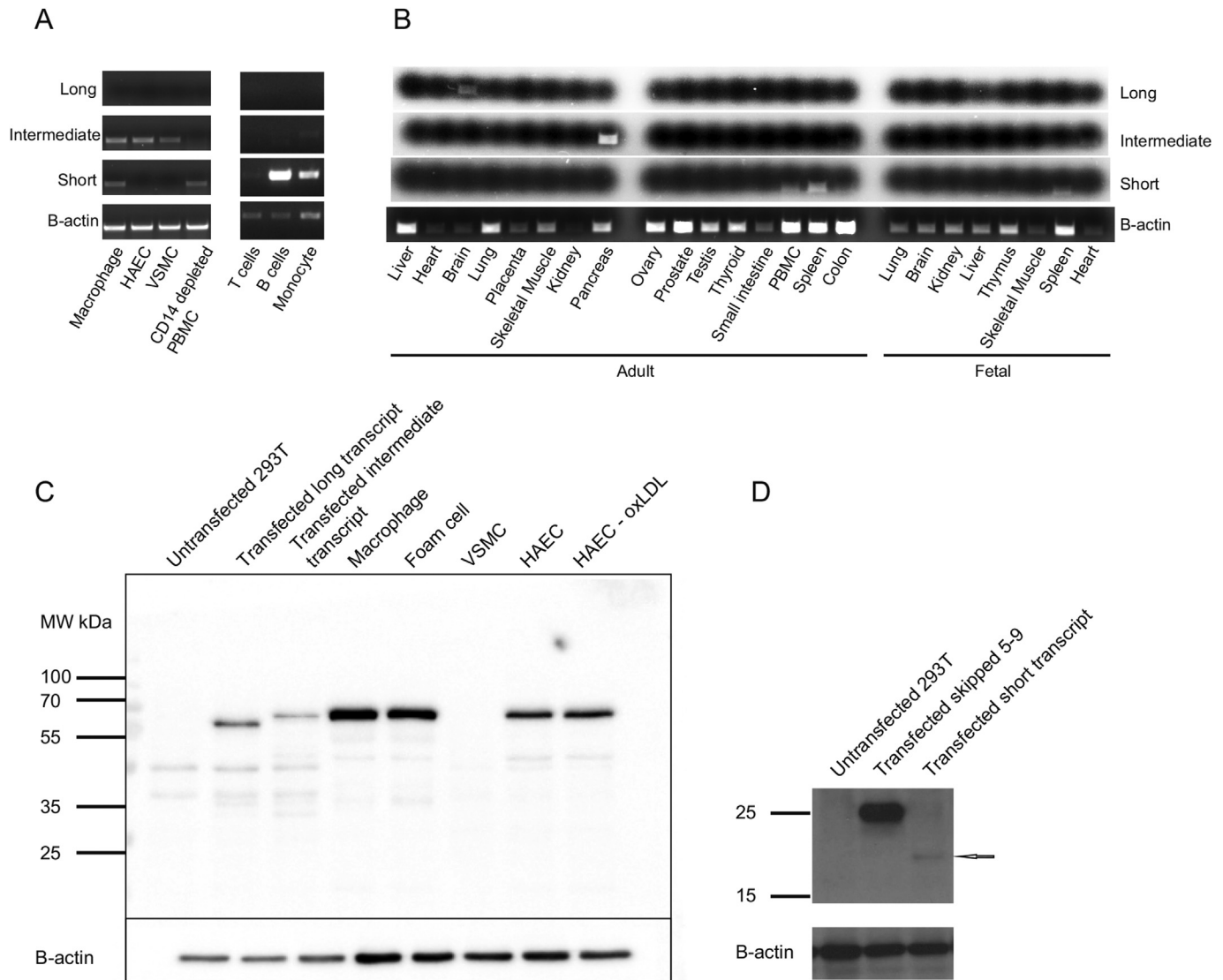


Fig. 3. PHACTR1 expression patterns in primary cells and tissues. (A) Expression in purified primary human cells showed that the long transcript was not expressed among the cells tested. By contrast, the intermediate transcript showed robust expression in human monocytes, macrophages, arterial endothelial cells (HAEC) and vascular smooth muscle cells (VSMC), but was not expressed by T cells and was barely detectable in B cells and PBMC-depleted of monocytes. (B) Analysis of a human tissue cDNA panel showed the long transcript exclusively expressed in adult brain. The intermediate transcript showed strong expression in the pancreas, but was also detectable in fetal heart. By contrast, the short transcript was present in PBMC and spleen, including fetal spleen. (C) Western blotting: PHACTR1 over-expression in 293T cells transfected with the long and intermediate isoforms compared to endogenous PHACTR1 in primary human cells. The shorter intermediate isoform has an apparently greater size than the long isoform in 293T cells. Macrophages, foam cells and HAECs expressed PHACTR1 with a similar size to the intermediate isoform. PHACTR1 protein was absent from VSMC. A band corresponding to endogenous expression of the short isoform was not seen in any cell type. The polyclonal rabbit anti-PHACTR1 antibody identifies a non-specific additional band at about 47 kDa even in untransfected cells, which do not express PHACTR1. (D) Western blotting with over-expression in 293T cells of the short transcript cloned from macrophages. A cloned form of the intermediate transcript skipping exons 5–9 is also shown for comparison.

PHACTR1 expression. In vascular smooth muscle cells, cholesterol loading had no effect on expression of the intermediate *PHACTR1* transcript.

3.6. Genetic regulation of *PHACTR1* expression by CAD SNP rs9349379

The rs9349379 SNP at the *PHACTR1* locus has been implicated by GWAS in the genetic risk of CAD [11,34]. The mechanism by which different alleles of this SNP influence CAD risk is unknown. Rs9349379 is in an intronic region of *PHACTR1* and is not in high linkage disequilibrium with any coding SNPs; it may therefore affect transcriptional regulation of *PHACTR1*. Using transcript-specific real time quantitative PCR, we tested for an association

between the alleles of rs9349379 and expression of *PHACTR1* in macrophages and CD14⁺-depleted PBMC. To do this we used primary cells derived from healthy non-smoking Caucasian individuals controlled for age (>18, <50 years), body mass index, cholesterol and triglyceride levels but with differing rs9349379 genotypes. There was a significant allelic effect on expression of the short transcript; the A allele was associated with higher expression (Fig. 5A). Macrophages from individuals homozygous for the protective A allele had 2.5-fold higher expression than macrophages from individuals homozygous for the G allele, which is the coronary artery disease risk-associated allele ($p = 0.0187$). There was a trend towards reduced expression of the intermediate transcript with the A allele, but this did not reach significance (Fig. 5B). In CD14⁺-depleted PBMCs no association was detectable

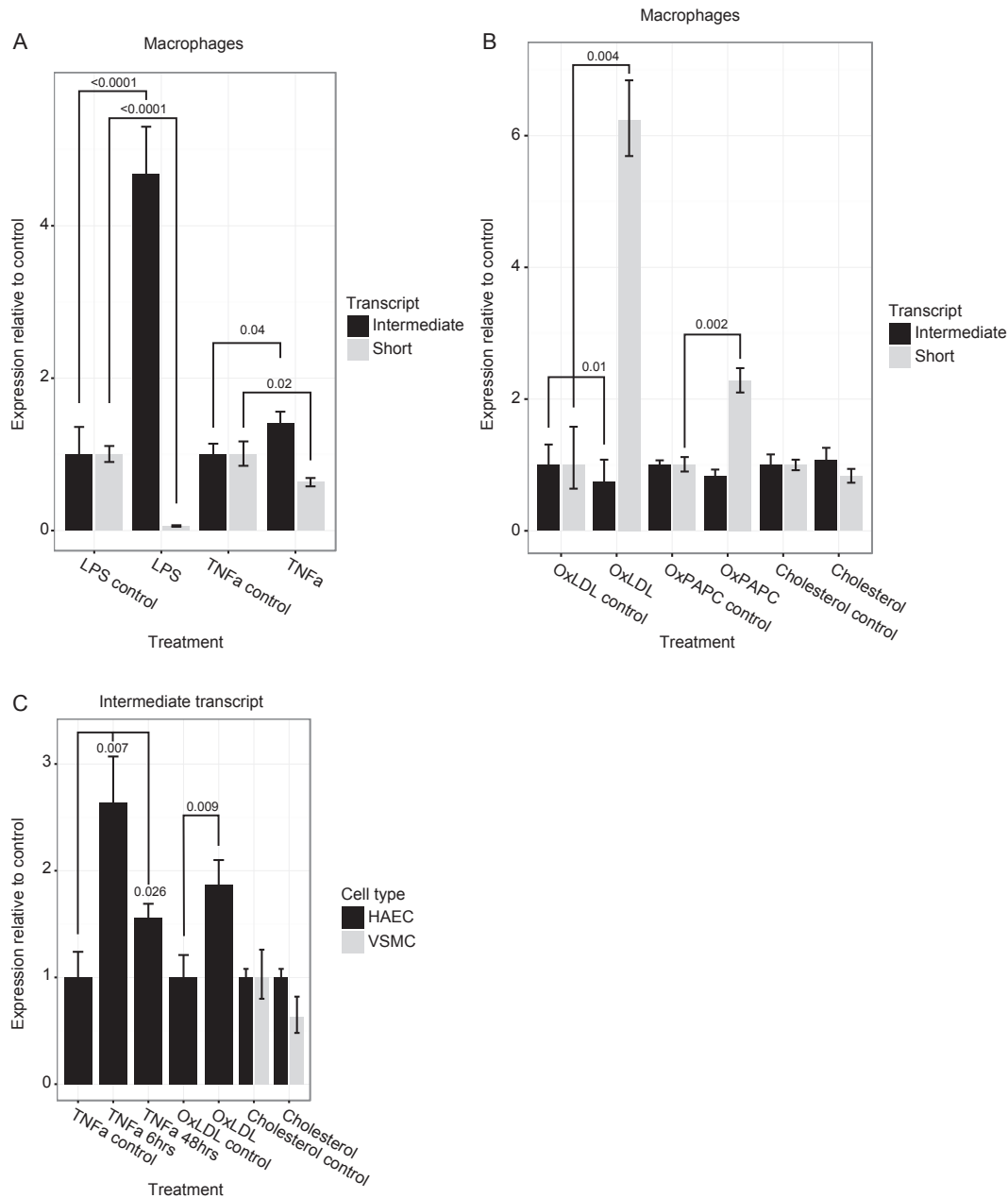


Fig. 4. *PHACTR1* transcripts are differentially regulated by atherogenic and inflammatory stimuli. (A) Effect of LPS (10 ng/ml) on mRNA expression in primary human macrophages. The intermediate transcript was upregulated, but the short transcript was suppressed. TNF α (10 ng/ml) produced a directionally similar effect. (B) Exposure of macrophages to oxLDL (50 mcg/ml) caused modest downregulation of the intermediate transcript and strong upregulation of the short transcript. OxPAPC (30 mcg/ml) caused a directionally similar effect to oxLDL on *PHACTR1* expression. (C) The intermediate transcript was upregulated by oxLDL in aortic endothelial cells. TNF α (10 ng/ml) caused upregulation of *PHACTR1* at 6 and 48-h time points in aortic endothelial cells. The short transcript was not expressed in aortic endothelial cells. Cyclodextrin-cholesterol (10 mcg/ml) loading of endothelial cells and vascular smooth muscle cells had no effect on *PHACTR1* expression. Significant differences are indicated by black lines with *p* values shown. The controls labelled on the x-axis represent the relevant buffer controls for each treatment.

between genotype and expression of either the short or intermediate transcript (Fig. 5C, D). Overall, these results demonstrate that rs9349379 is both a CAD risk-associated variant and an expression quantitative trait locus (eQTL) for *PHACTR1* in macrophages, but not in CD14⁺-depleted PBMC. Therefore, decreased expression of the short transcript in macrophages is associated with CAD. This genetic risk effect recapitulates the effect of LPS, which also caused suppression of the short transcript (Fig. 5A). Overall, these findings demonstrate that the CAD risk effect of genetic variants at rs9349379 may be operational in macrophages in atherosclerotic lesions.

4. Discussion

GWAS of complex diseases have identified genomic loci containing genetic variants that affect heritable disease risk [4]. For CAD, one of the most robust reported risk loci is at the *PHACTR1* gene locus where several studies have identified intronic SNPs that are associated with the risk of myocardial infarction, coronary stenosis and coronary calcification [9–11,34]. The same CAD risk SNP allele is associated with protection from a non-atherosclerotic form of vascular disease; cervical artery dissection. Unravelling the biological pathways underpinning the pleiotropic effect of the

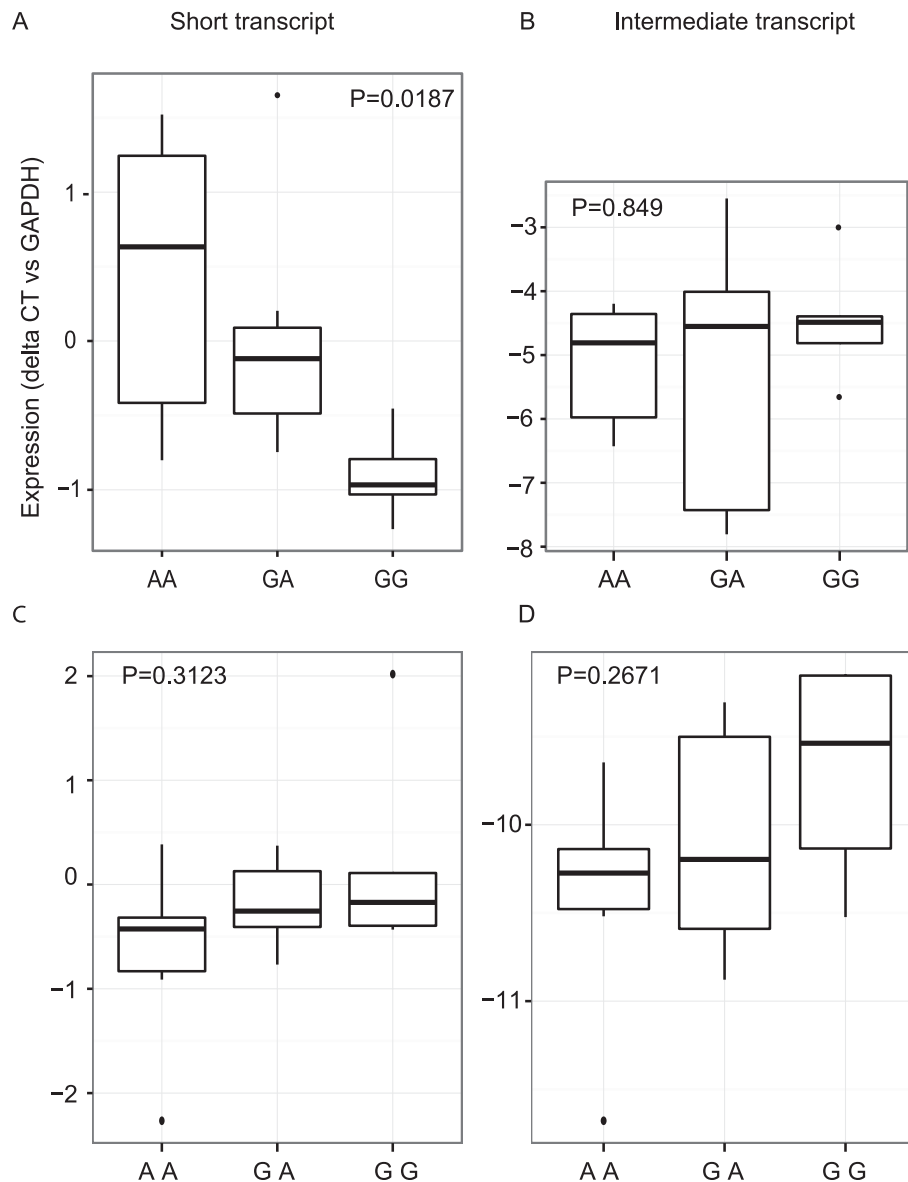


Fig. 5. Genotype of CAD risk SNP rs9349379 affects expression of the short *PHACTR1* transcript in macrophages. *PHACTR1* expression in primary human macrophages relative to *GAPDH*. (A) A significant genotype-expression level correlation was demonstrated in primary human macrophages ($n = 19$, AA = 6, AG = 8, GG = 5). Individuals with the AA genotype had the highest level of expression of the short *PHACTR1* transcript and subjects with the GG genotype had the lowest expression. (B) There was no significant correlation with the intermediate transcript ($n = 19$). (C, D) Primary human CD14⁺-depleted PBMCs did not show a significant correlation between expression and genotype of rs9349379 ($n = 20$, AA = 7, AG = 8, GG = 5).

PHACTR1 locus is critical to harnessing the genetic knowledge yielded by GWAS studies [14]. To better understand the role of *PHACTR1* in atherosclerosis, we investigated *PHACTR1* expression and the regulatory influence of inflammatory stimuli, atherogenic lipid and genetic variation.

Using immunohistochemistry, we demonstrated increased expression of *PHACTR1* in atherosclerotic lesions *in vivo*. Lipid-laden macrophages or foam cells were clearly the most abundant source of *PHACTR1* in plaques. The distribution was both nuclear and cytoplasmic in macrophages, foam cells and adventitial lymphocytes, whereas in secondary lymphoid tissue lymphocytes, *PHACTR1* was excluded from the nucleus. This differential intracellular localization may reflect the accumulation in atherosclerotic plaques of inflammatory lipids including lysophosphatidic acid (LPA), which activates RhoA signaling causing G- to F-actin polymerization, a mechanism known to promote *PHACTR1* nuclear

accumulation [35].

The previously reported *PHACTR1* transcript was isolated from brain and encodes a 580 amino acid protein [18,19]. We only found expression of this long *PHACTR1* transcript in brain. Previous molecular and functional studies have focused on this transcript, which may limit their relevance to understanding the role of *PHACTR1* beyond the nervous system. In cell types involved in atherosclerotic lesions, we found that *PHACTR1* was transcribed from a more distal TSS giving rise to the intermediate transcript. This intermediate transcript was the prevalent protein isoform in immune and endothelial cells and the encoded protein includes the 4 previously characterized actin-binding RPEL domains and NLS regions. In addition, we found transcription of a short transcript from an even more distal TSS. The encoded protein would only contain two RPEL motifs and lack the N-terminal NLS. The expression pattern of the short transcript differed markedly from

the longer forms and expression was only detected in immune cells, principally macrophages and B-lymphocytes. Macrophages were notable as the only cell type with high-level expression of the intermediate and short transcript. The limited distribution in immune cells is consistent with a role for the short transcript in immunity and inflammation. We were unable to detect protein expressed from the endogenous short transcript in the cell types studied despite using an antibody raised against peptides encompassing most of its coding region. This may indicate post-transcriptional effects, rapid turnover or technical limitations of the available antibodies. A striking finding was the absence of PHACTR1 protein in VSMC.

In macrophages, oxLDL exposure caused upregulation of the short transcript and suppression of the intermediate transcript whereas lipopolysaccharide and TNF α caused upregulation of the intermediate transcript and suppression of the short transcript. These findings demonstrate that PHACTR1 expression responds differently to inflammatory stimuli and to lipid or oxidative stress pathways activated by oxLDL. The effect of the lipid stimulus was not dependent on the presence of lipoproteins since synthetic oxidized phospholipids (oxPAPC) produced a similar response to oxLDL. A non-oxidized, non-lipoprotein stimulus, comprising pure cholesterol loading did not affect expression of PHACTR1 suggesting that oxidized lipids are required.

In endothelial cells only the intermediate transcript was expressed and this was upregulated by oxLDL but not by non-oxidized, non-lipoprotein cholesterol loading. In previous studies with umbilical vein endothelial cells TNF α had no effect on PHACTR1 expression [15,23]. However, we found that PHACTR1 was upregulated by TNF α in primary aortic arterial endothelial cells, indicating the difference between adult arterial endothelial cells and neonatal umbilical cord vein endothelial cells. These findings raise the possibility that oxLDL and TNF α could influence angiogenesis or atherosclerosis by affecting PHACTR1 expression in arterial endothelial cells. A previous study in HUVECs reported upregulation of PHACTR1 in response to the classic angiogenic factor, vascular endothelial growth factor [36].

The rs9349379 SNP is robustly associated with CAD and a fine mapping study suggested it was the likely causal variant in the region [15]. Previous eQTL studies have not taken account of the different PHACTR1 transcripts and so have not been informative about each transcript. A large-scale eQTL study in monocytes used microarray technology with a probe that would have detected multiple PHACTR1 transcripts [37]. An eQTL study of CAD candidate genes, did not detect significant PHACTR1 expression in PBMC, but the Taqman assay used would only assess the long transcript, which is not expressed in these cells [38]. An eQTL for PHACTR1 expression has recently been demonstrated in right coronary artery tissue, but the cellular origin of this effect was unclear due to the mixed cell populations in the tissue samples [15]. In addition, the different transcripts were not distinguished because the qPCR primers annealed to both the long and intermediate transcripts and the short transcript was not studied [15]. Using RNA-seq eQTLs were sought across a panel of human tissues and the rs9349379 SNP was found to be an eQTL for PHACTR1 in coronary artery, tibial artery and aortic artery tissue; the reference allele A was associated with higher PHACTR1 expression than the alternative G allele [16]. In this study, the composite nature of the tissue samples does not allow identification of the cellular origin of the effect and the transcript affected was not reported. The direction of genetic effect is the same as that of the short transcript in the present study. An eQTL study in endothelial cells using primers that would detect only the longer transcripts did not find a significant association with variants at the rs12526453 SNP, which is in linkage disequilibrium with rs9349379 (R^2 with rs9349379 = 0.32) [39]. We have

now established a new association in macrophages between expression of the previously uncharacterized short PHACTR1 transcript and genotype at rs9349379. In PBMCs-depleted of CD14⁺ cells (predominantly lymphocytes) there was a trend towards an eQTL effect in the opposite direction to that in macrophages. A previous study found that 4.4% of eQTL-associated SNPs produced opposite effects in different tissues [40].

Suppression of the short PHACTR1 transcript in macrophages by pro-inflammatory stimuli mirrors the effect of the CAD risk allele (rs9349379-G), which is associated with reduced expression of the short transcript compared to that seen with the protective allele (rs9349379-A). This suggests that having the risk genotype is, in this sense, equivalent to a persistent inflammatory stimulus, such as that caused by LPS. As the short transcript was only expressed in immune cells, these data imply that the short transcript could play a role in the macrophage inflammatory response in atherosclerotic plaques.

The rs9349379 SNP is located about 250 kb upstream from the short PHACTR1 transcript. At other risk loci chromatin looping has been established as a mechanism whereby regulatory elements containing the SNP are brought into proximity with the gene whose expression is altered. The *DEXI* gene, for example, is physically linked by chromatin looping to a type 1 diabetes risk SNP about 150 kb away and the SNP has an eQTL for *DEXI* [41]. We and others have found that GWAS risk SNPs alleles affect enhancer function by altering transcription factor binding affinity at regulatory elements [6,17]. We hypothesize that such a mechanism may be responsible for the effect of rs9349379 and the G allele is predicted to reduce binding of MEF2 transcription factors, that are expressed in macrophages [17,42–44].

Determining the mechanisms underlying the CAD risk conferred by GWAS-identified genetic risk variants is a major challenge for atherosclerosis research, especially for loci containing poorly characterized genes. We establish PHACTR1 as an important gene expressed in atherosclerotic lesions and as the candidate gene at the PHACTR1 CAD locus by demonstrating genetic regulation of a novel transcript and regulation by atherogenic stimuli. Previous molecular studies of PHACTR1 focused on a transcript which we show is not expressed in cells involved in atherosclerotic lesions. We have identified novel intermediate and short length transcripts as the dominant forms expressed in cells involved in atherosclerosis. Our genotype-expression analysis shows that the CAD-associated genetic variant is active in macrophages, the dominant cell type expressing PHACTR1 in human atherosclerotic plaques. The effects of the CAD risk genotype mirror those of an inflammatory stimulus.

