Using genetic and genomic approaches to understand haematopoietic cellular biology and dysregulation in disease

Alice Louise Mann

St John's College

Wellcome Sanger Institute

University of Cambridge

November 2017

This dissertation is submitted for the degree of Doctor of Philosophy

Using genetic and genomic approaches to understand haematopoietic cellular biology and dysregulation in disease

Alice Louise Mann
St John's College, University of Cambridge
Professor Nicole Soranzo

Summary

Genetic and genomic approaches have revolutionised the way we address disease aetiology, potential treatment and methods to understand fundamental biology. Many different approaches can be applied to attempt to resolve the mechanisms through which sequence variation disrupts downstream biological processes, which I discuss and apply in this thesis. Specifically, I use tractable haematopoietic cellular systems focusing mainly on neutrophils but also extending these analyses to monocytes and naïve CD4$^+$ cells. First, I introduce the fundamental principles of human genetic variation and associated challenges in resolving functional mechanisms. I then discuss how immune functions are dysregulated in classical autoimmune diseases and emerging evidence for the role of these cells in complex disorders not previously considered immune-mediated. I then integrate molecular phenotypes from resting monocytes, neutrophils and CD4$^+$ T cells with disease-risk loci. Molecular data have the advantage of enabling measurement in larger cohorts and have therefore been used in quantitative trait loci studies to identify variants influencing processes such as gene expression, histone modification or splicing. Using these data, I map molecular mechanisms acting at risk loci associated with a range of complex disorders.

Following this, I highlight recent efforts in applying systematic genome-wide association approaches to cellular and functional traits, many of which can represent intermediate processes disrupted by complex disease. I then apply such approaches to novel neutrophil functional phenotypes to ascertain whether such population-based approaches can be used to gain insight into neutrophil biology. Finally, I discuss studies of haematological blood cell count traits and immunophenotyping and apply a targeted recall-by-genotype study to dissect the relationship between these traits, specifically neutrophil count and surface receptor expression.

In summary, I demonstrate how describing biological mechanisms of genetic variants requires the integration of multiple and complementary datasets and offers insight into fundamental biology, disease risk and therapeutic utility.

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text. It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text

It does not exceed the prescribed word limit for the Biology Degree Committee.

A. L. Mann
November 2017

Acknowledgements

First, I would like to thank my supervisor, Nicole, for the opportunity of working in a stimulating field but also the unbelievable amount that I learnt from the mentorship and the advice. Also thank you for the time, space and patience that enabled me to become an independent scientist and a more confidence person all-round. Second, to the rest of Team Soranzo. There are no words that can properly express my gratitude for the help and advice over these past four years. It is incredible to think of how much I have learnt from those around me and this thesis is testament to what a fantastic team it really is. Particular thanks to Klaudia, who really taught me how to do computational and statistical analysis and for always answering questions even in the busy periods! Thank you to Louella, also for all the questions, the teaching and the unwavering positivity. The same goes for Valentina, Kousik, Lorenzo and Heather- for teaching me some great command line tricks. To Yumi for the enduring support both in work and in life. To Lisa for fantastic pointers on how to make figures look brilliant. Finally, thank you to Stephen for keeping the gene regulation interest going and for putting up with being my desk neighbour and my actual neighbour!

Thank you to all my collaborators. Kate, for enabling me to keep my lab skills going and for the mentorship. To Taco and Anton, for great neutrophil discussions, for the chance to visit Amsterdam and also for your hard work in organising all the replication studies and offering opportunities to follow up interesting results. Thank you to my thesis adviser, Edwin, for stimulating discussions and always a fresh perspective on the challenges. The same to my second adviser, Dan, for useful discussions and pointers. To Dirk, for general advice and also for feedback on my thesis. Thank you to the graduate programme committee, Julian, Carl, Annabel and Christina. Also to Paris, Carol and Sally for all the help and support. To everyone in pipelines at WTSI for helping everything run smoothly. I also thank and acknowledge my funders, the Medical Research Council and the Wellcome Trust.

One of the greatest things about being a PhD student at WTSI, is the fantastic student support network and the friends you make along the way, the PhD life would not have been the same without you all.

To my family, my biggest support network. Being a PhD student in my hometown meant that my family went through the same experiences that I did and were there to get me through it when I needed a little extra push.

To my husband, Tim, there would absolutely be no thesis without you, there are not enough ways for me to thank you for the past four years.

The work described in this thesis was funded by the Medical Research Council and the Wellcome Trust. I thank both of my funding bodies for enabling me to undertake my PhD research.

This thesis is dedicated to my husband,
Timothy Mann

Table of Contents

1 Introduction 1
1.1 Human genetic variation 1
1.2 Identification of trait-associated genetic variants using genome-wide association studies 2
1.3 Challenges in gaining functional insight from GWAS 4
1.4 Assigning function to genetic loci 5
1.4.1 Understanding the non-coding regulatory genome 5
1.4.1.1 Transcription initiation at promoters 5
1.4.1.2 Regulation of transcription by enhancers and other regulatory elements 6
1.4.1.3 Transcription factors 7
1.4.1.4 Transcription elongation and RNA processing 8
1.4.1.5 Chromatin structure 11
1.4.1.6 Higher-order chromatin structure 15
1.4.1.7 Non-coding RNA regulation 16
1.4.2 Quantitative trait loci studies with molecular phenotypes 19
1.5 Functional, cellular and immune phenotypes 23
1.6 Recall-by-genotype studies 24
1.7 Haematopoiesis as a paradigm for genetics 25
1.7.1 Genetics, correlation and causation 29
1.8 Aims of this thesis 31
2 Using immune molecular phenotypes to uncover biological mechanisms of disease-associated genetic loci 35
2.1 Introduction 35
2.1.1 Lessons from genetic and genomic analyses of autoimmune diseases 35
2.1.2 Expanding the complex disease repertoire for which immune phenotypes can resolve mechanisms 38
2.1.2.1 Advanced age-related macular degeneration (AMD) 38
2.1.2.2 Coronary artery disease (CAD) 41
2.1.2.3 Alzheimer’s disease (AD) 43
2.1.2.4 Lung function and chronic pulmonary obstructive disease (COPD) 44
2.1.3 Identification of disease-relevant cell types using enrichment approaches 45
2.1.4 Colocalisation methods evaluate shared genetics across different traits 47
2.1.5 Aims of this chapter 48
2.2 Materials and methods 49
2.3 Results 63
2.3.1 Functional enrichment in five diseases 63
2.3.2 Colocalisation of molecular traits with five diseases 66
2.3.3 Colocalisation of molecular traits reveals potential molecular mechanisms at disease risk loci 71
2.3.4 Therapeutic utility of colocalised gene targets 79
2.3.5 Complex regulatory mechanisms highlighted through integration of multiple molecular evidence 83
2.3.5.1 Cell-type specific regulatory activity at the CAD *SORT1* locus 83
2.3.5.2 In-depth dissection of the molecular mechanisms at the AMD *TNFRSF10A* disease locus 91
2.4 Discussion 107
3 Expanding genetic studies to cellular phenotypes: analytical exploration of neutrophil function phenotypes 116
3.1 Introduction 116
3.1.1 Neutrophil Biology and central role in immune responses 116
3.1.2 Neutrophils and disease 121
3.1.3 Functional phenotypes 122

3.1.4 Aims of this chapter: Investigating neutrophil functional responses 123
3.2 Materials and Methods 125
3.2.1 Neutrophil function data collection and experimental assays 125
3.2.2 Data interpretation and exploration 131
3.2.3 Genetic analysis and integration with epigenomic datasets 145
3.3 Results 148
3.3.1 Phenotype Correlation 148
3.3.2 Genetic variants associated with inter-individual variation in neutrophil function 150
3.3.3 Functional annotation of the PAF & fMLP respiratory burst rs116811177/rs115109232 locus 155
3.3.4 Replication of the rs116811177/rs115109232 locus in an independent cohort ... 158
3.4 Discussion 160

4 Dissecting the functional relationship between neutrophil count and surface expression of cellular receptors 166
4.1 Introduction 166
4.1.1 Aims of this chapter 169
4.2 Materials and Methods 170
4.2.1 Selection of receptor and genetic variant candidates for experimental follow up 170
4.2.1.1 G-CSF receptor 172
4.2.1.2 PLAUR 174
4.2.4 Study samples 177
4.2.5 Flow cytometry assessment of receptor surface expression 177
4.2.6 Phenotype processing and genetic association 180
4.3 Results 187
4.3.1 Measurement of surface expression level in a cohort of 70 individuals 187
4.3.2 Two independent genetic signals are associated with GCSFR surface expression levels 188
4.3.3 The relationship between GCSFR MFI and neutrophil count 192
4.3.4 Evidence of high molecular functionality in the *CSF3R* genic locus 194
4.3.5 Molecular QTL effects of the common GCSFR MFI association 195
4.3.6 Differential *CSF3R* transcript expression associated with rs3917924 196
4.3.7 Investigation of the rs4760, the PLAUR missense SNP 201
4.4 Discussion 203
4.5 Future work: Replication of the GCSFR MFI effect 206

5 Conclusion and outlook 210
5.1 The compromise between high-throughput and in-depth functional insights ... 210
5.2 The ever-increasing complexity of the epigenome and the regulatory code 213
5.3 Causal variants and causal relationships 214
5.4 Ongoing efforts and future goals in functional genomics 217

References 220

Supplementary Information 245

List of Figures

Figure 1.1: Alternative splicing mechanisms produces multiple distinct transcripts 10
Figure 1.2: Histone structure and modifications 12
Figure 1.3: Multiple layers of gene regulation... 14
Figure 1.4: Annotating genetic variants with epigenomic function 18
Figure 1.5: Haematopoiesis and the involvement of essential transcription factors 26
Figure 1.6 Mendelian randomization methodology and assumptions 30
Figure 1.7 Approaches to investigate functional mechanisms of genetic variants 32
Figure 2.1: Schematic of retinal structure and the effect of AMD pathology 39
Figure 2.2: Summary of the BLUEPRINT Epigenome variation project................................ 49
53
Figure 2.3: Summary of molecular QTL and disease locus colocalisation approach 53
Figure 2.4 Schematic of the four models evaluated by colocalisation methods describing the relationship between two trait associations within a locus 56
Figure 2.5: CD11b surface expression is increased on HL60 differentiation with DMSO or ATRA.. 60
Figure 2.6: Enrichment of molecular QTLs in six diseases.. 66
Figure 2.7: Colocalised gene and splicing QTLs per cell type and cell type combination across all diseases 70
Figure 2.8 Disease loci colocalised with multiple molecular features................................... 78
Figure 2.9: Regional association plot of the CAD *SORT1* locus 84
Figure 2.10: Genomic context of the CAD *SORT1* locus 87
Figure 2.11 C/EBPβ and PU.1 are bound directly over rs12740374 in differentiated HL60 cells 88
Figure 2.12: Transcription factor binding at the CAD *SORT1/PSRC1* locus in monocytes and neutrophils.. 90
Figure 2.13: TRAILR1/2 signalling pathways 92
Figure 2.14 Regional association plots for the *TNFRSF10A* locus 94
Figure 2.15: Different genetic signals across cell types at the *TNFRSF10A* locus................ 98
Figure 2.16: Correlation of molecular features at the *TNFRSF10A* locus 99
Figure 2.17: Epigenome characteristics of the *TNFRSF10A* locus.................................... 102
Figure 2.18: Genetic control of *TNFRSF10A* and *RP11-1149O23.3* is not shared across monocytes, T cells and neutrophils.. 104
Figure 3.1: The central role of neutrophils in the immune response 118
Figure 3.2: Neutrophil function study design 124
Fig 3.3: Neutrophil adhesion 127
Fig 3.4: The neutrophil respiratory burst response 129
Fig 3.5: Neutrophil degranulation 130
Fig 3.6: Exploratory data analysis: measurements by time 134
Fig 3.7: Raw data response distributions for respiratory burst and degranulation............... 137
Figure 3.8: Reagent batch effect has a substantial effect on adhesion data....................... 141
Figure 3.9: Reagent batch explains a high proportion of variation in neutrophil adhesion responses.. 143
Figure 3.10: Correction of known batch effects 144
Figure 3.11: Correlation of neutrophil function phenotypes 149
Figure 3.12: Association results for neutrophil function responses.................................... 151
Figure 3.13: QQ plot of the respiratory burst PAF and fMLP RFU at 6 cycles response..... 152
Figure 3.14: Genic location of the most significant signal associated with neutrophil respiratory burst.. 154
Figure 3.15: rs116811177/rs115109232 increases respiratory burst response after neutrophils are stimulated with PAF and fMLP... 154
Figure 3.16: Functional annotation of the rs116811177/rs115109232 (*MAP3K1*) locus 157
Figure 3.17: The rs116811177/rs115109232 (*MAP3K1*) locus in an independent cohort... 159
Figure 4.1: Molecular function of neutrophil count variant assigned genes........................ 171

Figure 4.2: Genetic variants associated with neutrophil count in and around the *CSF3R* gene 173
Figure 4.3: Genetic variants associated with neutrophil count in and around the *PLAUR* receptor gene 175
Figure 4.4: Gating strategy 181
Figure 4.5: MFI by time including outlier exclusion thresholds 182
Figure 4.6: Project and experimental design 188
Figure 4.7: G-CSF surface expression association results 190
Figure 4.8: Directions of effect of the two independent genetic signals associated with GCSFR surface expression and neutrophil count 193
Figure 4.9 Epigenetic and transcript architecture of the *CSF3R* locus 195
Figure 4.10: GCSFR transcript and protein structure 197
Figure 4.11: Relative abundance of all *CSF3R* transcripts, stratified by genotype of rs3917924, may suggest a marginal genetic splicing effect 200
Figure 4.12: Association of the common SNP, rs3917924, with *CSF3R* transcript expression levels 200
Figure 4.13: PLAUR surface expression association results 201
Figure 4.14 GCSFR MFI stratified by haplotype 207
Figure 5.1: Complex molecular regulatory mechanisms 216

List of Tables

Table 1.1: Summary of key blood quantitative trait loci studies 22

Table 1.2: Summary of the main haematological indices, measurement unit and related disorders... 28

Table 2.1: Antibodies used in ChIP-seq and flow cytometry experiments 62

Table 2.2: Number of colocalised disease loci per feature type... 67

Table 2.3: Colocalised genes that are also known drug targets 82

Table 2.4: Summary statistics of lead SNPs with gene expression of *TNFRSF10A* and *RP11-1149O23.3* expression, H3K27ac and H3K4me1 modification phenotypes in monocytes, neutrophils and T cells... 95

Table 2.5: Summary statistics of lead SNPs for H3K27ac with H3K4me1 modification phenotypes in monocytes, neutrophils and T cells... 96

Table 2.6: Association summary statistics of the lead SNP from Blueprint phase 2 genetic analysis 101

Table 2.7: Conditional causality analysis in the *TNFRSF10A* locus 106

Table 3.1: Examples of granule proteins, which granule(s) they are contained within and the function.. 120

Table 3.2: Functional traits and number of individuals for each trait across all three assays and stimuli used.. 138

Table 3.3: Function-associated variants across all assays.. 153

Table 4.1: Candidate variants selected for functional follow-up... 176

Table 4.2: Contents of each sample tube per donor... 178

Table 4.3: Antibodies used for each marker in flow cytometry assays............................... 179

Table 4.4: Reagents and beads 179

Table 4.5: 14 Significant variants associated with the G-CSF receptor surface levels........ 189

Table 4.6: Conditional analysis demonstrated the *CSF3R* locus contains two independent signals 191

Table 4.7: Haplotype frequencies of the Sanquin replication cohort.................................. 208

Abbreviations

AAT	Anti-inflammatory alpha-1-antitrypsin
AAV	Antineutrophil cytoplasmic antibody (ANCA)-associated vasculitis
ABCA1	ATP-binding cassette transporter
AID	Autoimmune disease
AMD	Age-related macular degeneration
AML	Acute myeloid leukaemia
ANCA	Antineutrophil cytoplasmic antibodies
APC	Allophycocyanin
APOE	Apolipoprotein E
APP	Amyloid precursor protein
ARHGEF26	Rho guanine nucleotide exchange factor 26
ATAC-seq	Assay for transposase-accessible chromatin using sequencing
ATP	Adenosine triphosphate
ATRA	All trans-retinoic acid
Aβ	Amyloid β
BAFF	B cell activating factor
BBB	Blood-brain barrier
BM	Bruch's membrane
BPI	Bactericidal/permeability-increasing protein
BRE	TFIIB recognition element
C/EBP	CCAAT/enhancer binding protein
CAD	Coronary artery disease
CANTOS	Canakinumab Antiinflammatory Thrombosis Outcome Study
CBR	Cambridge BioResource
CD	Crohn's disease
CDCV	Common disease-common variant
CEL	Celiac disease
CETP	Cholesterylester transfer protein
CFH	Complement factor H
CFI	Complement factor I
CHD	Coronary heart disease
ChIA-PET	Chromatin interaction analysis by paired-end tag sequencing
ChIP-seq	Chromatin immunoprecipitation with next-generation sequencing
CLP	Common lymphoid progenitor
CMP	Common myeloid progenitor
CNV	Choroidal neovascular membranes

COPD	Chronic obstructive pulmonary disorder
CR1	Complement factor 1
CRC	Colorectal cancer
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
CRP	C-reactive protein
CTCF	Transcriptional repressor CTCF
CVD	Cardiovascular disease
CytoB	Cytochalasin B
DC	Dendritic cell
DG	Diacylglycerol
DHS	Dnase I hypersensitive site
DMSO	Dimethyl sulfoxide
DNMT	DNA methyltransferase
DPE	Downstream promoter element
DTT	Dithiothreitol
EA	Effect allele
EAE	Experimental autoimmune encephalomyelitis
EAF	Effect allele frequency
eQTL	QTL for gene expression
ERK	Extracellular signal-related kinase
FACs	Fluorescence-activated cell sorting
FBC	Full blood count
FEV1	Forced expiratory volume
FITC	Fluorescein isothiocyanate
fMLP	N-formylmethionine-leucyl-phenylalanine
FPKM	Fragments per kilobase of transcript per million fragments sequenced
FS	Forward scatter
FVC	Forced vital capacity
GARFIELD	GWAS Analysis of Regulatory or Functional Information Enrichment with LD correction
GCSFR	Granulocyte colony-stimulating factor receptor
GM-CSF	Granulocyte-macrophage colony-stimulating factor
GMP	Granulocyte/macrophage progenitor
GPCR	G-protein-coupled receptors
GPI	Glycosyl phosphatidylinositol anchor
GTFs	General transcription factors
GWAS	Genome-wide association studies
HAT	Histone acetyltransferase

HDAC	Histone deacetylase
HDM	Histone demethylase
Hep3B	Human hepatocellular carcinoma cells
HFGP	Human Functional genomics project
HLA	Human Leukocyte antigen
hQTL	QTL for histone modification
HRP	Horseradish peroxidase
HSC	Haematopoietic stem cells
IBD	Inflammatory bowel disease
ICAM	Intercellular-adhesion molecules
IFN	Interferon
IGAP	International Genomics of Alzheimer's Project
IL	Interleukin
Inr	Initiator element
iPSC	Induced pluripotent stem cell
LBP	Lipopolysaccharide binding protein
LCL	Lymphoblastoid cell lines
lcRNA	Long non-coding RNA
LD	Linkage disequilibrium
LDL	Low-density lipoprotein
LDL-C	Low-density lipoprotein cholesterol
LFA-1	Lymphocyte function-associated antigen 1
LIPC	Lipase C
LOAD	Late-onset Alzheimer's disease
LPS	Lipopolysaccharides
MAC	Membrane attack complex
MAC-1	Macrophage-1 antigen
MAF	Minor allele frequency
MAPK	Mitogen-activated protein kinase
MD	Maximum difference (effect size estimate for isotype QTLs)
MDP	Muramyl dipeptide
MEP	Megakaryocyte/erythroid progenitor
MFI	Median fluorescence intensity
MHC	Major histocompatibility complex
miRNA	micro RNA
MPO	Myeloperoxidase
MR	Mendelian randomization
MS	Multiple sclerosis

NE-FSC	Neutrophil forward scatter parameter
NE-SFL	Neutrophil side fluorescence
NET	Neutrophil extracellular traps
NK cells	Natural killer cells
NOD2	Nucleotide-binding oligomerization domain-containing protein 2
nvAMD	Neovascular AMD
OA	Other allele
OR	Odds ratio
P-TEFb	Positive transcription elongation factor b
PAF	Platelet-activating factor
PBMC	Peripheral blood mononuclear cells
PBPC	Peripheral blood progenitor cells
PBPCT	Peripheral blood progenitor cells transplantation
PBS	Phosphate-buffered saline
PcHiC	Promoter-capture HiC
PD	Parkinson's disease
PDAC	Pancreatic ductal adenocarcinoma
PE	Phycoerythrin
PI3K	Phosphoinositide 3-kinase
PKC	Protein kinase C
PLAUR/uPAR	Urokinase receptor
PLC	Phospholipase C
PMA	Phorbol myristate acetate
PMNs	Polymorphonuclear leukocytes
Pol II	RNA polymerase II
PP	Posterior probability
PR3	Proteinase 3
PRCS	Peripheral retinal pigment epithelium/choroid/sclera
PSGL-1	P-selectin glycoprotein ligand-1
QTL	Quantitative trait locus
RA	Rheumatoid arthritis
RbG	Recall-by-genotype
RCT	Randomised clinical trial
RDW	Red cell distribution width
RFU	Relative fluorescence unit
RISC	RNA-induced silencing complex
RNAP	RNA polymerase
ROS	Reactive oxygen species

RPE	Retinal pigment epithelium
SCN	Severe congenital neutropenia
SD	Standard deviation
SE	Standard error
siRNA	Small interfering RNA
SJIA	Systemic juvenile idiopathic arthritis
SLE	Systemic lupus erythematosus
SNP	Single nucleotide polymorphism
SS	Side scatter
STZ	Serum-treated zymosan
T1D	Type 1 Diabetes
T2D	Type 2 diabetes
TAD	Topologically-associated domain
TF	Transcription factor
TLR	Toll-like receptor
TM	Transmembrane domain
TNF	Tumour necrosis factor
TNFRSF10A	Tumour necrosis factor receptor superfamily 10A
TRAIL	Tumour necrosis factor-related apoptosis-inducing ligand
TRAILR	Tumour necrosis factor-related apoptosis-inducing ligand receptor
T_{REG}	Regulatory T cells
TSS	Transcription start site
UTR	Untranslated region
VCM	Variable chromatin modules
VEGFA	Vascular endothelial growth factor A
VLDL	Very-low density lipoprotein
WGS	Whole-genome sequencing
YRI	Yoruba in Ibadan

Chapter 1

Introduction

1 Introduction

1.1 Human genetic variation

Genetic variation describes differences in DNA sequences across individuals that are inherited from maternal and paternal chromosomes. Variation also arises through factors such as errors in DNA replication, incomplete DNA repair, or through the controlled development of the highly variable immune receptor genes (MHC, T cell receptor) (Barnes and Lindahl, 2004, Shiina et al., 2009).

In studying population-level variation, we identify associations between the frequency of genetic variants and physiological differences. On a cellular level, we study how every cell in the human body contains the same DNA molecule yet different tissues carry out highly specialised functions. On a molecular level, sequence variation can affect gene expression and epigenetic functionality. Human genetics now encompasses the study of multiple layers of biological processes, which can represent intermediate steps through which variants ultimately affect organismal phenotypes.