Conflict of interest

The authors declared that they do not have anything to disclose regarding conflict of interest with respect to this manuscript.

Financial support

We thank the volunteers from the Oxford Biobank, NIHR Oxford Biomedical Research Centre, for their participation. The Oxford Biobank (www.oxfordbiobank.org.uk) is also part of the NIHR National Bioresearch which supported the recalling process of the volunteers. This work was supported by the Wellcome Trust (097089/Z/11/Z), the Medical Research Council (G116/165), the Novo Nordisk Foundation (Grant Number NNF15CC0018346) and the National Institute for Health Research Oxford Comprehensive Biomedical Research Centre Program.

Author contributions

MER and COC designed the study. MER, DL, AC and EJS performed the experiments. MER and COC wrote the manuscript. All authors analyzed the results and approved the final version of the manuscript.

Appendix A. Supplementary data

Supplementary data related to this article can be found at <http://dx.doi.org/10.1016/j.atherosclerosis.2016.04.025>.

References

- [1] C.K. Glass, J.L. Witztum, Atherosclerosis. the road ahead, *Cell* 104 (2001) 503–516.
- [2] G.K. Hansson, P. Libby, The immune response in atherosclerosis: a double-edged sword, *Nat. Rev. Immunol.* 6 (2006) 508–519.
- [3] S. Sayols-Baixeras, C. Lluis-Ganella, G. Lucas, R. Elosua, Pathogenesis of coronary artery disease: focus on genetic risk factors and identification of genetic variants, *Appl. Clin. Genet.* 7 (2014) 15–32.
- [4] C.A.D. Consortium, P. Deloukas, S. Kanoni, et al., Large-scale association analysis identifies new risk loci for coronary artery disease, *Nat. Genet.* 45 (2013) 25–33.
- [5] M. Nikpay, A. Goel, H.H. Won, et al., A comprehensive 1,000 Genomes-based genome-wide association meta-analysis of coronary artery disease, *Nat. Genet.* 47 (2015) 1121–1130.
- [6] K. Musunuru, A. Strong, M. Frank-Kamenetsky, et al., From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus, *Nature* 466 (2010) 714–719.
- [7] O. Harismendy, D. Notani, X. Song, et al., 9p21 DNA variants associated with coronary artery disease impair interferon-gamma signalling response, *Nature* 470 (2011) 264–268.
- [8] C.L. Miller, U. Haas, R. Diaz, et al., Coronary heart disease-associated variation in TCF21 disrupts a miR-224 binding site and miRNA-mediated regulation, *PLoS Genet.* 10 (2014) e1004263.
- [9] X. Lu, L. Wang, S. Chen, et al., Genome-wide association study in Han Chinese identifies four new susceptibility loci for coronary artery disease, *Nat. Genet.* 44 (2012) 890–894.
- [10] J. Hager, Y. Kamatani, J.B. Cazier, et al., Genome-wide association study in a Lebanese cohort confirms PHACTR1 as a major determinant of coronary artery stenosis, *PLoS One* 7 (2012) e38663.
- [11] C.J. O'Donnell, M. Kavousi, A.V. Smith, et al., Genome-wide association study for coronary artery calcification with follow-up in myocardial infarction, *Circulation* 124 (2011) 2855–2864.
- [12] Myocardial Infarction Genetics C, S. Kathiresan, B.F. Voight, et al., Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants, *Nat. Genet.* 41 (2009) 334–341.
- [13] R.S. Patel, A.A. Morris, Y. Ahmed, et al., A genetic risk variant for myocardial infarction on chromosome 6p24 is associated with impaired central hemodynamic indexes, *Am. J. Hypertens.* 25 (2012) 797–803.
- [14] S. Debette, Y. Kamatani, T.M. Metso, et al., Common variation in PHACTR1 is associated with susceptibility to cervical artery dissection, *Nat. Genet.* 47 (2015) 78–83.
- [15] M. Beaudoin, R.M. Gupta, H.H. Won, et al., Myocardial infarction-associated SNP at 6p24 interferes with MEF2 binding and associates with PHACTR1 expression levels in human coronary arteries, *Arterioscler. Thromb. Vasc. Biol.* 35 (2015) 1472–1479.
- [16] M. Mele, P.G. Ferreira, F. Reverter, et al., Human genomics. The human transcriptome across tissues and individuals, *Science* 348 (2015) 660–665.
- [17] M.E. Reschen, K.J. Gaulton, D. Lin, et al., Lipid-induced epigenomic changes in human macrophages identify a coronary artery disease-associated variant that regulates PPAP2B expression through Altered C/EBP-beta binding, *PLoS Genet.* 11 (2015) e1005061.
- [18] P.B. Allen, A.T. Greenfield, P. Svenningsson, D.C. Haspeslagh, P. Greengard, Phactrs 1–4: a family of protein phosphatase 1 and actin regulatory proteins, *Proc. Natl. Acad. Sci. U. S. A.* 101 (2004) 7187–7192.
- [19] T. Ota, Y. Suzuki, T. Nishikawa, et al., Complete sequencing and characterization of 21,243 full-length human cDNAs, *Nat. Genet.* 36 (2004) 40–45.
- [20] M. Wiazlak, J. Diring, J. Abella, et al., G-actin regulates the shuttling and PP1 binding of the RPEL protein Phactr1 to control actomyosin assembly, *J. cell Sci.* 125 (2012) 5860–5872.
- [21] G.B. Moorhead, L. Trinkle-Mulcahy, A. Ulke-Lemee, Emerging roles of nuclear protein phosphatases, *Nat. Rev. Mol. cell Biol.* 8 (2007) 234–244.
- [22] L. Korrodi-Gregorio, S.L. Esteves, M. Fardilha, Protein phosphatase 1 catalytic isoforms: specificity toward interacting proteins, *Transl. Res. J. laboratory Clin. Med.* 164 (2014) 366–391.
- [23] B. Allain, R. Jarray, L. Borriello, et al., Neuropilin-1 regulates a new VEGF-induced gene, Phactr-1, which controls tubulogenesis and modulates lamellipodial dynamics in human endothelial cells, *Cell Signal* 24 (2012) 214–223.
- [24] L. Favot, M. Gillingwater, C. Scott, P.R. Kemp, Overexpression of a family of RPEL proteins modifies cell shape, *FEBS Lett.* 579 (2005) 100–104.
- [25] K.J. Moore, F.J. Sheedy, E.A. Fisher, Macrophages in atherosclerosis: a dynamic balance, *Nat. Rev. Immunol.* 13 (2013) 709–721.
- [26] K.J. Moore, I. Tabas, Macrophages in the pathogenesis of atherosclerosis, *Cell* 145 (2011) 341–355.
- [27] G.K. Hansson, A. Hermansson, The immune system in atherosclerosis, *Nat. Immunol.* 12 (2011) 204–212.
- [28] R.J. Havel, H.A. Eder, J.H. Bragdon, The distribution and chemical composition of ultracentrifugally separated lipoproteins in human serum, *J. Clin. Invest.* 34 (1955) 1345–1353.
- [29] F. Consortium, the RP, Clst, et al., A promoter-level mammalian expression atlas, *Nature* 507 (2014) 462–470.
- [30] K.R. Rosenbloom, C.A. Sloan, V.S. Malladi, et al., ENCODE data in the UCSC Genome Browser: year 5 update, *Nucleic Acids Res.* 41 (2013) D56–D63.
- [31] A.D. Watson, J.A. Berliner, S.Y. Hama, et al., Protective effect of high density lipoprotein associated paraoxonase. Inhibition of the biological activity of minimally oxidized low density lipoprotein, *J. Clin. Invest.* 96 (1995) 2882–2891.
- [32] Y. Vengrenyuk, H. Nishi, X. Long, et al., Cholesterol loading reprograms the microRNA-143/145-myocardin axis to convert aortic smooth muscle cells to a dysfunctional macrophage-like phenotype, *Arterioscler. Thromb. Vasc. Biol.* 35 (2015) 535–546.
- [33] T. Devries-Seimon, Y. Li, P.M. Yao, et al., Cholesterol-induced macrophage apoptosis requires ER stress pathways and engagement of the type A scavenger receptor, *J. Cell Biol.* 171 (2005) 61–73.
- [34] Coronary Artery Disease Genetics C, A genome-wide association study in Europeans and South Asians identifies five new loci for coronary artery disease, *Nat. Genet.* 43 (2011) 339–344.
- [35] A. Schober, W. Siess, Lysophosphatidic acid in atherosclerotic diseases, *Br. J. Pharmacol.* 167 (2012) 465–482.
- [36] R. Jarray, S. Pavoni, L. Borriello, et al., Disruption of phactr-1 pathway triggers pro-inflammatory and pro-atherogenic factors: new insights in atherosclerosis development, *Biochimie* 118 (2015) 151–161.
- [37] T. Zeller, P. Wild, S. Szymczak, et al., Genetics and beyond—the transcriptome of human monocytes and disease susceptibility, *PLoS One* 5 (2010) e10693.
- [38] X. Zhang, A.D. Johnson, A.E. Hendricks, et al., Genetic associations with expression for genes implicated in GWAS studies for atherosclerotic cardiovascular disease and blood phenotypes, *Hum. Mol. Genet.* 23 (2014) 782–795.
- [39] A. Erbilgin, M. Civelek, C.E. Romanoski, et al., Identification of CAD candidate genes in GWAS loci and their expression in vascular cells, *J. Lipid Res.* 54 (2013) 1894–1905.
- [40] J. Fu, M.G. Wolfs, P. Deelen, et al., Unraveling the regulatory mechanisms underlying tissue-dependent genetic variation of gene expression, *PLoS Genet.* 8 (2012) e1002431.
- [41] L.J. Davison, C. Wallace, J.D. Cooper, et al., Long-range DNA looping and gene expression analyses identify DEXI as an autoimmune disease candidate gene, *Hum. Mol. Genet.* 21 (2012) 322–333.
- [42] L.D. Ward, M. Kellis, HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants, *Nucleic Acids Res.* 40 (2012) D930–D934.
- [43] A.P. Boyle, E.L. Hong, M. Hariharan, et al., Annotation of functional variation in personal genomes using RegulomeDB, *Genome Res.* 22 (2012) 1790–1797.
- [44] C. Aude-Garcia, V. Collin-Faure, H. Bausinger, D. Hanau, T. Rabilloud, C. Lemerrier, Dual roles for MEF2A and MEF2D during human macrophage terminal differentiation and c-Jun expression, *Biochem. J.* 430 (2010) 237–244.

Genetic fine mapping and genomic annotation defines causal mechanisms at type 2 diabetes susceptibility loci

We performed fine mapping of 39 established type 2 diabetes (T2D) loci in 27,206 cases and 57,574 controls of European ancestry. We identified 49 distinct association signals at these loci, including five mapping in or near *KCNQ1*. ‘Credible sets’ of the variants most likely to drive each distinct signal mapped predominantly to noncoding sequence, implying that association with T2D is mediated through gene regulation. Credible set variants were enriched for overlap with FOXA2 chromatin immunoprecipitation binding sites in human islet and liver cells, including at *MTNR1B*, where fine mapping implicated rs10830963 as driving T2D association. We confirmed that the T2D risk allele for this SNP increases FOXA2-bound enhancer activity in islet- and liver-derived cells. We observed allele-specific differences in NEUROD1 binding in islet-derived cells, consistent with evidence that the T2D risk allele increases islet *MTNR1B* expression. Our study demonstrates how integration of genetic and genomic information can define molecular mechanisms through which variants underlying association signals exert their effects on disease.

Genome-wide association studies (GWAS) of common variants, defined by minor allele frequency (MAF) $\geq 5\%$, have been successful in identifying loci contributing to T2D susceptibility^{1–5}. GWAS-identified loci are typically represented by a ‘lead’ SNP with the strongest signal of association in the region. However, lead SNPs may not directly influence disease susceptibility but instead be proxies for causal variants because of linkage disequilibrium (LD). Interpretation may be further complicated by the presence of more than one causal variant at a locus, possibly acting through the joint effects of alleles on the same haplotype. This complex genetic architecture would result in multiple ‘distinct’ association signals at the same locus, which could only be delineated, statistically, through conditional analyses.

With the exception of loci where the lead SNP is a protein-altering variant, including *PPARG*⁶, *KCNJ11-ABCC8* (ref. 7), *SLC30A8* (ref. 8) and *GCKR*⁹, the mechanisms by which associated alleles influence T2D susceptibility are largely unknown. At other loci, direct biological interpretation of the effect of genetic variation on T2D is more challenging because the association signals mostly map to noncoding sequence. Although recent reports have demonstrated a relationship between T2D-associated variants and transcriptional enhancer activity, particularly in human pancreatic islets, liver cells, adipose tissue and muscle^{10–14}, the DNA-binding proteins through which these effects are mediated remain obscure. Localization of noncoding causal variants may highlight the specific regulatory elements they perturb and potentially the genes through which they operate, providing valuable insights into the pathophysiological basis of T2D susceptibility at GWAS-identified loci.

To improve the localization of potential causal variants for T2D and to characterize the mechanisms through which these variants alter disease risk, we performed comprehensive fine mapping of 39 established susceptibility loci through high-density imputation into 27,206 cases and 57,574 controls from 23 studies of European ancestry

genotyped with the MetaboChip¹⁵ (Supplementary Tables 1 and 2). Within each locus, we aimed to (i) evaluate the evidence for multiple distinct association signals through conditional analyses; (ii) undertake fine mapping by defining credible sets of variants that account for $\geq 99\%$ of the probability of driving each distinct association signal; and (iii) interrogate credible sets for functional and regulatory annotation to provide insight into the mechanisms through which variants driving association signals influence disease risk.

RESULTS

Imputation into MetaboChip fine-mapping regions

The MetaboChip includes high-density coverage of 257 ‘fine-mapping regions’ that have previously been associated with 23 metabolic, cardiovascular and anthropometric traits¹⁵. SNPs in these regions were selected using reference data from the HapMap Project¹⁶ and the 1000 Genomes Project¹⁷. At design, 27 T2D susceptibility loci were selected for fine mapping. However, subsequent T2D GWAS efforts have identified additional loci that overlap 12 further fine-mapping regions that were initially selected for other traits (Supplementary Table 3). To enhance coverage of variation in the fine-mapping regions, we undertook imputation of the MetaboChip scaffold up to the 1000 Genomes Project phase 1 integrated reference panel (March 2012 release)¹⁸, including multi-ancestry haplotypes to reduce error rates¹⁹ (Online Methods).

The quality of imputation was variable across studies, particularly for variants with MAF $< 5\%$, and was dependent on the scaffold sample size (Supplementary Table 4). We defined variants to be ‘well imputed’ at widely used thresholds²⁰ of IMPUTEv2 (ref. 21) $\text{info} \geq 0.4$ or minimac²² $r^2 \geq 0.3$ in at least 80% of the total effective sample size ($N_e \geq 59,122$) across studies. With this definition, 99.4% and 89.0%, respectively, of common and low-frequency ($0.5\% \leq \text{MAF} < 5\%$) variants in 1000 Genomes Project European-ancestry haplotypes were

A full list of authors and affiliations appears at the end of the paper.

Received 1 December 2014; accepted 7 October 2015; published online 9 November 2015; doi:10.1038/ng.3437

Table 1 Established T2D susceptibility loci with multiple distinct signals of association at locus-wide significance in the GCTA joint regression model ($P_j < 1 \times 10^{-5}$)

Locus	Index variant	Chr.	Position (Build 37)	Risk allele	Other allele	MetaboChip GCTA joint model 27,206 cases and 57,574 controls			Validation GCTA joint model 19,662 cases and 115,140 controls			Combined GCTA joint model 46,868 cases and 172,714 controls		
						RAF	OR (95% CI)	P _j	RAF	OR (95% CI)	P _j	RAF	OR (95% CI)	P _j
DGKB	rs10276674	7	14,922,007	C	T	0.183	1.08 (1.04–1.11)	4.5 × 10 ^{−6}	0.216	1.09 (1.05–1.12)	1.3 × 10 ^{−6}	1.08 (1.06–1.11)	2.8 × 10 ^{−11}	
	rs1974620	7	15,065,467	T	C	0.519	1.06 (1.04–1.09)	1.6 × 10 ^{−6}	0.515	1.05 (1.03–1.08)	0.00014	1.06 (1.04–1.08)	1.0 × 10 ^{−9}	
CDKN2A	rs10811660	9	22,134,068	G	A	0.830	1.32 (1.27–1.38)	2.4 × 10 ^{−44}	0.817	1.21 (1.17–1.26)	2.6 × 10 ^{−21}	1.27 (1.23–1.30)	1.1 × 10 ^{−61}	
CDKN2B	rs10757283	9	22,134,172	T	C	0.437	1.14 (1.10–1.17)	7.4 × 10 ^{−18}	0.436	1.11 (1.07–1.14)	1.3 × 10 ^{−10}	1.12 (1.10–1.14)	3.6 × 10 ^{−26}	
KCNQ1	chr11:2692322:D	11	2,692,322	D	R	0.374	1.08 (1.05–1.10)	3.5 × 10 ^{−8}	0.413	1.09 (1.06–1.12)	1.2 × 10 ^{−8}	1.08 (1.06–1.10)	2.3 × 10 ^{−15}	
	rs2283220	11	2,755,548	A	G	0.661	1.06 (1.03–1.09)	0.000016	0.710	1.05 (1.02–1.08)	0.0031	1.06 (1.03–1.08)	2.4 × 10 ^{−7}	
rs2237895	rs2237895	11	2,857,194	C	A	0.428	1.08 (1.05–1.11)	6.6 × 10 ^{−7}	0.433	1.07 (1.03–1.10)	2.8 × 10 ^{−4}	1.07 (1.05–1.10)	5.3 × 10 ^{−10}	
	rs74046911	11	2,858,636	C	T	0.951	1.32 (1.24–1.40)	1.7 × 10 ^{−17}	0.943	1.25 (1.17–1.34)	4.8 × 10 ^{−10}	1.29 (1.23–1.35)	9.6 × 10 ^{−26}	
rs458069	rs458069	11	2,858,800	G	C	0.707	1.06 (1.03–1.10)	0.00026	0.707	1.07 (1.03–1.11)	0.00085	1.06 (1.04–1.09)	1.0 × 10 ^{−6}	
	rs1169288	12	121,416,650	C	A	0.334	1.10 (1.07–1.13)	5.4 × 10 ^{−10}	0.316	1.08 (1.05–1.12)	2.8 × 10 ^{−6}	1.09 (1.07–1.12)	8.1 × 10 ^{−15}	
rs1800574	rs1800574	12	121,416,864	T	C	0.027	1.21 (1.11–1.31)	5.2 × 10 ^{−6}	0.020	1.23 (1.12–1.35)	0.000026	1.22 (1.14–1.29)	5.1 × 10 ^{−10}	
	chr12:121440833:D	12	121,440,833	R	D	0.416	1.06 (1.03–1.09)	0.000028	0.382	1.08 (1.04–1.11)	2.5 × 10 ^{−6}	1.07 (1.05–1.09)	2.9 × 10 ^{−10}	
rs17066842	chr18:57739289:D	18	57,739,289	D	R	0.234	1.05 (1.02–1.09)	0.00079	0.254	1.07 (1.03–1.10)	0.000059	1.06 (1.04–1.08)	1.9 × 10 ^{−7}	
	rs17066842	18	58,040,624	G	A	0.961	1.13 (1.06–1.21)	0.00033	0.948	1.11 (1.04–1.19)	0.0012	1.12 (1.07–1.17)	1.4 × 10 ^{−6}	
rs4399645	rs4399645	19	46,166,073	T	C	0.395	1.07 (1.04–1.10)	4.4 × 10 ^{−7}	0.441	1.05 (1.01–1.08)	0.0046	1.06 (1.04–1.08)	1.4 × 10 ^{−8}	
	rs2238689	19	46,178,661	C	T	0.425	1.09 (1.07–1.12)	9.7 × 10 ^{−12}	0.424	1.07 (1.04–1.10)	9.0 × 10 ^{−6}	1.08 (1.06–1.11)	8.3 × 10 ^{−16}	
rs1800961	rs1800961	20	43,042,364	T	C	0.034	1.16 (1.09–1.24)	0.000011	0.041	1.16 (1.08–1.25)	0.000051	1.16 (1.10–1.22)	2.3 × 10 ^{−9}	

Each distinct association signal was represented by an index variant in the GCTA joint regression model based on (i) summary statistics from a combined meta-analysis of 46,868 cases and 172,714 controls of European ancestry and (ii) reference genotype data from GoDARTS (3,298 cases and 3,708 controls of European ancestry from the UK) to approximate LD across fine-mapping regions. Insertion-deletion alleles are coded as reference (R) or deletion (D). Chr., chromosome; RAF, risk allele frequency; OR, odds ratio for the risk allele; CI, confidence interval.

^aThe previously reported T2D GWAS SNP at the *HNF4A* locus (rs4812829) is not included in the fine-mapping region. However, the reported index variant, rs1800961, is independent of the GWAS-identified SNP and thus represents a new, distinct association signal at this locus.

well imputed and therefore retained for downstream association analyses. Within studies, imputation quality was consistent across loci, despite the differential priority of fine-mapping regions and their coverage of variation at design (**Supplementary Table 5**). Thus, 1000 Genomes Project imputation into the MetaboChip scaffold provides nearly complete coverage of common and low-frequency variation across the 39 T2D susceptibility loci and supports direct interrogation of the majority of variants with MAF $\geq 0.5\%$ in European-ancestry populations.

Distinct association signals at T2D susceptibility loci

The first step in fine mapping GWAS loci is to delineate distinct association signals arising from multiple causal variants in the same region, which can efficiently be achieved through approximate conditioning with GCTA²³. Within each T2D fine-mapping region, we identified distinct signals attaining ‘locus-wide’ significance (represented by an index variant with $P_j < 1 \times 10^{-5}$ in the joint association model) by applying GCTA in two stages (Online Methods). First, we selected index variants on the basis of fixed-effects meta-analysis across MetaboChip studies. Second, we performed *in silico* replication of the index variants in a validation meta-analysis of an additional 19,662 T2D cases and 115,140 controls from ten GWAS of European ancestry (**Supplementary Tables 1, 2 and 6**). Finally, because GCTA is only an approximation, we confirmed the association of each index variant through exact conditional analysis across MetaboChip studies (Online Methods and **Supplementary Table 7**).

The most dramatic delineation of distinct association signals was observed for the region flanking *KCNQ1*, where five non-coding index variants attained locus-wide significance (**Table 1** and **Supplementary Fig. 1**). Distinct association signals represented by three of the index variants have been reported in previous GWAS of European⁴ and East Asian²⁴ ancestry: rs74046911 ($P_j = 3.6 \times 10^{-26}$, $r^2 = 0.98$ with the East Asian lead SNP, rs2237897) and rs2237895 ($P_j = 2.1 \times 10^{-9}$, $r^2 = 0.75$ with one European lead SNP, rs163184), both of which map to a <50-kb intronic recombination interval of *KCNQ1*, and chr11:2692322:D ($P_j = 7.2 \times 10^{-16}$, $r^2 = 0.59$ with a second European lead SNP, rs231361), which resides in the *KCNQ1OT1* long noncoding RNA gene that controls regional imprinting²⁵. The two remaining distinct association signals at this locus are new. The first, indexed by rs458069 ($P_j = 3.2 \times 10^{-6}$), maps to the same <50-kb recombination interval as rs74046911 and rs2237895 but is in only weak LD with both ($r^2 = 0.02$ and 0.25 , respectively). The second, indexed by rs2283220 ($P_j = 2.2 \times 10^{-7}$), resides in a neighboring intron of *KCNQ1*, outside of the <50-kb recombination interval (**Supplementary Fig. 1**).

At the *HNF1A* locus, we observed three distinct association signals (**Table 1** and **Supplementary Fig. 2**), represented by index variants that are in only weak LD with the previously reported lead GWAS SNP, rs12427353. The index variants include two nonsynonymous variants, rs1169288 ($P_j = 4.4 \times 10^{-14}$, $r^2 = 0.09$, encoding HNF1A p.Ile27Leu) and rs1800574 ($P_j = 4.2 \times 10^{-10}$, $r^2 = 0.01$, encoding HNF1A p.Ala98Val), and one intergenic SNP, chr12:121440833:D ($P_j = 2.9 \times 10^{-10}$, $r^2 = 0.19$).

We also observed four loci that each had two distinct association signals (*CDKN2A-CDKN2B*, *DGKB*, *MC4R* and *GIPR*), with each locus represented by noncoding index variants (**Table 1** and **Supplementary Fig. 3**). The index variants at the *CDKN2A-CDKN2B* locus represent the known T2D haplotype association signal mapping to a 12-kb intergenic recombination interval^{26–28}. Previous European-ancestry GWAS meta-analyses⁴ have highlighted

Table 2 Distinct association signals at established T2D susceptibility loci for which the 99% credible set contains no more than ten variants

Locus	Index variant	Chr.	Position (Build 37)	Risk allele	Other allele	RAF	P value	OR (95% CI)	99% credible set		
									SNPs	Interval length (bp)	Interval position (bp)
<i>MTNR1B</i>	rs10830963	11	92,708,710	G	C	0.283	2.9×10^{-12}	1.10 (1.07–1.13)	1	1	92,708,710–92,708,710
<i>TCF7L2</i>	rs7903146	10	114,758,349	T	C	0.260	5.8×10^{-120}	1.39 (1.35–1.43)	3	4,279	114,754,071–114,758,349
<i>KCNQ1</i>	rs74046911	11	2,858,636	C	T	0.951	5.9×10^{-18}	1.33 (1.25–1.42)	3	197	2,858,440–2,858,636
<i>ZBED3</i>	rs7732130	5	76,435,004	G	A	0.278	6.4×10^{-10}	1.09 (1.06–1.12)	5	10,056	76,424,949–76,435,004
<i>CDKN2A- CDKN2B</i>	rs10757283	9	22,134,172	T	C	0.437	2.8×10^{-19}	1.14 (1.11–1.18)	5	1,007	22,133,645–22,134,651
<i>SLC30A8</i>	rs13266634	8	118,184,783	C	T	0.676	1.3×10^{-18}	1.13 (1.10–1.16)	6	33,133	118,184,783–118,217,915
<i>CDKN2A- CDKN2B</i>	rs10811660	9	22,134,068	G	A	0.830	7.0×10^{-43}	1.32 (1.27–1.37)	6	1,397	22,132,698–22,134,094
<i>HNF1B</i>	rs4430796	17	36,098,040	G	A	0.455	6.3×10^{-12}	1.09 (1.07–1.12)	7	5,791	36,097,775–36,103,565
<i>CDKAL1</i>	rs35261542	6	20,675,792	A	C	0.280	9.6×10^{-23}	1.15 (1.12–1.18)	8	30,073	20,673,880–20,703,952
<i>GLIS3</i>	chr9:4294707:I	9	4,294,707	I	R	0.360	6.5×10^{-8}	1.07 (1.05–1.10)	10	15,453	4,283,137–4,298,589

Association summary statistics and credible set construction are based on the meta-analysis of Metabochip studies in 27,206 cases and 57,574 controls of European ancestry. In loci with multiple distinct signals of association, results are presented from exact conditional analysis after adjusting for all other index variants in the fine-mapping region. In loci with a single signal of association, results are presented from unconditional analysis. Chr., chromosome; RAF, risk allele frequency; OR, odds ratio for the risk allele; CI, confidence interval.

a potential distinct association signal, located upstream of the recombination interval in the noncoding *CDKN2B-AS1* (*ANRIL*) transcript. However, our conditional analyses indicate that the association in this region can be fully explained by the two index SNPs in the recombination interval, which, when considered together, fully extinguish the *CDKN2B-AS1* signal (Supplementary Fig. 4). The index variants at *DGKB* and *MC4R* also confirm previously reported distinct association signals at these loci in European-ancestry GWAS meta-analyses⁴. At the *GIPR* locus, the two index variants (rs2238689, $P_j = 8.3 \times 10^{-16}$; rs4399645, $P_j = 1.4 \times 10^{-8}$) are not in strong LD with the previously reported⁴ lead SNP (rs8108269; $r^2 = 0.43$ with rs2238689 and 0.00 with rs4399645) but together can better explain the T2D association signal in this region.

Finally, we observed a new distinct association signal at the *HNF4A* locus, represented by the coding index variant rs1800961 ($P_j = 1.4 \times 10^{-9}$, encoding HNF4A p.Thr139Ile, referred to as p.Thr130Ile in some previous studies²⁹). Unfortunately, this fine-mapping region was included on the Metabochip for high-density lipoprotein cholesterol^{15,30} (Supplementary Table 3) and does not include the previously reported⁴ lead T2D SNP at this locus, rs4812829, precluding conditional analyses in these data. However, rs4812829 is not in LD with our index variant ($r^2 = 0.02$), suggesting that there are at least two distinct T2D association signals at the *HNF4A* locus.

Of the 49 distinct association signals achieving locus-wide significance across the T2D loci represented on the Metabochip (five at *KCNQ1*, three at *HNF1A*, two each at *CDKN2A-CDKN2B*, *DGKB*, *MC4R* and *GIPR*, and one each at the remaining loci), only three index variants are not common (Supplementary Fig. 5 and Supplementary Table 6): rs1800574 (MAF = 2.2%, odds ratio (OR) = 1.21) for one signal at the *HNF1A* locus; rs1800961 (MAF = 3.9%, OR = 1.16) at the *HNF4A* locus; and rs17066842 (MAF = 4.8%, OR = 1.12) for one signal at the *MC4R* locus.

Localizing variants driving T2D association signals

We used statistical evidence of association from the meta-analysis of Metabochip studies to construct 99% credible sets of variants²⁸ that are most likely to drive the 49 distinct signals (Online Methods, Supplementary Fig. 6 and Supplementary Table 8). For ten distinct association signals mapping to nine loci, the 99% credible set included no more than ten variants (Table 2 and Supplementary Table 9).

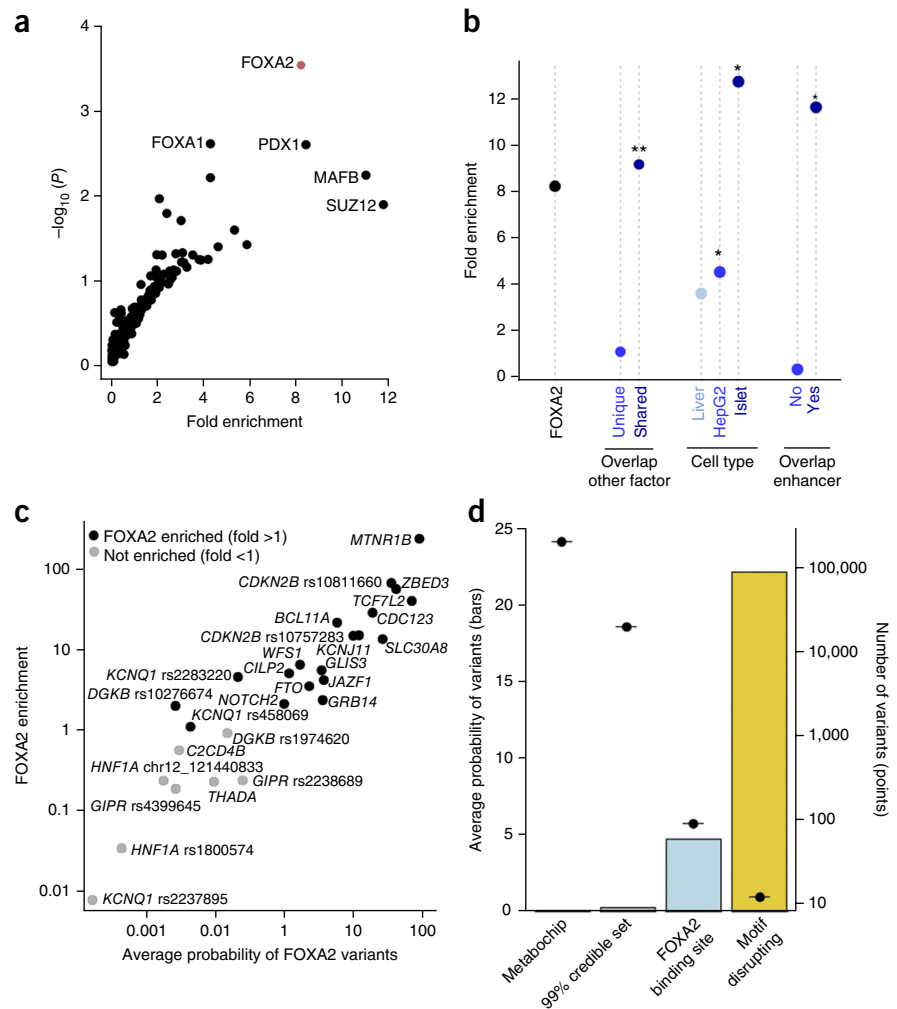
The greatest refinement was observed at the *MTNR1B* locus, where the credible set included only the index variant, rs10830963, accounting for more than 99.8% of the posterior probability of driving the association signal (π_C). Small credible sets were also observed for the association at *TCF7L2* (three variants, indexed by rs7903146, mapping to 4.3 kb) and one signal at *KCNQ1* (three variants, indexed by rs74046911, mapping to just 200 bp). The 99% credible sets for both distinct association signals at *CDKN2A-CDKN2B* together included just 11 variants in total and map to less than 2 kb.

We performed functional annotation of credible set variants to search for evidence that association signals are driven by coding alleles. Across the 49 signals, only nine coding variants attained $\pi_C > 1\%$ (Supplementary Table 10), including six previously reported nonsynonymous T2D risk alleles at *PPARG*⁶, *KCNJ11-ABCC8* (refs. 7,31,32), *SLC30A8* (refs. 8,33) and *GCKR*^{9,34}. The remaining three coding alleles were the index variants for association signals mapping to *HNF4A* (rs1800961, encoding p.Thr139Ile, $\pi_C = 97.4\%$) and *HNF1A* (rs1169288, encoding p.Ile27Leu, $\pi_C = 75.5\%$; rs1800574, encoding p.Ala98Val, $\pi_C = 34.0\%$). Our findings are supported by earlier studies, which reported nominal evidence for association of these three coding variants with T2D and defects in insulin secretion *in vivo* and demonstrated reduced transcriptional activity of *HNF1A* target genes using *in vitro* assays^{29,35}. These data provide strong evidence that *HNF4A* and *HNF1A* are T2D effector transcripts at these loci, a view further supported by the known impact of rare, loss-of-function mutations in these genes on maturity-onset diabetes of the young^{36,37}. Given the nearly complete coverage of common and low-frequency variants in fine-mapping regions after 1000 Genomes Project imputation, it is unlikely that additional distinct signals in the established T2D susceptibility loci represented on the Metabochip are driven by coding variation with MAF $\geq 0.5\%$, confirming reports that these associations are most likely mediated by effects on gene regulation^{10,13,14,38}.

Regulatory mechanisms underlying T2D association signals

We sought to understand the regulatory mechanisms through which variants at the 39 established T2D susceptibility loci influence disease by intersecting the 99% credible sets for each distinct association signal with chromatin immunoprecipitation sequence (ChIP-seq) data for 165 transcription factors, chromatin state maps from 12 cell types and long noncoding RNA transcripts from 25 cell types (Online

Figure 1 FOXA2-bound sites are a genomic marker of T2D risk variants. (a) Variants in ChIP-seq binding sites for 165 proteins were tested for enrichment of posterior probabilities of driving association signals in comparison to variants in shifted sites. Variants in FOXA2 ChIP-seq sites were significantly enriched ($P < 0.00030$). (b) FOXA2 ChIP-seq sites were partitioned on the basis of overlap with other genomic features. There was stronger enrichment for (i) FOXA2 sites overlapping a ChIP-seq site for another protein in comparison to unique sites; (ii) sites identified in primary islets in comparison to HepG2 or primary liver cells; and (iii) sites that overlapped islet enhancers in comparison to those that did not ($**P < 0.00030$, $*P < 0.05$). (c) Variants at each signal were tested for enrichment for FOXA2 binding sites. Nineteen signals had greater enrichment than expected in comparison to shifted sites; at 15 signals, this enrichment was nominally significant ($P < 0.05$). (d) FOXA2-bound variants disrupting recognition motifs have an increased probability of being causal.



Methods and **Supplementary Table 11**). We applied an enrichment procedure that compared the mean posterior probability of driving the association signal for credible set variants directly overlapping sites for each regulatory annotation with a null distribution obtained from randomly shifted site locations within 100 kb in either direction of the original element location.

We first applied this procedure to chromatin state and noncoding RNA elements using the 19,266 credible set variants for all 49 distinct association signals (**Supplementary Fig. 7**). Using Bonferroni correction for the 37 tested cell type annotations ($P < 0.0014$), variants in pancreatic islet enhancer elements¹⁴ had significantly higher posterior probability of driving association signals than expected from the null distribution (1.97-fold enrichment, $P = 0.00022$). We also observed nominal evidence for enrichment of the posterior probability of driving association signals among variants in human islet and hepatocellular carcinoma (HepG2) promoter elements^{10,14} ($P = 0.0052$ and 0.0064 , respectively). However, there was no corresponding enrichment of variants in regulatory elements for other cell types or in noncoding transcripts. These results are consistent with previous studies supporting a contribution of regulatory enhancer and promoter variants to T2D susceptibility in specific cell types^{11–14}.

We next sought to gain insight into the transcription factors these regulatory variants perturb and applied the same procedure to ChIP-seq binding data for 165 proteins (**Fig. 1** and **Supplementary Fig. 8**). Using Bonferroni correction for the 165 proteins tested ($P < 0.00030$), the 89 credible set variants overlapping 57 FOXA2 ChIP-seq binding sites, assayed in human HepG2 (ref. 10) and islet¹⁴ cells, had significantly higher posterior probability of driving association signals than expected from the null distribution (8.24-fold enrichment, $P = 0.00028$). The enrichment of FOXA2 ChIP-seq sites was exclusive to sites shared with at least one other factor (9.18-fold enrichment, $P = 0.00028$) in comparison to those that were not shared (1.12-fold enrichment, $P = 0.11$). Enrichment for FOXA2 binding was also more

pronounced among sites identified in pancreatic islets (15.43-fold enrichment, $P = 0.00045$) than in those identified in HepG2 cells (4.55-fold enrichment, $P = 0.011$). To exclude the possibility that the enrichment in HepG2 cells was driven by artifacts due to HepG2 being a cultured cell line, we compared the FOXA2 sites in HepG2 cells to those previously assayed in primary liver³⁹. We observed significant intersection of the HepG2 and liver FOXA2 sites that overlapped credible set variants ($P = 1.5 \times 10^{-9}$). Consequently, we detected similar FOXA2 binding site enrichment among sites detected in liver (3.63-fold enrichment, $P = 0.061$) to that observed in HepG2 cells. We also compared FOXA2 ChIP-seq sites, across the genome, from liver, HepG2 and islet cells (**Supplementary Fig. 9**). The number of sites varied across the cell types (8,023 for liver, 40,866 for HepG2 cells and 27,291 for islets), likely owing, in part, to technical differences, including in sequencing platform, depth and read length. However, the intersection of FOXA2 sites between each pair of cell types was highly significant ($P < 2.2 \times 10^{-16}$).

Given the preponderance of T2D-associated variants in islet enhancers, we next tested to what extent enrichment of FOXA2 binding is driven by colocalization of variants with these genomic features¹⁴. Variants in FOXA2-bound sites were not enriched for posterior probability of driving association signals after removing enhancer sites (0.36-fold enrichment, $P = 0.69$). Conversely, variants in islet enhancers remained nominally enriched when FOXA2-bound sites were removed (1.65-fold enrichment, $P = 0.014$). These results suggest that FOXA2 binding assayed by ChIP-seq, at a

Table 3 Motif-altering credible set variants in FOXA2 sites

Locus	Index variant	Motif-altering variant	Chr.	Position (Build 37)	Posterior probability (π_C)	Motif allele	Chromatin state
<i>MTNR1B</i>	rs10830963	rs10830963	11	92,708,710	0.998	G	Islet enhancer, HepG2 enhancer
<i>TCF7L2</i>	rs7903146	rs7903146	10	114,758,349	0.78	T	Islet enhancer
<i>SLC30A8</i>	rs13266634	rs13266634	8	118,184,783	0.29	T	Islet enhancer
<i>CDKN2B</i>	rs10811660	rs10811660	9	22,134,068	0.24	A	Islet enhancer
<i>CDC123</i>	rs11257658	rs11257655	10	12,307,894	0.21	T	Islet enhancer, HepG2 enhancer
<i>JAZF1</i>	rs1513272	rs849133	7	28,192,280	0.042	T	Islet enhancer
<i>KCNQ1</i>	rs2283220	rs231907	11	2,752,130	0.031	T	HepG2 enhancer
<i>FTO</i>	rs9927317	rs9940128	16	53,800,754	0.027	G	Islet enhancer, HepG2 enhancer
<i>FTO</i>	rs9927317	rs9939973	16	53,800,568	0.025	G	Islet enhancer, HepG2 enhancer
<i>KCNQ1</i>	rs458069	rs78688069	11	2,752,183	0.0006	A	HepG2 enhancer
<i>KCNQ1</i>	rs458069	rs190728714	11	2,813,084	0.00042	G	Islet enhancer
<i>DGKB</i>	rs10276674	rs7798360	7	15,055,972	0.00005	G	–

Chr., chromosome.

subset of enhancer element locations that are often shared by other proteins, is a genomic marker of variants with an increased posterior probability of driving T2D association signals.

Having demonstrated global over-representation for FOXA2 ChIP-seq binding by considering all loci simultaneously, we applied the same procedure to the 99% credible sets of each distinct association signal separately to identify those with the strongest evidence for local enrichment (Fig. 1). We observed over-representation of credible set variants in islet or HepG2 FOXA2-bound sites for 19 association signals, 15 of which attained nominal significance ($P < 0.05$). A total of 41 credible set variants at these 19 distinct association signals overlapped a FOXA2 ChIP-seq site in at least one of the two cell types (Supplementary Table 12). Of these, 12 variants were predicted to disrupt *de novo* recognition motifs (for FOXA2 and other factors) that were enriched in FOXA2-bound sequence (Table 3 and Supplementary Table 13). The mean posterior probability of driving the association (π_C) for these 12 variants was 22.0% on the basis of genetic fine mapping (Fig. 1), more than four times greater than for

those in FOXA2 ChIP-seq sites that were not motif disrupting at the same signals (mean $\pi_C = 5.2\%$, $P = 0.024$). Furthermore, 11 of these 12 variants also overlapped an enhancer element in islets (nine variants) or HepG2 cells (six variants), indicating that they are in transcriptionally active regions (Table 3). They include two variants with experimentally validated differences in regulatory activity: rs7903146 ($\pi_C = 77.6\%$) at *TCF7L2* (ref. 40) and rs11257655 ($\pi_C = 21.1\%$) at *CDC123* (ref. 41). They also include rs10830963, the index variant at the *MTNR1B* locus, which accounts for 99.8% of the posterior probability of driving the association signal on the basis of genetic fine mapping. These results suggest that FOXA2 binding patterns can be used to highlight specific variants that are potentially causal for T2D susceptibility through altered regulatory binding.

Altered regulatory activity of the *MTNR1B* credible variant

To demonstrate how local enrichment of FOXA2 binding can be used to highlight regulatory mechanisms through which credible set variants might influence T2D susceptibility, we focused on the *MTNR1B*

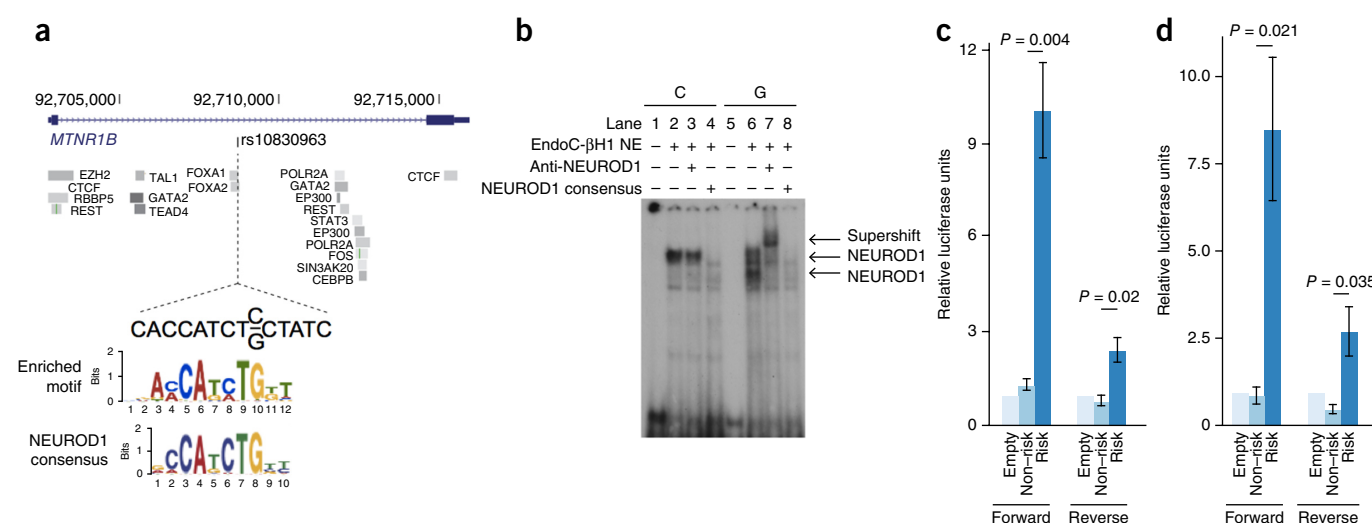


Figure 2 The lone variant in the 99% credible set at the *MTNR1B* locus affects FOXA2-bound enhancer activity. (a) The intronic variant rs10830963 has 99.8% posterior probability of driving the association signal at the *MTNR1B* locus. This variant overlaps a FOXA2 binding site, and the G risk allele is predicted to create a *de novo* recognition motif that closely matches the NEUROD1 consensus motif. (b) EMSA of 25-bp fragments encompassing either allele of rs10830963 in EndoC-βH1 cell extracts. Proteins were bound to both alleles. In the presence of an antibody to NEUROD1, only the signal corresponding to the risk allele was supershifted; in the presence of an unlabeled probe with the NEUROD1 consensus motif, the signal was competed away. NE, nuclear extract. (c,d) The 224-bp sequence surrounding each allele of rs10830963 was cloned into a luciferase reporter construct containing a minimal promoter and tested for luciferase activity in EndoC-βH1 (c) and HepG2 (d) cells ($n = 3$ independent experiments performed in triplicate for each cell type). Results are presented as means $\pm$ s.e.m. The risk allele had significantly increased enhancer activity over the protective allele in both forward and reverse orientations of the element in both cell types.

locus. Variants mapping to this region have among the strongest known effects on both T2D risk⁶ and fasting plasma glucose concentration⁴², and physiological data indicate an effect of *MTNR1B* on both insulin secretion and insulin action⁴³. The lone credible variant in the *MTNR1B* region, rs10830963, overlaps a FOXA2 ChIP-seq binding site, and the risk allele, G, is predicted to create a recognition motif that matches the consensus sequences of NEUROD1 and several other factors (Fig. 2 and Supplementary Table 13). We tested *in silico* predictions of protein binding at rs10830963 via electrophoretic mobility shift assay (EMSA) with 25-bp probe fragments centered on each allele in human pancreatic islet β cell (EndoC- β H1)⁴⁴ or human liver HepG2 cell extracts. We observed allele-specific binding patterns with extracts from both cell lines (Fig. 2 and Supplementary Fig. 10).