The most common type of genetic variation, and the focus of this thesis, is known as a single nucleotide polymorphism (SNP) where the type of nucleotide at one position varies across individuals. In humans, although there are four possible nucleotide combinations (A, T, G, C), in general only two of the possible four nucleotides are ever seen in a population, and one individual carries two copies (alleles) on each diploid chromosome (Casci, 2010, McDaniell et al., 2010). Variants are classified by the occurrence of the least frequent (minor) allele within a population. Common variants occur with minor allele frequency (MAF) ≥ 5% and rare variants are often defined as occurring with a MAF of less than 1%. A second class of variation is structural variation including insertions-deletions (indels), block substitutions, inversions and copy number variants (Frazer et al., 2009).

SNPs are not inherited independently but are correlated, resulting in the systematic association and correlation of alleles at nearby loci (Slatkin, 2008). This structure is known as linkage disequilibrium (LD) and is variable across populations of different ancestries. The International HapMap Project defined LD regions in 269 individuals of four different populations including Yoruba in Ibadan, Nigeria (YRI), Utah with northern and western European ancestry (CEU), Han Chinese in Beijing (CHB) and Japanese in Tokyo (JPT) (International HapMap Consortium, 2005). Alleles of SNPs within the same LD block are inherited more frequently together in the same haplotype. A set of highly correlated loci (high LD) is known as a haplotype block, the boundaries of which are associated with recombination hot spots. Within haplotype blocks, recombination is infrequent. In humans, haplotypes range in size from a few kb to over 100 kb (Wall and Pritchard, 2003, Daly et al., 2001). Despite the observation of a few large blocks, most European population haplotypes are smaller, between 5-20 kb (Wall and Pritchard, 2003). This discovery had wider implications for genetic association studies described in detail in Section 1.2.

1.2 Identification of trait-associated genetic variants using genome-wide association studies

Identification of LD patterns, the establishment of public databases containing millions of curated SNPs and emerging microarray technologies together transformed genetic studies (International HapMap Consortium, 2005, Sachidanandam et al., 2001). At the beginning of the GWAS era, genotyping arrays could be designed based on known LD structure to contain probes assaying approximately 500,000 “tag” SNPs, which captured the majority of common European variation without directly genotyping every variant (Barrett and Cardon, 2006). Later came the development of imputation methods, where high-quality haplotypes from reference populations were and still are used to estimate variant alleles that have not been directly genotyped (Huang et al., 2015). Using reference haplotypes such as those available from the UK10K, 1000 Genomes projects or both combined now enables association tests of tens of millions of variants (UK10K. Consortium et al., 2015, 1000 Genomes Project Consortium et al., 2015, Huang et al., 2015). With the falling costs of whole-genome sequencing, we are also moving to using next-generation sequencing technologies to sequence all sites, which vastly improves the accuracy of rare or private variant detection (Bomba et al., 2017).

Collectively these approaches are called genome-wide association studies (GWAS). For the analysis of diseases, GWAS identify discordant variant allele frequencies between cases and controls, where the association of a higher allele frequency with a disease suggests this is a risk factor. GWAS can also be applied to quantitative traits commonly using linear regression to test for association of the variant with increasing or decreasing trait values. In most studies, variants with additive effects are evaluated, where there is a linear and uniform increase in the trait value/disease risk with each copy of the effect allele (Bush and Moore, 2012).

For each variant, an independent statistical test is applied meaning that for a genome-wide approach, multiple tests are implemented. This greatly increases the probability of detecting false positive associations. When using a p value threshold of 0.05, there is a 5% probability of rejecting the null hypothesis by chance, which equates to a high number of observations if performing millions of tests. Therefore, it is advisable to use a more stringent p value threshold. Based on the International Hapmap Consortium estimation of the number of common (MAF ≥ 5%) independent variants across the genome in a European population, a significance p-value threshold of 5 x 10^{-08} was suggested to control for multiple testing in GWAS (International HapMap Consortium). Alternatively, for a specific cohort, the Bonferroni correction can be used, where the threshold of 0.05 is divided by the number of independent tests. Alternative methods are discussed in Chapter 2 and implemented in Chapter 4.

GWAS have transformed the study of complex traits and diseases by enabling the unbiased screening for significant genetic variants on a genome-wide scale. Hundreds of risk/trait-associated loci have now been identified. As of the 10th October 2017, the NHGRI-EB GWAS catalog contains 52,491 unique variant-trait associations (MacArthur et al., 2017). This high number reflects the genetic architecture of complex traits in that they are multifactorial and explained by many variants influencing genes and pathways that are biologically relevant to the trait (polygenic) (Visscher et al., 2017). However, the overall phenotypic-variation explained by the identified loci is low, suggesting we have not been able to identify all genetic factors that constitute pre-calculated heritability estimates (Visscher et al., 2017). This is referred to as the “missing heritability” problem, which is an important challenge in the field but not the focus of this thesis (Manolio et al., 2009).

Recently, an “omnigenic” model has been suggested in order to interpret the observation that trait heritability is spread across the whole genome, rather than clustered in key genes (Mumbach et al., 2017). This model posits that variants in highly relevant “core genes” directly affect the trait, but all genes (and variants within them) are highly interconnected through extensive networks, although a full knowledge of such connections is currently lacking (Mumbach et al., 2017). These multiple small effects cumulatively effect disease risk. The authors, however, acknowledge that GWAS provide important biological insights, such as identifying core genes and implicate pathways in which lead variants are enriched (Mumbach et al., 2017). Arguably, investigating cellular contexts of identified genes is still of value, particularly as the authors posit that these complex networks are also cell-type specific (Mumbach et al., 2017).

1.3 Challenges in gaining functional insight from GWAS

Despite the successes of GWAS in identifying many trait-associated variants, there remain some key challenges. This main focus of this thesis is in the functional interpretation of the frequency and effect size spectrum of loci that is currently detectable by GWAS. This includes mainly common variants with modest effect sizes or in some cases low-frequency variants with intermediate effects (McCarthy et al., 2008). Mechanistic interpretation represents a major bottleneck in the GWAS to function process. Biological hypotheses are more straightforward when genetic variants are located within coding regions, particularly if the gene function is known and relates to a relevant phenotype and the variation results in a change in amino acid sequence (non-synonymous) (Vasquez et al., 2016).

However, with the advent of GWAS, somewhat surprisingly, it became apparent that more than 90% of trait-associated SNPs were located in non-coding regions of the genome rather than within genic exons (Maurano et al., 2012, Vasquez et al., 2016). This complicates biological interpretation and linking of downstream consequences to the effect on the overall phenotypic trait.

In addition, whilst LD enabled early successes of GWAS by allowing the assessment of tag SNPs, it complicates a definitive identification of the causal SNP(s). Causal SNPs are those that underlie the true trait association and of all variants in the locus demonstrate the best model fit to the phenotype (Battle and Montgomery, 2014). Distinguishing the true causal variants from highly correlated proxy SNPs (those with an $r^2 > 0.8$) is extremely complex as these will likely fit the phenotype equally as well as the true causal variant (Battle and Montgomery, 2014). Larger sample sizes, high-density genotyping, imputation with a high-quality reference panel or whole-genome sequencing all increase the number of variants identified and therefore the likelihood of identifying the causal variant (Battle and Montgomery, 2014). However, even the various statistical approaches for fine-mapping causal variants are limited in cases of high correlation between variants (Chun et al., 2017). Ultimately, functional experiments are required to fully resolve such loci.

There are multiple approaches that attempt to address each of these challenges. This thesis will focus on those that aim to assign function to genetic loci, which can also aid identification of causal variants in some cases. I discuss the type of data and approaches in detail below.

1.4 Assigning function to genetic loci

1.4.1 Understanding the non-coding regulatory genome

Describing the biology of non-coding SNPs requires an understanding of the function of the regulatory genome. While we are unable to predict this function from DNA sequence, through the efforts of large-scale consortia such as ENCODE, ROADMAP and BLUEPRINT, we know now that much of the non-coding genome performs a regulatory function (Encode Project Consortium, 2012, Roadmap Epigenomics Consortium et al., 2015, Adams et al., 2012). There are multiple different layers of (epi)genomic function. The data made available through such consortia can be used to investigate the context of non-coding genetic variation. Below, I summarise our current knowledge of key concepts of epigenomics function and gene regulation.

1.4.1.1 Transcription initiation at promoters

Transcription is a highly regulated process where RNA polymerase (RNAP) enzymes generate an RNA molecule that is complementary to the sequence of DNA. Transcription is initiated at core promoters, which are DNA segments of between 50 and 100 bp (Roy and Singer, 2015). Here, the core transcription machinery including RNAP and general transcription factors (GTFs) assembles. There are various RNAP enzymes, RNA polymerase II (Pol II) transcribes protein-coding genes as well as the non-coding RNAs, small-nucleolar (sn)RNA and micro(mi)RNA (Guiro and Murphy, 2017). Studies utilising cell-free systems identified six GTFs, TFIIA, TFIIB, TFIID, TFIIE, TFIIF and TFIIH (Roeder, 1996, Roy and Singer, 2015). GTFs recognise specific elements of the core promoter through sequence-specific DNA binding. Classification of mammalian promoters based on canonical elements is complex as many do not contain such sequences, which include the TATA box, Initiator (Inr) element, the TFIIB recognition element (BRE) and downstream promoter element (DPE) (Roy and Singer, 2015). For example, only 5-7% of eukaryotic promoters contain a TATA box, and as such there are many cases of non-canonical core promoters (Roy and Singer, 2015). These can contain unmethylated CpG islands or ATG deserts (low occurrence of ATG trinucleotides). Particular chromatin modifications can also mark mammalian promoters, which I discuss in detail below.

Initiation is an important regulated step in transcription. Recently, the association of rs34481144 with severe risk of influenza in humans was shown to involve the disruption of promoter activity as a result of the change in one nucleotide from G (protective) to A (risk) (Allen et al., 2017). rs34481144 resides with the 5’ UTR of the interferon induced transmembrane protein 3 gene, *IFITM3*. Through a series of elegant experiments, the risk allele was shown to be associated with lower *IFITM3* gene expression, lower promoter activity and lower promoter binding of the innate immune interferon, IRF3 and disruption of a CpG methylation site in CD8$^+$ T cells, where reduced methylation increased binding of the insulator factor CTCF. Carriers of the risk allele had lower numbers of CD8$^+$ T cells in the airways during influenza infection, suggesting how reduced *IFITM3* expression (due to reduced promoter activity and demethylation) could increase susceptibility to severe infection and providing evidence for a role of *IFITM3* in the cellular response to infection (Allen et al., 2017). Therefore, sequence-specificity is important to the recruitment of factors required for promoter activity and can be affected by SNPs. This example also highlights a potential role for DNA methylation in regulating gene expression.

1.4.1.2 Regulation of transcription by enhancers and other regulatory elements

Transcription is also regulated by the activity of distal regulatory sequences located upstream or downstream of the promoter (Heinz et al., 2015). These cognate regulatory elements are known as enhancers that activate transcription (Roy and Singer, 2015). Enhancers were originally identified using plasmid-based assays as sequences of no more than 100 bp that could drive gene expression (Banerji et al., 1981, Banerji et al., 1983, Krijger and de Laat, 2016). Enhancer-gene interaction can be promiscuous but also selective and may not necessarily be between the nearest gene (Javierre et al., 2016, Mumbach et al., 2017, Krijger and de Laat, 2016). STARR-seq, a massively parallel reporter assay that enables the assessment of all genome-wide candidate enhancers through the ability of these sequences to drive transcription, was used to show that there were two different clusters of enhancer sequences that separately activated housekeeping genes and developmental genes (Zabidi et al., 2015).

Silencers have similar properties to enhancers but instead act to inhibit transcription. Insulators are boundary elements that inhibit the spreading of transcription and chromatin interactions between neighbouring genomic regions (Gaszner and Felsenfeld, 2006, Ali et al., 2016). CTCF is a key factor in mediating insulation (Ali et al., 2016). Therefore, the spatial and temporal control of gene expression by distal regulators represents another layer of regulation and functionality of the non-coding genome (Ong and Corces, 2011).

The enrichment of SNPs in enhancer regions is now well established and commonly used as a method to assign functionality to non-coding SNPs (Farh et al., 2015, Huang et al., 2017b, Musunuru et al., 2010, Chen et al., 2016a). Multiple examples of SNPs modifying enhancer activity are discussed throughout this thesis and my investigation into disease risk loci in Chapter 2 adds further examples to the many already demonstrated.

1.4.1.3 Transcription factors

Transcription factors regulate gene expression through the sequence-directed binding to DNA at either promoters or regulatory elements such as enhancers. Multiple transcription factors bound at enhancers interact with components such as the Mediator complex or the general TF, TFIID to help recruit RNA polymerase II (Kagey et al., 2010). Looping out of intervening DNA enables interaction between enhancers and promoters. Other factors, such as the cohesin complex can act as scaffold proteins to ensure the stability of these interactions (Kagey et al., 2010, Schmidt et al., 2010). A study that assayed the binding of over 100 transcription factors in colorectal cancer (CRC) LoVo cells found that TFs were bound in clusters across the genome; 75% of the TF peaks were localised in 0.8% of the genome, consistent with previous observations that TF act combinatorially (Yan et al., 2013). Almost all clusters were formed around cohesin, demonstrating the importance of the cohesin complex in enabling complex TF binding (Yan et al., 2013).

The initial selection of enhancers during the differentiation of specific cell lineages is controlled by pioneer transcription factors such as the haematopoietic-specific master regulator, PU.1 (Heinz et al., 2015). Pioneer factors can bind to their cognate motifs prior to any transcriptional activity or chromatin modification and at sites of DNase I inaccessibility (Heinz et al., 2010, Pham et al., 2013). Although PU.1 is an important factor for multiple haematopoietic cell types, PU.1 binding was shown to be cell-type specific (Pham et al., 2013, Heinz et al., 2010). Cooperative binding of PU.1 with other collaborative transcription factors together establish the cell-type specific transcriptional signatures that support lineage-specific differentiation (Heinz et al., 2015, Pham et al., 2013, Adams and Workman, 1995). For example, PU.1 is required for the generation of the general myeloid progenitor and the common lymphoid progenitor but different co-factors are associated with PU.1 at cognate binding sites between macrophages and B cells (Heinz et al., 2010). For example, C/EBP and AP-1 motifs were highly enriched within macrophage-specific distal PU.1 sites whilst E2A, EBF, Oct and NF-κB motifs were enriched in B cell specific PU.1 sites. These additional TFs both had roles in macrophage and B cell differentiation respectively (Heinz et al., 2010). In a PU.1 deficient myeloid progenitor cell line, the absence of PU.1 resulted in a reduced genome-wide C/EBPβ binding pattern. No corresponding PU.1 motifs were found in the C/EBPβ binding sites that remained. Restoration of PU.1 expression in this cell line using a fusion protein, increased PU.1 binding and the number of induced C/EBPβ-bound sites, 75% of which were now co-bound by both TFs and enriched for the PU.1 motif (Heinz et al., 2010). The importance of combinatorial TF binding was confirmed by evaluating the effects of naturally occurring motif mutations in PU.1 and C/EBPα between two different mouse strains (Heinz et al., 2013). Loss of binding of one TF as a result of motif disruption led to the corresponding loss of the second TF and vice versa (Heinz et al., 2013). It is suggested that co-binding of these TFs enables competition with nucleosomes to maintain open chromatin and establish the required cell-type specific binding (Heinz et al., 2010).

Some enhancers require additional co-factors to become fully activated, particularly in response to external or internal signals. Cell type-specific responses to the same stimuli can be achieved through the collaboration between pioneer factors, which first select enhancer sites in the respective cell types and open chromatin (Mullen et al., 2011, Heinz et al., 2015). Following this, a second tier of signal-dependent TFs can bind to these previously established enhancers ensuring that a specific subset of regulatory elements is activated in different cell types (Mullen et al., 2011, Heinz et al., 2010, Ghisletti et al., 2010). Multiple studies have provided evidence for a relatively small number of TFs that interact and bind with pioneer factors to determine cell type specific differentiation and signalling responses by directing the genes to which signalling TFs bind. For example, Mullen *et al.* (2011) used ChIP-seq to show that TGFβ signalling is mediated by Smad2/3, but only 1% of Smad3 binding sites were occupied in more than one cell type between embryonic stem cells, pro-B cells and myotubes (Mullen et al., 2011). Further, they showed that cell-type specific signalling responses were the result of Smad2/3 co-occupying distinct sites with cell-type specific master/pioneer TFs; Oct4 in ES cells, PU.1 in pro-B cells and Myod1 in myotubes (Mullen et al., 2011). Similar cooperative interactions were shown *in vivo* where 61% of NF-κB binding sites in strain-specific mice were already bound by PU.1 and CEBPα before Toll-like receptor 4 (TLR4) stimulation (Heinz et al., 2013).

In summary, cooperative binding of a relatively small and defined group of TFs establishes cell-type specificity of gene expression, lineage differentiation and response to external and internal stimuli. Given the importance of TF in these processes, it is often investigated whether a SNP disrupts TF binding motifs, many examples of which are discussed throughout this thesis.

1.4.1.4 Transcription elongation and RNA processing

Transcriptional regulation is not restricted to initiation. For many mammalian genes, high levels of transcription initiation were observed, but this was not correlated with a high level of gene expression (Guenther et al., 2007). This is due to post-initiation regulation where negative elongation factors can cause Pol II promoter-proximal pausing. Pol II can be released by factors such as the Positive transcription elongation factor (P-TEFb) (Rahl et al., 2010, Zhou et al., 2012). This mechanism is thought to enable fine-tuning in transcription to produce the optimal level of cellular gene transcription as some genes will progress to productive elongation but not all (Zhou et al., 2012). Regulation at this stage also influences processes that can be coupled to transcription such as 5' mRNA capping, splicing and 3' cleavage and polyadenylation (Zhou et al., 2012).

Splicing is the removal of introns within genes to produce a mature processed RNA. Alternative splicing is widespread, occurring with up to 94% multiexonic human genes (Chen et al., 2014). The process can generate multiple transcripts from a single gene as a result of exon skipping, alternative 3' acceptors, alternative 5' donor sites or intron retention (Figure 1.1) (Chen et al., 2014, Nilsen and Graveley, 2010). Splicing can be tissue and developmental-stage specific and is important in disease, with 15% of disease-causing mutations being located in splice sites (Chen et al., 2014). Mutations in splicing factor genes occur at high frequency in haematological cancers (Chen et al., 2014). Extensive transcript diversity as a result of alternative splicing was recently shown in haematopoietic progenitor and precursor cell populations, where 7,881 novel splice junctions were discovered as well as 2,301 alternative splicing events (Chen et al., 2014). In many cases transcript changes were not associated with detectable changes in gene expression, showing that increasing cell-type commitment during lineage differentiation involves the use of alternative transcript isoforms. Therefore, a full understanding of development diversity requires an assessment of all transcriptome effects not just those at the gene level (Chen et al., 2014).

Two methods to quantify splicing events are summarised in Figure 1.1. Both of these methods were used in the BLUEPRINT consortium and as such as used in the analysis of variant function throughout this thesis. Accurate splicing quantification requires RNA-seq data. This is a technique that uses next-generation sequencing to quantify genome-wide gene expression profiles, where high gene expression is represented by an increased number of reads mapping to the corresponding gene location in the reference genome (Marioni et al., 2008). Reads across splicing junctions can also be counted, as is employed in the splicing annotation method, referred to as percent splice in (Figure 1.1) (Chen et al., 2016a). Alternatively, the relative expression levels of all known and annotated transcripts, as defined by GENCODE for example, can be estimated using RNA-seq reads across the gene body (Figure 1.1) (Chen et al., 2016a).

Splicing and donor-acceptor sites are highly sequence specific and therefore could be disrupted by genetic variants (Figure 1.1). In addition, branch points, exonic and intronic splicing enhancers/silencers and mRNA secondary structures can also be influenced by SNPs and result in splicing changes (Hiller et al., 2006). For example, the multiple sclerosis risk SNP, rs17612638 (G) abrogates an exonic splicing silencer, which normally functions to repress the use of a 5' splice site of exon 4 of the *PTPRC* gene (Lynch and Weiss, 2001). This gene encodes a receptor of the protein tyrosine phosphatase family, also known as

CD45, which expressed all nucleated haematopoietic cells (Lynch and Weiss, 2001, Nakano et al., 1990). The immune-related function of this gene suggests that disruption of the tightly regulated exon 4 and resultant alternative transcripts may underlie the observed MS risk (Lynch and Weiss, 2001).

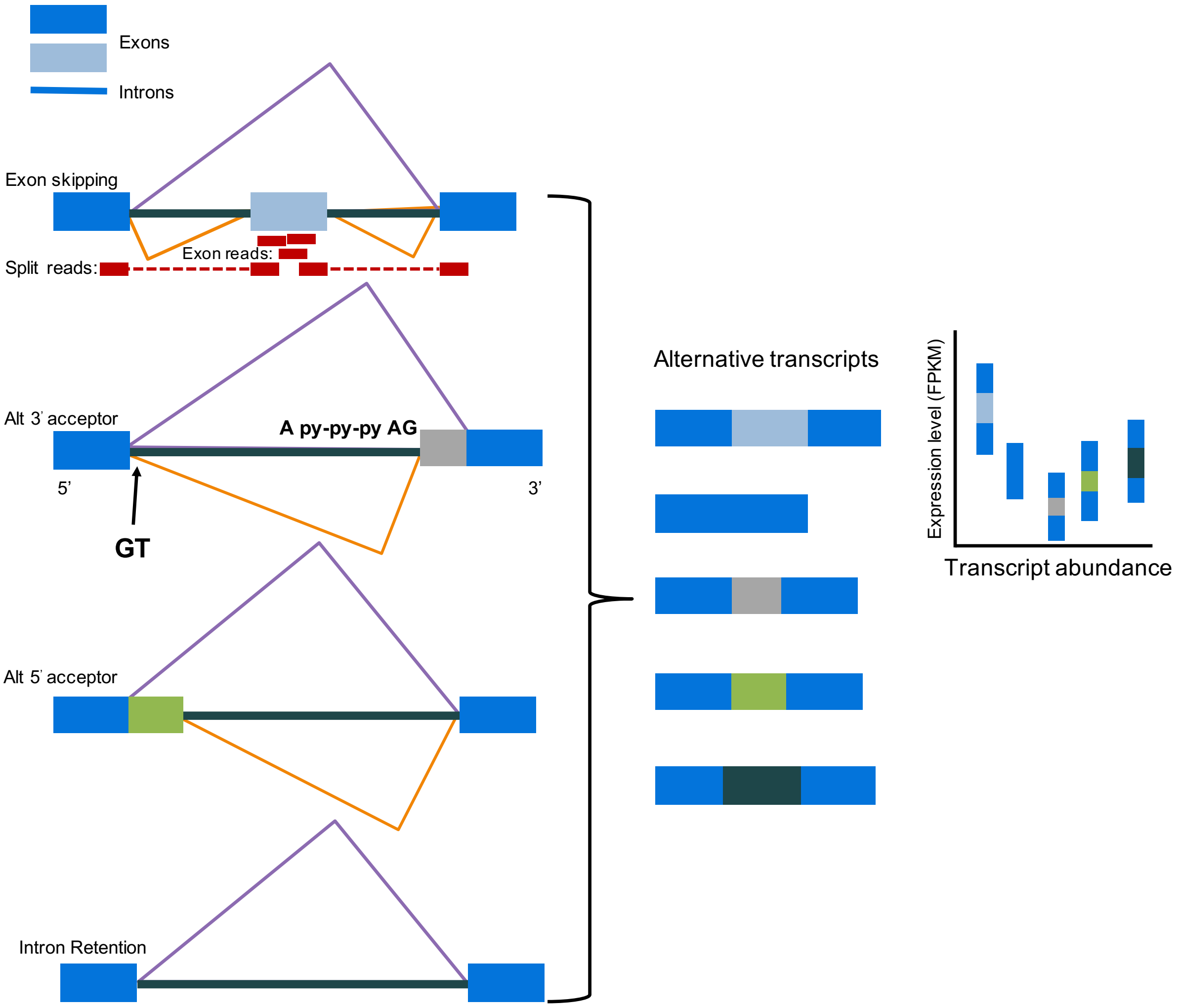


Figure 1.1: Alternative splicing mechanisms produces multiple distinct transcripts
Schematic summarises the different molecular changes involved in alternative splicing and the different possible RNA transcripts. The cognate vertebrate splicing donor site contained in the 5' intron sequence (GT) is also shown along with the 3' splicing acceptor site (AG). The polypyrimidine tract (py-py-py) is a region high in C and T/U pyrimidines. Upstream of this tract is the branch point, which includes an A nucleotide and is important in the splicing molecular mechanism. RNA-seq can be used to quantify the reads (shown in red) across the splicing junctions. The examples above show split reads across two introns and reads within an exon, which both support exon inclusion. Splicing can also be assessed by quantifying the expression of the known alternative transcripts (right) by counting reads expressed in fragments per kilobase of transcript per million fragments sequenced (FPKM). Adapted from (Chen et al., 2014, Nilsen and Graveley, 2010). The percent splice-in method portrayed above is similar to that described by Geuvadis consortium and the information in the figure above was adapted from the (Geuvadis, 2010) webpage listed in the references.