To determine the specific protein(s) bound at each allele, we then performed supershift experiments using antibodies directed against NEUROD1, FOXA2 and three other factors (TAL1, PTF1A and YY1) whose consensus binding sequences resemble the recognition motif (Online Methods). We observed a shift in the presence of antibody to NEUROD1 for the signal corresponding to the risk allele in EndoC- β H1 extracts, which could be competed away by an excess of unlabeled probe with the NEUROD1 consensus sequence (Fig. 2). None of the antibodies tested (including to NEUROD1) shifted the signal corresponding to the risk allele in HepG2 cell extracts (Supplementary Fig. 10). These results demonstrate that, *in vitro*, the risk allele of rs10830963 preferentially binds NEUROD1 in islet-derived cells and binds a protein not identified from known recognition motifs in liver-derived cells.

To relate allelic differences in protein binding to genomic activity at this site, we cloned the 224-bp region surrounding rs10830963 into a luciferase reporter vector containing a minimal promoter and tested its enhancer activity in the EndoC- β H1 and HepG2 cell lines. Consistent with *in silico* predictions, we observed a significant ($P < 0.05$) increase in luciferase expression for the element with the risk allele in comparison to the one with the protective allele in both cell lines (Fig. 2). Furthermore, RNA sequencing (RNA-seq) data reported from human islets have linked the T2D risk allele of rs10830963 to increased expression of *MTNR1B*^{45,46}. Taken together, these results suggest that the G allele of rs10830963 increases T2D risk through increased FOXA2-bound enhancer activity, potentially mediated through NEUROD1 binding in islets, and consequently higher expression of *MTNR1B*.

Candidate effector genes at FOXA2-enriched T2D signals

We hypothesized that the locus-specific effects in mouse transcription factor knockout models would mimic patterns of binding enrichment at human disease-associated loci. We thus attempted to relate FOXA2 binding at the 19 FOXA2-enriched association signals (Fig. 1) to target effector genes using previously reported pancreatic islet expression profiles from wild-type and *Foxa1* and *Foxa2* double-null mice⁴⁷ (Online Methods). Syntenic genes mapping within 500 kb of the credible sets for the 19 FOXA2-enriched signals were significantly downregulated (45.2% decrease) in *Foxa1* and *Foxa2* double-knockout mice (Supplementary Fig. 11) in comparison to genes across the genome (0.021% increase; $P = 0.012$), whereas those mapping within 500 kb of the other 30 T2D association signals were not significantly downregulated (2.25% decrease; $P = 0.20$). We observed a consistent downregulation (39.6% decrease) when considering only the genes mapping closest to each FOXA2-enriched signal, in comparison to those across the genome (0.021% increase; $P = 0.0021$). Thus, data related to altered

gene expression in *Foxa1* and *Foxa2* double-knockout mice support patterns of FOXA2 binding site enrichment in humans.

We next identified specific genes at the 19 FOXA2-enriched association signals whose mouse counterparts were downregulated in *Foxa1* and *Foxa2* double-knockout mice, as these might represent effector transcripts for these loci (Supplementary Table 14). Several of these genes have previously been implicated as likely effector transcripts in humans, including *TCF7L2* (refs. 48,49) (57% decrease), *KCNJ11* (refs. 7,50) (38% decrease) and *SLC30A8* (ref. 51) (135% decrease). These data also implicate new candidate effector genes at FOXA2-enriched association signals (Supplementary Table 14). For example, in *Foxa1* and *Foxa2* double-knockout mice, there was marked downregulation of *Reg4* (1,415% decrease), which maps to a syntenic region of the FOXA2-enriched *NOTCH2* GWAS locus, highlighting *REG4* as a likely effector transcript in humans. Additional examples of candidate effector genes include *IGF2* at the *KCNQ1* locus (135% decrease) and *CAMK1D* at the *CDC123* locus (81% decrease). Together, these results provide additional support for the importance of FOXA2 binding at a subset of T2D susceptibility loci and further highlight specific genes through which regulatory variants in these regions may operate.

DISCUSSION

We have undertaken comprehensive fine mapping of 39 established T2D susceptibility loci in 27,206 cases and 57,574 controls of European ancestry and have demonstrated that the existence of multiple distinct association signals in these regions is a common phenomenon. Index variants for just three of the 49 distinct association signals are not common, despite nearly complete coverage of variation with $MAF \geq 0.5\%$ in fine-mapping regions after 1000 Genomes Project imputation. Although we cannot evaluate the impact of rare variation ($MAF < 0.5\%$) in established T2D susceptibility loci without large-scale resequencing, our data strongly argue against a role for low-frequency variants of large effect via synthetic association⁵². We have demonstrated that seven distinct association signals, mapping to six T2D susceptibility loci represented on the Metabochip, are likely to be driven by coding alleles, including new index variants mapping to *HNF1A* and *HNF4A*. Outside of these regions, our fine mapping confirms previous reports that T2D association signals are primarily driven by noncoding alleles, with effects that are mediated through gene regulation^{10,13,14,38}.

We have demonstrated, by genomic annotation and functional assays, that FOXA2 binding assayed by ChIP-seq can be used to pinpoint candidate causal regulatory elements, providing routes to understanding the biology of specific T2D susceptibility loci. These elements highlight variants and effector transcripts through which association signals are mediated, via altered binding of either FOXA2, directly, or another transcription factor. For example, at the *MTNR1B* locus, the risk allele of the lone credible set variant, rs10830963, preferentially binds NEUROD1 in islet-derived cells *in vitro* and increases FOXA2-bound enhancer activity in human islet and liver-derived cells. These data are consistent with previous reports correlating the risk allele with higher *MTNR1B* expression^{45,46} and not loss of function⁵³, and they suggest that altered NEUROD1 binding in islets contributes to T2D susceptibility at this locus. Further experiments will be required to confirm our *in vitro* findings regarding NEUROD1 binding *in vivo*. However, our attempts to perform ChIP-seq in primary islet samples of the defined *MTNR1B* genotype were repeatedly unsuccessful, owing to lack of a suitable antibody to NEUROD1. These studies are further complicated by the limited availability of

primary human islets, and the slow division rate of human islet-derived cell lines is an impediment to the implementation of genome editing technologies.

FOXA2 is a pioneer factor that binds native chromatin and book-marks genomic regions for transcriptional activity⁵⁴, and it is involved in pancreatic and hepatic development^{55,56}. FOXA2 is also expressed in other T2D-relevant cell types, such as adipocytes. Future studies will be required to elucidate the extent to which FOXA2 binding events across cell types influence disease risk. *Foxa2*-null mice have impaired insulin secretion⁴⁷, and common variants at the FOXA2 locus are associated with fasting plasma glucose concentrations^{42,57}. Our findings are thus consistent with the involvement of FOXA2 in maintaining normal glucose homeostasis. Common T2D-associated variants at FOXA2 have also been reported in South Asians⁵⁸, although these variants did not attain genome-wide significant association in the largest GWAS for the disease from multiple ancestry groups^{1–5} and therefore require further replication. Enrichment of FOXA2 binding has also been reported within genomic intervals containing GWAS signals for endocrine, neuropsychiatric, cardiovascular and cancer traits⁵⁹. Our study has the advantage that we consider only the FOXA2 sites that directly overlap variants driving association signals by first fine mapping GWAS loci, thereby providing more targeted credible sets for functional enrichment. Nevertheless, the results of these studies, taken together, suggest a possible role for FOXA2 across a broad spectrum of complex human phenotypes.

In conclusion, we have highlighted that FOXA2 binding patterns can be used to inform future hypothesis-driven investigation of the variants, genes and molecular mechanisms underlying T2D association signals mapping to noncoding sequence. Continued identification of the effector transcripts at these noncoding association signals will require the use of expression quantitative trait locus (eQTL) data and knockout models, in combination with high-throughput experimental data derived from chromatin conformation capture techniques such as capture C. Our findings support the use of transcription factor binding events as a means to partition susceptibility loci, potentially residing in distinct pathways, within disease-relevant cell types. Finally, our study demonstrates the usefulness of fine mapping through integration of genetic and genomic information from relevant tissues and cellular models to elucidate the pathophysiology of complex human diseases, thus offering a promising avenue for the translation of GWAS findings into clinical use.

URLs. DIAGRAM Consortium, <http://diagram-consortium.org/>; Endocells, <http://www.endocells.com/>.

METHODS

Methods and any associated references are available in the [online version of the paper](#).

Accession codes. Association summary statistics are available for download from the DIAGRAM Consortium website (see URLs).

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

Funding for the research undertaken in this study has been received from the Academy of Finland (including grants 77299, 102318, 10493, 118065, 123885, 124243, 129293, 129680, 136895, 139635, 211119, 213506, 251217 and 263836); Agence National de la Recherche; Association de Langue Française pour l'Etude du Diabète et des Maladies Métaboliques; Association Diabète Risque Vasculaire; Association Française des Diabétiques; the Association of Danish Pharmacies; the Augustinus Foundation; the Becket Foundation; the British Diabetes

Association (BDA) Research; the British Heart Foundation; the Central Norway Health Authority; the Central Finland Hospital District; the Center for Inherited Disease Research (CIDR); the City of Kuopio; the City of Leutkirch; Copenhagen County; the Danish Centre for Evaluation and Health Technology Assessment; the Danish Council for Independent Research; the Danish Heart Foundation; the Danish Research Councils; Deutsche Forschungsgemeinschaft (including project ER 155/6-2); the Diabetes Research Foundation; Diabetes UK; the Doris Duke Charitable Foundation; Erasmus Medical Center; Erasmus University; the Estonian government (SF0180142s08); the European Commission (including ENGAGE HEALTH-F4-2007-201413, FP7-201413, FP7-245536, EXGENESIS LSHM-CT-2004-005272, FP6 LSHM_CT_2006_037197, LSHM-CT-2007-037273, Directorate C-Public Health 2004310, DG XII); the European Regional Development Fund; the Federal Ministry of Education and Research, Germany (including FKZ 01GI1128 and FKZ 01EO1001); the Federal Ministry of Health, Germany; the Finnish Diabetes Association; the Finnish Diabetes Research Foundation; the Finnish Foundation for Cardiovascular Research; the Finnish Medical Society; the Folkhalsan Research Foundation; the Foundation for Life and Health in Finland; the Foundation for Old Servants; the Fredrick och Ingrid Thuring Foundation; the French region of Nord-Pas-de-Calais (Contrat de Projets Etat-Région); the German Center for Diabetes Research; the German Research Council (including grant GRK1041); the German National Genome Research Network; Groupe d'Etude des Maladies Métaboliques et Systémiques; the Health Care Centers in Vasa, Närpes and Korsholm, Finland; the Health Foundation; the Heinz Nixdorf Foundation; Helmholtz Zentrum München; the Helsinki University Central Hospital Research Foundation; the Hospital District of Southwest Finland; the Ib Henriksens Foundation; IngaBritt and Arne Lundberg's Research Foundation (including grant 359); Karolinska Institutet; the Knut and Alice Wallenberg Foundation (including grant KAW 2009.0243); Kuopio University Hospital; the Lundbeck Foundation; the Magnus Bergvall Foundation; the Medical Faculty of University Duisburg-Essen; the Medical Research Council, UK (including grants G0000649 and G0601261); the Ministry for Health, Welfare and Sports, the Netherlands; the Ministry of Education and Culture, Finland (including grants 722 and 627; 2004-2011); the Ministry of Education, Culture and Science, the Netherlands; the Ministry of Health and Prevention, Denmark; the Ministry of Social Affairs and Health, Finland; the Ministry of Innovation, Science, Research and Technology of North Rhine-Westphalia, Germany; the Munich Center of Health Sciences; the Municipal Health Care Center and Hospital in Jakobstad, Finland; the municipality of Rotterdam, the Netherlands; the Närpes Health Care Foundation; the National Health Screening Service of Norway; the National Heart, Lung, and Blood Institute, USA (including grant numbers/contracts HHSN268201100005C, HHSN268201100006C, HHSN268201100007C, HHSN268201100008C, HHSN268201100009C, HHSN268201100010C, HHSN268201100011C, HHSN268201100012C, N01HC25195, N02HL64278, R01HL087641, R01HL59367 and R01HL086694); the National Human Genome Research Institute, USA (including grant numbers/contracts U01HG004402 and N01HG65403); the National Institute for Diabetes and Digestive and Kidney Diseases, USA (including grants R01DK078616, U01DK085526, K24DK080140 and R01DK073490); the National Institute for Health and Welfare, Finland; the National Institutes of Health, USA (including grant numbers/contracts HHSN268200625226C, UL1RR025005, R01DK062370, R01DK072193, 1Z01HG000024, AG028555, AG08724, AG04563, AG10175, AG08861, U01HG004399, DK58845, CA055075, DK085545 and DK098032); the Netherlands Genomics Initiative; the Netherlands Organisation for Health Research and Development; the Netherlands Organisation of Scientific Research NOW Investments (including grants 175.010.2005.011, 911-03-012 and 050-060-810); the Nord-Trøndelag County Council; the Nordic Center of Excellence in Disease Genetics; the Norwegian Institute of Public Health; the Norwegian Research Council; the Novo Nordisk Foundation; the Ollquist Foundation; the Oxford National Institute for Health Research (NIHR) Biomedical Research Centre; the Paavo Nurmi Foundation; the Paivikki and Sakari Sohlberg Foundation; the Perklen Foundation; the Pirkanmaa Hospital District, Finland; Programme Hospitalier de Recherche Clinique; Programme National de Recherche sur la Diabète; the Research Institute for Diseases in the Elderly (including grant 014-93-015); the Robert Dawson Evans Endowment, Department of Medicine, Boston University School of Medicine and Boston Medical Center; the Royal Swedish Academy of Sciences; Sarstedt, Germany; the Signe and Ane Gyllenberg Foundation; the Sigrid Juselius Foundation; the Slottery Machine Association, Finland; the Social Insurance Institution of Finland; the South Ostrobothnia Hospital District; the state of Baden-Württemberg, Germany; the Stockholm County Council (including grant 560183); the Swedish Cultural Foundation, Finland; the Swedish Diabetes Foundation; the Swedish e-science Research Center; the Swedish Foundation for Strategic Research; the Swedish Heart-Lung Foundation; the Swedish Research Council (including grants SFO EXODIAB 2009-1039, 521-2010-3490, 521-2007-4037, 521-2008-2974, ANDIS 825-2010-5983, LUDC 349-2008-6589 and 8691); the Swedish Society of Medicine; the Tore

Nilsson Foundation; the Torsten and Ragnar Soderbergs Stiftelser (including grant MT33/09); University Hospital Essen; University of Tromsø; the University College London NIHR Biomedical Research Centre; the UK NIHR Cambridge Biomedical Research Centre; Uppsala University; Uppsala University Hospital; the Vaasa Hospital District; the Velux Foundation; and the Wellcome Trust (including the Biomedical Collections Grant GR072960 and grants 076113, 083948, 090367, 090532, 083270, 086596, 098017, 095101, 098051 and 098381). We are grateful to R. Scharfmann (INSERM U1016, Cochin Institute Paris) for the gift of EndoC β H1 cells and for providing technical support with their maintenance. We thank P. Johnson and the Oxford NIHR Biomedical Research Centre-funded Islet Isolation facility for providing human islets for this study. Detailed acknowledgments are provided in the **Supplementary Note**.

AUTHOR CONTRIBUTIONS

Writing group. K.J.G., T.F., Y. Lee, A.R., R.M., M.E.R., A.L.G., D.A., M. Boehnke, T.M.T., M.I.M., A.P.M. **Central meta-analysis group.** K.J.G., T.F., Y. Lee, R.M., A. Mahajan, A. Locke, N.W.R., N.R., T.M.T., M.I.M., A.P.M. **Annotation and functional analysis group.** K.J.G., A.R., M.E.R., S.K.T., J.K.R., N.L.B., M.v.d.B., A.C., I.D., E. Birney, L. Pasquali, J. Ferrer, C.A.O'C., A.L.G., M.I.M. **Validation meta-analysis group.** R.M., R.A.S., I.P., L.J.S., A.P.M. **MetaboChip cohort-level primary analysis.** Y. Lee, T.G., T.S., D.T., L.Y., H.G., S. Wahl, M.F., R.J.S., H. Kestler, H. Chheda, L.E., S.G., T.M.T., A.P.M. **Validation cohort-level primary analysis.** V. Steinthorsdottir, G.T., L.Q., L.C.K., E.M.v.L., S.M.W., M. Li, H. Chen, C. Fuchsberger, P. Kwan, C.M., M. Linderman, Y. Lu. **MetaboChip design.** H.M.K., B.F.V. **Cohort sample collection, genotyping, phenotyping or additional analysis.** B.F.V., G.R.A., P.A., D.B., B.B., R.B., M. Blüher, H.B., L.L.B., E.P.B., N.P.B., J.C., G.C., P.S.C., M.C.C., D.J.C., A.T.C., R.M.v.D., A.S.F.D., M.D., S.E., J.G.E., T.E., E.E., J. Fadista, J. Flannick, P. Fontanillas, C. Fox, P.W.F., K.G., C.G., B.G., O.G., G.B.G., N.G., C.J.G., M.H., C.T.H., C.H., O.L.H., A.B.H., S.E.H., D.J.H., A.U.J., A.J., M.E.J., T.J., W.-H.L.K., N.D.K., L.K., N.K., A.K., P. Kovacs, P. Kraft, J. Kravic, C. Langford, K.L., L. Liang, P.L., C.M.L., E.L., A. Linneberg, C.-T.L., S.L., J.L., V.L., S. Männistö, O. McLeod, J.M., E.M., G.M., T.W.M., M.M.-N., C.N., M.M.N., N.N.O., K.R.O., D.P., S.P., L. Peltonen, J.R.B.P., C.G.P.P., M.R., D. Ruderfer, D. Rybin, Y.T.v.d.S., B.S., G. Sigurðsson, A.S., G. Steinbach, P.S., K. Strauch, H.M.S., Q.S., B.T., E. Tikkanen, A.T., J. Trakalo, E. Tremoli, T.T., R.W., S. Wiltshire, A.R.W., E.Z. **Validation cohort principal investigators.** R.J.E.L., J.D., J.C.F., E. Boerwinkle, J.S.P., C.v.D., E.S., J.B.M., F.B.H., U.T., K. Stefansson, P.D., P.J.D., T.M.F., A.T.H., I.B., C. Langenberg, N.J.W., M. Boehnke, M.I.M. **MetaboChip cohort principal investigators.** T.A.L., R.R., M.S., N.L.P., L. Lind, S.M.K.-K., E.K.-H., T.E.S., J.S., J. Kuusisto, M. Laakso, A. Metspalu, R.E., K.-H.J., S. Moebus, S.R., V. Salomaa, E.I., B.O.G., R.N.B., F.S.C., K.L.M., H. Koistinen, J. Tuomilehto, K.H., I.N., P.D., P.J.D., T.M.F., A.T.H., U.d.F., A.H., T.I., A.P., S.C., R.S., P. Froguel, O.P., T.H., A.D.M., C.N.A.P., S.K., O. Melander, P.M.N., L.C.G., I.B., C. Langenberg, N.J.W., D.A., M. Boehnke, M.I.M. **Project management.** K.J.G., A.L.G., D.A., M. Boehnke, T.M.T., M.I.M., A.P.M. **DIAGRAM Consortium management.** D.A., M. Boehnke, M.I.M.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the [online version of the paper](#).

Reprints and permissions information is available online at <http://www.nature.com/reprints/index.html>.

- Kooner, J.S. *et al.* Genome-wide association study in individuals of South Asian ancestry identifies six new type 2 diabetes susceptibility loci. *Nat. Genet.* **43**, 984–989 (2011).
- Cho, Y.S. *et al.* Meta-analysis of genome-wide association studies identifies eight new loci for type 2 diabetes in East Asians. *Nat. Genet.* **44**, 67–72 (2012).
- Voight, B.F. *et al.* Twelve type 2 diabetes susceptibility loci identified through large scale association analysis. *Nat. Genet.* **42**, 579–589 (2010).
- Morris, A.P. *et al.* Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nat. Genet.* **44**, 981–990 (2012).
- Mahajan, A. *et al.* Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. *Nat. Genet.* **46**, 234–244 (2014).
- Altshuler, D. *et al.* The common PPAR γ Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes. *Nat. Genet.* **26**, 76–80 (2000).
- Gloyn, A.L. *et al.* Large-scale association studies of variants in genes encoding the pancreatic-cell K_{ATP} channel subunits Kir6.2 (*KCNJ11*) and SUR1 (*ABCC8*) confirm that the *KCNJ11* E23K variant is associated with type 2 diabetes. *Diabetes* **52**, 568–572 (2003).
- Sladek, R. *et al.* A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature* **445**, 881–885 (2007).
- Dupuis, J. *et al.* New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes. *Nat. Genet.* **42**, 105–116 (2010).
- ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
- Elbein, S.C. *et al.* Genetic risk factors for type 2 diabetes: a trans-regulatory genetic architecture? *Am. J. Hum. Genet.* **91**, 466–477 (2012).
- Trynka, G. *et al.* Chromatin marks identify critical cell types for fine-mapping complex trait variants. *Nat. Genet.* **45**, 124–130 (2013).
- Parker, S.C.J. *et al.* Chromatin stretch enhancer states drive cell-specific gene regulation and harbour human disease risk variants. *Proc. Natl. Acad. Sci. USA* **110**, 17921–17926 (2013).
- Pasquali, L. *et al.* Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants. *Nat. Genet.* **46**, 136–143 (2014).
- Voight, B.F. *et al.* The MetaboChip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. *PLoS Genet.* **8**, e1002793 (2012).
- International HapMap Project Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851–861 (2007).
- 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
- 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).
- Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.* **11**, 499–511 (2010).
- Winkler, T.W. *et al.* Quality control and conduct of genome-wide association meta-analyses. *Nat. Protoc.* **9**, 1192–1212 (2014).
- Howie, B.N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* **5**, e1000529 (2009).
- Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G.R. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat. Genet.* **44**, 955–959 (2012).
- Yang, J. *et al.* Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat. Genet.* **44**, 369–375 (2012).
- Unoki, H. *et al.* SNPs in *KCNQ1* are associated with susceptibility to type 2 diabetes in East Asian and European populations. *Nat. Genet.* **40**, 1098–1102 (2008).
- Fitzpatrick, G.V., Soloway, P.D. & Higgins, M.J. Regional loss of imprinting and growth deficiency in mice with a targeted deletion of *KvDMR1*. *Nat. Genet.* **32**, 426–431 (2002).
- Zeggini, E. *et al.* Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science* **316**, 1336–1341 (2007).
- Shea, J. *et al.* Comparing strategies to fine-map the association of common SNPs at chromosome 9p21 with type 2 diabetes and myocardial infarction. *Nat. Genet.* **43**, 801–805 (2011).
- Maller, J.B. *et al.* Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nat. Genet.* **44**, 1294–1301 (2012).
- Jafar-Mohammadi, B. *et al.* A role for coding functional variants in *HNF4A* in type 2 diabetes susceptibility. *Diabetologia* **54**, 111–119 (2011).
- Teslovich, T.M. *et al.* Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466**, 707–713 (2010).
- Florez, J.C. *et al.* Haplotype structure and genotype-phenotype correlations of the sulfonylurea receptor and the islet ATP-sensitive potassium channel gene. *Diabetes* **53**, 1360–1368 (2004).
- Hamming, K.S. *et al.* Co-expression of the type 2 diabetes susceptibility gene variants *KCNJ11* E23K and *ABCC8* S1369A alter the ATP and sulfonylurea sensitivities of the ATP-sensitive K⁺ channel. *Diabetes* **58**, 2419–2424 (2009).
- Nicolson, T.J. *et al.* Insulin storage and glucose homeostasis in mice null for the granule zinc transporter ZnT8 and studies of the type 2 diabetes-associated variants. *Diabetes* **58**, 2070–2083 (2009).
- Beer, N.L. *et al.* The P446L variant in *GCKR* associated with fasting plasma glucose and triglyceride levels exerts its effect through increased glucokinase activity in liver. *Hum. Mol. Genet.* **18**, 4081–4088 (2009).
- Holmkvist, J. *et al.* Common variants in *HNF-1 α* and risk of type 2 diabetes. *Diabetologia* **49**, 2882–2891 (2006).
- Yamagata, K. *et al.* Mutations in the hepatocyte nuclear factor-1 α gene in maturity-onset diabetes of the young (MODY3). *Nature* **384**, 455–458 (1996).
- Yamagata, K. *et al.* Mutations in the hepatocyte nuclear factor-4 α gene in maturity-onset diabetes of the young (MODY1). *Nature* **384**, 458–460 (1996).
- Gusev, A. *et al.* Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. *Am. J. Hum. Genet.* **95**, 535–552 (2014).
- Soccio, R.E. *et al.* Species-specific strategies underlying conserved functions of metabolic transcription factors. *Mol. Endocrinol.* **25**, 694–706 (2011).
- Gaulton, K.J. *et al.* A map of open chromatin in human pancreatic islets. *Nat. Genet.* **42**, 255–259 (2010).
- Fogarty, M.P., Cannon, M.E., Vadlamudi, S., Gaulton, K.J. & Mohlke, K.L. Identification of a regulatory variant that binds FOXA1 and FOXA2 at the *CDC123/CAMK1D* type 2 diabetes GWAS locus. *PLoS Genet.* **10**, e1004633 (2014).
- Manning, A.K. *et al.* A genome-wide approach accounting for body-mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance. *Nat. Genet.* **44**, 659–669 (2012).
- Dimas, A.S. *et al.* Impact of type 2 diabetes susceptibility variants on quantitative glycemic traits reveals mechanistic heterogeneity. *Diabetes* **63**, 2158–2171 (2014).

44. Ravassard, P. *et al.* A genetically engineered human pancreatic β cell line exhibiting glucose-inducible insulin secretion. *J. Clin. Invest.* **121**, 3589–3597 (2011).
45. Fadista, J. *et al.* Global genomic and transcriptomic analysis of human pancreatic islets reveals novel genes influencing glucose metabolism. *Proc. Natl. Acad. Sci. USA* **111**, 13924–13929 (2014).
46. Lyssenko, V. *et al.* Common variant in *MTNR1B* associated with increased risk of type 2 diabetes and impaired early insulin secretion. *Nat. Genet.* **41**, 82–88 (2009).
47. Gao, N. *et al.* Foxa1 and Foxa2 maintain the metabolic and secretory features of the mature β -cell. *Mol. Endocrinol.* **24**, 1594–1604 (2010).
48. Zhou, Y. *et al.* TCF7L2 is a master regulator of insulin production and processing. *Hum. Mol. Genet.* **23**, 6419–6431 (2014).
49. Lyssenko, V. *et al.* Mechanisms by which common variants in the *TCF7L2* gene increase risk of type 2 diabetes. *J. Clin. Invest.* **117**, 2155–2163 (2007).
50. Gloy, A.L. *et al.* Activating mutations in the gene encoding the ATP-sensitive potassium-channel subunit Kir6.2 and permanent neonatal diabetes. *N. Engl. J. Med.* **350**, 1838–1849 (2004).
51. Flannick, J. *et al.* Loss-of-function mutations in *SLC30A8* protect against type 2 diabetes. *Nat. Genet.* **46**, 357–363 (2014).
52. Dickson, S.P., Wang, K., Krantz, I., Hakonarson, H. & Goldstein, D.B. Rare variants create synthetic genome-wide associations. *PLoS Biol.* **8**, e1000294 (2010).
53. Bonnefond, A. *et al.* Rare *MTNR1B* variants impairing melatonin receptor 1B function contribute to type 2 diabetes. *Nat. Genet.* **44**, 297–301 (2012).
54. Zaret, K.S. & Carroll, J.S. Pioneer transcription factors: establishing a competence for gene expression. *Genes Dev.* **25**, 2227–2241 (2011).
55. Gao, N. *et al.* Dynamic regulation of *Pdx1* enhancers by Foxa1 and Foxa2 is essential for pancreas development. *Genes Dev.* **22**, 3435–3448 (2008).
56. Lee, C.S., Friedman, J.R., Fulmer, J.T. & Kaestner, K.H. The initiation of liver development is dependent on Foxa transcription factors. *Nature* **435**, 944–947 (2005).
57. Scott, R.A. *et al.* Large-scale association analyses identify new loci influencing glycaemic traits and provide insight into the underlying biological pathways. *Nat. Genet.* **44**, 991–1005 (2012).
58. Tabassum, R., Chavali, S., Dwivedi, O.P., Tandon, N. & Bharadwaj, D. Genetic variants of *FOXA2*: risk of type 2 diabetes and effect on metabolic traits in North Indians. *J. Hum. Genet.* **53**, 957–965 (2008).
59. Johnson, M.E., Schug, J., Wells, A.D., Kaestner, K.H. & Grant, S.F. Genome-wide analyses of ChIP-Seq derived FOXA2 DNA occupancy in liver points to genetic networks underpinning multiple complex traits. *J. Clin. Endocrinol. Metab.* **99**, E1580–E1585 (2014).

Kyle J Gaulton^{1,2,161}, Teresa Ferreira^{1,161}, Yeji Lee^{3,161}, Anne Raimondo^{4,161}, Reedik Mägi^{5,161}, Michael E Reschen^{6,161}, Anubha Mahajan¹, Adam Locke³, N William Rayner^{1,4,7}, Neil Robertson^{1,4}, Robert A Scott⁸, Inga Prokopenko⁹, Laura J Scott³, Todd Green¹⁰, Thomas Sparso¹¹, Dorothee Thuillier¹², Loic Yengo¹², Harald Grallert^{13–15}, Simone Wahl^{13–15}, Mattias Frånberg^{16–18}, Rona J Strawbridge¹⁶, Hans Kestler^{19,20}, Himanshu Chheda²¹, Lewin Eisele²², Stefan Gustafsson²³, Valgerdur Steinthorsdottir²⁴, Gudmar Thorleifsson²⁴, Lu Qi^{25–28}, Lennart C Karssen²⁹, Elisabeth M van Leeuwen²⁹, Sara M Willems^{8,29}, Man Li³⁰, Han Chen^{31,32}, Christian Fuchsberger³, Phoenix Kwan³, Clement Ma³, Michael Linderman³³, Yingchang Lu³⁴, Soren K Thomsen⁴, Jana K Rundle⁴, Nicola L Beer^{1,4}, Martijn van de Bunt^{1,4}, Anil Chalisey⁶, Hyun Min Kang³, Benjamin F Voight³⁵, Gonçalo R Abecasis³, Peter Almgren³⁶, Damiano Baldassarre^{37,38}, Beverley Balkau^{39,40}, Rafn Benediktsson^{41,42}, Matthias Blüher^{43,44}, Heiner Boeing⁴⁵, Lori L Bonnycastle⁴⁶, Erwin P Bottinger⁴⁷, Noël P Burt¹⁰, Jason Carey¹⁰, Guillaume Charpentier⁴⁸, Peter S Chines⁴⁶, Marilyn C Cornelis⁴⁹, David J Couper⁵⁰, Andrew T Crenshaw¹⁰, Rob M van Dam^{26,51}, Alex S F Doney^{52,53}, Mozghan Dorkhan⁵⁴, Sarah Edkins⁷, Johan G Eriksson^{55–58}, Tonu Esko^{5,59,60}, Elodie Eury⁶¹, João Fadista³⁶, Jason Flannick¹⁰, Pierre Fontanillas¹⁰, Caroline Fox^{62,63}, Paul W Franks^{26,36,64,65}, Karl Gertow¹⁶, Christian Gieger^{13,14}, Bruna Gigante⁶⁶, Omri Gottesman⁴⁷, George B Grant¹⁰, Niels Grarup¹¹, Christopher J Groves⁴, Maija Hassinen⁶⁷, Christian T Have¹¹, Christian Herder^{68,69}, Oddgeir L Holmen⁷⁰, Astradur B Hreidarsson⁴², Steve E Humphries⁷¹, David J Hunter^{25–27,72}, Anne U Jackson³, Anna Jonsson³⁶, Marit E Jørgensen⁷³, Torben Jørgensen^{74–76}, Wen-Hong L Kao^{30,160}, Nicola D Kerrison⁸, Leena Kinnunen⁵⁵, Norman Klopp^{13,77}, Augustine Kong²⁴, Peter Kovacs^{43,44}, Peter Kraft^{25,32,72}, Jasmina Kravic³⁶, Cordelia Langford⁷, Karin Leander⁶⁶, Liming Liang^{25,32}, Peter Lichtner⁷⁸, Cecilia M Lindgren^{1,10}, Eero Lindholm³⁶, Allan Linneberg^{74,79,80}, Ching-Ti Liu³¹, Stéphane Lobbens⁶¹, Jian'an Luan⁸, Valeriya Lyssenko^{36,73}, Satu Männistö⁵⁵, Olga McLeod¹⁶, Julia Meyer⁸¹, Evelin Mihailov⁵, Ghazala Mirza⁸², Thomas W Mühleisen^{83–85}, Martina Müller-Nurasyid^{81,86–88}, Carmen Navarro^{89–91}, Markus M Nöthen^{83,84}, Nikolay N Oskolkov³⁶, Katharine R Owen^{4,92}, Domenico Palli⁹³, Sonali Pechlivanis²², Leena Peltonen^{7,10,21,55,160}, John R B Perry⁸, Carl G P Platou^{70,94}, Michael Roden^{68,69,95}, Douglas Ruderfer⁹⁶, Denis Rybin⁹⁷, Yvonne T van der Schouw⁹⁸, Bengt Sennblad^{16,17}, Gunnar Sigurdsson^{42,99}, Alena Stančáková¹⁰⁰, Gerald Steinbach¹⁰¹, Petter Storm³⁶, Konstantin Strauch^{81,87}, Heather M Stringham³, Qi Sun^{26,27}, Barbara Thorand^{14,15}, Emmi Tikkanen^{21,102}, Anke Tonjes^{43,44}, Joseph Trakalo¹, Elena Tremoli^{37,38}, Tiinamaija Tuomi^{21,58,103,104}, Roman Wennauer¹⁰⁵, Steven Wiltshire^{1,160}, Andrew R Wood¹⁰⁶, Eleftheria Zeggini⁷, Ian Dunham¹⁰⁷, Ewan Birney¹⁰⁷, Lorenzo Pasquali^{108–110}, Jorge Ferrer^{111,112}, Ruth J F Loos^{8,34,47,113}, Josée Dupuis^{31,62}, Jose C Florez^{60,114–116}, Eric Boerwinkle^{117,118}, James S Pankow¹¹⁹, Cornelia van Duijn^{29,120}, Eric Sijbrands¹⁰⁵, James B Meigs^{114,121}, Frank B Hu^{25–27}, Unnur Thorsteinsdottir^{24,41}, Kari Stefansson^{24,41}, Timo A Lakka^{67,122,123}, Rainer Rauramaa^{67,123}, Michael Stumvoll^{43,44}, Nancy L Pedersen¹²⁴, Lars Lind¹²⁵, Sirkka M Keinänen-Kiukaanniemi^{126,127}, Eeva Korpi-Hyövähti¹²⁸, Timo E Saaristo^{129,130}, Juha Saltevo¹³¹, Johanna Kuusisto¹⁰⁰, Markku Laakso¹⁰⁰, Andres Metspalu^{5,132}, Raimund Erbel¹³³, Karl-Heinz Jöckel¹²², Susanne Moebus²², Samuli Ripatti^{7,21,102,134}, Veikko Salomaa⁵⁵, Erik Ingelsson^{1,23}, Bernhard O Boehm^{135,136}, Richard N Bergman¹³⁷, Francis S Collins⁴⁶, Karen L Mohlke¹³⁸, Heikki Koistinen^{55,139,140}, Jaakko Tuomilehto^{55,141–143}, Kristian Hveem⁷⁰, Inger Njølstad¹⁴⁴, Panagiotis Deloukas^{7,145}, Peter J Donnelly^{1,146},

Timothy M Frayling¹⁰⁶, Andrew T Hattersley¹⁴⁷, Ulf de Faire⁶⁶, Anders Hamsten¹⁶, Thomas Illig^{13,77}, Annette Peters^{14,15,88}, Stephane Cauchi¹², Rob Sladek^{148,149}, Philippe Froguel^{9,12,61}, Torben Hansen^{11,150}, Oluf Pedersen¹¹, Andrew D Morris¹⁵¹, Collin N A Palmer^{52,53}, Sekar Kathiresan^{10,115,152}, Olle Melander³⁶, Peter M Nilsson³⁶, Leif C Groop^{21,36}, Inês Barroso^{7,153,154}, Claudia Langenberg⁸, Nicholas J Wareham⁸, Christopher A O'Callaghan⁶, Anna L Gloyn^{1,4,92,162}, David Altshuler^{10,114–116,155,156,162}, Michael Boehnke^{3,162}, Tanya M Teslovich^{3,162}, Mark I McCarthy^{1,4,92,162} & Andrew P Morris^{1,5,157,158,162} for the DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium¹⁵⁹

¹Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK. ²Department of Genetics, Stanford University, Stanford, California, USA. ³Department of Biostatistics, University of Michigan, Ann Arbor, Michigan, USA. ⁴Oxford Centre for Diabetes, Endocrinology and Metabolism, University of Oxford, Oxford, UK. ⁵Estonian Genome Center, University of Tartu, Tartu, Estonia. ⁶Centre for Cellular and Molecular Physiology, Nuffield Department of Medicine, University of Oxford, Oxford, UK. ⁷Wellcome Trust Sanger Institute, Hinxton, UK. ⁸Medical Research Council (MRC) Epidemiology Unit, University of Cambridge School of Clinical Medicine, Institute of Metabolic Science, Cambridge Biomedical Campus, Cambridge, UK. ⁹Genomics of Common Disease, Imperial College London, London, UK. ¹⁰Broad Institute of Harvard and MIT, Cambridge, Massachusetts, USA. ¹¹Novo Nordisk Foundation Center for Basic Metabolic Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark. ¹²Lille Institute of Biology, European Genomics Institute of Diabetes, Lille, France. ¹³Research Unit of Molecular Epidemiology, Helmholtz Zentrum München–German Research Center for Environmental Health, Neuherberg, Germany. ¹⁴Institute of Epidemiology II, Helmholtz Zentrum München–German Research Center for Environmental Health, Neuherberg, Germany. ¹⁵German Center for Diabetes Research, Neuherberg, Germany. ¹⁶Atherosclerosis Research Unit, Department of Medicine Solna, Karolinska Institutet, Stockholm, Sweden. ¹⁷Science for Life Laboratory, Stockholm, Sweden. ¹⁸Department for Numerical Analysis and Computer Science, Stockholm University, Stockholm, Sweden. ¹⁹Leibniz Institute for Age Research, Fritz Lipmann Institute, Jena, Germany. ²⁰Medical Systems Biology, Ulm University, Ulm, Germany. ²¹Finnish Institute for Molecular Medicine (FIMM), Helsinki, Finland. ²²Institute for Medical Informatics, Biometry and Epidemiology, University Hospital of Essen, Essen, Germany. ²³Department of Medical Sciences, Molecular Epidemiology and Science for Life Laboratory, Uppsala University, Uppsala, Sweden. ²⁴deCODE Genetics/Amgen, Inc., Reykjavik, Iceland. ²⁵Department of Epidemiology, Harvard School of Public Health, Boston, Massachusetts, USA. ²⁶Department of Nutrition, Harvard School of Public Health, Boston, Massachusetts, USA. ²⁷Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts, USA. ²⁸Department of Epidemiology, School of Public Health and Tropical Medicine, Tulane University, New Orleans, Louisiana, USA. ²⁹Department of Epidemiology, Erasmus University Medical Center, Rotterdam, the Netherlands. ³⁰Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, USA. ³¹Department of Biostatistics, Boston University School of Public Health, Boston, Massachusetts, USA. ³²Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts, USA. ³³Icahn Institute for Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York, New York, USA. ³⁴Genetics of Obesity and Related Metabolic Traits Program, Icahn School of Medicine at Mount Sinai, New York, New York, USA. ³⁵Department of Pharmacology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA. ³⁶Lund University Diabetes Centre, Department of Clinical Science Malmö, Scania University Hospital, Lund University, Malmö, Sweden. ³⁷Centro Cardiologico Monzino, Istituto di Ricovero e Cura a Carattere Scientifico (IRCCS), Milan, Italy. ³⁸Dipartimento di Scienze Farmacologiche e Biomolecolari, Università di Milano, Milan, Italy. ³⁹INSERM Centre de Recherche Épidémiologie et Santé des Populations (CESP) U1018, Villejuif, France. ⁴⁰University Paris Sud 11, UMR S 1018, Villejuif, France. ⁴¹Faculty of Medicine, University of Iceland, Reykjavik, Iceland. ⁴²Landspítali University Hospital, Reykjavik, Iceland. ⁴³Integrated Treatment and Research (IFB) Center for Adiposity Diseases, University of Leipzig, Leipzig, Germany. ⁴⁴Department of Medicine, University of Leipzig, Leipzig, Germany. ⁴⁵German Institute of Human Nutrition, Potsdam-Rehbrücke, Germany. ⁴⁶National Human Genome Research Institute, US National Institutes of Health, Bethesda, Maryland, USA. ⁴⁷Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, New York, USA. ⁴⁸Endocrinology-Diabetology Unit, Corbeil-Essonnes Hospital, Corbeil-Essonnes, France. ⁴⁹Department of Preventive Medicine, Northwestern University Feinberg School of Medicine, Chicago, Illinois, USA. ⁵⁰Collaborative Studies Coordinating Center, Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA. ⁵¹Saw Swee Hock School of Public Health, National University of Singapore, Singapore. ⁵²Diabetes Research Centre, Biomedical Research Institute, University of Dundee, Ninewells Hospital, Dundee, UK. ⁵³Pharmacogenomics Centre, Biomedical Research Institute, University of Dundee, Ninewells Hospital, Dundee, UK. ⁵⁴Lund University Diabetes Centre, Department of Clinical Science Malmö, Novo Nordisk Scandinavia, Malmö, Sweden. ⁵⁵Department of Chronic Disease Prevention, National Institute for Health and Welfare, Helsinki, Finland. ⁵⁶Department of General Practice and Primary Health Care, University of Helsinki, Helsinki, Finland. ⁵⁷Unit of General Practice, Helsinki University General Hospital, Helsinki, Finland. ⁵⁸Folkhalsan Research Center, Helsinki, Finland. ⁵⁹Division of Endocrinology, Children's Hospital, Boston, Massachusetts, USA. ⁶⁰Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, Massachusetts, USA. ⁶¹CNRS UMR 8199, Institute of Biology and Lille 2 University, Pasteur Institute, Lille, France. ⁶²National Heart, Lung, and Blood Institute's Framingham Heart Study, Framingham, Massachusetts, USA. ⁶³Division of Endocrinology and Metabolism, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts, USA. ⁶⁴Department of Clinical Sciences, Lund University, Malmö, Sweden. ⁶⁵Department of Public Health and Clinical Medicine, Umeå University, Umeå, Sweden. ⁶⁶Division of Cardiovascular Epidemiology, Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden. ⁶⁷Kuopio Research Institute of Exercise Medicine, Kuopio, Finland. ⁶⁸Institute for Clinical Diabetology, German Diabetes Center, Leibniz Center for Diabetes Research at Heinrich Heine University Düsseldorf, Düsseldorf, Germany. ⁶⁹German Center for Diabetes Research, partner site Düsseldorf, Düsseldorf, Germany. ⁷⁰Nord-Trøndelag Health Study (HUNT) Research Center, Department of Public Health and General Practice, Norwegian University of Science and Technology, Levanger, Norway. ⁷¹Cardiovascular Genetics, British Heart Foundation (BHF) Laboratories, Institute of Cardiovascular Sciences, University College London, London, UK. ⁷²Program in Genetic Epidemiology and Statistical Genetics, Harvard School of Public Health, Boston, Massachusetts, USA. ⁷³Steno Diabetes Center, Gentofte, Denmark. ⁷⁴Research Centre for Prevention and Health, Capital Region of Denmark, Copenhagen, Denmark. ⁷⁵Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark. ⁷⁶Faculty of Medicine, University of Aalborg, Aalborg, Denmark. ⁷⁷Hannover Unified Biobank, Hannover Medical School, Hannover, Germany. ⁷⁸Institute of Human Genetics, Helmholtz Zentrum München–German Research Center for Environmental Health, Neuherberg, Germany. ⁷⁹Copenhagen University Hospital, Rigshospitalet, Copenhagen, Denmark. ⁸⁰Department of Clinical Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark. ⁸¹Institute of Genetic Epidemiology, Helmholtz Zentrum München–German Research Center for Environmental Health, Neuherberg, Germany. ⁸²Biomedical Research Centre Genomics Core Facility, Guy's and St Thomas' National Health Service (NHS) Foundation Trust, Guy's & St Thomas' Hospital, London, UK. ⁸³Institute of Human Genetics, University of Bonn, Bonn, Germany. ⁸⁴Department of Genomics, Life and Brain Center, University of Bonn, Bonn, Germany. ⁸⁵Institute of Neuroscience and Medicine (INM-1), Research Centre Jülich, Jülich, Germany. ⁸⁶Department of Medicine I, University Hospital Grosshadern, Ludwig Maximilians Universität, Munich, Germany. ⁸⁷Institute of Medical Informatics, Biometry and Epidemiology, Chair of Genetic Epidemiology, Ludwig Maximilians Universität, Neuherberg, Germany. ⁸⁸DZHK (German Center for Cardiovascular Research), partner site Munich Heart Alliance, Munich, Germany. ⁸⁹Department of Epidemiology, Murcia Regional Health Council, Instituto Murciano de Investigación Biosanitaria Virgen de la Arrixaca (IMIB-Arrixaca), Murcia, Spain. ⁹⁰Centro de Investigación Biomédica en Red de Epidemiología y Salud Pública (CIBERESP), Madrid, Spain. ⁹¹Department of Health and Social Sciences, Universidad de Murcia, Murcia, Spain. ⁹²Oxford National Institute for Health Research Biomedical Research Centre, Churchill Hospital, Oxford, UK. ⁹³Cancer Research and Prevention Institute (ISPO), Florence, Italy. ⁹⁴Department of Internal Medicine, Levanger Hospital, Nord-Trøndelag Health Trust, Levanger, Norway. ⁹⁵Department of Endocrinology and Diabetology, University Hospital Düsseldorf, Düsseldorf, Germany. ⁹⁶Division of Psychiatric Genomics, Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, New York, USA. ⁹⁷Boston University Data Coordinating Center, Boston, Massachusetts, USA. ⁹⁸University Medical Center Utrecht, Utrecht, the Netherlands. ⁹⁹Icelandic Heart Association, Kopavogur, Iceland. ¹⁰⁰Department of Medicine, University of Eastern Finland and Kuopio University Hospital, Kuopio, Finland. ¹⁰¹Department of Clinical Chemistry and Central Laboratory, University of Ulm, Ulm, Germany. ¹⁰²Department of Public Health, Hjelt Institute, University of Helsinki, Helsinki, Finland. ¹⁰³Department of Endocrinology, Abdominal Center, Helsinki University Hospital, Helsinki, Finland. ¹⁰⁴Research Program for Diabetes and Obesity, University of Helsinki, Helsinki, Finland. ¹⁰⁵Department of Internal