1.4.1.5 Chromatin structure

The regulatory processes described above do not navigate a simple linear DNA sequence, but a complex three-dimensional structure known as chromatin. For DNA to fit into an approximate 10 μm-diameter nucleus it is highly condensed in a nucleoprotein complex (Nieto Moreno et al., 2015). 147 bp of DNA is wrapped 1.7 turns around the histone protein octamer, which is known as a nucleosome (Figure 1.2) (Luger et al., 1997). Octamers comprise two H3-H4 and two H2A-H2B dimers and histone H1 (Figure 1.2) (Luger et al., 1997). Nucleosomes are repeating units (Figure 1.3) and this structure allows further supercoiling and condensation into functional structural domains (Lavelle, 2014). Chromatin remodellers disassemble local compacted nucleosomes to allow access for Pol II and other cofactors, which is essential for active gene expression. This is a state generally referred to as “open chromatin”, whereas “closed chromatin” generally refers to genes and regulatory elements that are inaccessible due to the compact structure (Figure 1.3) (Bannister and Kouzarides, 2011). Further compaction beyond this leads to the formation of constitutively closed heterochromatin containing repressed genes.

Chemical modification of the core histone proteins or protruding amino-terminal tails is also an important regulatory mechanism and confers function to chromatin. Histone modifications are chemical groups that are added to specific residues in the histone protein sequence by chromatin modifying proteins (Figure 1.2). Possible modifications include histone phosphorylation, acetylation, methylation and ubiquitylation. The charges associated with certain modifications, such as the negatively charged phosphorylation, can affect the interactions between histones and, it has been suggested, with the negatively charged DNA phosphate backbone changing the local compaction of DNA (Bannister and Kouzarides, 2011). In addition, these groups can act as molecular “flags” for the binding of histone chaperones, other functional cofactors or additional chromatin remodellers. These proteins contain domains which can recognise modifications, for example, CHD1 binds to H3K4me3 through the chromodomain and the heterochromatin protein, HP1, binds to methylated lysine 9 on histone H3 (Flanagan et al., 2005, Bannister et al., 2001). Proteins containing bromodomains bind to acetyl-lysine modifications and subsequently initiate transcription, therefore targeting these domains offers an attractive potential for specific therapeutics in inflammation, viral infection and in regulating oncogene expression (Filippakopoulos and Knapp, 2014).

Histone modifications are dynamic, can be altered in response to intracellular and extracellular stimuli, and regulate multiple processes beyond chromatin structure and transcription including DNA repair, replication and recombination (Bannister and Kouzarides, 2011). Chromatin structure within genic regions can also influence alternative splicing (as discussed above) through kinetic coupling with transcription whereby nucleosomes act as obstacles, promoting Pol II pausing and influencing exon inclusion/exclusion (Kadener et al., 2001, Schor et al., 2009, Bintu et al., 2012).

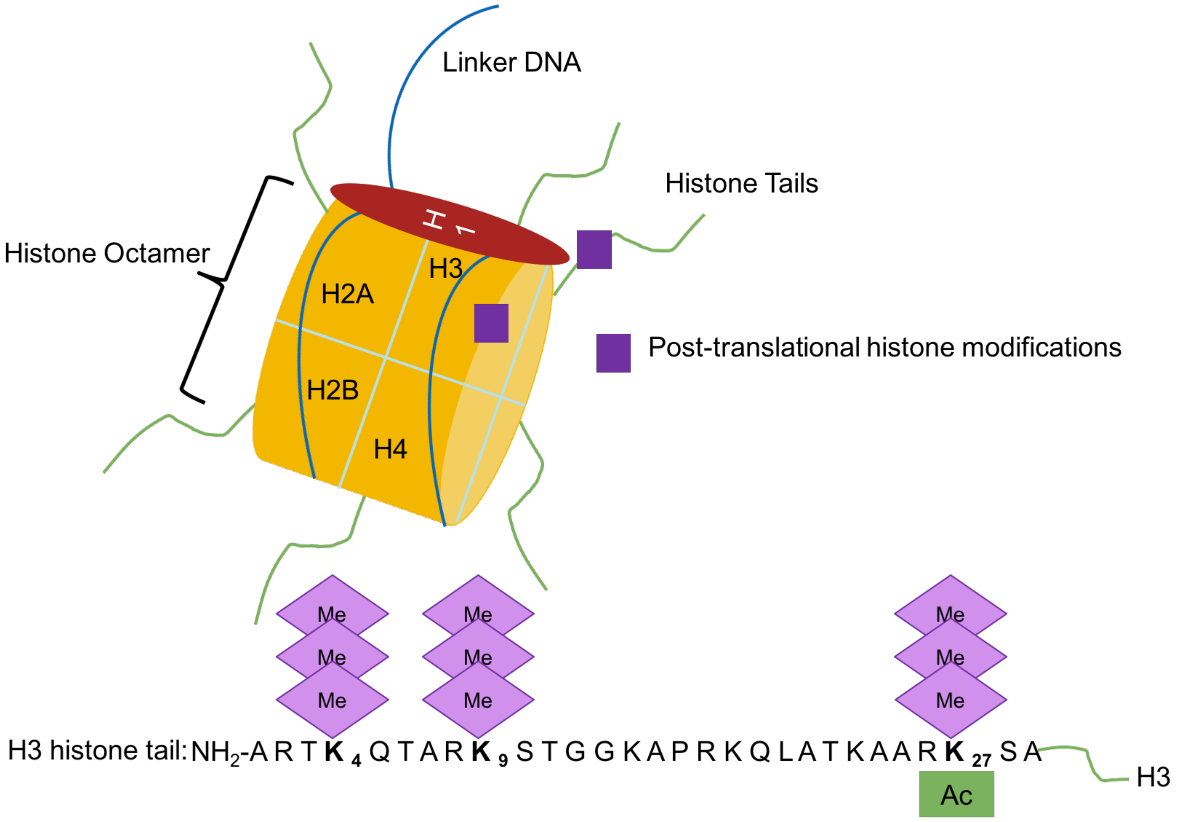

Linker DNA
Histone Tails
Histone Octamer
H1
H2A
H3
H2B
H4
Post-translational histone modifications
Me
Me
Me
Me
Me
Me
Me
Me
Me
H3 histone tail: NH$_2$-A R T **K**$_4$ Q T A R **K**$_9$ S T G G K A P R K Q L A T K A A R **K**$_{27}$ S A
Ac
H3

Figure 1.2: Histone structure and modifications
Nucleosomes are protein structure units consisting of approximately 147 bp of DNA (dark blue) wrapped around the octameric protein structure containing two copies of each of the core histones H2A, H2B, H3 and H4 (yellow). Histone H1 is a linker histone that stabilises higher order structure of chromatin and protects the DNA from nuclease digestion. Most histone modifications (dark purple) occur on the N-terminal histone tails (green). Modifications considered in this thesis are shown below for the histone tail of the H3 core histone. The notation of, for example, H3K4me3 refers first to the histone H3, then to the lysine residue that is fourth in the sequence counting from the N-terminus and then to the chemical modification itself, here a tri-methylation of the lysine residue. Modifications also occur within the core globular protein structure. Adapted from (Fullgrabe et al., 2011).

Genome-wide profiling of histone-bound regions indicated that specific histone modifications are associated with specialised functional genomic regions including promoters or enhancers. As such, these approaches have transformed the way we now identify functional genomic regions (Barski et al., 2007, Hon et al., 2009). To identify these regions the technique, chromatin immunoprecipitation followed by next-generation sequencing (ChIP-seq), uses antibodies specific to a histone modification (or transcription factor) to enrich crosslinked protein-DNA fragments for bound-regions, which are then sequenced (Barski et al., 2007, Schmidt et al., 2009). Bound genomic regions are identified by pile-ups of sequence reads (referred to as "peaks"), which provide a quantitative measurement of genome-wide protein binding (Figure 1.4).

Insights from these genome-wide profiles include the observation that H3K4me3 preferentially associates with promoters and marks regions of active transcription (Hon et al., 2009). Chromatin signatures at promoters were found to be similar across cell types but in contrast, H3K4me1 associated with cell-type specific enhancers (Heintzman et al., 2009 2009). However, many H3K4me1-associated enhancer regions were later found to be inactive when tested in reporter assays, leading to the discovery that active enhancers are marked by a combination of H3K4me1 and H3K27ac (Figure 1.2-1.3) (Creyghton et al., 2010). Instead, H3K4me1 alone marks poised enhancers that may not necessarily be active but could reflect molecular 'memory' of previous activation (Heinz et al., 2015, Creyghton et al., 2010). For example, many inactive haematopoietic stem cell developmental genes were found to be regulated by distal enhancers enriched with H3K4me1 (Creyghton et al., 2010, Cui et al., 2009). H3K27ac, which is deposited by both p300 and CREB binding protein (CBP) can also mark active promoters, when not in conjunction with H3K4me1 (Creyghton et al., 2010).

Clearly the context of chromatin functional state has important consequences for molecular function. For example, using STARR-seq, it was observed that although many sequences possessed the capacity to act as enhancers, many were endogenously repressed (Zabidi et al., 2015, Krijger and de Laat, 2016). The multiple layers of transcriptional regulation and chromatin context are summarised in Figure 1.3. Also shown is the high levels of 5-methyl cytosine (5mC) in closed chromatin, contributing to gene repression (Figure 1.3) (Jones, 2012). Recent advances in genome-wide DNA methylation mapping techniques have highlighted the varied roles of this epigenetic mark depending on the genomic context and interpretation of the functional effect requires appreciation of multiple genomic factors (Jones, 2012).

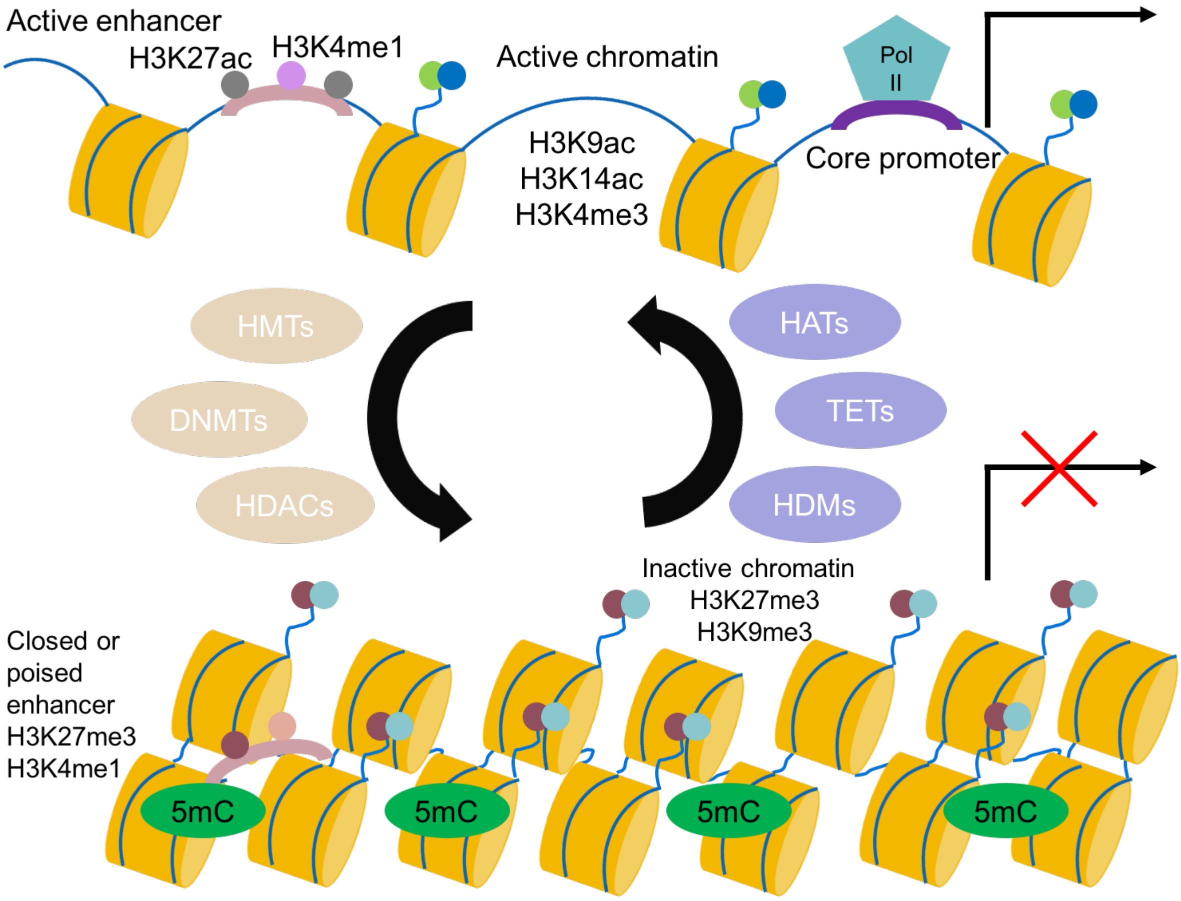


Figure 1.3: Multiple layers of gene regulation
This schematic summarises the many molecular processes that control transcription. The level of DNA compaction controls access of DNA-binding cofactors. In the bottom panel, DNA is highly compacted and hypermethylated at cytosine residues (5mC) preventing access to transcriptional cofactors and repressing gene expression. Histone remodelling proteins (purple) can open chromatin allowing access to other cofactors (top panel). This leads to activation of RNA polymerase II and transcription initiation at the core promoter. Enhancer-bound cofactors can also influence transcription of distal genes through long-range interactions as a result of DNA looping and clustering. DNMT = DNA methyltransferase. HAT = histone acetyltransferase. HDAC = histone deacetylase. HDM = histone demethylase. HMT = histone methyltransferase. TET = ten-eleven translocation. Adapted from (Greco and Condorelli, 2015).

1.4.1.6 Higher-order chromatin structure

With the advent of chromatin conformation capture techniques came the ability to study the three-dimensional spatial genomic structure on a global scale, showing that regulatory loops are widespread and provide another mechanism for transcriptional regulation (Dekker et al., 2002). Chromatin conformation capture (3C) and adaptations of this approach including 4C, 5C, Hi-C, ChIA-PET and promoter-capture HiC (PcHiC), identify long-range interactions by formaldehyde cross-linking of genomic regions located close in physical space (de Wit and de Laat, 2012). Similar to ChIP-seq, these fragments are sequenced and mapped to the reference genome, thereby identifying fragments connecting distally located elements. Chromatin conformation techniques differ by the resolution of interactions detected. For example, genome-wide approaches such as HiC revealed chromatin loops on a larger scale (100kb to 5Mb) referred to as topologically associated domains (TADs) (Lieberman-Aiden et al., 2009, Dixon et al., 2012, Krijger and de Laat, 2016). TADs are more likely to be tissue-invariant but sub-TADs (median size of ~185 kb) and regulatory loops that form within TADs are more tissue-specific and dynamic (Dixon et al., 2012, Phillips-Cremins et al., 2013, Krijger and de Laat, 2016). Stabilisation of TADs requires CTCF and cohesin whereas regulatory loops also require additional tissue-specific TFs (Krijger and de Laat, 2016, Phillips-Cremins et al., 2013, Kagey et al., 2010).

The physical partitioning of the genome into these architectural domains correlates well with genomic function including actively transcribed or repressed genes (Symmons et al., 2014). A definitive causal relationship between promoter-enhancer chromatin looping and gene expression was demonstrated by inducing looping between the beta-globin gene and corresponding super enhancer (locus control region), which resulted in significantly upregulated beta-globin gene expression (Deng et al., 2014).

Connections between distal enhancers and gene targets complicate assignment of genes to regulatory SNPs. HiC data can be used to identify target genes of distal regulatory SNPs. PcHiC is used predominantly in this thesis and achieves higher resolution in comparison to HiC by enriching fragments for genome-wide promoter-mediated interactions using an array with promoter-probes of all cellular genes (Mifsud et al., 2015). This approach was recently used to identify the interacting regions of 31,253 promoters in 17 primary human haematopoietic cells (Javierre et al., 2016). Interactions were found to be highly cell-type specific, recapitulating the haematopoietic tree and interacting regions were enriched in GWAS disease variants (Javierre et al., 2016). Using this data, the 6q23 locus, associated with RA and psoriasis, was found to interact with the promoter of the most proximal gene, *TNFAIP3*, but also with the promoter of *IL20RA,* located 680 kb upstream (McGovern et al., 2016). The risk allele of the likely causal SNP in this locus, rs6927172, correlated with increased gene expression of *IL20RA*, increased binding of both enhancer-associated histone marks and the TF, NFκB (McGovern et al., 2016). In this case, monoclonal therapy against IL-20 has been shown to be effective for both diseases (McGovern et al., 2016). On a genome-wide scale, an independent but similar capture approach, HiChIP, was used to map disease SNP target genes (Mumbach et al., 2017). Instead of focusing on promoter interactions, HiChIP is a protein-centric technique that was recently used with H3K27ac as a bait to assay interactions in T cell populations (Mumbach et al., 2017). Using H3K27ac interaction maps, 2,597 target genes were identified for 684 autoimmune disease variants (Mumbach et al., 2017, Trynka, 2017). Only 14% of the mapped target genes represented the closest gene to the GWAS variant. This demonstrates the utility of interaction data to identify target genes, which is important in the translation of GWAS to the clinic.

Capture techniques can be used to identify SNP target genes, but long-range interactions could themselves be disrupted by these variants. Disruption of TF binding has long been suggested as the predominant mechanism underlying regulatory variation (Pai et al., 2015). However, only a minority, 10-20%, of GWAS SNPs were found to be located within TF binding motifs (of 823 variants assessed), suggesting other regulatory mechanisms may underlie genetic associations (Farh et al., 2015). Evidence of allele-specific interactions has been observed, using ChIA-PET of CTCF and Pol II in different human cell lines. For example, 50 loci showed allele-specific tandem loops (loops coordinated by two CTCF motifs positioned in a tandem manner) that contained phased SNPs within the gene body (Tang et al., 2015). 44% of these loci displayed allele-specific expression (Tang et al., 2015). The authors also showed that the asthma-associated SNP, rs12936231, disrupted a CTCF motif and CTCF binding further abrogating looping and chromatin topology, which they postulated could represent the primary molecular event underlying the locus (Tang et al., 2015). Similar observations have been made combining H3K27ac HiChIP interaction data from primary human cells with available genome phasing (Mumbach et al., 2017). The authors observed 4.2% of loops exhibited allelic bias (FDR < 0.05) where risk alleles either disrupted or increased enhancer-gene interactions (Mumbach et al., 2017). Thorough examination of the allelic bias of chromatin interactions in a larger population-scale cohort is needed to establish this as a widespread disease-relevant regulatory mechanism.

1.4.1.7 Non-coding RNA regulation

90% of the genome is transcribed into non-coding RNAs including ribosomal, transfer-RNAs, long non-coding RNAs and microRNAs, compared to 2-3% transcribed to protein (Roy and Singer, 2015, Lee, 2012). miRNAs are short (19-24 nucleotides) and function to cleave or repress complementary mRNA post-transcriptionally where binding is mediated by the RNA-induced silencing complex (RISC) (Hrdlickova et al., 2014). The translation of more than half of protein-coding genes is regulated by miRNAs (Hrdlickova et al., 2014). Many lncRNAs, which consist of a heterogeneous group of RNAs more than 200 nucleotides, are thought to regulate expression of protein-coding genes (Harrow et al., 2012, Hrdlickova et al., 2014). lncRNAs exhibit cell-type specific expression and widespread regulatory functions through interaction with DNA, RNA or protein enabling the control of processes such as gene silencing, RNA maturation and transport, protein production and chromatin remodelling (Derrien et al., 2012, Hrdlickova et al., 2014).

Non-coding RNAs have been implicated in a range of neurodegenerative, cardiovascular and autoimmune diseases as well as cancer (Hrdlickova et al., 2014). Disease SNPs have been shown to confer risk by disrupting the function of non-coding RNAs, for example, by altering RNA expression or by changing binding sites in target genes. rs57095329 is associated with systemic lupus erythematosus (SLE) and located in the promoter of microRNA, miR-146a (Luo et al., 2011, Hrdlickova et al., 2014). Increased SLE risk is associated with lower miR-146a expression levels, observed in peripheral blood leukocytes (Luo et al., 2011). Upregulated type I interferon pathway activity is known to occur in SLE pathogenesis and miR-146a functions as a negative regulator of this activity, explaining how a decreased miRNA expression could increase disease risk (Luo et al., 2011, Tang et al., 2009). Non-coding RNA function and target gene interaction is another important regulatory function to consider in genetic function studies. Figure 1.4 summarises how all of the described epigenomic data can be used to annotate function of trait-associated variants and in part aid the prediction of putative causal SNPs.