Medicine, Erasmus University Medical Center, Rotterdam, the Netherlands. ¹⁰⁶Genetics of Complex Traits, University of Exeter Medical School, University of Exeter, Exeter, UK. ¹⁰⁷European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Hinxton, UK. ¹⁰⁸Division of Endocrinology, Germans Trias i Pujol University Hospital and Research Institute, Badalona, Spain. ¹⁰⁹Josep Carreras Leukaemia Research Institute, Badalona, Spain. ¹¹⁰CIBER de Diabetes y Enfermedades Metabólicas Asociadas (CIBERDEM), Barcelona, Spain. ¹¹¹Department of Medicine, Imperial College London, London, UK. ¹¹²Institut d'Investigacions Biomèdiques August Pi i Sunyer, Centre Esther Koplowitz, Barcelona, Spain. ¹¹³Mindich Child Health and Development Institute, Icahn School of Medicine at Mount Sinai, New York, New York, USA. ¹¹⁴Department of Medicine, Harvard Medical School, Boston, Massachusetts, USA. ¹¹⁵Center for Human Genetic Research, Massachusetts General Hospital, Boston, Massachusetts, USA. ¹¹⁶Diabetes Research Center, Diabetes Unit, Massachusetts General Hospital, Boston, Massachusetts, USA. ¹¹⁷Human Genetics Center, University of Texas Health Science Center at Houston, Houston, Texas, USA. ¹¹⁸Human Genome Sequencing Center at Baylor College of Medicine, Houston, Texas, USA. ¹¹⁹Division of Epidemiology and Community Health, University of Minnesota, Minneapolis, Minnesota, USA. ¹²⁰Netherlands Genomics Initiative, Netherlands Consortium for Healthy Ageing and Center for Medical Systems Biology, Rotterdam, the Netherlands. ¹²¹General Medicine Division, Massachusetts General Hospital, Boston, Massachusetts, USA. ¹²²Institute of Biomedicine/Physiology, University of Eastern Finland, Kuopio, Finland. ¹²³Department of Clinical Physiology and Nuclear Medicine, Kuopio University Hospital, Kuopio, Finland. ¹²⁴Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden. ¹²⁵Department of Medical Sciences, Uppsala University Hospital, Uppsala, Sweden. ¹²⁶Faculty of Medicine, Institute of Health Sciences, University of Oulu, Oulu, Finland. ¹²⁷Unit of General Practice, Oulu University Hospital, Oulu, Finland. ¹²⁸South Ostrobothnia Central Hospital, Seinäjoki, Finland. ¹²⁹Finnish Diabetes Association, Tampere, Finland. ¹³⁰Pirkanmaa District Hospital, Tampere, Finland. ¹³¹Department of Medicine, Central Finland Central Hospital, Jyväskylä, Finland. ¹³²Institute of Molecular and Cell Biology, University of Tartu, Tartu, Estonia. ¹³³Clinic of Cardiology, West German Heart Centre, University Hospital of Essen, University Duisburg-Essen, Essen, Germany. ¹³⁴Public Health Genomics Unit, National Institute for Health and Welfare, Helsinki, Finland. ¹³⁵Division of Endocrinology and Diabetes, Department of Internal Medicine, University Medical Centre Ulm, Ulm, Germany. ¹³⁶Lee Kong Chian School of Medicine, Imperial College London and Nanyang Technological University, Singapore. ¹³⁷Diabetes and Obesity Research Institute, Cedars-Sinai Medical Center, Los Angeles, California, USA. ¹³⁸Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA. ¹³⁹Division of Endocrinology, Department of Medicine, Helsinki University Central Hospital, Helsinki, Finland. ¹⁴⁰Minerva Foundation Institute for Medical Research, Helsinki, Finland. ¹⁴¹Instituto de Investigación Sanitaria del Hospital Universitario La Paz, Madrid, Spain. ¹⁴²Centre for Vascular Prevention, Danube University Krems, Krems, Austria. ¹⁴³Diabetes Research Group, King Abdulaziz University, Jeddah, Saudi Arabia. ¹⁴⁴Department of Community Medicine, Faculty of Health Sciences, University of Tromsø, Tromsø, Norway. ¹⁴⁵William Harvey Research Institute, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, London, UK. ¹⁴⁶Department of Statistics, University of Oxford, Oxford, UK. ¹⁴⁷Institute of Biomedical and Clinical Science, University of Exeter Medical School, Exeter, UK. ¹⁴⁸Montreal Diabetes Research Center, Centre de Recherche du Centre Hospitalier de l'Université de Montréal, Montreal, Quebec, Canada. ¹⁴⁹McGill University and Génome Québec Innovation Centre, Montreal, Quebec, Canada. ¹⁵⁰Faculty of Health Sciences, University of Southern Denmark, Odense, Denmark. ¹⁵¹Usher Institute of Population Health Sciences and Informatics, University of Edinburgh, Edinburgh, UK. ¹⁵²Cardiovascular Research Center, Massachusetts General Hospital, Boston, Massachusetts, USA. ¹⁵³University of Cambridge Metabolic Research Laboratories, Wellcome Trust-MRC Institute of Metabolic Science, Cambridge, UK. ¹⁵⁴National Institute for Health Research Cambridge Biomedical Research Centre, Cambridge, UK. ¹⁵⁵Department of Genetics, Harvard Medical School, Boston, Massachusetts, USA. ¹⁵⁶Department of Molecular Biology, Harvard Medical School, Boston, Massachusetts, USA. ¹⁵⁷Department of Biostatistics, University of Liverpool, Liverpool, UK. ¹⁵⁸Department of Molecular and Clinical Pharmacology, University of Liverpool, Liverpool, UK. ¹⁵⁹A list of members and affiliations appears in the **Supplementary Note**. ¹⁶⁰Deceased. ¹⁶¹These authors contributed equally to this work. ¹⁶²These authors jointly supervised this work. Correspondence should be addressed to K.J.G. (kgaulton@gmail.com), M.I.M. (mark.mccarthy@dr1.ox.ac.uk) or A.P.M. (apmorris@liverpool.ac.uk).

ONLINE METHODS

Ethics statement. All human research was approved by the relevant institutional review boards and conducted according to the Declaration of Helsinki. All participants provided written informed consent.

Metabochip imputation and association analysis. We considered a total of 27,206 T2D cases and 57,574 controls from 23 studies from populations of European ancestry (Supplementary Table 1), all genotyped with the Metabochip. Sample and variant quality control was performed within each study (Supplementary Table 2). To improve the quality of the genotype scaffold in each study, variants were subsequently removed if (i) allele frequencies differed from those for European-ancestry haplotypes from the 1000 Genomes Project Consortium phase 1 integrated reference panel (March 2012 release)¹⁸ by more than 20%; AT/GC variants had MAF >40% because of potential undetected errors in strand alignment; or (iii) MAF was <1%, because of difficulties in calling rare variants. Each scaffold was then imputed up to the phase 1 integrated reference panel (all ancestries; March 2012 release) from the 1000 Genomes Project Consortium¹⁸, using IMPUTEv2 (ref. 21) or minimac²². Within each study, well-imputed variants (IMPUTEv2 info >0.4 or minimac $r^2 > 0.3$) were tested for T2D association under an additive model after adjustment for study-specific covariates (Supplementary Table 2), including principal components to adjust for population structure. Association summary statistics for each variant for each study were corrected for residual population structure using the genomic control inflation factor⁶⁰ obtained from 3,598 independent ($r^2 < 0.05$) QT-interval variants, which were not expected to be associated with T2D⁴ (Supplementary Table 2). We then combined association summary statistics for each variant across studies via fixed-effects inverse variance-weighted meta-analysis. The results of the meta-analysis were subsequently corrected by a second round of QT-interval genomic control ($\lambda_{QT} = 1.18$) to account for structure between studies. Variants were excluded from downstream analyses if they were reported in less than 80% of the total effective sample size, defined as $N_e = 4 \times N_{\text{cases}} \times N_{\text{controls}} / (N_{\text{cases}} + N_{\text{controls}})$, thus removing those that were not well imputed in the majority of studies.

Identification of distinct association signals in established GWAS loci. We used GCTA²³ to select index variants in each of the 39 established loci represented on Metabochip with nominal evidence of association ($P_j < 0.001$) with T2D in an approximate joint regression model. The GCTA model made use of (i) summary statistics from the fixed-effects meta-analysis Metabochip studies and (ii) genotype data for 3,298 T2D cases and 3,708 controls of UK ancestry from GoDARTS as a reference for LD across each fine-mapping region. For comparison, we also obtained association summary statistics for the selected index variants from the GCTA joint regression model on the basis of genotype data from an alternative reference consisting of 4,435 T2D cases and 5,757 controls of Finnish ancestry from FUSION (Supplementary Fig. 12 and Supplementary Table 15). Selected index variants were then carried forward for *in silico* follow-up in validation meta-analysis.

The validation meta-analysis consisted of 19,662 T2D cases and 115,140 controls from ten GWAS from populations of European ancestry, genotyped with a range of genome-wide arrays (Supplementary Table 1). Sample and variant quality control was performed within each study (Supplementary Table 2). Each scaffold was then imputed up to the phase 1 integrated reference panel (all ancestries; March 2012 release) from the 1000 Genomes Project Consortium¹⁸, using IMPUTEv2 or minimac. Within each study, well-imputed variants (IMPUTEv2 info ≥ 0.4 or minimac $r^2 \geq 0.3$) were tested for T2D association under an additive model after adjustment for study-specific covariates (Supplementary Table 2), including principal components to adjust for population structure. Association summary statistics for each variant for each study were corrected for residual population structure using the genomic control inflation factor⁶⁰ (Supplementary Table 2). We then combined association summary statistics for each variant across studies via fixed-effects inverse variance-weighted meta-analysis.

Association summary statistics for the selected index variants from the Metabochip and validation meta-analyses were next combined via fixed-effects inverse variance-weighted meta-analysis. In each of the 39 established loci represented on the Metabochip, GCTA²³ was used to select index variants with locus-wide evidence of association ($P_j < 1 \times 10^{-5}$) in the approximate joint regression model on the basis of (i) summary statistics from the combined

meta-analysis and (ii) genotype data for 3,298 T2D cases and 3,708 controls from GoDARTS as a reference for LD across in each fine-mapping region.

For established loci with multiple index variants selected at locus-wide significance from the GCTA approximate joint regression model in combined meta-analysis, we performed exact conditioning within each Metabochip study (Supplementary Table 7). To obtain the association signal attributed to a specific index variant, high-quality variants (IMPUTEv2 info >0.4 or minimac $r^2 > 0.3$) were tested for T2D association under an additive model after adjustment for study-specific covariates (Supplementary Table 2) and genotypes at other selected index variants in the fine-mapping region. Association summary statistics for each study were corrected for residual population structure using the QT-interval genomic control inflation factor obtained in the Metabochip meta-analysis. For each association signal, summary statistics for each variant were then combined across discovery studies via fixed-effects inverse variance meta-analysis and subsequently corrected by a second round of QT-interval genomic control ($\lambda_{QT} = 1.18$).

Credible set construction. In an ideal fine-mapping experiment, we would calculate the posterior probability of driving each distinct association signal for all variants mapping to a locus. However, the posterior probability is determined by the association signal effect size of the variant and the corresponding standard error, which is also affected by the quality of imputation across studies, among other factors. To minimize the impact of imputation quality on fine mapping, we therefore retained only variants that were directly typed and/or well imputed in at least 80% of the total effective sample size. Assuming that the variant driving an association signal meets these quality criteria, the probability that it would be contained within the 99% credible set would be ~0.99.

For each distinct signal, we first calculated the posterior probability, π_{Cj} , that the j th variant is driving the association, given by

$$\pi_{Cj} = \frac{\Lambda_j}{\sum_k \Lambda_k}$$

where the summation is over all retained variants in the fine-mapping region. In this expression, Λ_j is the approximate Bayes factor⁶¹ for the j th variant, given by

$$\Lambda_j = \sqrt{\frac{V_j}{V_j + \omega}} \exp \left[\frac{\omega \beta_j^2}{2V_j(V_j + \omega)} \right]$$

where β_j and V_j denote the estimated allelic effect (log(OR)) and corresponding variance from the meta-analysis across Metabochip studies. In loci with multiple distinct signals of association, results are presented from exact conditional meta-analysis after adjusting for all other index variants in the fine-mapping region. In loci with a single association signal, results are presented from unconditional meta-analysis. The parameter ω denotes the prior variance in allelic effects, taken here to be 0.04 (ref. 61). The 99% credible set²⁹ for each signal was then constructed by (i) ranking all variants according to their Bayes factor Λ_j and (ii) including ranked variants until their cumulative posterior probability of driving the association attained or exceeded 0.99.

Genomic annotation data and enrichment analyses. We obtained genomic annotation data for transcription factor binding sites assayed through ChIP experiments from multiple sources. We used sites from the ENCODE Project Consortium¹⁰ for 161 proteins available from the UCSC human genome browser. We also obtained raw ChIP and input sequence data for additional factors assayed in primary pancreatic islets¹⁴. We then processed these additional factors using protocols employed by the ENCODE Project Consortium¹⁰. First, sequence reads were aligned to the human genome (hg19) using the Burrows-Wheeler Aligner (BWA)⁶² with sex-specific references and were then converted to BAM files using SAMtools⁶³ after removing duplicate reads and those not uniquely mapped. Binding sites were called from reads of each replicate, as well as reads pooled across all replicates, using SPP⁶⁴. The raw sites from each replicate of a protein were compared using an irreproducible

discovery rate⁶⁵ (IDR) threshold of 0.02. The resulting number of sites passing this IDR threshold was then used to filter the pooled sites of a protein. The set of sites was further filtered for artifacts using a blacklist of genomic regions from the ENCODE Project Consortium. Sites from all sources for each protein, including ENCODE, were then combined. The complete set of 165 proteins employed in these analyses is presented in **Supplementary Table 11**. In addition, we obtained FOXA2 ChIP-seq sites that were previously identified in human liver³⁹ and converted their positions to hg19 coordinates.

We obtained annotation data for five histone modifications (H3K4me1, H3K4me3, H3K27ac, H3K36me3 and H3K27me3) and CTCF binding assayed from ChIP experiments. We used data from nine cell types from ENCODE¹⁰ (GM12878, K562, HepG2, HSMM, HUVEC, NHEK, NHLF, H1-hESC and HMEC); we also obtained raw ChIP data assayed in primary pancreatic islets¹⁴ and premature and mature human adipose stromal cells⁶⁶. We mapped reads to the hg19 reference genome using BWA⁶² and used the resulting mapped reads from these 12 cell types as input to ChromHMM⁶⁷. We assigned states on the basis of the following chromatin signatures: active promoter (H3K4me3 and H3K27ac); strong enhancer 1 (H3K4me3, H3K27ac and H3K4me1); strong enhancer 2 (H3K27ac and H3K4me1); weak enhancer (H3K4me1); poised promoter (H3K27me3, H3K4me3 and H3K4me1); repressed (H3K27me3); insulator (CTCF); and transcription (H3K36me3). For each cell type, we pooled the three enhancer states into one enhancer category and the two promoter states into one promoter category. We also identified long noncoding RNA data from the Human Body Map (UCSC Genome Browser) and from pancreatic islets⁶⁸.

For each genomic annotation, we tested for overall enrichment of the posterior probability that overlapping variants in the 99% credible sets are driving distinct association signals (π_C). We first calculated the mean posterior probability (mean π_C) over the set of variants overlapping a given annotation. We then generated a null distribution of the mean posterior probability (mean π_C) by (i) shifting the genomic locations of binding sites a random distance within 100 kb in either direction; (ii) recalculating the mean posterior probability for 99% credible set variants overlapping shifted sites; and (iii) repeating this procedure 100,000 times. We estimated the fold enrichment of each overlap by calculating the expected null posterior probability and dividing the observed probability by the expected probability. We calculated a *P* value for the enrichment by the proportion of permutations for which the expected posterior probability of driving the association signal was greater than or equal to that observed. We considered cell type annotations to be significantly enriched if the *P* value was less than $0.05/37 = 0.0014$ (Bonferroni correction for 37 annotations). We considered transcription factor binding site annotations to be significantly enriched if the *P* value was less than $0.05/165 = 0.00030$ (Bonferroni correction for 165 factors). We next partitioned binding sites into those that are shared with another factor (where the genomic interval intersects a site for at least one other factor) and those that are unique. We also partitioned binding sites on the basis of overlap with islet enhancer elements. For each factor with significant enrichment across all credible sets (FOXA2), we applied the same enrichment analysis but restricted to credible set variants for each distinct association signal, separately.

We assessed the evidence for intersection in FOXA2 ChIP-seq sites from islets¹⁴, HepG2 cells¹⁰ and liver³⁹, across the genome and overlapping credible set variants, using BEDtools⁶⁹.

Motif analysis. We conducted recognition motif enhancement analyses for the set of FOXA2 ChIP-seq binding sites. First, we obtained the repeat-masked genomic sequence underlying each site using the UCSC human genome browser. We scanned sequences for enrichment in these motifs using MEME-ChIP⁷⁰, which uses up to 100 bp surrounding the midpoint of each site. This resulted in 198 enriched motifs with an *E* value (expected number of hits) less than 0.05 (**Supplementary Table 16**). We compared each motif to those known from JASPAR⁷¹, ENCODE¹⁰ and Homer⁷² using Tomtom⁷³.

Second, we identified variants in FOXA2 ChIP-seq sites predicted to disrupt an enriched recognition motif by (i) scanning the 25 bp of sequence flanking each variant allele using FIMO⁷⁴ (*P* < 0.0001) and (ii) retaining variants in highly conserved positions (entropy less than 0.5). For the 12 variants at FOXA2-enriched signals disrupting at least one recognition motif (**Table 3** and **Supplementary Table 14**), we compared their posterior probabilities of driving

the association (π_C) with those for non-disrupting variants in FOXA2 ChIP-seq sites at the same signals using a two-sided Wilcoxon rank-sum test.

Electrophoretic mobility shift assays. EMSA was performed using nuclear extracts from human HepG2 and EndoC- β H1 cells. HepG2 cells were the generous gift of the Ratcliffe laboratory⁷⁵ and were authenticated by genotyping in the major histocompatibility complex (MHC) region. Endo- β H1 cells were obtained from Endocells and have been previously authenticated⁴⁴. Both cell lines were tested and found negative for mycoplasma contamination. Nuclear extracts were incubated with [γ -³²P]ATP end-labeled double-stranded DNA probes (PerkinElmer). The forward-strand probe sequences used are presented in **Supplementary Table 17**.

For each lane of the EMSA, 5 μ g of nuclear extract was incubated with 100 fmol labeled probes in a 10- μ l binding reaction containing 10 mM Tris-HCl, pH 7.5, 4% glycerol, 1 mM MgCl₂, 0.5 mM EDTA, 0.5 mM DTT, 50 mM NaCl and 1 μ g poly(dI-dC). For competition assays, unlabeled probe at 100-fold excess was added to the binding reaction before addition of labeled probes. For supershift assays, the nuclear extract was preincubated with 1 μ g of antibody for 30 min on ice before the probe was added. The following antibodies were used: anti-NEUROD1 (sc-1084X, Santa Cruz Biotechnology), anti-PTF1A (sc-98612X, Santa Cruz Biotechnology), anti-HNF3B (FOXA2) (sc-6554X, Santa Cruz Biotechnology), anti-YY1 (sc-281X, Santa Cruz Biotechnology), anti-TAL1 (sc-12984X, Santa Cruz Biotechnology), normal rabbit immunoglobulin (sc-2027, Santa Cruz Biotechnology) and normal goat immunoglobulin (sc-2028, Santa Cruz Biotechnology).

Luciferase activity. We synthesized 224-bp sequences containing either the risk or protective allele of the *MTNR1B* enhancer sequence at rs10830963 in either the forward or reverse orientation by GeneArt (Life Technologies). Complementary single-stranded oligonucleotides were then annealed and subcloned into the minimal promoter-driven luciferase vector pGL4.23 (Promega) using the NheI and XhoI restriction sites. Isolated clones were verified by sequencing.

For luciferase assays, HepG2 human liver cells and EndoC- β H1 human β cells⁴⁵ were counted and seeded into 24-well trays (Corning) at 1.5×10^5 (HepG2) or 1.4×10^5 (EndoC- β H1) cells/well. Transfections were performed in triplicate with either Lipofectamine 2000 (HepG2) or FuGENE 6 (EndoC- β H1) according to the manufacturer's instructions. Cells were transfected with 700 ng of pGL4.23 DNA containing the protective or risk *MTNR1B* enhancer sequence in either the forward or reverse orientation or with an equivalent amount of empty vector DNA, plus 10 ng of pRL-SV40 DNA (Promega) as a transfection control, per well. Cells were lysed 48 h after transfection and analyzed for firefly and *Renilla* luciferase activities using the Dual-Luciferase Assay System (Promega) according to the manufacturer's instructions, in half-volume 96-well tray format on an Enspire Multimode Plate Reader (PerkinElmer). Firefly luciferase activity was normalized to *Renilla* luciferase activity for each well, and the results were expressed as a mean normalized activity relative to cells transfected with empty vector. All experiments were performed three times in triplicate. A two-sided unpaired *t* test was used to compare luciferase activity between alleles.

Mouse gene expression analysis. We obtained fold changes in pancreatic islet gene expression in wild-type compared to *Foxa1* and *Foxa2* double-null mice⁴⁷. We used Ensembl to map mouse genes to human orthologs. We filtered for human genes annotated as protein coding in GENCODE. This filtering resulted in 4,629 human protein-coding genes for analysis.

First, we calculated the genomic interval spanned by the variants in each credible set. We expanded this interval for 500 kb on either side and identified the set of genes overlapping this region using BEDtools⁶⁹. To account for syntenic differences in gene order between species, we retained only genes that were (i) on the same chromosome and (ii) in exactly the same relative order in both the mouse and human genomes. At the *GIPR* locus, one of the genes was ordered differently and thus removed from the analysis. At two loci, *KCNJ11* and *HNF1A*, at least one of the genes was located on a different part of the same chromosome, and, at another locus, *GCK*, the genes were located on different chromosomes. For these three loci, we retained only the genes that were at the same chromosomal location to the interval covered by

the credible set for the association signal for that locus (by lifting over from hg19 to mouse build mm10). Second, for each distinct association signal, we identified the closest gene to the index variant using BEDtools⁶⁹. We then partitioned distinct association signals into those with evidence for enriched FOXA2 binding (fold enrichment >1) and those without, counting each gene only once in a given group. For each analysis, we converted the fold changes to percentages and compared the percent change in expression using a one-sided Wilcoxon rank-sum test between genes in each partition and all 4,629 protein-coding genes.

60. Devlin, B. & Roeder, K. Genomic control for association studies. *Biometrics* **55**, 997–1004 (1999).
61. Wakefield, J. Bayesian measure of the probability of false discovery in genetic epidemiology studies. *Am. J. Hum. Genet.* **81**, 208–227 (2007).
62. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
63. Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
64. Kharchenko, P.V., Tolstorukov, M.Y. & Park, P.J. Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat. Biotechnol.* **26**, 1351–1359 (2008).
65. Li, Q., Brown, J.B., Huang, H. & Bickel, P.J. Measuring reproducibility of high-throughput experiments. *Ann. App. Stat.* **5**, 1752–1779 (2011).
66. Mikkelsen, T.S. *et al.* Comparative epigenomic analysis of murine and human adipogenesis. *Cell* **143**, 156–169 (2010).
67. Ernst, J. & Kellis, M. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat. Biotechnol.* **28**, 817–825 (2010).
68. Morán, I. *et al.* Human β cell transcriptome analysis uncovers lncRNAs that are tissue-specific, dynamically regulated, and abnormally expressed in type 2 diabetes. *Cell Metab.* **16**, 435–448 (2012).
69. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
70. Machanick, P. & Bailey, T.L. MEME-ChIP. Motif analysis of large DNA datasets. *Bioinformatics* **27**, 1696–1697 (2011).
71. Mathelier, A. *et al.* JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **42**, D142–D147 (2014).
72. Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B-cell identities. *Mol. Cell* **38**, 576–589 (2010).
73. Bailey, T.L. *et al.* MEME Suite: tools for motif discovery and searching. *Nucleic Acids Res.* **37**, W202–W208 (2009).
74. Grant, C.E., Bailey, T.L. & Noble, W.S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017–1018 (2011).
75. Pugh, C.W., Tan, C.C., Jones, R.W. & Ratcliffe, P.J. Functional analysis of an oxygen-regulated transcriptional enhancer lying 3' to the mouse erythropoietin gene. *Proc. Natl. Acad. Sci. USA* **88**, 10553–10557 (1991).

Bibliography

1. Murray, C. J. & Lopez, A. D. Measuring the global burden of disease. *New England Journal of Medicine* **369**, 448–457 (2013).
2. Mendis, S., Davis, S. & Norrving, B. Organizational update: The World Health Organization Global Status Report on Noncommunicable Diseases 2014; one more landmark step in the combat against stroke and vascular disease. *Stroke* **46**, e121–e122 (2015).
3. Bentzon, J. F., Otsuka, F., Virmani, R. & Falk, E. Mechanisms of plaque formation and rupture. *Circulation Research* **114**, 1852–1866 (2014).
4. Sakakura, K. *et al.* Pathophysiology of atherosclerosis plaque progression. *Heart, Lung and Circulation* **22**, 399–411 (2013).
5. Sary, H. C. *et al.* A definition of the intima of human arteries and of its atherosclerosis-prone regions. A report from the Committee on Vascular Lesions of the Council on Arteriosclerosis, American Heart Association. *Circulation* **12**, 391–405 (1992).
6. Fernandez, A. Z., Siebel, A. L. & El-Osta, A. Atherogenic factors and their epigenetic relationships. *International Journal of Vascular Medicine* **2010**, 1–7 (2010).
7. Steinbrecher, U. P., Parthasarathy, S., Leake, D. S., Witztum, J. L. & Steinberg, D.

- 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**, 3883–3887 (1984).
8. Jang, Y., Lincoff, A., Plow, E. F. & Topol, E. J. Cell adhesion molecules in coronary artery disease. *Journal of the American College of Cardiology* **24**, 1591–1601 (1994).
 9. Linton, M. F. & Fazio, S. Class A scavenger receptors, macrophages, and atherosclerosis. *Current Opinion in Lipidology* **12**, 489–495 (2001).
 10. Hansson, G. K. Inflammation, atherosclerosis, and coronary artery disease. *New England Journal of Medicine* **352**, 1685–1695 (2005).
 11. Ross, R. Atherosclerosis – an inflammatory disease. *New England Journal of Medicine* **340**, 115–126 (1999).
 12. Swirski, F. K. *et al.* Monocyte accumulation in mouse atherogenesis is progressive and proportional to extent of disease. *Proceedings of the National Academy of Sciences* **103**, 10340–10345 (2006).
 13. Mallat, Z., Taleb, S., Ait-Oufella, H. & Tedgui, A. The role of adaptive T cell immunity in atherosclerosis. *Journal of Lipid Research* **50**, S364–S369 (2008).
 14. Chen, S., Crother, T. R. & Arditi, M. Emerging role of IL-17 in atherosclerosis. *Journal of Innate Immunity* **2**, 325–333 (2010).
 15. Galis, Z. S. *et al.* Cytokine-stimulated human vascular smooth muscle cells synthesize a complement of enzymes required for extracellular matrix digestion. *Circulation Research* **75**, 181–189 (1994).
 16. Sasaguri, T. *et al.* A role for interleukin 4 in production of matrix metalloproteinase 1 by human aortic smooth muscle cells. *Atherosclerosis* **138**, 247–253 (1998).

17. Bond, M. Inhibition of transcription factor NF-kappaB reduces matrix metalloproteinase-1, -3 and -9 production by vascular smooth muscle cells. *Cardiovascular Research* **50**, 556–565 (2001).
18. Willis, A. I., Pierre-Paul, D., Sumpio, B. E. & Gahtan, V. Vascular smooth muscle cell migration: Current research and clinical implications. *Vascular and Endovascular Surgery* **38**, 11–23 (2004).
19. Strickland, D. K., Kounnas, M. Z. & Argaves, W. S. LDL receptor-related protein: A multiligand receptor for lipoprotein and proteinase catabolism. *The FASEB Journal* **9**, 890–898 (1995).
20. Esterbauer, H., Gebicki, J., Puhl, H. & Jürgens, G. The role of lipid peroxidation and antioxidants in oxidative modification of LDL. *Free Radical Biology and Medicine* **13**, 341–390 (1992).
21. Chisolm, G. M. & Steinberg, D. The oxidative modification hypothesis of atherogenesis: An overview. *Free Radical Biology and Medicine* **28**, 1815–1826 (2000).
22. Itabe, H. *et al.* Oxidized phosphatidylcholines that modify proteins: ANALYSIS by monoclonal antibody against oxidized low density lipoprotein. *Journal of Biological Chemistry* **271**, 33208–33217 (1996).
23. Kohno, H. *et al.* Simple and practical sandwich-type enzyme immunoassay for human oxidatively modified low density lipoprotein using antioxidized phosphatidylcholine monoclonal antibody and antihuman apolipoprotein-B antibody. *Clinical Biochemistry* **33**, 243–253 (2000).
24. Fang, J. C. Circulating autoantibodies to oxidized LDL correlate with impaired coronary endothelial function after cardiac transplantation. *Arteriosclerosis, Thrombosis, and Vascular Biology* **22**, 2044–2048 (2002).

25. Ehara, S. *et al.* Elevated levels of oxidized low density lipoprotein show a positive relationship with the severity of acute coronary syndromes. *Circulation* **103**, 1955–1960 (2001).
26. Toshima, S.-i. *et al.* Circulating oxidized low density lipoprotein levels : A biochemical risk marker for coronary heart disease. *Arteriosclerosis, Thrombosis, and Vascular Biology* **20**, 2243–2247 (2000).
27. Tsimikas, S. *et al.* Oxidized phospholipids, Lp(a) lipoprotein, and coronary artery disease. *New England Journal of Medicine* **353**, 46–57 (2005).
28. Holvoet, P., Vanhaecke, J., Janssens, S., Werf, F. V. de & Collen, D. Oxidized LDL and malondialdehyde-modified LDL in patients with acute coronary syndromes and stable coronary artery disease. *Circulation* **98**, 1487–1494 (1998).
29. Gaut, J. Mechanisms for oxidizing low-density lipoprotein insights from patterns of oxidation products in the artery wall and from mouse models of atherosclerosis. *Trends in Cardiovascular Medicine* **11**, 103–112 (2001).
30. Parthasarathy, S., Steinberg, D. & Witztum, J. L. The role of oxidized low-density lipoproteins in the pathogenesis of atherosclerosis. *Annual Review of Medicine* **43**, 219–225 (1992).
31. Hiramatsu, K., Rosen, H., Heinecke, J. W., Wolfbauer, G. & Chait, A. Superoxide initiates oxidation of low density lipoprotein by human monocytes. *Arteriosclerosis, Thrombosis, and Vascular Biology* **7**, 55–60 (1987).
32. Gordon, S. The role of the macrophage in immune regulation. *Research in Immunology* **149**, 685–688 (1998).
33. Daugherty, A., Dunn, J. L., Rateri, D. L. & Heinecke, J. W. Myeloperoxidase, a catalyst for lipoprotein oxidation, is expressed in human atherosclerotic lesions.

Journal of Clinical Investigation **94**, 437–444 (1994).

34. Moncada, S., Palmer, R. M. & Higgs, E. A. Nitric oxide: Physiology, pathophysiology, and pharmacology. *Pharmacological reviews* **43**, 109–142 (1991).

35. Heinecke, J. W. Oxidants and antioxidants in the pathogenesis of atherosclerosis: Implications for the oxidized low density lipoprotein hypothesis. *Atherosclerosis* **141**, 1–15 (1998).

36. Parthasarathy, S., Raghavamenon, A., Garelnabi, M. O. & Santanam, N. in *Methods in molecular biology* **610**, 403–417 (Humana Press, 2009).

37. Levitan, I., Volkov, S. & Subbaiah, P. V. Oxidized LDL: Diversity, patterns of recognition, and pathophysiology. *Antioxidants & Redox Signaling* **13**, 39–75 (2010).

38. Steinbrecher, U. Receptors for oxidized low density lipoprotein. *Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids* **1436**, 279–298 (1999).

39. Gieseg, S. P., Crone, E. & Amit, Z. in *Endogenous toxins: Targets for disease treatment and prevention* (eds. O'Brien, P. P. J. & Bruce, P. W. R.) 619–645 (Wiley-VCH Verlag GmbH & Co. KGaA, 2010). doi:10.1002/9783527628100

40. Henriksen, T., Evensen, S. A. & Carlander, B. Injury to human endothelial cells in culture induced by low density lipoproteins. *Scandinavian Journal of Clinical and Laboratory Investigation* **39**, 361–368 (1979).

41. Baird, S. K., Reid, L., Hampton, M. B. & Gieseg, S. P. OxLDL induced cell death is inhibited by the macrophage synthesised pterin, 7,8-dihydroneopterin, in U937 cells but not THP-1 cells. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research* **1745**, 361–369 (2005).

42. Hessler, J. R., Robertson, A. L. & Chisolm, G. M. LDL-induced cytotoxicity and its inhibition by HDL in human vascular smooth muscle and endothelial cells in

- culture. *Atherosclerosis* **32**, 213–229 (1979).
43. Jessup, W. & Kritharides, L. Metabolism of oxidized LDL by macrophages. *Current Opinion in Lipidology* **11**, 473–481 (2000).
44. Dimmeler, S., Haendeler, J., Galle, J. & Zeiher, A. M. Oxidized low-density lipoprotein induces apoptosis of human endothelial cells by activation of CPP32-like proteases : A mechanistic clue to the ‘response to injury’ hypothesis. *Circulation* **95**, 1760–1763 (1997).
45. Baird, S. K., Hampton, M. B. & Gieseg, S. P. Oxidized LDL triggers phosphatidylserine exposure in human monocyte cell lines by both caspase-dependent and -independent mechanisms. *FEBS Letters* **578**, 169–174 (2004).
46. Maxeiner, H. *et al.* Complementary roles for scavenger receptor A and CD36 of human monocytederived macrophages in adhesion to surfaces coated with oxidized low-density lipoproteins and in secretion of H₂O₂. *The Journal of Experimental Medicine* **188**, 2257–2265 (1998).
47. Pirillo, A., Norata, G. D. & Catapano, A. L. LOX-1, OxLDL, and atherosclerosis. *Mediators of Inflammation* **2013**, 1–12 (2013).
48. Asmis, R., Begley, J. G., Jelk, J. & Everson, W. V. Lipoprotein aggregation protects human monocyte-derived macrophages from oxLDL-induced cytotoxicity. *Journal of Lipid Research* **46**, 1124–1132 (2005).
49. Sukhanov, S. Novel effect of oxidized low-density lipoprotein: Cellular ATP depletion via downregulation of glyceraldehyde-3-phosphate dehydrogenase. *Circulation Research* **99**, 191–200 (2006).
50. Li, R. *et al.* Angiopoeitin-2 modulates Survivin expression in OxLDL-induced endothelial cell apoptosis. *Biochemical and Biophysical Research Communications*

417, 619–622 (2012).

51. Robbesyn, F., Salvayre, R. & Negre-Salvayre, A. Dual role of oxidized LDL on the NF-kappaB signaling pathway. *Free Radical Research* **38**, 541–551 (2004).

52. Sánchez-Quesada, J. L., S, S. B. & Ordóñez-Llanos, J. Electronegative low-density lipoprotein. *Current Opinion in Lipidology* **15**, 329–335 (2004).

53. Ishikawa, K., Navab, M., Leitinger, N., Fogelman, A. M. & Lusis, A. J. Induction of heme oxygenase-1 inhibits the monocyte transmigration induced by mildly oxidized LDL. *Journal of Clinical Investigation* **100**, 1209–1216 (1997).

54. Bochkov, V. N. & Leitinger, N. Anti-inflammatory properties of lipid oxidation products. *Journal of Molecular Medicine* **81**, 613–626 (2003).

55. Deanfield, J. E., Halcox, J. P. & Rabelink, T. J. Endothelial function and dysfunction: Testing and clinical relevance. *Circulation* **115**, 1285–1295 (2007).

56. Caterina, R. D. *et al.* Nitric oxide decreases cytokine-induced endothelial activation. Nitric oxide selectively reduces endothelial expression of adhesion molecules and proinflammatory cytokines. *Journal of Clinical Investigation* **96**, 60–68 (1995).

57. Matsushita, K. *et al.* Nitric oxide regulates exocytosis by S-nitrosylation of N-ethylmaleimide-sensitive factor. *Cell* **115**, 139–150 (2003).

58. Pober, J. S. & Sessa, W. C. Evolving functions of endothelial cells in inflammation. *Nature Reviews Immunology* **7**, 803–815 (2007).

59. Ghosh, S. & Karin, M. Missing pieces in the NF-κB puzzle. *Cell* **109**, S81–S96 (2002).

60. Moncada, S. & Erusalimsky, J. D. Does nitric oxide modulate mitochondrial energy generation and apoptosis? *Nature Reviews Molecular Cell Biology* **3**, 214–220 (2002).

61. Li, D. & Mehta, J. L. Antisense to LOX-1 inhibits oxidized LDL-mediated upregulation of monocyte chemoattractant protein-1 and monocyte adhesion to human coronary artery endothelial cells. *Circulation* **101**, 2889–2895 (2000).
62. Li, D. Statins modulate oxidized low-density lipoprotein-mediated adhesion molecule expression in human coronary artery endothelial cells: Role of LOX-1. *Journal of Pharmacology and Experimental Therapeutics* **302**, 601–605 (2002).
63. Khaidakov, M. *et al.* Large impact of low concentration oxidized LDL on angiogenic potential of human endothelial cells: A microarray study. *PLoS ONE* **7**, e47421 (2012).
64. Li, D. LOX-1, an oxidized LDL endothelial receptor, induces CD40/CD40L signaling in human coronary artery endothelial cells. *Arteriosclerosis, Thrombosis, and Vascular Biology* **23**, 816–821 (2003).
65. Lievens, D., Eijgelaar, W., Biessen, E., Daemen, M. & Lutgens, E. The multifunctionality of CD40L and its receptor CD40 in atherosclerosis. *Thrombosis and Haemostasis* **102**, 206–214 (2009).
66. Liu, P., Sun, M. & Sader, S. Matrix metalloproteinases in cardiovascular disease. *Canadian Journal of Cardiology* **22**, 25B–30B (2006).
67. Li, D. LOX-1 mediates oxidized low-density lipoprotein-induced expression of matrix metalloproteinases in human coronary artery endothelial cells. *Circulation* **107**, 612–617 (2003).
68. Li, L. & Renier, G. The oral anti-diabetic agent, gliclazide, inhibits oxidized LDL-mediated LOX-1 expression, metalloproteinase-9 secretion and apoptosis in human aortic endothelial cells. *Atherosclerosis* **204**, 40–46 (2009).
69. Dandapat, A., Hu, C., Sun, L. & Mehta, J. L. Small concentrations of oxLDL induce capillary tube formation from endothelial cells via LOX-1 dependent redox-

- sensitive pathway. *Arteriosclerosis, Thrombosis, and Vascular Biology* **27**, 2435–2442 (2007).
70. Brown, M. S. & Goldstein, J. L. Lipoprotein metabolism in the macrophage: Implications for cholesterol deposition in atherosclerosis. *Annual Review of Biochemistry* **52**, 223–261 (1983).
71. Ouimet, M. & Marcel, Y. L. Regulation of lipid droplet cholesterol efflux from macrophage foam cells. *Arteriosclerosis, Thrombosis, and Vascular Biology* **32**, 575–581 (2012).
72. Kennedy, M. A. *et al.* ABCG1 has a critical role in mediating cholesterol efflux to HDL and preventing cellular lipid accumulation. *Cell Metabolism* **1**, 121–131 (2005).
73. Smith, J. D. *et al.* ABCA1 mediates concurrent cholesterol and phospholipid efflux to apolipoprotein A-I. *Journal of Lipid Research* **45**, 635–644 (2004).
74. Phillips, M. C., Johnson, W. J. & Rothblat, G. H. Mechanisms and consequences of cellular cholesterol exchange and transfer. *Biochimica et Biophysica Acta (BBA) - Reviews on Biomembranes* **906**, 223–276 (1987).
75. Jian, B. *et al.* Scavenger receptor class B type I as a mediator of cellular cholesterol efflux to lipoproteins and phospholipid acceptors. *Journal of Biological Chemistry* **273**, 5599–5606 (1998).
76. Maor, I. & Aviram, M. Oxidized low density lipoprotein leads to macrophage accumulation of unesterified cholesterol as a result of lysosomal trapping of the lipoprotein hydrolyzed cholesteryl ester. *Journal of lipid research* **35**, 803–819 (1994).
77. Witztum, J. L. & Steinberg, D. Role of oxidized low density lipoprotein in atherogenesis. *Journal of Clinical Investigation* **88**, 1785–1792 (1991).
78. Jessup, W., Mander, E. L. & Dean, R. T. The intracellular storage and turnover

- of apolipoprotein B of oxidized LDL in macrophages. *Biochimica et Biophysica Acta (BBA) - Lipids and Lipid Metabolism* **1126**, 167–177 (1992).
79. Liao, F., Andalibi, A., deBeer, F. C., Fogelman, A. M. & Lusis, A. J. Genetic control of inflammatory gene induction and NF-kappa B-like transcription factor activation in response to an atherogenic diet in mice. *Journal of Clinical Investigation* **91**, 2572–2579 (1993).
80. Nagy, L., Tontonoz, P., Alvarez, J. G., Chen, H. & Evans, R. M. Oxidized LDL regulates macrophage gene expression through ligand activation of PPAR γ . *Cell* **93**, 229–240 (1998).
81. Tontonoz, P., Nagy, L., Alvarez, J. G., Thomazy, V. A. & Evans, R. M. PPAR γ promotes monocyte/macrophage differentiation and uptake of oxidized LDL. *Cell* **93**, 241–252 (1998).
82. Carr, A. C., McCall, M. R. & Frei, B. Oxidation of LDL by myeloperoxidase and reactive nitrogen species: Reaction pathways and antioxidant protection. *Arteriosclerosis, Thrombosis, and Vascular Biology* **20**, 1716–1723 (2000).
83. Cyrus, T. *et al.* Disruption of the 12/15-lipoxygenase gene diminishes atherosclerosis in apo Edeficient mice. *Journal of Clinical Investigation* **103**, 1597–1604 (1999).
84. Park, Y. M., Febbraio, M. & Silverstein, R. L. CD36 modulates migration of mouse and human macrophages in response to oxidized LDL and may contribute to macrophage trapping in the arterial intima. *Journal of Clinical Investigation* **119**, 136–145 (2008).
85. Park, Y. M. *et al.* Oxidized LDL/CD36 interaction induces loss of cell polarity and inhibits macrophage locomotion. *Molecular Biology of the Cell* **23**, 3057–3068 (2012).

86. Stiko-Rahm, A., Hultgardh-Nilsson, A., Regnstrom, J., Hamsten, A. & Nilsson, J. Native and oxidized LDL enhances production of PDGF AA and the surface expression of PDGF receptors in cultured human smooth muscle cells. *Arteriosclerosis, Thrombosis, and Vascular Biology* **12**, 1099–1109 (1992).
87. Kohno, M. *et al.* Induction by lysophosphatidylcholine, a major phospholipid component of atherogenic lipoproteins, of human coronary artery smooth muscle cell migration. *Circulation* **98**, 353–359 (1998).
88. Mitra, S., Goyal, T. & Mehta, J. L. Oxidized LDL, LOX-1 and atherosclerosis. *Cardiovascular Drugs and Therapy* **25**, 419–429 (2011).
89. Shen, C.-M., Mao, S. J., Huang, G., Yang, P.-C. & Chu, R.-M. Stimulation of smooth muscle cell proliferation by oxLDL- and acetyl LDL-induced macrophage-derived foam cells. *Life Sciences* **70**, 443–452 (2001).
90. Maiolino, G. *et al.* The role of oxidized low-density lipoproteins in atherosclerosis: The myths and the facts. *Mediators of Inflammation* **2013**, 1–13 (2013).
91. Chang, M.-K. *et al.* Apoptotic cells with oxidation-specific epitopes are immunogenic and proinflammatory. *The Journal of Experimental Medicine* **200**, 1359–1370 (2004).
92. Puijvelde, G. van *et al.* Induction of oral tolerance to oxidized low-density lipoprotein ameliorates atherosclerosis. *Circulation* **114**, 1968–1976 (2006).
93. Stemme, S. *et al.* T lymphocytes from human atherosclerotic plaques recognize oxidized low density lipoprotein. *Proceedings of the National Academy of Sciences* **92**, 3893–3897 (1995).
94. Meier, P., Spertini, F., Blanc, E. & Burnier, M. Oxidized low-density lipoproteins activate CD4+ T cell apoptosis in patients with end-stage renal disease through Fas

- engagement. *Journal of the American Society of Nephrology* **18**, 331–342 (2006).
95. Richmond, T. J. & Davey, C. A. The structure of DNA in the nucleosome core. *Nature* **423**, 145–150 (2003).
96. Luger, K., Mäder, A. W., Richmond, R. K., Sargent, D. F. & Richmond, T. J. Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature* **389**, 251–260 (1997).
97. Bednar, J. *et al.* Nucleosomes, linker DNA, and linker histone form a unique structural motif that directs the higher-order folding and compaction of chromatin. *Proceedings of the National Academy of Sciences* **95**, 14173–14178 (1998).
98. Conaway, J. W. Introduction to theme "Chromatin, epigenetics, and transcription". *Annual Review of Biochemistry* **81**, 61–64 (2012).
99. Bird, A. P. CpG-rich islands and the function of DNA methylation. *Nature* **321**, 209–213 (1986).
100. Burdge, G. C. & Lillycrop, K. A. Nutrition, epigenetics, and developmental plasticity: Implications for understanding human disease. *Annual Review of Nutrition* **30**, 315–339 (2010).
101. Nigris, F. de, Lerman, L. O. & Napoli, C. New insights in the transcriptional activity and coregulator molecules in the arterial wall. *International Journal of Cardiology* **86**, 153–168 (2002).
102. Goll, M. G. & Bestor, T. H. Eukaryotic cytosine methyltransferases. *Annual Review of Biochemistry* **74**, 481–514 (2005).
103. Bannister, A. J. & Kouzarides, T. Regulation of chromatin by histone modifications. *Cell Research* **21**, 381–395 (2011).

104. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
105. Carthew, R. W. & Sontheimer, E. J. Origins and mechanisms of miRNAs and siRNAs. *Cell* **136**, 642–655 (2009).
106. Wang, Z. *et al.* Transcriptional and epigenetic regulation of human microRNAs. *Cancer Letters* **331**, 1–10 (2013).
107. Duursma, A. M., Kedde, M., Schrier, M., Sage, C. le & Agami, R. miR-148 targets human DNMT3b protein coding region. *RNA* **14**, 872–877 (2008).
108. Bianchi, M., Renzini, A., Adamo, S. & Moresi, V. Coordinated actions of MicroRNAs with other epigenetic factors regulate skeletal muscle development and adaptation. *International Journal of Molecular Sciences* **18**, 840 (2017).
109. Stenvinkel, P. *et al.* Impact of inflammation on epigenetic DNA methylation - a novel risk factor for cardiovascular disease? *Journal of Internal Medicine* **261**, 488–499 (2007).
110. Sharma, P. *et al.* Detection of altered global DNA methylation in coronary artery disease patients. *DNA and Cell Biology* **27**, 357–365 (2008).
111. Chao, C.-L. *et al.* The graded effect of hyperhomocysteinemia on the severity and extent of coronary atherosclerosis. *Atherosclerosis* **147**, 379–386 (1999).
112. Kim, Y.-R. *et al.* Epigenetic upregulation of p66shc mediates low-density lipoprotein cholesterol-induced endothelial cell dysfunction. *American Journal of Physiology - Heart and Circulatory Physiology* **303**, H189–H196 (2012).
113. Napoli, C. *et al.* Deletion of the p66Shc longevity gene reduces systemic and tissue oxidative stress, vascular cell apoptosis, and early atherogenesis in mice fed

- a high-fat diet. *Proceedings of the National Academy of Sciences* **100**, 2112–2116 (2003).
114. Kumar, A. *et al.* Histone and DNA methylation-mediated epigenetic down-regulation of endothelial kruppel-like factor 2 by low-density lipoprotein cholesterol. *Arteriosclerosis, Thrombosis, and Vascular Biology* **33**, 1936–1942 (2013).
115. Mitra, S. *et al.* Prior exposure to oxidized low-density lipoprotein limits apoptosis in subsequent generations of endothelial cells by altering promoter methylation. *American Journal of Physiology-Heart and Circulatory Physiology* **301**, H506–H513 (2011).
116. Hastings, N. E., Simmers, M. B., McDonald, O. G., Wamhoff, B. R. & Blackman, B. R. Atherosclerosis-prone hemodynamics differentially regulates endothelial and smooth muscle cell phenotypes and promotes pro-inflammatory priming. *American Journal of Physiology - Cell Physiology* **293**, C1824–C1833 (2007).
117. Zampetaki, A. *et al.* Histone deacetylase 3 is critical in endothelial survival and atherosclerosis development in response to disturbed flow. *Circulation* **121**, 132–142 (2009).
118. Cao, Q. *et al.* Histone deacetylase 9 represses cholesterol efflux and alternatively activated macrophages in atherosclerosis development. *Arteriosclerosis, Thrombosis, and Vascular Biology* **34**, 1871–1879 (2014).
119. N’Guessan, P. D. *et al.* Statins control oxidized LDL-mediated histone modifications and gene expression in cultured human endothelial cells. *Arteriosclerosis, Thrombosis, and Vascular Biology* **29**, 380–386 (2009).
120. Choi, J.-H. Trichostatin A exacerbates atherosclerosis in low density lipoprotein receptor-deficient mice. *Arteriosclerosis, Thrombosis, and Vascular Biology* **25**, 2404–

2409 (2005).