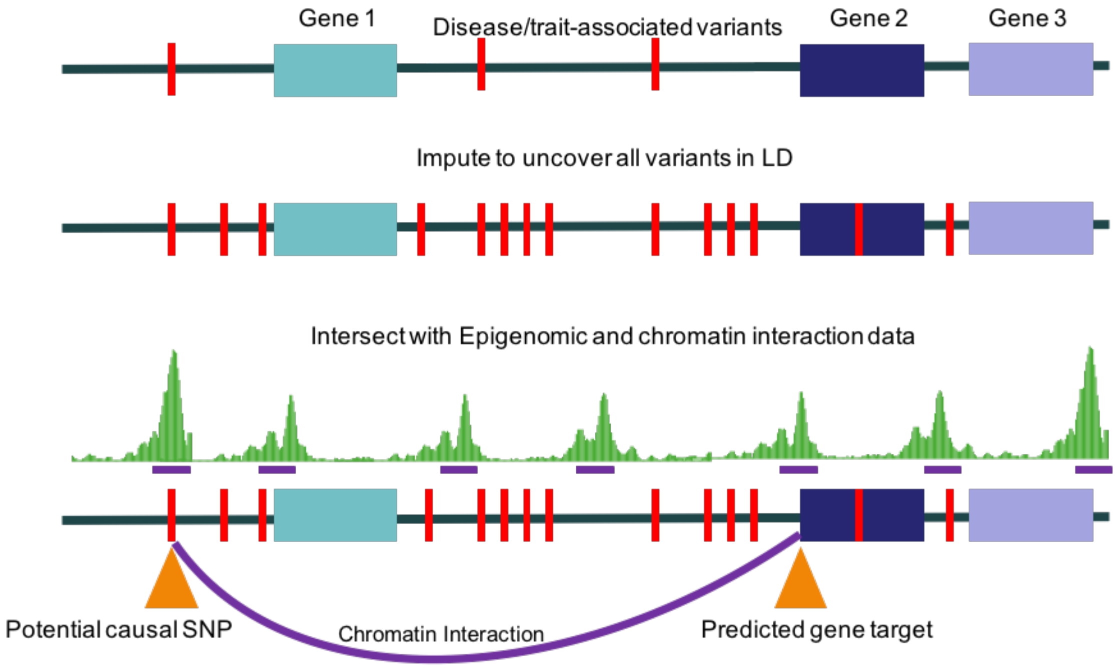

Gene 1
Disease/trait-associated variants
Gene 2
Gene 3
Impute to uncover all variants in LD
Intersect with Epigenomic and chromatin interaction data
Potential causal SNP
Chromatin Interaction
Predicted gene target

Figure 1.4: Annotating genetic variants with epigenomic function
Schematic summarises initial steps in predicting molecular mechanisms of trait-associated SNPs. Imputation, targeting genotyping or use of whole-genome sequencing data identifies all variants in LD. Disease-associated SNPs are intersected with epigenomic regions such as chromatin modification or transcription factor binding (ChIP-seq binding peaks in green). Combined with high-resolution chromatin interaction data, putative target genes can be identified. Further techniques to identify function such as quantitative trait studies are discussed below. Figure based on (Krijger and de Laat, 2016).

1.4.2 Quantitative trait loci studies with molecular phenotypes

Annotating the genome with epigenomic data (Figure 1.4), while helping to highlight molecular function, is prone to chance overlaps. Alternatively, using epigenomic data as a quantitative trait in association mapping can identify, with statistical confidence, specific variants (and those in high LD) associated with disrupting a molecular function. If a genomic locus is associated with both a disease or complex trait and with a molecular phenotype such as gene expression, this is a strong indicator of possible causal mechanism (Nica and Dermitzakis, 2013).

Variation in gene expression can arise from environmental factors, epigenetic effects, random biological noise and genetic effects. QTL mapping uncovers the genetic basis of variation in quantitative phenotypes in a similar approach to GWAS. Smaller cohorts can reduce power and therefore, rather than genome-wide, the number of variants tested in a QTL study is constrained within a genomic window surrounding each molecular feature. These QTLs are referred to as cis-QTLs, which are SNPs that act locally to the feature being investigated (Nica and Dermitzakis, 2013). The definition of "local" can vary between studies; a window of 1 Mb either side of the start and end of the feature was used in the Chen *et al.* (2016) study. This approach limits the burden of multiple testing if all genome-wide variants were assessed. Depending on the assay, the expression of all genes (~22,000) can be tested for cis-QTLs.

Early studies showed heritability of gene expression, chromatin modifications and transcription factor binding and identified that eQTLs (SNPs associated with gene expression variation) were fairly widespread, with some observations of up to 30% of genes having an eQTL in lymphoblastoid cell lines (LCL) (Stranger et al., 2007, Price et al., 2011, Grundberg et al., Pickrell et al., 2010, Montgomery and Dermitzakis, 2011, McDaniell et al., 2010, Pai et al., 2015). With increasing sample sizes and denser genotypes or sequenced data, the number of discovered eQTLs has increased. For example, the latest G. TEx analysis of RNA-seq gene expression across 44 tissues with 449 donors identified 152,869 cis-eQTLs for 19,725 genes corresponding to 50.3% and 86.1% of all known lincRNA and protein-coding genes respectively (G. TEx Consortium, 2017).

Cis-eQTLs are enriched at gene start sites and variants upstream of the TSS are observed to have greater effect sizes than those in gene bodies, suggesting that SNPs regulating transcription have a larger impact than those that may regulate post-transcriptional processes (G. TEx Consortium, 2017). However, splice site QTLs or those that introduce a stop codon do have a high impact on downstream consequences (G. TEx Consortium, 2017). Early eQTL studies demonstrated high cell-type specificity, Dimas *et al.* (2009)

identified that 69-80% of eQTLs across three cell types, LCLs, primary fibroblasts and umbilical T cells were cell type specific (N = 75) (Dimas et al., 2009). Similar tissue specificity has been later confirmed in primary cell types (Chen et al., 2016a). Cell-type specificity can also manifest as opposing direction of effects of the same QTLs in different contexts. For example, Raj *et al.* (2014) identified 7000 shared eQTLs between monocytes and T cells (Raj et al., 2014). The effect size for most eQTLs, defined as the most significant SNP per gene, was concordant across the two cell types but for 42 genes, the most significant SNP had opposing directions where the allele with increased expression in one cell and decreased in the other (Raj et al., 2014). QTL studies in stimulated cell types have shown that context specificity not only applies to different cell types but also to different active states. Specific QTLs were only detected in activated immune cells when stimulated by, for example, bacterial components (LPS) or inflammatory cytokines (IFN-γ) (Fairfax et al., 2014, Naranbhai et al., 2015, Kim-Hellmuth et al., 2017, Alasoo et al., 2017).

eQTLs are often used to integrate with GWAS SNPs to identify gene targets. Zhu *et al* (2016) used a Mendelian randomization method adapted for summary statistics to analyse complex trait and disease GWAS and blood eQTL data (N = 5311) and subsequently identified 126 loci for which there was evidence of pleiotropy between gene expression and complex trait variance (Zhu et al., 2016). Here, pleiotropy describes genetic loci associated with two traits that may not be linked via a causal mechanism where the variant affects a phenotype through an endophenotype such as gene expression. Importantly, for approximately 60% of the colocalised cases, the regulated gene target, as identified by an eQTL, was not the nearest gene to the sentinel GWAS SNP (Zhu et al., 2016). Therefore, identifying gene targets based on proximity may lead to incorrect assignment.

QTL studies also allow the integrated study of genetic effects on gene expression, chromatin and TF binding, which has provided many insights into the mechanism of gene regulation. 55% of eQTLs in LCLs overlapped with DNase I hypersensitivity QTLs marking open chromatin, suggesting that a subset of eQTLs may influence gene expression through disruption of chromatin modification or transcription factor binding (Degner et al., 2012). Three studies measuring chromatin state, modification, TF binding and Pol II occupancy, provided initial evidence of high variability in enhancer function as well as suggesting that TF binding was the primary mechanism underlying modification of regulatory chromatin (Table 1.1) (Kasowski et al., 2013, McVicker et al., 2013, Kilpinen et al., 2013). These observations were confirmed and expanded by two recent studies in LCLs that assayed genome-wide binding of histone modifications, PU.1 and Pol II binding (Grubert et al., 2015, Waszak et al., 2015). Both studies showed extensive local correlation of molecular features in defined genomic windows (< 1Mb), which Waszak *et al.* (2015) referred to as variable chromatin modules (VCMs). Interestingly, SNP-mediated changes in the local chromatin state were also correlated with those observed more distally in regions located up to 200 kb away (Denker and de Laat, 2015). This coordination was shown to result from physical interaction; Grubert *et al.* (2015) showed that 15% of proximal hQTLs were associated with changes at distal histone modifications that were connected by long-range chromatin interactions by using HiC and ChIA-PET data. Distal hQTLs were enriched within TADs and the majority of local-distal QTL pairs occurred between different enhancers (Grubert et al., 2015, Koch, 2015). Both studies provided evidence that TF activity underpinned chromatin variation, which in turn correlated with gene expression, in 99% of cases positively (Waszak et al., 2015). A single genetic variant could, therefore, propagate to multiple correlated features, perhaps explaining why a degree of chromatin variation cannot be correlated with proximal effects (Waszak et al., 2015). Single disease SNPs could therefore disrupt an entire coordinated molecular system, supporting the use of epigenomic data including chromatin interactions in identifying disease mechanisms and target genes and thus demonstrating the power of QTL studies to provide medically relevant insights as well as improve our understanding of genomic regulation (Koch, 2015, Denker and de Laat, 2015).

Chen *et al.* (2016) extended these efforts by assaying gene expression, splicing, DNA methylation, H3K4me1 and H3K27ac QTLs across multiple primary human cell types; monocytes, CD4$^+$ T cells and neutrophils (Chen et al., 2016a). Of the 20,403 genes assessed across the three cell types between 33.9-39.3% of genes had an eQTL. Here, an average of 9.89% of methylation probes, 25.7% of H3K4me1 peaks and 11.5% of H2K27ac peaks had at least one QTL. Confirming previous observations, there was a high degree of cell-type specificity to all marks (Dimas et al., 2009, Chen et al., 2016a). Particularly, hQTLs were highly cell specific, as expected for enhancer function. By considering lead SNPs and those in high LD ($r^2 \geq 0.8$), ~43.4% of eQTLs were also hQTLs, confirming previous observations of high correlation between chromatin and gene expression. For 18.4% of the genes, a splicing QTL effect was identified, but these were largely independent of eQTLs, shown by a low concordance of lead QTLs for the respective traits ($r^2 < 0.1$). There was also a high degree of colocalisation with autoimmune disease, which is discussed in detail in Chapter 2. The details of key QTL studies are summarised in Table 1.1.

In summary, observations from QTL studies and genome-wide approaches discussed above both support the role of key TFs underpinning chromatin state effects and gene expression, at least for a subset of sites. For regulatory QTLs that cannot be explained by TF binding or correlated with gene expression effects, it remains to be shown whether these effects could be explained by disruption of long-range interactions or whether there is extensive redundancy between enhancers removing downstream consequences of genetic disruption.

Author	Cell type	Stimulated/Resting	Molecular Trait	Trait Assay	Cohort
(Dimas et al., 2009)	LCL	Resting	Gene expression	Microarray	75
(Kasowski et al., 2010)	LCL	Resting	NF-kB, Pol II	ChIP-seq	10
(Maranville et al., 2011)	LCLs	Glucocorticoids	Gene expression	Microarray	114
(Degner et al., 2012)	YRI LCL	Resting	Open chromatin	DNase-seq	70
(Barreiro et al., 2012)	Dendritic cells	*M.tuberculosis*	Gene expression	Microarray	65
(Westra et al., 2013)	Whole blood	Resting	Gene expression	Microarray	5311
(Battle et al., 2014)	Whole blood	Resting	Gene expression	RNA-seq	922
(Lappalainen et al., 2013)	LCL	Resting	Gene expression, miRNA	RNA-seq	452-462
(Kasowski et al., 2013)	LCL	Resting	H3K27ac, H3K4me1, H3K4me3, H3K36me3 and H3K27me3, CTCF, SA1	ChIP-seq	19
(Kilpinen et al., 2013)	LCL	Resting	H3K4me1, H3K4me3, H3K27ac, H3K27me3, TFIIB, Pu.1, MYC, Pol II	ChIP-seq	8 + 2 trios
(McVicker et al., 2013)	YRI LCL	Resting	H3K4me1, H3K4me3, H3K27ac, H3K27me3, Pol II	ChIP-seq	10
(Ding et al., 2014)	CEU LCL	Resting	CTCF	ChIP-seq	51
(Raj et al., 2014)	CD4$^+$ T cell, Monocytes	Resting	Gene expression	Microarray	461
(Fairfax et al., 2014)	Monocytes	LPS (2h), LPS (24h), IFNγ (24h)	Gene expression	Microarray	262-414
(Lee et al., 2014)	Dendritic cells	LPS (5hr), influenza (10hr), IFNβ (6.5hr)	Gene expression	Microarray	534
(Naranbhai et al., 2015)	Neutrophils	Resting	Gene expression	Microarray	101
(Kumasaka et al., 2016)	CEU LCL	Resting	Open chromatin	ATAC-seq	24
(Caliskan et al., 2015)	PBMCs	Rhinovirus	Gene expression	Microarray	98
(Waszak et al., 2015)	CEU LCL	Resting	PU.1, Pol II, H3K4me1, H3K4me3, H3K27ac	ChIP-seq	47
(Chen et al., 2016a) BLUEPRINT	Monocytes, neutrophils, CD4$^+$ T cells	Resting	H3K27ac, H3K4me1, gene, splicing, methylation	ChIP-seq, RNA-seq, 450K	Up to 197
(Joehanes et al., 2017)	Whole blood	Resting	Gene and exon expression	Microarray	5257
(Kim-Hellmuth et al., 2017)	Monocytes	LPS, MDP, 5'-ppp-dsRNA (90min, 1 h)	Gene expression	Microarray	134
(Alasoo et al., 2017) (preprint)	iPSC differentiated macrophages	IFNγ (18h), *Salmonella* (5h), IFNγ + *Salmonella*	Gene expression, chromatin accessibility/open chromatin	RNA-seq, ATAC-seq	86, 42
(G. TEx Consortium, 2017)	44 Multiple tissues	*post mortem*	Gene expression	RNA-seq	449
Watt et al., 2018 (in preparation)	Neutrophils	Resting	H3K4me3, H3K27me3, PU.1, CEBPB, CTCF	ChIP-seq	22-110

Table 1.1: Summary of key blood quantitative trait loci studies

1.5 Functional, cellular and immune phenotypes

Beyond molecular phenotypes, heritable genetic variation has been observed in cellular and functional phenotypes. Examples include the levels of a broad range of blood cell types and surface receptor expression levels quantified using FACs-based immunophenotyping (Orru et al., 2013, Roederer et al., 2015) as well as cytokine production and circulating cytokine levels (Brodin et al., 2015, Ahola-Olli et al., 2017). These additional phenotypes allow comprehensive insights into immune functions and disease risk.

The Human Functional Genomics Project (HFGP) has collated an array of deeply phenotyped individuals with information such as microbiome composition, immune responses against human pathogens and disease status (autoimmune, diabetes, Lyme's disease, gout) (Netea et al., 2016, Li et al., 2016b). Li *et al.* (2016) demonstrated how host genetics plays a major role in the variation of immune cell cytokine responses from either whole blood, peripheral blood mononuclear cells (PBMCs) or macrophages stimulated *ex vivo* in a healthy population (Li et al., 2016b). Interestingly, the authors observed that the cytokine with the strongest inter-individual variation was IL6. Variants in this pathway have been previously associated with a multitude of diseases (Chapter 2). This further supports the functional importance of this cytokine in immune responses. In total, 17 novel genome-wide significant QTLs were associated with the production of mostly monocyte- or T cell-specific cytokines. cQTLs were enriched in regions under selective pressure, in ENCODE monocyte-specific enhancers, in infectious disease SNPs (for monocyte-derived cytokine QTLs) and in autoimmune disease SNPs (for T cell-derived cQTLs) (Li et al., 2016b). Similar autoimmune disease- and complex trait- loci enrichments were identified using 27 SNPs associated with circulating levels of 41 different cytokines from an independent GWAS in a large healthy cohort of up to 8,293 Finnish individuals (Ahola-Olli et al., 2017). Continuing on the efforts to measure protein-level traits, 38 variants were associated with immunoglobulin levels (IgA, IgG, IgM), which are effector molecules of the adaptive immune system (Jonsson et al., 2017). Similarly, these variants also had known roles in autoimmune diseases and haematopoietic malignancies.

An exemplary study demonstrated how the combination of multiple pieces of genetic, molecular and functional evidence can resolve complex autoimmune disease risk loci, in this case, the *TNFSF13B* gene locus encoding the cytokine B cell activating factor (BAFF) (Steri et al., 2017). An indel variant was associated with multiple sclerosis and systemic lupus erythematosus in a Sardinian cohort, as well as with 18 different endophenotypes including B cell and monocyte counts (Steri et al., 2017). The variant produces an alternative polyadenylation site and a 3' UTR truncated transcript, which resulted in both a gene expression and protein translation effect, the latter due to the presence of fewer miRNA binding sites. Ultimately this culminated in an increased level of soluble BAFF. Elevated BAFF levels were observed prior to disease diagnosis in separate preclinical samples, which was clear evidence of the causal relationship between higher BAFF protein levels and autoimmune disease (Steri et al., 2017). This clearly shows the power of combining functional and molecular phenotypes with longitudinal and clinical datasets when evaluating causal relationships between functional and disease phenotypes.

There are more and more studies recognising the importance of multiple phenotypes in facilitating functional interpretation of GWAS loci and in providing basic biological insights. Very recently, the Hi-HOST Phenome Project have generated a catalog cellular GWAS associations using 79 phenotypes in response to live pathogens in 528 LCLs and identified 17 genome-wide significant loci (Wang et al., 2017a). The cellular phenotypes measured included readouts of endocytosis, endosomal trafficking, cell signalling, cell death, cytokine production as well as the molecular readouts of transcriptional regulation (Wang et al., 2017a). In addition, the Enhancing GTEx (eGTEx) project was recently announced, wherein a bid to describe the effect of variation from “molecule to individual”, other intermediate measurements such as protein expression and telomere length will be assayed in the wide range of tissue types from this project in addition to gene expression and molecular phenotypes (eGTEx Project, 2017).

In future, similar efforts will likely be extended to multiple primary cell types and greater sample sizes providing rich resources for functionally annotating genetic loci.

1.6 Recall-by-genotype studies

Recall-by-genotype (RbG) studies are genotyped-directed experimental phenotyping investigations representing downstream hypothesis-driven approaches to investigate functional mechanisms (Corbin et al., 2017). They have emerged as the primary choice for designing experiments to further investigate the function of observations first identified in large-scale genetic studies. They allow greater functional resolution with smaller sample sizes compared to hypothesis-free GWAS (Figure 1.7).

RbG test a small number of predicted causal variants (between 1 and 10) selected from the integration of GWAS-associated variants, functional studies and statistical methods such as fine-mapping. Similar to GWAS, RbG studies have the advantage of utilising genetic variants that have arisen from the random allocation of alleles at conception, which cannot, in turn, be influenced by the traits of interest (Section 1.7.1). A further advantage of RbG studies are that they are designed to query causal relationships in selected stratified groups based on previously observed biological associations. This increases the precision of functional insight in a cost-effective, efficient manner (Corbin et al., 2017). I demonstrate the implementation and utility of a RbG study in Chapter 4.

1.7 Haematopoiesis as a paradigm for genetics

Haematopoiesis is the production of all mature blood cell types including thrombocytes (platelets), erythrocytes (red blood cells), myeloid cells (monocytes, macrophages, neutrophils) and lymphocytes (B cells and T cells) (Figure 1.5). Self-renewing haematopoietic stem cells (HSC) in the bone marrow differentiate to lineage-committed progenitor cells, which further differentiate into mature cells (Orkin and Zon, 2008, Vasquez et al., 2016). Chromatin regulation is important in this differentiation process and mutations in factors mediating histone modification and chromatin architecture result in myeloid malignancies (Woods and Levine, 2015). Chromatin was recently shown to be highly dynamic during lineage specification with 17,035 enhancers established *de novo* mainly after commitment of the first lineage progenitor (Lara-Astiaso et al., 2014). TFs are key to the activity of these enhancers, full activation of which preceded lineage-specific gene expression programmes (Lara-Astiaso et al., 2014). Therefore, haematopoiesis represents a model system for the study of all stages of stem cell development as well as chromatin formation, transcription factory activity and the cell-type specificity of these processes.

Mature haematopoietic cells perform vital biological roles including oxygen transport (red blood cells), blood clotting (platelets) and immune responses (myeloid and lymphoid cells). Sustained haematopoiesis occurs under homeostatic conditions as well as during infection, (Orkin and Zon, Amulic et al., 2012). Dysregulated blood cell function is a known factor in the aetiology of a wide variety of diseases. Understanding the biological context of disease-dysregulated processes can highlight important haematopoietic pathways and novel genes in haematopoiesis and mature cell function. The role of these cells in disease and function is discussed in detail in Chapter 2 and 3 of this thesis.

Haematopoiesis and mature blood cells are both relatively experimentally tractable. Whole blood is easily accessible from a high number of individuals and from this specific cell populations can be isolated with high purity and relative technical ease. The evolutionary conservation of haematopoiesis also facilitates study in model organisms. As a result, haematopoiesis is one of the best-characterised mammalian cellular differentiation systems.

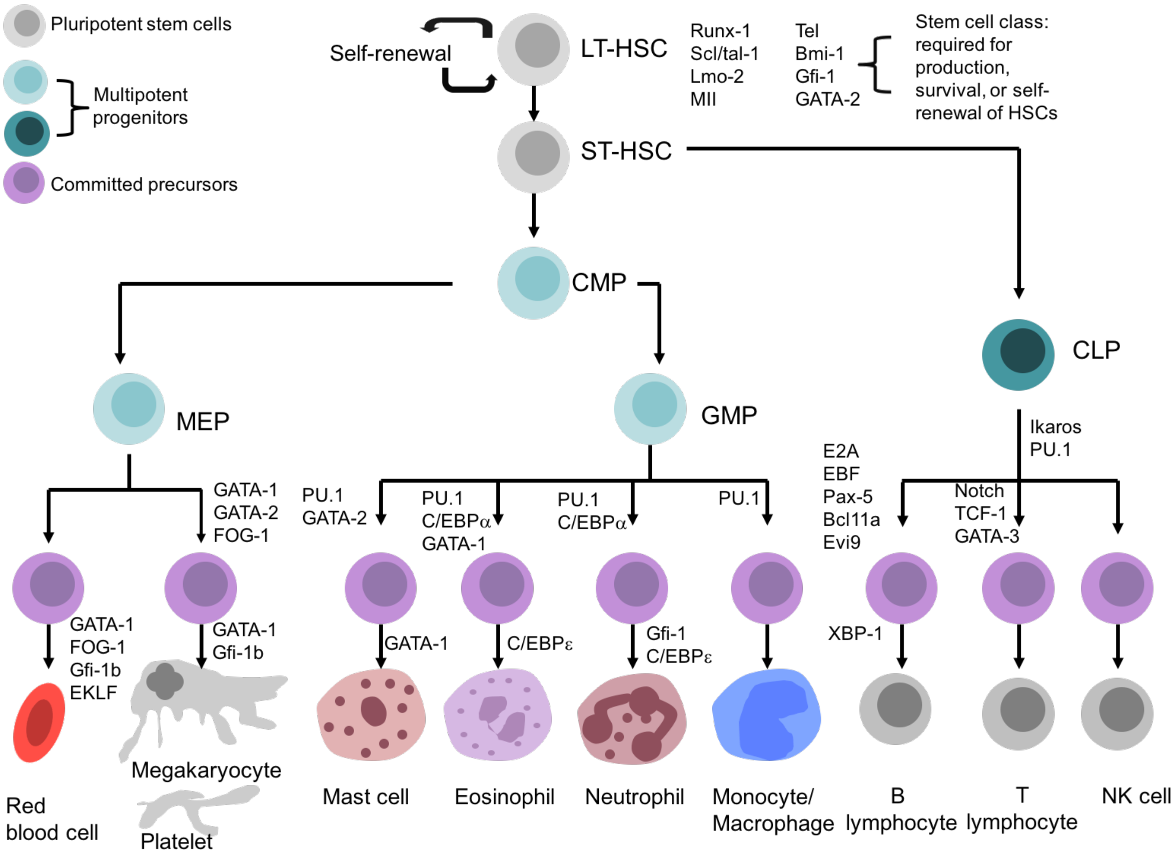


Figure 1.5: Haematopoiesis and the involvement of essential transcription factors
Differentiation of self-renewing haematopoietic stem cells to form all mature cells (red blood cell, platelet, mast cell, eosinophil, neutrophil, monocyte and macrophage, B and T lymphocytes, NK cells). The transcription factors required for each stage were discovered using conventional gene knockouts that resulted in a blockage of haematopoietic differentiation. LT-HSC: long-term haematopoietic stem cell, ST-HSC: short-term haematopoietic stem cell, CMP: common myeloid progenitor; CLP: common lymphoid progenitor, MEP: megakaryocyte/erythroid progenitor, GMP: granulocyte/macrophage progenitor. Additional TFs, not shown here, were predicted using a highly sensitive ChIP-seq protocol to be involved in 16 differentiation stages (Lara-Astiaso et al., 2014). Figure adapted from (Orkin and Zon, 2008).

Blood cell phenotypes such as full blood counts (FBC), are also readily measured by automated haematology analysers (Chami and Lettre, 2014, Astle et al., 2016). The deviation from normal size, physical characteristics or number of blood cells is diagnostic for human disease such as infection, anaemia, thrombotic diseases or haematological disorders (Table 1.2) (Vasquez et al., 2016, Soranzo et al., 2009). FBC is therefore routinely measured as part of clinical diagnosis and assessment of general health (Chami and Lettre, 2014). Table 1.2 summarises the full range of phenotypes that can be measured with recent analysers such as the Sysmex system (Astle et al., 2016, Vasquez et al., 2016, Sysmex Corporation).

Blood cell traits vary across healthy individuals and part of this variation is due to genetic factors (Pilia et al., 2006, Evans et al., 1999, Garner et al., 2000, Chami and Lettre, 2014). Therefore, studying naturally occurring genetic variation of circulating mature blood cell counts is a common and successful strategy used to gain insight into the regulation of haematopoiesis (Table 1.2). This approach has yielded many insights, not only in identifying novel haematopoietic regulators but also for the wider field of human genetics. For example, blood GWAS has been successful in identifying novel regulators of haematopoiesis (Gieger et al., 2011, van der Harst et al., 2012, Bielczyk-Maczynska et al., 2014). Previously unknown genes identified from GWAS of RBCs and platelets displayed haematopoietic phenotypes in model organisms (Vasquez et al., 2016).

Up until 2016, blood GWAS only explained a fraction of variation in the population (4-10%) and high-powered cohorts for studying myeloid and lymphoid parameters were lacking (Vasquez et al., 2016, Gieger et al., 2011, van der Harst et al., 2012). The recent large GWAS using data from the UK biobank cohort (N = 173,480) investigated a high number of traits, 36 in total (Table 1.2) (Astle et al., 2016). 2,706 independent variants were identified, representing a ten-fold increase in the number of known loci that included hundreds of rare variants with high effects sizes (Vasquez et al., 2016, Kim-Hellmuth and Lappalainen, 2016). Most of the sentinel variants were highly specific across red blood cell, white cell and platelet traits and enriched in corresponding cell-type specific enhancers. Coding variants were enriched with Mendelian disease mutations, a demonstration of how important clinical insight can be gleaned from large-scale GWAS. Plausible molecular mechanisms were identified through integration with the BLUEPRINT QTL data for 276 blood trait variants that colocalised with at least one molecular QTL (Astle et al., 2016). It was estimated that a higher proportion of variance in the blood indices was explained by the common autosomal genotypes from this study, for example between 5-21% of variance in white cell traits (Astle et al., 2016). The full UK Biobank cohort of 500,000 individuals could identify further significant variants explaining trait variance (Collins, 2012).

	Trait [Units]	Description	Determination	Example Diseases/disorders
RBC	Red blood cell count [per pL]	Count of RBCs per unit volume of blood	Impedance (measured)	Anaemia, polycythemia vera
HGB	Haemoglobin concentration [g/dl]	Concentration of Hb per unit volume of blood	Light absorbance (measured)	
HCT	Hematocrit [%]	Volume fraction of blood occupied by red cells	Impedance (measured)	
MCV	Mean corpuscular haemoglobin concentration [fL]	Mean volume of RBCs	(HCT/RBC)×10 (derived)	
RDW	Red cell distribution width [fL]	Coefficient of variation of red cell volume distribution	CV of impedance measured red cell volume distribution (measured)	
MCH	Mean corpuscular haemoglobin [pg]	Average mass of Hb per red cell	(HGB/RBC)×10 (derived)	
MCHC	Mean corpuscular haemoglobin concentration [g/dL]	Concentration of Hb per unit of volume occupied by red cells	(HGB/HCT)×100 (derived)	
PLT	Platelet count [per nL]	Count of platelets per unit volume of blood	Impedance (measured)	Essential thrombocythemia, thrombotic Thrombocytopenic purpura
MPV	Mean platelet volume [fL]	Mean volume of platelets	(PCT/PLT)×10000 (derived)	
PDW	Platelet distribution width [fL]	Spread of the platelet volume distribution (PDV)	Impedance: Coefficient of variation of PDV (measured)	
PCT	Plateletcrit [%]	Volume fraction of blood occupied by platelets	Impedance (measured)	
WBC	White blood cell count [per nL]	Aggregate count of white cells per unit volume of blood	Impedance (measured)	Autoimmune/immunological, infection, inflammation, leukaemia
NEU	Neutrophil count [per nL]	Count of neutrophils per unit volume of blood	(NEUT%×WBC)/100% (derived)	Myelodysplasia, bacterial infections
LYM	Lymphocyte count [per nL]	Aggregate count of lymphoid cells per unit volume of blood	(LYMPH%×WBC)/100% (derived)	Lymphoma, viral infections
MON	Monocyte count [per nL]	Count of monocytes per unit volume of blood	(MONO%×WBC)/100% (derived)	Myelomonocytic leukaemia, chronic infections (tuberculosis)
EOS	Eosinophil count [per nL]	Count of eosinophils per unit volume of blood	(EO%×WBC)/100% (derived)	Allergies, asthma, parasitic infections
BAS	Basophil count [per nL]	Count of basophils per unit volume of blood	(BASO%×WBC)/100% (derived)	Hyperthyroidism, myeloproliferation disorders

Table 1.2: Summary of the main haematological indices, measurement unit and related disorders
Adapted from (Vasquez et al., 2016, Astle et al., 2016). Additional traits were also tested in the Astle *et al.* (2016) GWAS, that included for example the percentage of granulocytes that is made up by neutrophils. I list the main traits measuring mature blood cell counts here that are routinely measured and have been explored in previous studies.

1.7.1 Genetics, correlation and causation

Correlation between blood indices and increased risk of certain diseases such as obesity, stroke and cardiovascular diseases has been observed (del Zoppo, 1998, Poitou et al., 2011, Ensrud and Grimm, 1992, Hoffman et al., 2004, Boos and Lip, 2007). However, correlation does not necessarily show causation as epidemiological and observational relationships can be subject to confounding factors, measurement error, bias or reverse causation (where the disease state influences the endophenotype such as blood indices).

Genetics, with the exception of somatic mutations, is pre-determined at birth where variants are segregated randomly and independently of other traits (Evans and Davey Smith, 2015). In this way, confounding and reverse causation are both reduced as genetics precedes any biological effect or outcome (Evans and Davey Smith, 2015). We can also measure genetic variants with high precision, reducing measurement error that can occur in observational studies. Approaches have therefore been developed that use genetic variants (instrumental variables) that are known to influence a biological intermediate (exposure), which itself affects disease risk. In this case, the studied variants should also be related to the risk of the disease. This approach can assess both the causality of biological intermediates and quantify the size of the causal effect and is referred to as Mendelian Randomization (Evans and Davey Smith, 2015). There are certain assumptions that must not be violated in these analyses, which in some cases can be challenging to definitively confirm. These are summarised in Figure 1.6.

This approach was implemented by Astle *et al.* (2016) to test for causal relationships between blood indices and each of a group of six autoimmune, three cardiometabolic and five neuropsychiatric diseases. Positive correlations were found between eosinophil count and rheumatoid arthritis and asthma, with a weaker effect between neutrophil indices and asthma. Interestingly, there was a reduced likelihood for causality between red blood cell, white blood cell, granulocyte and neutrophil counts and risk of coronary heart disease (CHD), despite previously reported correlations (Wheeler et al., 2004, Astle et al., 2016).

Overall, studying the process of haematopoiesis and mature cell function increases our understanding of basic biology. Concomitantly, it also offers the potential to use blood cell traits as disease biomarkers and tractable intermediate phenotypes in genetic studies and functional follow-ups.

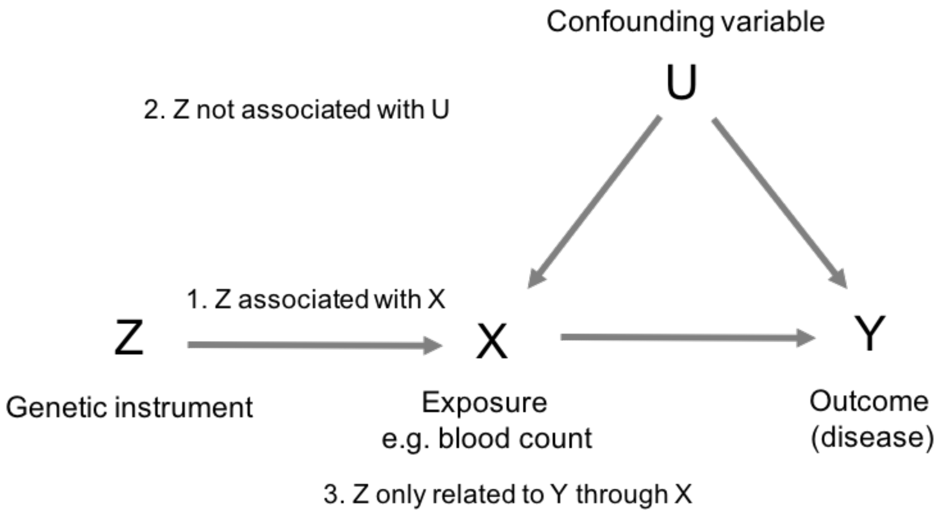


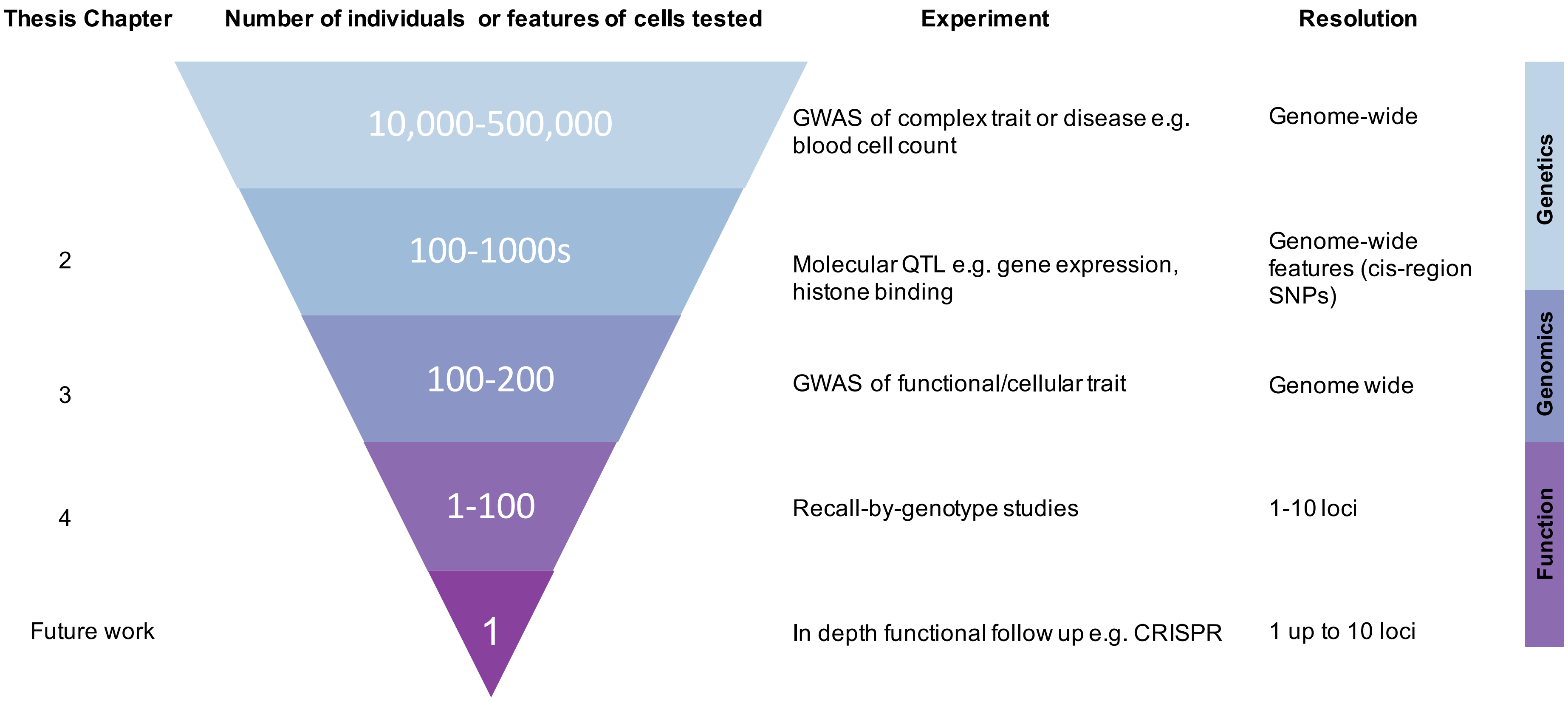
Figure 1.6 Mendelian randomization methodology and assumptions
Schematic summarising a causal relationship between an exposure and an outcome/disease assessed by using genetic variants (Z) that are associated with the exposure and under causality are also associated with disease. Causal relationships are depicted with arrows. The three assumptions are also given. Adapted from (Evans and Davey Smith, 2015).

1.8 Aims of this thesis

In this thesis, I use a combination of genetic and genomics approaches I have discussed to resolve functional consequences of genetic variation whilst also understanding the biology of haematopoietic cells. These are summarised in Figure 1.7.

In Chapter 2, I discuss how these approaches have increased our understanding of autoimmune diseases. I apply the lessons learnt from these studies to diseases that are not traditionally classified as immune-mediated. I use epigenomic phenotypes to resolve mechanisms of risk loci and also explore potential insight into pathways or genes that could provide future therapeutic avenues for these diseases. I demonstrate that the combination of genomic and genetic approaches provides hypothesis-free identification of genes and pathways dysregulated in disease, representing an early step in identifying new therapeutic avenues.

Following from this, I apply GWAS to novel neutrophil phenotypes with an overall aim of expanding the phenotype repertoire by providing additional functional datasets with which to annotate trait- or disease- associated loci. Finally, I implemented a recall-by-genotype study to perform an in-depth investigation into two genetic loci where there was previous evidence of an association with neutrophil count. Throughout my thesis, I demonstrate the application of varied but complementary approaches in gaining biological insight from genetic associations.



Thesis Chapter	Number of individuals or features of cells tested	Experiment	Resolution
	10,000-500,000	GWAS of complex trait or disease e.g. blood cell count	Genome-wide
2	100-1000s	Molecular QTL e.g. gene expression, histone binding	Genome-wide features (cis-region SNPs)
3	100-200	GWAS of functional/cellular trait	Genome wide
4	1-100	Recall-by-genotype studies	1-10 loci
Future work	1	In depth functional follow up e.g. CRISPR	1 up to 10 loci

Figure 1.7 Approaches to investigate functional mechanisms of genetic variants
Schematic summarises the type of experiments, number of individuals required, resolution of variants investigated and the chapters of this thesis where the techniques are used

Chapter 2

Using immune molecular phenotypes to uncover biological mechanisms of disease-associated genetic loci

Collaboration Note

The custom colocalisation gwas-pw pipeline was designed by Louella Vasquez, which I adapted for my analysis here. Kousik Kundu (Wellcome Sanger Institute) developed custom scripts for visualisation of epigenomic signals such as in Figure 2.10 and performed the analysis with the updated phase 2 of the BLUEPRINT cohort data. Tao Jiang (Department of Public Health and Primary Care) and Klaudia Walter (Wellcome Sanger Institute) advised on the gene-specific regression analyses and Valentina Iotchkova (Weatherall Institute of Molecular Medicine, formerly Wellcome Sanger Institute) on GARFIELD implementation and enrichment analyses. Stephen Watt (Wellcome Sanger Institute) provided the monocyte and neutrophil transcription factor datasets and advised on peak selection and investigation. All other analyses were performed by myself.

We thank all authors and consortia for access to GWAS summary statistics, further details of access can be found in the Methods section and Supplementary information. UKBiLEVE GWAS statistics were obtained through application to the UK Biobank. SLE GWAS was accessed from Immunobase, Coronary artery disease from the CARDIoGRAMplusC4D consortium and Type 2 diabetes summary statistics from the DIAGRAM consortium website. We also thank the International Genomics of Alzheimer's Project (IGAP) for providing summary results data for these analyses (see the Supplementary Information for full information of author contributions and Grants for this study). We also thank the IAMDGC consortium for sharing the summary statistics and effect sizes of advanced age-related macular degeneration GWAS from Fritsche *et al.* (2016). Applications and queries can be made through the following contacts: goncalo@umich.edu, larsf@umich.edu. Further information on AMD statistics: http://amdgenetics.org

2 Using immune molecular phenotypes to uncover biological mechanisms of disease-associated genetic loci

2.1 Introduction

2.1.1 Lessons from genetic and genomic analyses of autoimmune diseases

The study of autoimmune diseases (AID) has generated many important biological insights including demonstrating the central role for the function of multiple immune cell types (Farh et al., 2015, Glinos et al., 2017). Overall, there is a 4.5% prevalence of the 81 identified AID in the general population, which is higher for women (6.4%) than men (2.7%) (Hayter and Cook, 2012). The importance of genetic factors and shared environment is demonstrated by the familial clustering of autoimmune diseases (Gutierrez-Arcelus et al., 2016). Initial linkage studies identified some of these genomic risk regions that had large effect sizes by looking for markers that co-segregated with the disease phenotype. These included the MHC, encoding the major histocompatibility complex, with diseases such as Type 1 diabetes (T1D) (Rich et al., 1984) and systemic lupus erythematosus (SLE) (Gaffney et al., 1998) and the nucleotide-binding oligomerisation domain containing 2 (NOD2) gene with Crohn's disease (Hugot et al., 2001, Gutierrez-Arcelus et al., 2016, de Lange and Barrett, 2015). Strong associations in the MHC region, which contains many immune-related genes, are now well established for a wide range of diseases, such as coeliac disease (CEL), rheumatoid arthritis (RA) and multiple sclerosis (MS) (Sollid et al., 1989, Nepom, 1998, Hollenbach and Oksenberg, 2015). These associations implicate a role for MHC-antigen presentation in triggering the immune response as a general phenomenon in autoimmune disease pathogenesis. Other important genes were also identified through candidate gene studies, which test for association with alleles of genes selected *a priori* (Gutierrez-Arcelus et al., 2016). One example is the *CTLA4* locus, which was associated with T1D and later with autoantibody-positive RA (Nistico et al., 1996, Plenge et al., 2005). *CTLA4* encodes an immunoglobulin superfamily protein expressed on the surface of T helper cells that negatively regulates T cell activation (Nistico et al., 1996, Gutierrez-Arcelus et al., 2016).

The advent of GWAS enabled systematic and unbiased genome-wide searches leading to the identification of hundreds of AID risk loci, many of which are shared between different immune disorders (Gutierrez-Arcelus et al., 2016). The majority of these signals are common (MAF > 5%) with small to moderate effect sizes (OR < 1.6) (Gutierrez-Arcelus et al., 2016). MS risk loci, excluding the MHC region, have odds ratios between 1.1 and 1.6, where an OR of 1 signifies no difference in odds of diseases between cases and controls for that allele (Gutierrez-Arcelus et al., 2016, ImmunoBase, 2017). Such observations supported the common disease-common variant (CDCV) hypothesis, first proposed by Risch and Merikangas in 1996, which suggests that complex disease risk is a result of the accumulation of multiple, low-effect risk factors (Risch and Merikangas, 1996). However, larger sample sizes and higher-powered studies are required to detect rare variants, therefore future efforts may discover that variants with a MAF < 1% also play a role in common diseases (Bomba et al., 2017).

Despite the large number of AID variants now discovered (more than 300 loci (Gutierrez-Arcelus et al., 2016)), a limited degree of the estimated heritability is explained by non-HLA loci (Glinos et al., 2017). Heritability is the proportion of observable phenotypic variation that can be attributed to genetics, which can be estimated from twin or sibling studies (Selmi et al., 2012). A recent GWAS of systemic lupus erythematosus (SLE) including 15,991 controls and 7,219 cases, estimated the heritability explained by 43 identified risk alleles to be 15.3%, with a total estimated heritability of 66% (Bentham et al., 2015). This 'missing heritability' may be due to limitations in study power precluding detection of the full effect size and frequency spectrum of variants, particularly rare variants (Vasquez et al., 2016). The combination of multiple studies to increase sample sizes and application of improved imputation methods have been utilised by, for example, the International IBD Genetics Consortium (IIBDGC). These efforts enabled increased loci discovery allowing identification of novel pathways implicated in IBD risk such as cytokine signalling, innate defence and lymphocyte activation (de Lange and Barrett, 2015, Jostins et al., 2012).

Investigation of the biological consequences of known variants has provided important paradigms in the functional interpretation of GWAS SNPs. Haematopoietic cell types have long been known to play key roles in immune responses to infection, homeostatic clearance of cell debris and in regulating the balance between reacting to non-self-antigen; not self-antigen (Vasquez et al., 2016). Genomic data from haematopoietic cells are therefore ideally suited for mechanistic interpretation at AID risk loci. Early studies indicated that AID SNPs affect gene expression in whole blood and PBMCs, for example over half of coeliac GWAS variants were also eQTLs (Dubois et al., 2010, Glinos et al., 2017). These observations have been expanded to a wide-range of AIDs and multiple primary immune cell types and additional regulatory elements such as H3K27ac and TF binding (Farh et al., 2015, Tehranchi et al., 2016, Chen et al., 2016a). RNA splicing can also represent a gene expression-independent regulatory mechanism in genetic disease (Li et al., 2016c, Chen et al., 2016a). Li *et al.* (2016) showed that splicing (s)QTLs independent of eQTLs were enriched in gene bodies, in most cases within the target introns. In addition, the sQTLs were also enriched in AID even when compared to eQTLs (Li et al., 2016c).

Clinical insight can be gleaned from combining GWAS SNPs with immune molecular or functional phenotypes (Barrett et al., 2015). Selection of drug targets based genetic evidence provides promising therapeutic possibilities twice as often as those selected without such prior information (Nelson et al., 2015, Barrett et al., 2015). GWAS of genetic variants pre-determined at birth simulates a randomised clinical trial, where randomisation ensures a balance of all confounders (Evans and Davey Smith, 2015). The advantage of GWAS is that drug administration is not required and individuals have been "exposed" across a lifetime rather than for the length of an RCT (Evans and Davey Smith, 2015, Finan et al., 2017). Integration of LPS-stimulated monocyte eQTL data and GWAS SNPs can aid therapeutic insight demonstrated recently where five IBD risk variants were to found increase gene expression of the integrin genes *ITGA*, *ITGAL*, *ICAM* and *ITGB8* (de Lange et al., 2017). Integrins mediate leucocyte adhesion to inflamed endothelial tissues. Therefore, increased surface levels could contribute to the pro-inflammatory environment observed in IBD patients (de Lange et al., 2017, de Lange and Barrett, 2015). Targeting integrins, for example using monoclonal antibodies, has already shown promising therapeutic results in the context of IBD (de Lange et al., 2017). Clearly, discovering novel common associations and their associated mechanisms can still provide additional clinical insight.