121. Okamoto, H. *et al.* Trichostatin A, an inhibitor of histone deacetylase, inhibits smooth muscle cell proliferation via induction of p21(WAF1). *Journal of Atherosclerosis and Thrombosis* **13**, 183–191 (2006).

122. Bekkering, S. *et al.* Oxidized low-density lipoprotein induces long-term proinflammatory cytokine production and foam cell formation via epigenetic reprogramming of monocytes. *Arteriosclerosis, Thrombosis, and Vascular Biology* **34**, 1731–1738 (2014).

123. Holdt, L. M. *et al.* ANRIL expression is associated with atherosclerosis risk at chromosome 9p21. *Arteriosclerosis, Thrombosis, and Vascular Biology* **30**, 620–627 (2010).

124. Motterle, A. *et al.* Functional analyses of coronary artery disease associated variation on chromosome 9p21 in vascular smooth muscle cells. *Human Molecular Genetics* **21**, 4021–4029 (2012).

125. Holdt, L. M. *et al.* Alu elements in ANRIL non-coding RNA at chromosome 9p21 modulate atherogenic cell functions through trans-regulation of gene networks. *PLoS Genetics* **9**, e1003588 (2013).

126. Ji, R. *et al.* MicroRNA expression signature and antisense-mediated depletion reveal an essential role of MicroRNA in vascular neointimal lesion formation. *Circulation Research* **100**, 1579–1588 (2007).

127. Wang, M. *et al.* MicroRNA-21 regulates vascular smooth muscle cell function via targeting tropomyosin 1 in arteriosclerosis obliterans of lower extremities. *Arteriosclerosis, Thrombosis, and Vascular Biology* **31**, 2044–2053 (2011).

128. Zhou, J. *et al.* MicroRNA-21 targets peroxisome proliferators-activated receptor- α in an autoregulatory loop to modulate flow-induced endothelial inflammation. *Pro-*

- ceedings of the National Academy of Sciences* **108**, 10355–10360 (2011).
129. Loyer, X. *et al.* Inhibition of MicroRNA-92a prevents endothelial dysfunction and atherosclerosis in mice. *Circulation Research* **114**, 434–443 (2013).
130. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods* **10**, 1213–1218 (2013).
131. Betteridge, D. *Lipids: Current perspectives*. (Taylor & Francis, 1996). at <<https://books.google.co.uk/books?id=M5kEA7JiAzgC>>
132. Feher, M. & Richmond, W. *Lipids and Lipid Disorders*. (Mosby International Limited, 2000).
133. Hoff, H. F. *et al.* Modification of low density lipoprotein with 4-hydroxynonenal induces uptake by macrophages. *Arteriosclerosis, Thrombosis, and Vascular Biology* **9**, 538–549 (1989).
134. Boullier, A. Minimally oxidized LDL offsets the apoptotic effects of extensively oxidized LDL and free cholesterol in macrophages. *Arteriosclerosis, Thrombosis, and Vascular Biology* **26**, 1169–1176 (2006).
135. Havel, R. J., Eder, H. A. & Bragdon, J. H. The distribution and chemical composition of ultracentrifugally separated lipoproteins in human serum. *Journal of Clinical Investigation* **34**, 1345–1353 (1955).
136. Duggan, S., Rait, C., Gebicki, J. & Giese, S. Inhibition of protein oxidation by the macrophage-synthesised antioxidant 7,8-dihydroneopterin. *Redox Report* **6**, 188–190 (2001).
137. Takei, A. Expression of adhesion molecules by human endothelial cells exposed

to oxidized low density lipoprotein: Influences of degree of oxidation and location of oxidized LDL. *Atherosclerosis* **154**, 79–86 (2001).

138. Cominacini, L. *et al.* Antioxidants inhibit the expression of intercellular cell adhesion molecule-1 and vascular cell adhesion molecule-1 induced by oxidized LDL on human umbilical vein endothelial cells. *Free Radical Biology and Medicine* **22**, 117–127 (1997).

139. Hoff, H. F. & O’Neil, J. Lesion-derived low density lipoprotein and oxidized low density lipoprotein share a lability for aggregation, leading to enhanced macrophage degradation. *Arteriosclerosis, Thrombosis, and Vascular Biology* **11**, 1209–1222 (1991).

140. Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: Ten years of next-generation sequencing technologies. *Nature Reviews Genetics* **17**, 333–351 (2016).

141. Pareek, C. S., Smoczynski, R. & Tretyn, A. Sequencing technologies and genome sequencing. *Journal of Applied Genetics* **52**, 413–435 (2011).

142. Soon, W. W., Hariharan, M. & Snyder, M. P. High-throughput sequencing for biology and medicine. *Molecular Systems Biology* **9**, 640–640 (2014).

143. Pearson, W. R. & Lipman, D. J. Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences* **85**, 2444–2448 (1988).

144. Cock, P. J. A., Fields, C. J., Goto, N., Heuer, M. L. & Rice, P. M. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Research* **38**, 1767–1771 (2009).

145. Andrews, S. FastQC A Quality Control tool for High Throughput Sequence Data. *Babraham Bioinformatics* at <<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>>

146. Gordon, A. & Hannon, G. Fastx-toolkit: FASTQ/a short-reads preprocessing tools. (2010). at <http://hannonlab.cshl.edu/fastx_toolkit>
147. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10 (2011).
148. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology* **10**, R25 (2009).
149. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: Discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105–1111 (2009).
150. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: A fast spliced aligner with low memory requirements. *Nature Methods* **12**, 357–360 (2015).
151. Li, H. & Durbin, R. Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
152. Liao, Y., Smyth, G. K. & Shi, W. The subread aligner: Fast, accurate and scalable read mapping by seed-and-vote. *Nucleic Acids Research* **41**, e108–e108 (2013).
153. Anders, S., Pyl, P. T. & Huber, W. HTSeq—a python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2014).
154. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2013).
155. Quinlan, A. R. BEDTools: The swiss-army tool for genome feature analysis: BEDTools: The swiss-army tool for genome feature analysis. *Current Protocols in Bioinformatics* **47**, 11121–111234 (2014).

156. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods* **14**, 417–419 (2017).
157. Feng, J., Liu, T., Qin, B., Zhang, Y. & Liu, X. S. Identifying ChIP-seq enrichment using MACS. *Nature Protocols* **7**, 1728–1740 (2012).
158. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2009).
159. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15**, 1 (2014).
160. Ritchie, M. E. *et al.* Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research* **43**, e47–e47 (2015).
161. Lun, A. T. & Smyth, G. K. Cseq: A bioconductor package for differential binding analysis of ChIP-seq data using sliding windows. *Nucleic Acids Research* **44**, e45–e45 (2015).
162. Gentleman, R. & Carey, V. in *Encyclopedia of genetics, genomics, proteomics and bioinformatics* **5**, R80 (John Wiley & Sons, Ltd, 2005).
163. Smeds, L. & Künstner, A. ConDeTri - a content dependent read trimmer for Illumina data. *PLoS ONE* **6**, e26314 (2011).
164. Salmela, L. Correction of sequencing errors in a mixed set of reads. *Bioinformatics* **26**, 1284–1290 (2010).
165. Schmieder, R., Lim, Y., Rohwer, F. & Edwards, R. TagCleaner: Identification and removal of tag sequences from genomic and metagenomic datasets. *BMC Bioinformatics* **11**, 341 (2010).

166. Chen, S., Zhou, Y., Chen, Y. & Gu, J. Fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
167. Ewels, P., Magnusson, M., Lundin, S. & Käller, M. MultiQC: Summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**, 3047–3048 (2016).
168. Sonesson, C., Love, M. I. & Robinson, M. D. Differential analyses for RNA-seq: Transcript-level estimates improve gene-level inferences. *F1000Research* **4**, 1521 (2015).
169. Luo, W., Friedman, M. S., Shedden, K., Hankenson, K. D. & Woolf, P. J. GAGE: Generally applicable gene set enrichment for pathway analysis. *BMC Bioinformatics* **10**, 161 (2009).
170. Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and b cell identities. *Molecular Cell* **38**, 576–589 (2010).
171. McLean, C. Y. *et al.* GREAT improves functional interpretation of cis-regulatory regions. *Nature Biotechnology* **28**, 495–501 (2010).
172. Gu, Z. *RGREAT: Client for great analysis.* (2018).
173. Purcell, S. *et al.* PLINK: A tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics* **81**, 559–575 (2007).
174. Welter, D. *et al.* The NHGRI GWAS catalog, a curated resource of SNP-trait associations. *Nucleic Acids Research* **42**, D1001–D1006 (2013).
175. Pers, T. H., Timshel, P. & Hirschhorn, J. N. SNPsnap: A web-based tool for identification and annotation of matched SNPs. *Bioinformatics* **31**, 418–420 (2014).

176. Kalari, K. R. *et al.* MAP-RSeq: Mayo analysis pipeline for RNA sequencing. *BMC Bioinformatics* **15**, 224 (2014).
177. Boria, I., Boatti, L., Saggese, I. & Mignone, F. in *Methods in molecular biology* **14**, 243–256 (Springer New York, 2014).
178. D’Antonio, M. *et al.* RAP: RNA-seq analysis pipeline, a new cloud-based NGS web application. *BMC Genomics* **16**, S3 (2015).
179. Afgan, E. *et al.* The galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Research* **44**, W3–W10 (2016).
180. Varet, H., Brillet-Guéguen, L., Coppée, J.-Y. & Dillies, M.-A. SARTools: A DESeq2- and EdgeR-based R pipeline for comprehensive differential analysis of RNA-seq data. *PLOS ONE* **11**, e0157022 (2016).
181. Moulos, P. & Hatzis, P. Systematic integration of RNA-seq statistical algorithms for accurate detection of differential gene expression patterns. *Nucleic Acids Research* **43**, e25–e25 (2014).
182. Cornwell, M. *et al.* VIPER: Visualization pipeline for RNA-seq, a snakemake workflow for efficient and complete RNA-seq analysis. *BMC Bioinformatics* **19**, (2018).
183. Wang, D. hppRNA - a Snakemake-based handy parameter-free pipeline for RNA-Seq analysis of numerous samples. *Briefings in Bioinformatics* bbw143 (2017). doi:10.1093/bib/bbw143
184. Wolfien, M. *et al.* TRAPLINE: A standardized and automated pipeline for RNA sequencing data analysis, evaluation and annotation. *BMC Bioinformatics* **17**, (2016).
185. Zhao, S. *et al.* QuickRNASeq lifts large-scale RNA-seq data analyses to the next level of automation and interactive visualization. *BMC Genomics* **17**, (2016).

186. Backman, T. W. H. & Girke, T. systemPipeR: NGS workflow and report generation environment. *BMC Bioinformatics* **17**, (2016).
187. Qin, Q. *et al.* ChiLin: A comprehensive ChIP-seq and DNase-seq quality control and analysis pipeline. *BMC Bioinformatics* **17**, (2016).
188. Ji, H., Jiang, H., Ma, W. & Wong, W. H. Using cisgenome to analyze chip-chip and chip-seq data. *Current protocols in bioinformatics* **Chapter 2**, Unit2.13 (2011).
189. Carroll, T. S., Liang, Z., Salama, R., Stark, R. & Santiago, I. de. Impact of artifact removal on ChIP quality metrics in ChIP-seq and ChIP-exo data. *Frontiers in Genetics* **5**, (2014).
190. Ou, J. *et al.* ATACseqQC: A bioconductor package for post-alignment quality assessment of ATAC-seq data. *BMC Genomics* **19**, (2018).
191. Yu, G., Wang, L.-G. & He, Q.-Y. CHIPseeker: An r/bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics* **31**, 2382–2383 (2015).
192. Schmidt, E. M. *et al.* GREGOR: Evaluating global enrichment of trait-associated variants in epigenomic features using a systematic, data-driven approach. *Bioinformatics* **31**, 2601–2606 (2015).
193. Chen, L. & Qin, Z. S. traseR: An r package for performing trait-associated SNP enrichment analysis in genomic intervals: Table 1. *Bioinformatics* **32**, 1214–1216 (2015).
194. Barr, T. L. *et al.* Genomic biomarkers and cellular pathways of ischemic stroke by RNA gene expression profiling. *Neurology* **75**, 1009–1014 (2010).
195. Pedrotty, D. M., Morley, M. P. & Cappola, T. P. Transcriptomic biomarkers of cardiovascular disease. *Progress in Cardiovascular Diseases* **55**, 64–69 (2012).

196. Perisic, L. *et al.* Profiling of atherosclerotic lesions by gene and tissue microarrays reveals PCSK6 as a novel protease in unstable carotid atherosclerosis. *Arteriosclerosis, Thrombosis, and Vascular Biology* **33**, 2432–2443 (2013).
197. Vemuganti, R. & Dempsey, R. Carotid atherosclerotic plaques from symptomatic stroke patients share the molecular fingerprints to develop in a neoplastic fashion: A microarray analysis study. *Neuroscience* **131**, 359–374 (2005).
198. Sukhanov, S., Song, Y. H. & Delafontaine, P. Global analysis of differentially expressed genes in oxidized LDL-treated human aortic smooth muscle cells. *Biochemical and Biophysical Research Communications* **306**, 443–449 (2003).
199. Hu, Y.-W., Zhao, J.-Y., Li, S.-F., Wang, Q. & Zheng, L. Genome-wide profiling to analyze the effects of Ox-LDL induced THP-1 macrophage-derived foam cells on gene expression. *Genomics Data* **2**, 328–331 (2014).
200. Laguna-Fernández, A., Novella, S., Bueno-Betí, C., Marrugat, J. & Hermenegildo, C. Endothelial transcriptomic changes induced by oxidized low density lipoprotein disclose an up-regulation of Jak–Stat pathway. *Vascular Pharmacology* **73**, 104–114 (2015).
201. Mattaliano, M. D. *et al.* LOX-1-dependent transcriptional regulation in response to oxidized LDL treatment of human aortic endothelial cells. *American Journal of Physiology - Cell Physiology* **296**, C1329–C1337 (2009).
202. Minta, J., Yun, J. & St-Bernard, R. Microarray analysis of oxidized low-density lipoprotein (ox-LDL)-regulated genes in human coronary artery smooth muscle cells. *Cell Biology International Reports* **17**, (2010).
203. Chen, T. *et al.* MicroRNA-125a-5p partly regulates the inflammatory response, lipid uptake, and ORP9 expression in oxLDL-stimulated monocyte/macrophages.

- Cardiovascular Research* **83**, 131–139 (2009).
204. Qin, B. *et al.* MicroRNAs expression in ox-LDL treated HUVECs: MiR-365 modulates apoptosis and bcl-2 expression. *Biochemical and Biophysical Research Communications* **410**, 127–133 (2011).
205. Hammad, S. M. *et al.* Oxidized LDL immune complexes and oxidized LDL differentially affect the expression of genes involved with inflammation and survival in human U937 monocytic cells. *Atherosclerosis* **202**, 394–404 (2009).
206. Damián-Zamacona, S. *et al.* Early transcriptomic response to ldl and oxLDL in human vascular smooth muscle cells. *PLOS ONE* **11**, e0163924 (2016).
207. Virgili, F. *et al.* Effect of oxidized low-density lipoprotein on differential gene expression in primary human endothelial cells. *Antioxidants and Redox Signaling* **5**, 237–247 (2003).
208. Deng, D. *et al.* Molecular signatures determining coronary artery and saphenous vein smooth muscle cell phenotypes: Distinct responses to stimuli. *Atherosclerosis Supplements* **7**, 230 (2006).
209. Hirose, K. *et al.* Different responses to oxidized low-density lipoproteins in human polarized macrophages. *Lipids in Health and Disease* **10**, 1 (2011).
210. Reschen, M. E. *et al.* Lipid-induced epigenomic changes in human macrophages identify a coronary artery disease-associated variant that regulates PPAP2B Expression through Altered C/EBP-beta binding. *PLOS Genetics* **11**, e1005061 (2015).
211. Kopylova, E., Noé, L. & Touzet, H. SortMeRNA: Fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics* **28**, 3211–3217 (2012).
212. Griffiths-Jones, S. Annotating non-coding RNAs with Rfam. *Current protocols in bioinformatics* **Chapter 12**, Unit 12.5 (2005).

213. Quast, C. *et al.* The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Research* **41**, D590–D596 (2012).
214. Zhao, S., Fung-Leung, W.-P., Bittner, A., Ngo, K. & Liu, X. Comparison of RNA-seq and microarray in transcriptome profiling of activated T cells. *PLoS ONE* **9**, e78644 (2014).
215. Dirx, A. E. M. *et al.* Tumor angiogenesis modulates leukocyte-vessel wall interactions in vivo by reducing endothelial adhesion molecule expression. *Cancer research* **63**, 2322–2329 (2003).
216. Schlesinger, M. & Bendas, G. Vascular cell adhesion molecule-1 (VCAM-1)-an increasing insight into its role in tumorigenicity and metastasis. *International Journal of Cancer* **136**, 2504–2514 (2014).
217. Alloza, I. *et al.* RNAseq based transcriptomics study of SMCs from carotid atherosclerotic plaque: BMP2 and IDs proteins are crucial regulators of plaque stability. *Scientific Reports* **7**, (2017).
218. Zhang, M. *et al.* Increased plasma BMP-2 levels are associated with atherosclerosis burden and coronary calcification in type 2 diabetic patients. *Cardiovascular Diabetology* **14**, 64 (2015).
219. Shang, Y. *et al.* Transcriptional corepressors HIPK1 and HIPK2 control angiogenesis via TGF-beta-TAK1 dependent mechanism. *PLoS Biology* **11**, e1001527 (2013).
220. Nardinocchi, L. *et al.* Transcriptional regulation of hypoxia-inducible factor 1-alpha by HIPK2 suggests a novel mechanism to restrain tumor growth. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research* **1793**, 368–377 (2009).
221. Han, S.-R. Enzymatically modified LDL induces Cathepsin H in human monocytes:

- Potential relevance in early atherogenesis. *Arteriosclerosis, Thrombosis, and Vascular Biology* **23**, 661–667 (2003).
222. Ganguly, P. & Alam, S. F. Role of homocysteine in the development of cardiovascular disease. *Nutrition Journal* **14**, (2015).
223. Cheng, F. *et al.* Impact of glutathione peroxidase-1 deficiency on macrophage foam cell formation and proliferation: Implications for atherogenesis. *PLoS ONE* **8**, e72063 (2013).
224. Blanchette, M. Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression. *Genome Research* **16**, 656–668 (2006).
225. Jeziorska, D. M., Jordan, K. W. & Vance, K. W. A systems biology approach to understanding cis-regulatory module function. *Seminars in Cell & Developmental Biology* **20**, 856–862 (2009).
226. Hindorff, L. A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences* **106**, 9362–9367 (2009).
227. Remenyi, A., Schöler, H. R. & Wilmanns, M. Combinatorial control of gene expression. *Nature Structural & Molecular Biology* **11**, 812–815 (2004).
228. Maston, G. A., Evans, S. K. & Green, M. R. Transcriptional regulatory elements in the human genome. *Annual Review of Genomics and Human Genetics* **7**, 29–59 (2006).
229. Kolovos, P., Knoch, T. A., Grosveld, F. G., Cook, P. R. & Papantonis, A. Enhancers and silencers: An integrated and simple model for their function. *Epigenetics*

Chromatin **5**, 1 (2012).

230. Kato, M. *et al.* TGF-beta induces acetylation of chromatin and of Ets-1 to alleviate repression of miR-192 in diabetic nephropathy. *Science Signaling* **6**, ra43–ra43 (2013).

231. Boyle, A. P. *et al.* High-resolution mapping and characterization of open chromatin across the genome. *Cell* **132**, 311–322 (2008).

232. Song, L. *et al.* Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. *Genome Research* **21**, 1757–1767 (2011).

233. Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43–49 (2011).

234. Lizio, M. *et al.* Gateways to the FANTOM5 promoter level mammalian expression atlas. *Genome Biology* **16**, 22 (2015).

235. Jadhav, K. & Zhang, Y. Activating transcription factor 3 in immune response and metabolic regulation. *Liver Research* **1**, 96–102 (2017).

236. Shi, H. *et al.* YTHDF3 facilitates translation and decay of m⁶-methyladenosine-modified RNA. *Cell Research* **27**, 315–328 (2017).

237. Nguyen, A. T. & Zhang, Y. The diverse functions of dot1 and h3k79 methylation. *Genes & Development* **25**, 1345–1358 (2011).

238. He, J., Xia, M., Tsang, W. H., Chow, K. L. & Xia, J. ICA1L forms BAR-domain complexes with PICK1 and is crucial for acrosome formation in spermiogenesis. *Journal of Cell Science* **128**, 3822–3836 (2015).

239. Miller, C. L. *et al.* Integrative functional genomics identifies regulatory mechanisms at coronary artery disease loci. *Nature Communications* **7**, 12092 (2016).

240. Hogan, N. T. *et al.* Transcriptional networks specifying homeostatic and inflammatory programs of gene expression in human aortic endothelial cells. *eLife* **6**, (2017).
241. Bassuk, A. G. & Leiden, J. M. A direct physical association between ets and ap-1 transcription factors in normal human t cells. *Immunity* **3**, 223–237 (1995).
242. Toma, I. & McCaffrey, T. A. Transforming growth factor- β and atherosclerosis: Interwoven atherogenic and atheroprotective aspects. *Cell and Tissue Research* **347**, 155–175 (2011).
243. Tsunawaki, S., Sporn, M., Ding, A. & Nathan, C. Deactivation of macrophages by transforming growth factor- β . *Nature* **334**, 260–262 (1988).
244. Jager, S. C. de *et al.* Growth differentiation factor 15 deficiency protects against atherosclerosis by attenuating CCR2-mediated macrophage chemotaxis. *The Journal of Experimental Medicine* **208**, 217–225 (2011).
245. Ahuja, S. S., Paliogianni, F., Yamada, H., Balow, J. E. & Boumpas, D. T. Effect of transforming growth factor-beta on early and late activation events in human t cells. *Journal of immunology (Baltimore, Md. : 1950)* **150**, 3109–3118 (1993).
246. Robertson, A.-K. L. *et al.* Disruption of TGF-beta signaling in t cells accelerates atherosclerosis. *Journal of Clinical Investigation* **112**, 1342–1350 (2003).