AID GWAS alone have also successfully identified genes and pathways that are current drug targets. For example, the *IL6R* pathway, which contains rheumatoid arthritis risk variants, is targeted by the humanised monoclonal antibody therapy, Tocilizumab (Okada et al., 2014, Law et al., 2014). The same RA GWAS also identified novel risk genes not currently targeted by RA therapies but were used for treating other diseases, offering the potential for the repurposing of licensed drugs (Okada et al., 2014). To further capitalise on the therapeutic potential of GWAS, a new genotyping array was designed to include genes encoding druggable proteins, targets with bioactivity and those with clinical indications of any licensed therapeutics (Finan et al., 2017). GWAS with such an array will enable direct association of variants with druggable genes (Finan et al., 2017).

2.1.2 Expanding the complex disease repertoire for which immune phenotypes can resolve mechanisms

Inflammation has also been shown to be important in disorders not traditionally classified as immune-mediated such as Parkinson's disease (Tufekci et al., 2012) and schizophrenia (Muller et al., 2015), suggesting functional and clinical insight could be gleaned from the application of similar approaches described above. Below, I discuss previous evidence for the role peripheral immune function in the pathogenesis of five diseases, which I focus on in this thesis.

2.1.2.1 Advanced age-related macular degeneration (AMD)

AMD is the leading cause of irreversible blindness later in life in the developed world. AMD affects the central part of the macular and is classified into early, intermediate or advanced based on severity (Pennington and DeAngelis, 2016). The hallmark of AMD is the accumulation of lipid-rich, protein-containing drusen deposits between the retinal pigment epithelium (RPE) and Bruch's membrane (BM) in the retina (Figure 2.1). The RPE forms part of the blood-ocular barrier and performs many important functions including nutrient transport, cytokine release and phagocytosis of fragments released from photoreceptors (Tan et al., 2016). Towards the end stages of the disease, the RPE eventually disintegrates leading to loss of photoreceptors and vision (Figure 2.1). Ordinarily, the RPE and sub-retinal regions are devoid of blood vessels, but in the neovascular ("wet") form of the disease, abnormal growth of blood vessels from the choroid spreads into these regions (Pennington and DeAngelis, 2016) (Figure 2.1). Most current therapies target this growth by inhibiting the angiogenesis-promoting vascular endothelial growth factor A (VEGFA) (Pennington and DeAngelis, 2016). However, disease progression continues for most patients, requiring further treatment.

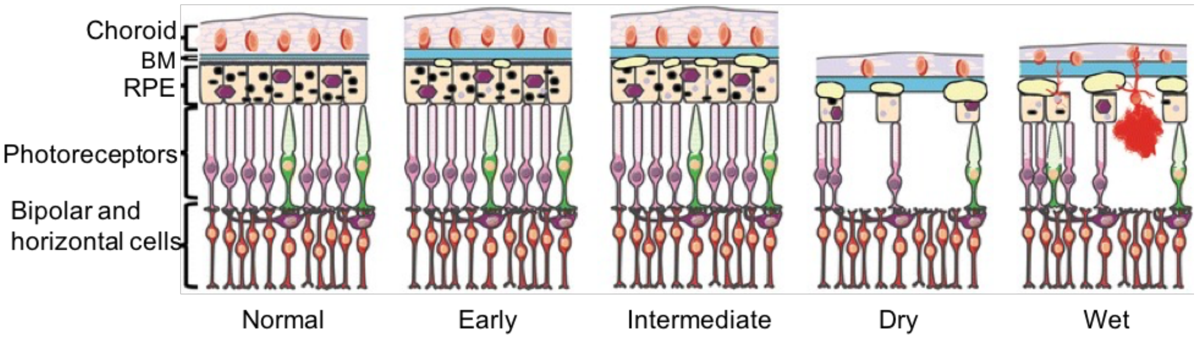


Figure 2.1: Schematic of retinal structure and the effect of AMD pathology
This schematic shows the outer layers of the central retina in normal conditions (left) and the different types of AMD classifications. Small drusen deposits accumulate in the retinal pigment epithelium (RPE) in early AMD, and as the disease progresses, the Bruch's membrane (BM) becomes thicker and additional drusen deposits form. In the later stages of AMD, dry and wet, there is extensive accumulation of drusen deposits, loss of photoreceptors and damage to RPE integrity. The subretinal space refers to the space between the RPE and photoreceptors. In the wet form, choroidal neovascularisation (CNV) occurs. This figure was adapted from (Tan et al., 2016) under the CC license (http://creativecommons.org/licenses/by/4.0/).

AMD is multifactorial with a substantial genetic component, although environmental factors such as smoking and age also contribute (Seddon et al., 2005). Genetic studies have revealed many associated loci and have provided evidence for the involvement of immune components. For example, almost 60% of AMD risk can be explained by variants located near the complement genes *CFH*, *C2/CFB*, *C3*, *CFI* and *C9* (Tan et al., 2016). The risk to AMD also increases with rare alleles. For example, the highly penetrant missense mutation in the complement factor I gene (*CFI*), which functions to inactivate complement pathways, corresponds to an odds ratio of 22.20 (95% CI = 2.98-164.49) (van de Ven et al., 2013, Tan et al., 2016). The p.Gly119Arg mutant protein is expressed and secreted at lower levels (van de Ven et al., 2013).

A recent large-scale GWAS from the International AMD Genomics Consortium (IAMDGC) that included 16,144 advanced AMD patients and 17,832 controls identified 52 independently associated variants in 34 loci (Fritsche et al., 2016). This study confirmed the involvement of inflammation and complement genes (*VTN, CFH, C2/CFH, C3, CFI, C9*) as well as lipid pathway genes (*CETP, LIPC, APOE, ABCA1)* (Fritsche et al., 2016). The complement pathway is an important part of innate immunity that functions to amplify immune responses ultimately resulting in the formation of a membrane attack complex (MAC) at the surface of the pathogen causing cell lysis (Tan et al., 2016). Increased MAC in the retina and drusen has been observed in AMD patients (Tan et al., 2016, Hageman et al., 2001). Although the majority of the complement system is synthesised by the liver, there is also synthesis in the RPE and choroid (Tan et al., 2016, Luo et al., 2013).

Despite initially being considered “immunologically privileged”, the RPE has been shown to contain specialised resident immunocompetent cells including microglia, dendritic cells and perivascular macrophages (Parmeggiani et al., 2012). Para-inflammatory (prolonged inflammation in response to damage)-associated modifications can result in damage to the blood-retinal barrier in AMD as well as microglial activation and recruitment of macrophages (Parmeggiani et al., 2012, Kauppinen et al., 2016). Indeed, immunocompetent cells such as lymphocytes and macrophages have been observed in AMD retinal tissues and isolated mouse bone marrow-derived M1 and M2b macrophages stimulated RPE cells to induce inflammatory cytokine expression and complement factors C3 and CFB (Parmeggiani et al., 2012, Lopez et al., 1991, Luo et al., 2013).

Systemic immune alterations such as increased serum complement components and pro-inflammatory cytokines (IL-1α, IL-1β and IL-17) have been observed in AMD patients (Lechner et al., 2015). Higher numbers of circulating neutrophils were observed in neovascular AMD (nvAMD) patients (Lechner et al., 2015). Also, similarly in nvAMD patients, there was an increased inflammatory transcriptome signature in peripheral blood monocytes, and in an independent study monocytes expressed higher levels of chemokine receptors CCR1, CCR2 and CX3CR1, HLA-DR and phosphorylated STAT3 (Grunin et al., 2016, Grunin et al., 2012, Chen et al., 2016b). Recently, in a study of 161 nvAMD patients and 43 controls, stimulated PBMCs, particularly monocytes, secreted higher levels of the IL8, CCL2 and VEGF compared to controls (Lechner et al., 2017). The pro-inflammatory IL8 and CCL2 promote the recruitment of neutrophils and monocytes and lymphocytes respectively (Lechner et al., 2017). Additional molecular mechanistic insight into monocyte-RPE interactions was shown using a coculture of human CD14^{+} blood monocytes and primary porcine RPE cells (Mathis et al., 2017). OTX2 is a key TF regulating retinal genes such as retinol dehydrogenase 5 (RDH5), which re-isomerises all-trans-retinal into 11-cis-retinal. It was shown that TNFα, secreted from activated monocytes, mediates the downregulation of *OTX2* and *RDH5* (Mathis et al., 2017).

In summary, immune dysregulation in AMD could be the result of the combined action of resident and infiltrating immune cells as a result of a switch from clearance (of RPE-debris) and immunosuppressive environment to a proinflammatory milieu (Nussenblatt and Ferris, 2007). Genetic variants could disrupt this balance, thereby influencing risk. Definitive assessment of the causality of peripheral immune factors and disease pathogenesis is required, but the observations of systemic immune activation in AMD patients suggests that using the more accessible peripheral immune cells to identify functional mechanisms may provide further insight into the pathogenic process. Some success, for example, in an early pilot phase I/II randomized study of suppression of systemic immune activity, was observed for wet AMD (Nussenblatt et al., 2010). However, other attempts, including the use of eculizumab to inhibit complement have been less successful (Yehoshua et al., 2014). Elucidating the exact mechanisms of peripheral immune involvement may help these efforts.

2.1.2.2 Coronary artery disease (CAD)

Coronary artery disease (CAD) is the most common type of heart disease (Khera and Kathiresan, 2017). Familial history and therefore, a genetic component, has been implicated as an important risk factor (Framingham Heart Study (Watkins and Farrall, 2006), PROCAM study (Assmann et al., 2002), INTERHEART study (Yusuf et al., 2004)). Common CAD is multifactorial and polygenic (Won et al., 2015). Low-density lipoprotein (LDL) cholesterol levels, lipoprotein(a) and BMI have also been demonstrated to be causal risk factors and are themselves under genetic control (Watkins and Farrall, 2006, Do et al., 2013, Clarke et al., 2009, Voight et al., 2012, Khera and Kathiresan, 2017). The INTERHEART study showed that adjustment for the known classical risk factors only marginally reduced the risk (OR from 1.55 to 1.45), suggesting there are other genetic factors involved (Yusuf et al., 2004, Watkins and Farrall, 2006).

There is a well-established causal association of the inflammatory IL6 cytokine pathway with CAD risk (Interleukin-6 Receptor Mendelian Randomisation Analysis Consortium et al., 2012, Il R. Genetics Consortium Emerging Risk Factors Collaboration et al., 2012). The missense SNP, rs2228145, increases the soluble form of the IL6 receptor by improving the efficiency of membrane-bound receptor proteolytic cleavage while also increasing the unique transcript encoding the soluble receptor (Ferreira et al., 2013). Reduced IL6R membrane expression on monocytes and CD4$^+$ T cells results in impaired signalling and IL6 response and reduces CAD risk (also that of RA), clearly demonstrating the pathogenic role of inflammation in disease progression (Ferreira et al., 2013).

Monocytes play a key role in pathological plaque deposition in the coronary arteries, known as atherosclerosis (Ghattas et al., 2013). After recruitment to atherosclerotic lesions, monocytes mature into macrophages (Meeuwsen et al., 2017). Phagocytosis of oxidised LDL stimulates macrophage to form foam cells, which are a constituent of atherosclerotic plaques (Meeuwsen et al., 2017). Recruitment of other immune cells such as neutrophils, mast cells and lymphocytes also contributes to plaque destabilisation eventually leading to plaque rupture (Meeuwsen et al., 2017).

Supporting the importance of leukocyte function, the recent UK Biobank CAD GWAS with 4,831 cases and 115,455 controls identified 15 novel associations including the *ARHGEF26* locus, which is involved in transendothelial migration of leukocytes and encodes the rho guanine nucleotide exchange factor 26 (Klarin et al., 2017). The novel *ARHGEF26* locus was not associated with established risk factors, but previous mouse work did demonstrate a role in atherosclerosis (Klarin et al., 2017). Endogenous siRNA-mediated *ARHGEF26* knock-down decreased leukocyte adhesion to endothelial cells and transendothelial migration (Klarin et al., 2017). Overexpression of the exogenous mutant (Leu29) not only rescued the phenotype but was increased compared to wild-type, which is consistent with a gain of function ARHGEF26 effect associated with rs12493885 (Val29Leu) and increased CAD risk (Klarin et al., 2017). These observations are evidence that dysregulation of leukocyte function is associated with risk of CAD and disease prognosis.

Understanding the contribution of immune processes to CAD risk has important clinical implications. Currently, clinical management of CAD-associated events has improved leading to more than a 50% decrease in age-adjusted mortality rate in the United States (Khera and Kathiresan, 2017). Despite this, CAD is still the biggest cause of death worldwide, and there is a 12% mortality rate within six months of the first coronary event (Meeuwsen et al., 2017). Many available therapies target lipid and thrombosis reduction, but some inflammation-modulating processes are being investigated (Fernandez-Ruiz, 2016). Promising results were reported from a recent clinical trial targeting inflammation called the Canakinumab Antiinflammatory Thrombosis Outcome Study (CANTOS) (Ridker et al., 2017, Harrington, 2017, Couzin-Frankel, 2017). The trial included more than 10,000 heart attack patients with elevated levels of C-reactive protein (CRP). The drug tested, canakinumab, is a human monoclonal antibody that targets the inflammatory cytokine interleukin-1β and is already used in the treatment of systemic juvenile idiopathic arthritis (Harrington, 2017). For patients receiving canakinumab as four infusions each year over 3.5 years, the risk of a second cardiovascular event decreased from 4.5% to 3.86% and the likelihood of angioplasty or cardiac bypass surgery decreased by 30% (Couzin-Frankel, 2017). However, an increased number of deaths from infection among patients who received more doses was observed (Harrington, 2017). While promising and offering a proof of concept for inflammation playing a key role in disease pathogenesis, further research is needed to provide more targeted immune therapeutics which may reduce the risk of mortality related to infection (Harrington, 2017, Couzin-Frankel, 2017). Providing detailed descriptions of the pathways and cell types which may be involved in CAD risk could, therefore, represent an early step in improving therapeutic options for CAD patients.

2.1.2.3 Alzheimer's disease (AD)

Alzheimer's disease is the most common form of neurodegenerative dementia characterised by accumulation in the brain of amyloid β (Aβ) plaques and hyper-phosphorylated tau protein aggregates (Pimenova et al., 2017). This chapter focuses on late-onset Alzheimer's disease (LOAD), which constitutes 99% of AD cases. LOAD is multifactorial with a strong but highly complex genetic component (Pimenova et al., 2017, Gatz et al., 2006).

Involvement of immune components was highlighted in a large-scale meta-analysis of 74,046 individuals (Lambert et al., 2013). For example, a significant intronic SNP was identified in complement factor 1 (CR1). CR1 is expressed on blood cells and specialised brain-resident immune cells known as microglia (Pimenova et al., 2017, Wyss-Coray and Rogers). Other immune loci identified include *ZCWPW1/PILRA/PILRB,* which are monocyte and neutrophil immune infiltration receptors, the *HLA-DRB1/HLA-DRB5* locus of the MHCII region, *CD33* and the *MS4A* gene family, which are both expressed on microglia and myeloid cells (Pimenova et al., 2017, Lambert et al., 2013).

Brain-resident microglia are active immune cells, but there may also be a role for systemic immune responses in disease pathogenesis (Heneka et al., 2015, Czirr and Wyss-Coray, 2012). There are multiple lines of evidence that show that systemic inflammation detrimentally affects brain function and contributes to AD progression (Heneka et al., 2015). Prolonged LPS challenge in amyloid precursor protein (APP) transgenic mice, which contain a mutant APP associated with familiar forms of human AD, has been shown to result in cognitive impairment through increased amyloid deposition (Lee et al., 2008, Czirr and Wyss-Coray, 2012). Increased cognitive decline was observed in AD patients with infection, which correlated with infection TNF levels (Heneka et al., 2015). However, the functionality of peripheral immune cells in AD pathogenesis and their interaction with the brain are not yet clear. There is some evidence that the integrity of the blood-brain barrier (BBB) can be compromised in AD, which could allow infiltration of peripheral monocytes (Zenaro et al., 2017, Heneka et al., 2015). Compelling evidence from both mouse models and from human brain tissue showed that neutrophils could migrate to the central nervous system, which was dependent on the LFA-1 integrin, in turn, triggered by the $A\beta_{42}$ peptide (Zenaro et al., 2015). Depletion of neutrophils resulted in memory improvements in mice. This study suggests that neutrophils could contribute to inflammatory conditions in AD and also damage the BBB (Zenaro et al., 2015).

Integration with genomics data from the Immune Variation project with AD risk loci demonstrated enrichment among monocyte eQTLs, but not T cells, suggesting using monocyte data may help dissect mechanism underlying genetic susceptibility (Raj et al., 2014). Further demonstration of the importance of myeloid cells came from a recent association analysis of age of onset of Alzheimer's disease-defined survival that identified protective rs1057233 (G) associated with reduced *SPI1* expression in human myeloid cells (Huang et al., 2017c). *SPI1* encodes the myeloid master transcription factor, PU.1, which is critical for myeloid lineage differentiation and function (Huang et al., 2017c). PU.1 levels correlated with mouse microglial phagocytic activity and reduced *SPI1* expression corresponded to reduced AD risk (Huang et al., 2017c). Global enrichment of AD heritability in myeloid and B-lymphoid H3K4me1-designated enhancers and in the PU.1 cistrome (i.e. genome-wide binding regions as assayed by ChIP-seq) was also demonstrated (Huang et al., 2017c). Circulating blood monocytes are easily accessed in high numbers compared to brain-derived tissue and share transcriptional patterns with brain-resident microglia (Raj et al., 2014). Therefore, although there is evidence for a role of peripheral immune cells in AD, these cells are also useful as highly related proxies for specialised brain cells (Raj et al., 2014, Proitsi et al., 2014).

Taken together these observations provide evidence for the role of immune cells in AD pathogenesis, suggesting that integrative genomic approaches using immune data may provide insight into pathogenic mechanisms.

2.1.2.4 Lung function and chronic pulmonary obstructive disease (COPD)

Chronic obstructive pulmonary disease (COPD) is an inflammatory airway disease and is the third leading cause of global mortality (Wain et al., 2015). The ratio of two lung measurements, FEV_1/FVC, is used to evaluate airflow obstruction and diagnose COPD. FEV_1 measures how much air is exhaled in one second after maximal inhalation. FVC measures the volume of air exhaled after a maximal inhalation and then a six-second maximal forced exhalation (Weiss, 2010).

Smoking and indoor pollution are major risk factors for disease pathogenesis, but there is also a strong genetic component underlying smoking behaviour and disease risk (Wain et al., 2015, Hukkinen et al., 2011). Evaluation of these lung ratios as quantitative traits has allowed the identification of genetic determinants of lung function in large populations, enabling greater power to detect associations. Ten loci were found to be associated with extremes of the lung ratio FEV_1 in never smokers, and one was located in the MHC region, *HLA-DQB1/HLA-DQA2* (Wain et al., 2015). This important immune locus potentially implicates an immunological role in extremes of lung function and as this locus was also associated with COPD also highlights immune factors in this disorder.

A hallmark of COPD is chronic inflammation but whether this and the role of immune cells are causal independent risk factors for COPD remains to be definitively demonstrated (Brusselle et al., 2011). Cigarette smoke activates innate immune cells, which then stimulate an adaptive immune response (Brusselle et al., 2011). Viral and bacterial infections exacerbate COPD and potentiate the inflammatory environment in the lung (Brusselle et al., 2011). Dysregulation of peripheral immune cells could contribute to the pathogenesis of the disease, either due to the extrapulmonary effects or through infiltration into the lung (Brusselle et al., 2011). Neutrophil infiltration has been linked to tissue destruction and disease progression in COPD patients (Huang et al., 2017a). Understanding potential immune-mediated molecular disruptions associated with the disease can help the design of treatments with the aim to reduce serious symptoms and complications (Brusselle et al., 2011).

2.1.3 Identification of disease-relevant cell types using enrichment approaches

Disease-relevant cell types are not always known and have traditionally been identified through immunology studies, patient observations and mouse models (Glinos et al., 2017). For example, inflammatory cells were identified in the brain lesions of patients and autoreactive T cells responding to myelin antigens were identified in the mouse model for multiple sclerosis (MS), experimental autoimmune encephalomyelitis (EAE) (Fletcher et al., 2010).

With the advent of high-throughput genetic association studies and increased availability of cell-type specific epigenome data from consortia such as the Encode Project Consortium (2012), Roadmap Epigenomics Consortium et al. (2015) and IHEC (Stunnenberg et al., 2016), novel statistical enrichment approaches were developed to assess genomic evidence for the relevance of cell types in complex disease and annotate the putative function of non-coding GWAS variants. These approaches evaluate the statistical significance of a quantified overlap between GWAS SNPs and regulatory annotations. Annotations can take of the form of ChIP-seq binding/signal regions, open chromatin regions or chromatin regulatory states all denoted with the genomic start and end positions of the mapped genomic region (referred to as a peak). An assessment of the significance of enrichment is required given that a large degree of the genome can be bound by these regulatory features, leading to spurious functional assignment occurring by chance if only a simple overlap is applied (Iotchkova et al., 2016).

Earlier functional enrichment approaches, for example as implemented by Maurano *et al.* (2012) demonstrated an enrichment (40%) of significant GWAS variants from 207 diseases and 447 quantitative traits in open chromatin (DHS sites) assayed in 349 tissues, compared to frequency and genomic-location matched SNPs from 1000 Genomes Project (Maurano et al., 2012). This high enrichment highlighted that underlying regulatory function of non-coding SNPs, demarcated by open chromatin, could underpin many complex trait associations. In addition, the authors demonstrated the ability of functional enrichment studies in the *de novo* identification of relevant cell types by comparing the enrichment of progressively more significant trait-associated variants in cell-type specific DHS sites to the proportion of all SNPs from the summary statistics that also overlapped the same DHSs. Higher enrichment for Crohn's disease variants was observed in Th17, and Th1 T cell open chromatin than other immune cell types and multiple sclerosis variants were enriched in B and T cell open chromatin (Maurano et al., 2012). The basis of predicting disease relevant cell types lies in the known cell specificity of regulatory element activity. Therefore, preferential enrichment in regulatory data from certain cell types suggests these variants are more likely functional in those cell types. This approach enables efficient identification of relevant cell types without complex patient, animal or *in vitro* studies by leveraging known disease risk loci and experimentally derived epigenomic data.

However, the above method did not account for linkage disequilibrium (LD) between significant variants (Iotchkova et al., 2016). Chance overlaps are also more likely to occur in regions of highly correlated variants, where high LD between variants can lead to overlaps that are not truly functional (Alasoo et al., 2017). This can inflate enrichment values and generate false positives. Later methods accounting for variant correlation made other interesting observations, that enrichment of disease SNPs in cell specific H3K4me3 regions associated with active transcription was not driven by gene proximity, and could, therefore, highlight potentially causal variants (Trynka et al., 2013). Enrichment of trait-associated SNPs in cell type-specific DHSs was confirmed by an independent method, fgwas, and also observed depletion in repressed chromatin (Pickrell, 2014). Predicted causal SNPs of a variety of autoimmune diseases were shown to be enriched in cell type specific H3K27ac enhancer regions and genetic heritability is also highly enriched in regulatory regions (Farh et al., 2015, Finucane et al., 2015). Collectively, these enrichments suggest that a high proportion of disease risk is mediated by regulatory changes that could ultimately generate variation in gene expression (Chun et al., 2017).

Evidence of enrichment below the genome-wide significance threshold (p values $\leq 5 \times 10^{-08}$) has been observed (Maurano et al., 2012), suggesting that appropriately evaluating enrichment at genome-wide suggestive thresholds (commonly p value $\leq 1 \times 10^{-05}$) could provide biologically relevant observations particularly when limited power precludes discovery of all true associations (Maurano et al., 2012). To address the potential confounding issues and to provide a robust assessment of enrichment at all GWAS significance thresholds, a novel method was developed within the Soranzo team by Valentina Iotchkova, known as GARFIELD (GWAS Analysis of Regulatory or Functional Information Enrichment with LD correction) (Iotchkova et al., 2016) (described in detail in

Materials and Methods). GARFIELD annotates independent variants, accounting for LD as well as for genomic location and minor allele frequency. The odds ratio is calculated for multiple GWAS thresholds, allowing for assessment of informative enrichment at suggestive thresholds, which can highlight novel findings given limited study power. The method confirmed enrichment of Crohn's disease variants in blood DHS (Iotchkova et al., 2016) and demonstrated highly significant enrichment of blood cell indices in corresponding cell-type specific enhancers (Astle et al., 2016).

Recently, the application of the GARFIELD method has been extended to assess the enrichment of GWAS SNPs in molecular QTLs where QTLs were used as regulatory annotations. This approach demonstrated enrichment of neutrophil, monocyte and T cell molecular features in autoimmune diseases including IBD, RA, T1D and MS (Chen et al., 2016a). An independent study of naïve and stimulated iPSC-derived macrophage eQTL and chromatin accessibility QTLs also used GARFIELD to reveal enrichment in autoimmune diseases, Alzheimer's disease and schizophrenia (Alasoo et al., 2017).

2.1.4 Colocalisation methods evaluate shared genetics across different traits

Enrichment methods do not assess whether variation in regulatory function and disease risk or trait variance can be attributed to a single shared variant or whether these are driven by independent effects in the same locus, known as pleiotropy (Chun et al., 2017). Therefore, once enrichment approaches have highlighted relevant cell types, robust approaches are required to identify loci where there is evidence for shared genetic control of multiple traits. This can be achieved by using Bayesian colocalisation methods (described in detail in Materials and Methods) (Pickrell et al., 2016, Guo et al., 2015).

Robust evaluations are particularly important given that eQTLs are widely present across the genome, and the majority of expressed genes are likely to be associated with at least one cis-eQTL (Pai et al., 2015). Guo *et al.* (2015) developed and implemented a Bayesian colocalisation method (*coloc*) using ten immune disease-associated variants (595 in total from 154 regions) and eQTLs from resting and stimulated monocytes and naïve B cells (Guo et al., 2015). They identified 125 eQTLs that overlapped with disease SNPs, but only six genes had evidence for colocalisation (two traits share the same causal variant). A higher proportion, 21% across all AID loci tested colocalised with LCLs, CD4$^+$ T cells and CD14$^+$ monocytes eQTLs (FDR < 5%), identified using an independent method (Chun et al., 2017). For some diseases, this proportion was higher, for example 60% for ulcerative colitis loci colocalised with at least one eQTL (Chun et al., 2017). Overall, across all diseases, 75% of the disease-eQTL pairs were identified as pleiotropic where independent genetic variants within the locus were associated with each trait (Chun et al., 2017). Therefore, the previous usage of colocalisation methods has demonstrated the need for appropriate evaluation of shared trait and QTL loci, demonstrating that simple overlap approaches are prone to false positives.

2.1.5 Aims of this chapter

We previously demonstrated the enrichment of molecular phenotypes in AID variants and described many disease loci where there is evidence of shared molecular mechanisms (Chen et al., 2016a). It is therefore well established that while GWAS provide unbiased systematic identification of disease-associated loci, functional insight must be gleaned through the combination of intermediate phenotypes QTLs. Detailed “omic” data collected from healthy individuals can be used as a powerful tool in understanding disease mechanisms, as these cohorts enable association of variants with intermediate phenotypes which themselves have not been affected by disease status. This limits the confounding factors and possible reverse causation whereby a dysregulated disease state could cause molecular changes rather than those changes being risk factors for the disease. The combination of genomics and functional experiments have also demonstrated how the disruption in the function of peripheral immune cells can contribute to the pathogenesis of a wide-range of diseases.

To expand this analysis, I used the same molecular QTLs from the BLUEPRINT project, but applied them to a collection of four non-autoimmune disorders and one prototypic AID, SLE (Chen et al., 2016a). I used the GARFIELD method to evaluate significant enrichment of GWAS variants in immune molecular QTLs. Following this, I implemented statistical colocalisation methods using gene expression, histone and splicing-associated QTLs in neutrophils, monocytes and T cells with GWAS SNPs. I then further evaluated whether there is evidence that immune phenotypes can, at least in part, aid the development of mechanistic hypotheses underpinning genetic risk at disease loci. I also thoroughly investigated biological mechanisms to provide functional hypotheses that will aid the design of further experimental dissection of disease loci. I demonstrated that this approach required the integration of multiple data sources and analytical approaches in order to provide in-depth molecular insight into a specific disease locus.

2.2 Materials and methods

Quantitative trait loci (QTL) data: Molecular phenotypes from the BLUEPRINT study were used to assign function to disease loci (Chen et al., 2016a). The study design, summarised in Figure 2.2, included generation of a mean read depth of ~7X-whole-genome sequences, transcriptomes (RNA-seq), histone ChIP-seq data (H3K4me1, H3K27ac) and DNA methylation probes (450K array) in population sample of up to 197 healthy individuals. These data were collected in three cell types, CD66b^{+}CD16^{+} neutrophils, CD14^{+}CD16^{-} monocytes and CD4^{+}CD45RA^{+} T cells.

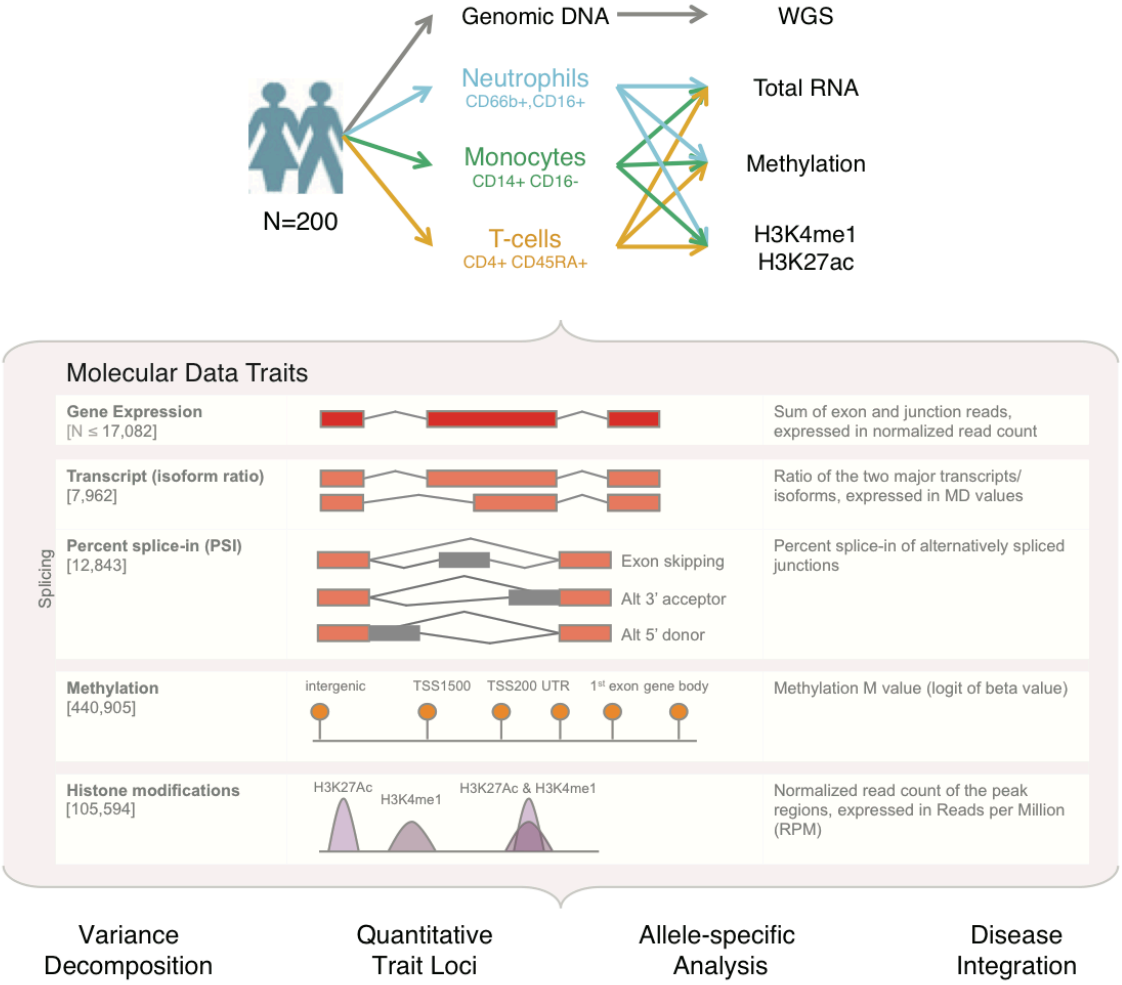


Figure 2.2: Summary of the BLUEPRINT Epigenome variation project
Overview of the study design and molecular traits investigated in BLUEPRINT. Figure reproduced from (Chen et al., 2016a) under the CC license (http://creativecommons.org/licenses/by/4.0/).

For investigating genetic functional mechanisms in this chapter, I focused on gene expression, splicing events (exon skipping or alternative acceptors or donors, Figure 2.2), H3K27ac and H3K4me1 QTLs from all three cell types. I also evaluated disease variant enrichment in methylation QTLs. Full analysis methods are detailed in Chen *et al.* (2016) but briefly, phenotype values of histone signal from ChIP-seq and expression from RNA-seq data were corrected for sequencing centre batch effects. For ChIP-seq data, phenotypes were normalised by the total number of million reads per individual (RPM) and normalised by taking log2. Gene expression was normalised in DESEQ and expressed in units proportional to $\log_2$FPKM, corrected for gene length and sequencing depth. These phenotype values were used for the whole cohort for visualisation and further testing described in this chapter. For QTL association, unknown sources of non-genetic covariation were removed by correcting for the first ten PEER factors (Chen et al., 2016a). Each feature was tested for association with variants within the feature and 1Mb upstream and downstream of the start and end position. The p value of association was corrected for increased false positives due to testing multiple variants by calculating the qvalue. This method controls for the false discovery rate, which is the proportion of false positives generated by testing multiple hypotheses (Bass JDSwcfAJ, 2015). Variants with a qvalue less than 5% were designated significantly associated with the phenotype. For visualising modified histone regions with more resolution, as in Figure 2.10 and 2.17, aligned reads expressed in $\log_2$RPM were used for gene expression and histone signals and calculated across 50 bp non-overlapping sliding windows across the genome.

Disease GWAS datasets: To perform colocalisation methods, I collected summary statistics from GWAS with European cohorts of relatively high sample sizes and power. All GWAS were annotated with the genome build hg19. Z-score was calculated using the supplied effect size estimate divided by the standard error. Advanced age-related macular degeneration summary statistics including beta and standard error estimates were provided by kind permission of the International AMD Genomic Consortium (Fritsche et al., 2016). The study consisted of 12,023,830 variants, 16,144 cases and 17,832 controls. Alzheimer’s disease summary statistics were obtained from the International Genomics of Alzheimer’s Project (IGAP) stage 1 meta-analysis from 2013 (Lambert et al., 2013). Stage 1 of the study consisted of 7,055,881 SNPs, 17,008 cases and 37,154 controls. Coronary artery disease (CAD) summary statistics were obtained from the CARDIoGRAMplusC4D consortium website from 2015 study (Nikpay et al., 2015). The study consisted of 9,455,778 variants, 60,801 CAD cases and 123,504 controls and the additive model associations were used in this chapter. Summary statistics for systemic lupus erythematosus were obtained from the immunobase website and accessed in October 2016 (ImmunoBase, 2017). The new SLE GWAS consisted of 7,219 cases and 15,991 European controls (Bentham et al., 2015). The summary statistics of extremes of the lung ratio, FEV_1, from never smokers were obtained from the UKBiLEVE project as part of the UK Biobank access (Wain et al., 2015). The initial study included 50,008 individuals, which were further stratified into never or heavy smokers and further into range of FEV_1, for example, 9745 never smokers with low FEV_1, 9827 never smokers with average FEV_1 and 4902 never smokers with high FEV_1 (Wain et al., 2015). Type 2 diabetes summary statistics (Morris et al., 2012) were also used to compare enrichments of other traits, these were accessed from the DIAGRAM consortium website (Stage 1 GWAS) in 2016. Further details of the GWAS summary statistics and website links are detailed in the Supplementary Information.

GWAS enrichment (GARFIELD): GARFIELD (Iotchkova et al., 2016), was implemented to assess significant enrichment of GWAS SNPs with molecular QTLs. In this thesis, the version of the method utilises genome-wide summary statistics to calculate odds ratios for association between an overlap (SNP-QTL annotation) and disease status (significant p-value for the disease/complex trait), as was used in the Astle *et al.* (2016) and Chen *et al.* (2016) studies. The QTL summary statistics are formatted into annotations by selecting all significant QTLs (qvalue < 5%) for each QTL type and each cell type. Where there are duplicated QTLs, due to association with multiple features, the lowest p value is used. The method greedily prunes input GWAS disease/trait-associated variants, retaining the most significant variant and removing variants with LD $r^2 \geq 0.1$. LD tags are pre-calculated using 1Mb windows and the combined UK10K and 1000 genomes Phase 3 panel (Europeans). The process is repeated until no significant variants remain. Independent variants are then overlapped with the annotations of interest by matching genomic positions. SNPs in high LD with the independent variant are also annotated as overlaps ($r^2 \geq 0.8$). Odds ratios are calculated at various GWAS thresholds (1 x 10^{-08}, 1 x 10^{-07}, 1 x 10^{-06}, 1 x 10^{-05}, 1 x 10^{-04}, 1 x 10^{-03}, 1 x 10^{-02}, 1 x 10^{-01}, 1). The significance of the odds ratio at each GWAS threshold is calculated using a generalized linear model in a logistic regression approach that controls for LD (number of variant proxies), minor allele frequency and local gene density (variant distance to the TSS) input as categorical variables. The method corrects for multiple annotations tested by applying the Bonferroni correction using the effective number of annotations, which is calculated based on the correlation between annotation-SNP overlap matrices. In this thesis, there was an increased representation of rare variants in the specially designed respiratory array used in the FEV_1 UKBiLeve summary statistics and rare variants were also included in the AMD and CAD studies. In the BLUEPRINT molecular data, the analysis focused on common variants (MAF ≥ 1%). Therefore, variants with a MAF < 1% were filtered from these summary statistics before evaluating enrichment.

Generation of regions for colocalisation input: I performed a locus pre-selection step to test for colocalisation as assessing all regions across the whole genome would constitute a high computational burden given the vast number of QTL associations. I generated a list of regions where significant disease SNPs and index molecular QTLs overlapped, based on the hypothesis that genomic regions associated with molecular feature(s) and a complex trait were more likely to share underlying genetic causes. Specifically, starting from the BLUEPRINT QTL summary statistics, significant QTLs (qvalue ≤ 5%) were selected. For each disease-QTL combination, overlaps were annotated if any lead QTL per feature (gene, splicing, methylation probe, H3K27ac or H3K4me1 peak) and proxy variants in high LD ($r^2 \geq 0.8$), were also significant in the GWAS summary statistics (p value $\leq 5 \times 10^{-08}$). If the overlap occurred with lead QTL proxies, the corresponding lead QTL information was retained. For each lead QTL, all features associated with the unique lead QTLs that overlapped were identified. These were referred to as feature-QTL pairs and represent different regions tested for colocalisation (Figure 2.3). For each pair, the genomic region assessed in colocalisation was defined by the BLUEPRINT QTL testing region (SNPs within the feature and 1 Mb upstream and downstream). Only variants that overlap between the QTL and disease study are evaluated in colocalisation.

Colocalisation of disease and molecular trait-associated loci: To perform colocalisation, I implemented the method, gwas-pw (Pickrell et al., 2016). This is a Bayesian method that assesses whether the overlap observed between a GWAS SNP and molecular trait QTL is due to a sharing of the genetic effect, i.e. the same genetic variant is associated with both the GWAS trait and the molecular trait. The method estimates a probability that the association evidence in a given genomic region falls into one of four models. Model 1 and 2 indicate that the locus contains a single variant associated with the first trait or the second trait only, i.e. there is one association. Model 3 indicates colocalisation where a single variant is associated with both traits. Finally, model 4 indicates that two independent associations exist, where the variant is associated with the first trait and a second, independent variant is associated with the second trait (Figure 2.4).

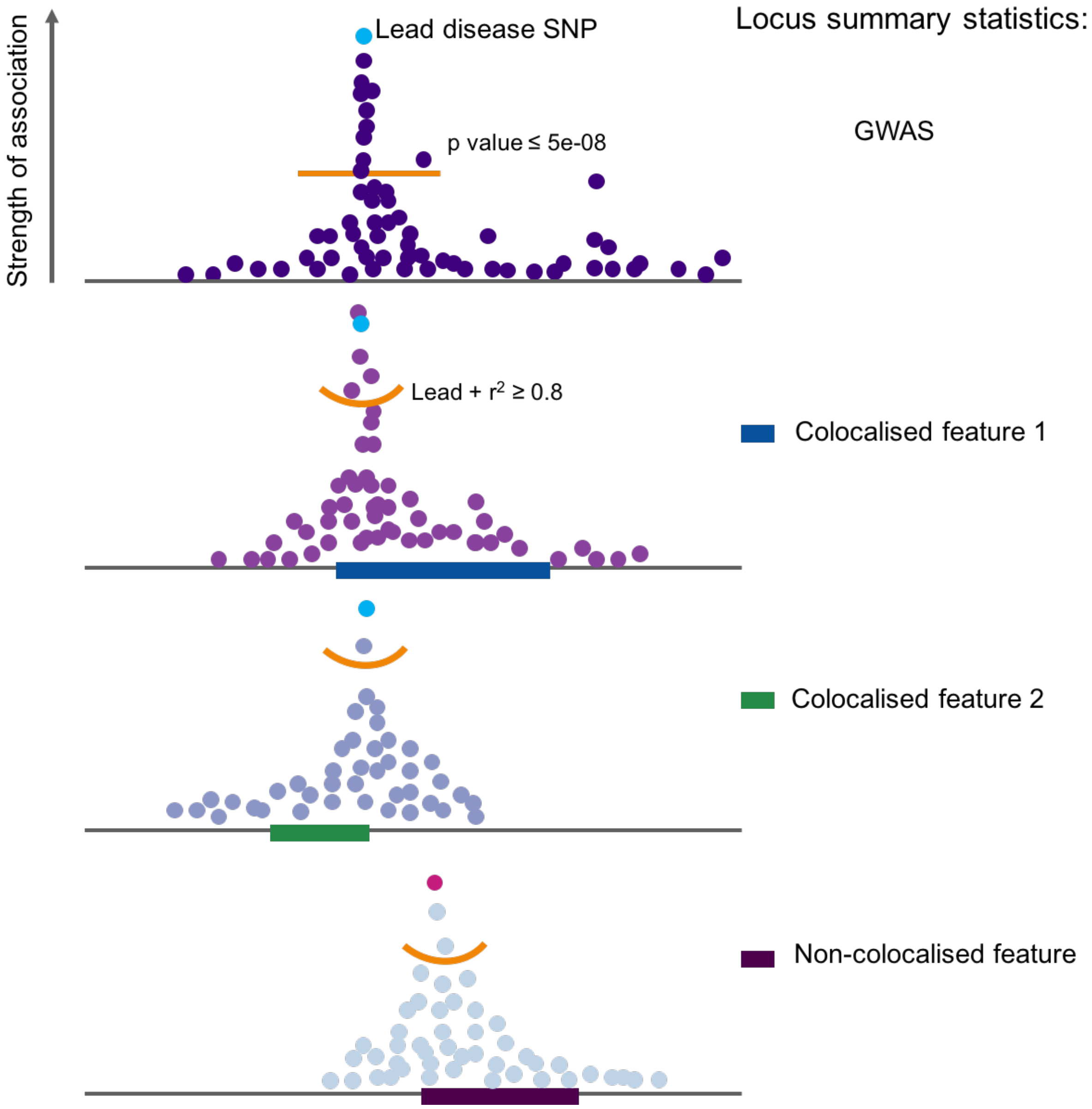


Figure 2.3: Summary of molecular QTL and disease locus colocalisation approach
This schematic summarises the selection of overlapped QTL-SNP regions and assignment of the disease SNP after colocalisation on individually tested regions. Variants tested for association in each set of summary statistics are represented by a coloured dot and the schematics represent manhattan plots with the higher the variant, the more significantly associated with the specified trait. Molecular features, such as genes or histone binding regions are represented by coloured blocks. For each identified GWAS-feature pair, lead QTLs and proxies associated with molecular features (above the curved orange line) are selected and an overlap is called if any of these SNPs are significantly associated with the GWAS trait (p value ≤ 5 x 10^{-08}) (above the straight orange line). For each overlap, all features significantly associated (qvalue ≤ 5%) with this lead QTL are assessed for colocalisation (above as the blue and green features). All SNPs within the feature and the cis-genomic region, defined as 1Mb upstream and downstream, are input to colocalisation if shared with the GWAS study. Where many features colocalised with one disease locus, the lead disease SNP in the test window (light blue) is used to assign the previously reported locus. The bottom panel represents a scenario where an overlap may have been detected between a molecular feature SNP and a significant GWAS SNP, but colocalisation was not detected with the disease GWAS signal.

Full details are listed in the Pickrell *et al.* (2016) publication but I briefly summarised the main points and equations from the method below. Here, we assume that the BLUEPRINT and disease cohorts are independent and there is no overlap or correlation. First, the method calculates a Bayes factor, which corresponds to the association evidence for three alternative models (below). Bayes factors for each SNP are approximated from Wakefield [2008]. Equation 1, below, gives the Wakefield approximate Bayes factor for model 1 where the SNP is associated with trait 1.

$$WABF_1 = \sqrt{1 - r_1}\exp\left[\frac{Z_1^2}{2} r_1\right] \qquad (1)$$

$$BF^{(1)} = WABF_1 \qquad (2)$$

$$BF^{(2)} = WABF_2 \qquad (3)$$

$$BF^{(3)} = WABF_1 WABF_2 \, , \qquad (4)$$

where Z_1 is the Z score estimate (the maximum likelihood estimate of beta divided by standard error, $\sqrt{V_1}$) of each SNP with the trait of interest and $r_1 = \frac{W_1}{V_1 + W_1}$. Bayes factors are averaged over computations with varying W. Therefore, in this method, the Z score for a SNP in the region for each trait is used to evaluate evidence for colocalisation. The three approximate Bayes factors above relate to three models; where the SNP is associated with the first trait (equation 2), second where the SNP is associated with the second trait (equation 3) and the third where a SNP is associated with two traits (equation 4).

Next support for an association is evaluated in a given genomic region, r, for all SNPs in the region. The regional Bayes factor (RBF_r) is evaluated for each model against the null model of no association in the region. The RBF_r evaluates the integral sum of the posterior probability (PP) for all SNPs in the region where the PP is the product of the Bayes factor and the prior probability of the variant being causal in the locus. In the model, all SNPs have equal prior probability of being causal. The method assumes one casual variant and if this is missing, the power to detect a shared genetic effect is reduced (model 3).

$$RBF_r^{(1)} = \sum_{i=1}^{K} \pi_i^{(1)} BF_i^{(1)}$$

$$RBF_r^{(2)} = \sum_{i=1}^{K} \pi_i^{(2)} BF_i^{(2)}$$

$$RBF_r^{(3)} = \sum_{i=1}^{K} \pi_i^{(3)} BF_i^{(3)}$$

$$RBF_r^{(4)} = \sum_{i=1}^{K} \sum_{j=1}^{K} \pi_i^{(1)} \pi_j^{(2)} BF_i^{(1)} BF_j^{(2)} I[i \ \neq j],$$

where $\pi_i^{(1)}$is the prior probability that SNP i is the causal one under model 1, $\pi_i^{(2)}$, is the prior probability that SNP i is causal under model 2 and $\pi_i^{(3)}$ is the prior probability that SNP i is causal under model 3. $RBF_r^{(4)}$assumes there are two causal SNPs that independently influence the two traits. K refers to the number of SNPs. The SNP priors are set as follows $\pi_i^{(1)} = \pi_i^{(2)} = \pi_i^{(3)} = \frac{1}{K}$.

Next, the prior probability of the regional models is calculated by the method to maximise the log-likelihood function of all SNPs in a region, over all four models. In our case, this is calculated per locus, which is pre-defined as a region with an overlapping molecular QTL and GWAS SNP (Figure 2.3) not genome-wide, given the tendency for QTL testing regions (2 Mb regions) to overlap. Finally, a posterior probability for each model per locus is calculated by multiplying the corresponding model prior probability by the RBF_r. For each locus, four posterior probabilities are generated for model 1, 2, 3 and 4. The PP for model 3 is used to evaluate whether there is evidence for a shared genetic effect.

There are some limitations in this method. It is difficult to distinguish between model 3 and 4 if there is high LD between the lead variants of each of the two traits ($r^2 \geq 0.8$). The model does not provide an estimation and direction of causality between the molecular trait and disease.

The threshold for calling a region colocalised was PP ≥ 0.99 for model 3, as this gives high confidence that there is statistical evidence for an underlying shared genetic signal between two traits. Only regions where there were equal or more than 20 SNPs shared between the disease and molecular datasets were considered as colocalised loci. Locus zoom plots were generated using custom scripts and used to provide visual evidence for colocalisation. The p-value supplied in the GWAS datasets was used in plotting as well as the raw p-value for the molecular QTLs (not 5% qvalue). Only SNPs that were shared between the disease and molecular QTLs were plotted in the locus zoom plot (as these were the input into gwas-pw). Final results excluded the MHC region, which was defined as the region on chr6 between the positions 20000000 and 40000000 based on previous investigations of the region (Trowsdale and Knight, 2013).

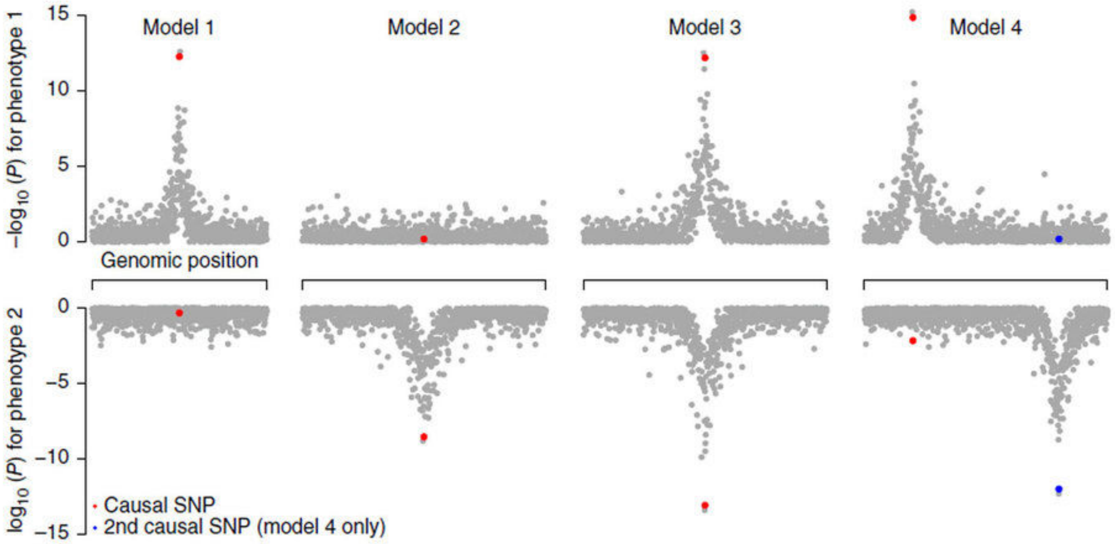


Figure 2.4 Schematic of the four models evaluated by colocalisation methods describing the relationship between two trait associations within a locus
In this analysis, the genomic region above represents a molecular QTL testing region overlapping a GWAS SNP. Model 1 and 2 are single association models, model 3 represents a colocalised region where two traits have shared genetic signals and model 4 is where two independent variants are associated with different traits. Figure reproduced from Pickrell *et al*, 2016 (Pickrell et al., 2016).

Assigning a unique disease locus: Multiple molecular features colocalised with some disease loci. In order to evaluate all possible molecular consequences of each disease loci, a lead disease SNP was assigned to each feature-QTL pair, each assessed independently for colocalisation. Molecular features sharing the same disease SNP were aggregated. The SNP with the lowest p-value in the testing region (+/- 1Mb) was selected from the GWAS summary statistics. The assigned SNP was compared to the reported disease lead SNP. In most cases, there was an exact match between the assigned and reported disease SNP. However, for SLE and AD, summary statistics from the full meta-analysis were not publicly available, instead for both summary statistics from the stage 1 GWAS were available. Reported lead SNPs in the study were based on meta-analysis. In these cases, LD (r^2 1000G) between the assigned and reported disease SNP was used to assign the published locus. For SLE, three regions significant in stage 1, which colocalised with features were removed from the final results as no corresponding region could be identified in the published meta-analysis. The FEV_1 GWAS contained a complex region of extended LD that was classified as the *KANSL1* inversion locus, defined by one conditionally significant genetic signal. The extended LD made it difficult to assign a lead disease SNP in my analyses. I found that features colocalised with FEV_1 disease SNPs rs111907488, rs62060763 and rs2532349 were all located in this region, therefore I combined all of these features into the *KANSL1* locus defined by the study reported lead SNP, rs2532349. Importantly, this method assigns the most significant lead SNP per locus, but the colocalisation could occur between secondary or further independent signals in either GWAS or BLUEPRINT associations. However, the lead SNP is assigned for ease of comparison to previous findings. In depth investigation of each locus is required to assess whether colocalisation occurs with the primary signal.

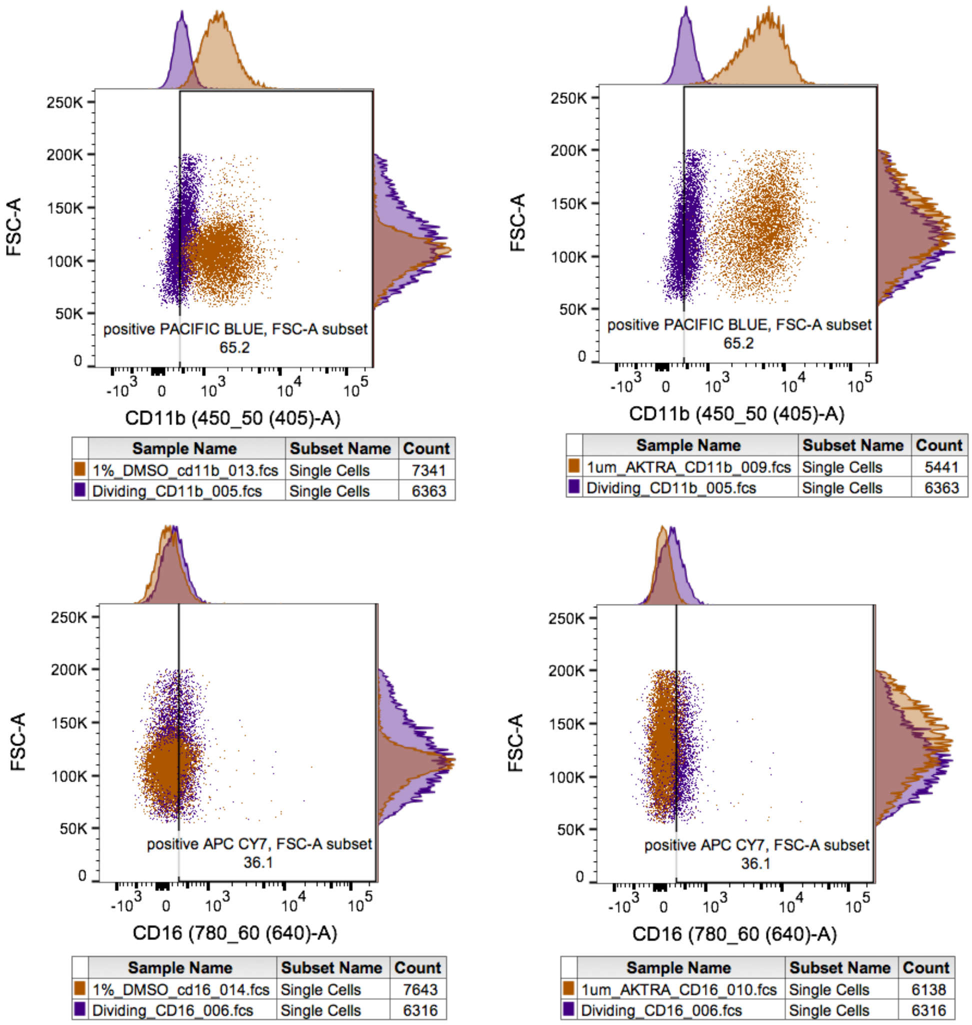
Conditional SNP analysis within phenotype: To gain a better understanding of molecular signals and whether these were shared between cell types, I used GCTA conditional analysis to estimate conditionally independent signals based on association statistics and LD between the variants in the locus. GCTA version 1.25.2 with the --cojo-cond option was implemented using QTL summary statistics from the BLUEPRINT study (Chen et al., 2016a, Yang et al., 2011, Yang et al., 2012). Summary statistics for each feature were input into GCTA. For LD estimation, genotype information from the BLUEPRINT cohort was used in plink hard call format using the --bfile option. Iterative conditional analysis was performed if any SNPs were associated with the phenotype after conditioning on the lead SNP. To evaluate this the q value, which represents the p value corrected for the number of SNPs tested in the region, was calculated using the qvalue R package version 1.43.0 (Bass JDSwcfAJ, 2015). GCTA was then implemented on output summary statistics conditioning on the SNP with the lowest qvalue from the previous conditional analysis. In most cases, the conditional analysis was confirmed by using a simple linear regression model with individual genotype data in R.

Linear regression analysis: For specific loci, as part of an in-depth investigation, various linear regression models were implemented using the lm() function in R. Phenotypes as described above were extracted from the full matrices from the Chen *et al.* (2016) study for features of interest. Units were $\log_2$RPM for the histone signal or normalised expression values, which are proportional to log2FPKM. SNP genotypes were input for every individual where 0 denotes homozygous reference, 1 heterozygous and 2 homozygous alternative, the latter was the defined effect allele for the Chen *et al* (2016) BLUEPRINT study. Genotypes were input as numeric to test for a trend across the 0,1,2 genotypes rather than as a factor where the difference between each genotype level is evaluated. Genotypes and phenotype values were matched using the unique study ID. To evaluate the causal relationships at the AMD *TNFRSF10A* locus, I used phenotype values for 158 individuals for which all phenotype data was available (gene, H3K27ac, H3K4me1). I implemented a two-stage approach to test for causality, first removing the effects of particular phenotypes by including these as covariates in a linear model. I used the Shapiro Wilko test to confirm normality of residuals after correction and used these residuals to test for remaining association with the locus lead SNP genotype.

R^2 and goodness of fit: The model fit was evaluated using the R^2 generated from the lm() function. Briefly R^2 or the coefficient of determination measures the proportion of y that is explained by x, the predictor variable, where a value of 1 demonstrates that x explains all of the variation in y. To evaluate whether fitting of additional covariates improved the model fit, the anova() function in R was used to perform a Chi-squared test to compare nested linear model 1 and linear model 2, related by the inclusion of additional covariates into model 2 (see Section 2.3.5.2). The chi-squared test evaluates whether the reduction in the residual sum of squares is statistically significant or not.

Linkage disequilibrium calculation: LD between variants was assessed using the 1000 genomes panel via the HaploReg resource and denoted as 1000G (Ward and Kellis, 2012). Alternatively, where indicated, LD was calculated from the Astle *et al.* (2016) cohort using PLINK v2 with the flags --ld and --bfile for the input imputed hard call files generated as part of the main GWAS analysis (Astle et al., 2016).

HL60 differentiation model: Additional functional data was required in some cases to further assign mechanism to disease loci. There are known limitations with access to primary human neutrophils and technical difficulties associated with using genomics approaches in these cells. The cell line, HL60, are commonly used as a model for neutrophils but resemble an early population of promyelocytes (Birnie, 1988). To address this and provide an additional high-quality granulocytic dataset, I implemented a well-established method for differentiating the HL60 cell line into a more mature neutrophil-like state and functional phenotype, with the addition of all-trans retinoic acid (ATRA) or dimethyl sulfoxide (DMSO) (Breitman et al., 1980, Chang et al., 2006). HL60 cells were grown at 37^{0}C in RPMI medium supplemented with 10% FBS and penicillin and streptomycin. Cells were passaged at a cell density of less than 1 x 10^6 cells per mL media. HL60 cells were seeded at a density of 10 x 10^6 cells/mL and incubated with either 1 µM ATRA or 1% DMSO for 96 hours. Cells without addition or either ATRA or DMSO were grown for 96hours without media change as a control. Every 24 hours, cells were counted for viability using a C-chip counting chamber and 1:1 dilution of Tryphan blue. After 96 hours, cells were harvested and fixed with 1% formaldehyde. To assess the success of differentiation, cell viability as well as the surface expression of neutrophil marker CD11b were measured. For flow cytometry measurements, 1 x 10^6 cells were harvested, spun at 1200 rpm for 5 minutes and pellets were resuspended in 100 µL FACS buffer (2% BSA, 0.001% EDTA in D-PBS). Cells were washed and incubated with 5 µL of Fc receptor blocking solution (Human TruStain FcX, BioLegend 422301) for 10 minutes on ice and afterwards washed with FACs buffer (1200 rpm for 5 minutes). 2 µL of the relevant antibody or isotype control was added for 30 minutes at 4^{0}C (Table 2.1). The stained cells were then washed three times. Unstained and isotype controls were also prepared. Samples were analysed using a BD LSR Fortessa Cell Analyser. In addition, gene expression of candidate genes known to vary in the differentiation process was also evaluated. RNA was extracted using a QIAGEN RNeasy mini kit and treated for DNase using TurboDNase (Life Technologies). Oligo(dT) primers and Superscript II (Invitrogen) were added for 2 hours at 42^{0}C for reverse transcription. HL60 genomic DNA was used as a control for the RT-PCR standard curve. CT values were calculated and compared to two reference genes, actin and C/EBPβ.



	Sample Name	Subset Name	Count
🟧	1%_DMSO_cd11b_013.fcs	Single Cells	7341
🟪	Dividing_CD11b_005.fcs	Single Cells	6363

	Sample Name	Subset Name	Count
🟧	1um_AKTRA_CD11b_009.fcs	Single Cells	5441
🟪	Dividing_CD11b_005.fcs	Single Cells	6363

	Sample Name	Subset Name	Count
🟧	1%_DMSO_cd16_014.fcs	Single Cells	7643
🟪	Dividing_CD16_006.fcs	Single Cells	6316

	Sample Name	Subset Name	Count
🟧	1um_AKTRA_CD16_010.fcs	Single Cells	6138
🟪	Dividing_CD16_006.fcs	Single Cells	6316

Figure 2.5: CD11b surface expression is increased on HL60 differentiation with DMSO or ATRA

Dot plots and histograms of the fluorescent signal of either CD11b (top panel) or CD16 (bottom panel) measured using flow cytometry. CD11b surface expression is increased on differentiated HL60 cells (DMSO and ATRA) compared to dividing HL60 (undifferentiated). CD16 expression is largely unchanged upon differentiation as has been previously observed (Jacob et al., 2002). Reduced proliferation and known gene expression changes (Lee et al., 2002) were also demonstrated (Supplementary Figures 2.1-2.2).

Chromatin immunoprecipitation followed by next-generation sequencing (ChIP-seq): ChIP experiments were adapted from a previously published protocol (Schmidt et al., 2009), with some modifications. 20 x 10^6 cells were used for TF immunoprecipitation (IP) and 5 x 10^6 cells for histone modifications. Cell pellets were fixed by incubation with 1% formaldehyde for 10 minutes at RT, followed by five minutes with 2.5M glycine. Cells were pelleted and washed with PBS, flash frozen in dry ice and stored at -80˚C. Cells were sonicated for eight cycles with 30 seconds on and 45 seconds off at 4˚C using a Diagenode PicoRuptor biorupter. Sonication efficiency (150-500bp fragments) was verified using an Agilent DNA bioanalyser. 2.5μg of each antibody was bound to Protein A Dynabeads (Invitrogen) in a 4-hour incubation. Sonicated lysate was added to the antibody-bead mix and incubated overnight at 4˚C. The bound-beads were then washed with cold RIPA buffer ten times. Crosslinks were reversed with incubation at 65˚C for five hours to elute DNA. Samples were then incubated with 2μl RNase at 37˚C for 30 minutes followed by incubation with Proteinase K at 55˚C for one hour. Ampure beads (Beckman Coulter, A63881) were added to the DNA in a 1:1.8 ratio and samples were washed twice in cold 70% ethanol. DNA was then eluted in 50 μl elution buffer. Samples were stored at -20˚C prior to Illumina library preparation, which was carried out according to the Illumina TruSeq ChIP sample kit protocol. An additional step of enriching fragments through PCR prior to gel purification was added to avoid amplification of contaminants. RNA index sequences were ligated to ChIP-enriched DNA fragments (200-500bp in size) for multiplexed libraries. Libraries were submitted for single-end sequencing with a read length of 50bp using a HiSeq2000. For analysis of sequencing data, reads were first aligned to the human reference genome (hg19 CRCh37) using BWA version 0.6.1 with default parameters (Sanger pipelines). ChIP-seq data analysis was performed using an in-house pipeline developed by Louella Vasquez that implemented standard analysis procedures, described here. Duplicate reads were removed (Picard MarkDuplicates v1.103) and reads with a zero-mapping source were removed. Peaks were called using MACs v2.0.10.20131216 (Zhang et al., 2008) with default parameters using the estimated fragment size evaluated by PhantomPeakQualTools vr18. Narrow flags were used for all factors apart from H3K4me1 for which the broad flag was used. For the background control, sonicated input DNA was used from the respective ATRA or DMSO treatments. Encode quality control metrics (Phantom Quality Tools) were used to evaluate the success of IP as well as visual inspection in the UCSC genome browser. Significant peaks were selected if 1% FDR or less.

Antibody	Supplier	Source/Clone
Anti-PU.1	Santa Cruz, sc22805	Rabbit polyclonal
Anti-CEBPβ	Santa Cruz 150 X	Rabbit polyclonal
Anti-trimethyl histone H3(lys4)	Diagenode C15410003	Rabbit polyclonal
Anti-H3K27acetyl	Abcam ab4729	Rabbit polyclonal
Anti-monomethyl histone H3(lys4)	Diagenode C15410194	Rabbit polyclonal
PE Mouse Anti-Human CD11b	BD Pharmingen 557321	ICRF44
PE Mouse IgG1, K isotype control	BD Pharmingen 556650	MOPC-21
Pacific Blue Anti-Human CD11b	BD 558123	ICRF44
Pacific Blue Mouse IgG1, isotype control	BD 558120	MOPC21
APC CY7 Anti-Human CD16	BD 557758	3G8

Table 2.1: Antibodies used in ChIP-seq and flow cytometry experiments
Antibodies against specific proteins studied, the supplier and reference as well as source or clone for flow cytometry experiments.

2.3 Results

2.3.1 Functional enrichment in five diseases

First, I aimed to assess the enrichment of non-autoimmune disease SNPs in immune molecular QTLs. I used the GARFIELD method (Materials and Methods) to evaluate significant enrichment by calculating the odds ratio (OR) for enrichment of GWAS variants in molecular features such as gene expression, RNA splicing, histone marked regions and lastly DNA methylation from three cell types (Chen et al., 2016a). Higher OR estimates indicated increased odds that an overlap occurs with a significant GWAS variant as opposed to an overlap with a non-significant GWAS variant.

I identified significant enrichment of AD, CAD, FEV_1, AMD and SLE variants associated at the GWAS p value threshold of 1 x 10^{-05} in a number of molecular feature types, ranging from four significant features for AD to all features with significant enrichment for SLE. This is in contrast to the complete lack of significant enrichment of immune QTLs with Type 2 diabetes (T2D) SNPs. I included T2D-associated SNPs as a negative control as despite recent links to inflammation, it is currently considered that a disordered metabolic state in Type 2 diabetic patients then leads to immune dysregulation (Hameed et al., 2015). This is supported by low enrichment of T2D variants in immune molecular features, as observed previously in two independent studies, and further confirmed here in my analysis (Figure 2.6) (Chen et al., 2016a, Alasoo et al., 2017).

The highest and some of the most significant enrichments were observed with neutrophil, monocyte and T cell splicing QTLs and FEV_1-associated variants, with a mean OR of 12.520 across all three cell types and p values ranging from 6.404 x 10^{-18} for T cell splicing QTLs to 2.264 x 10^{-24} for neutrophil splicing QTLs. Significant enrichment in splicing QTLs was also observed for other traits. AMD-associated variants showed the highest enrichment in splicing QTLs from all three cell types (mean OR = 4.541), with the highest OR also in neutrophils. SLE-associated variants also showed the highest enrichment in neutrophil splicing regions (OR = 5.247, p value = 8.505 x 10^{-13}).

However, there were large OR confidence limits for the FEV_1 splicing enrichment, which indicated a higher error in this measurement. This could be due to a lower number of variants overlapping these annotations when in comparison to some (but not all) of the disease GWAS. For FEV_1, 24 variants were annotated as overlapping monocyte splicing QTLs at the GWAS p value threshold of 1 x 10^{-05}, 24 overlapping T cell splicing QTLs and 24 with neutrophil splicing QTLs, but for AMD-GWAS, 39, 36 and 40 variants overlapped monocyte, neutrophil and T cell splicing QTLs respectively at the same GWAS 1 x 10^{-05} p value threshold.

SLE-associated variants were significantly enriched across all molecular QTL types, whereas other trait variants demonstrated more variable enrichment patterns. The high and significant enrichment across the majority of QTL and cell types was previously observed for other autoimmune diseases including Crohn's disease, rheumatoid arthritis and multiple sclerosis (Chen et al., 2016a). The ubiquitous effect may reflect extensive cross-talk between different immune cell populations or that increased power in both GWAS and QTL studies are required to fully resolve finer immune cell population signatures.

Interestingly, CAD was the only trait for which there appears a possible cell-type specific pattern of enrichment with a consistent significant enrichment for monocyte gene, H3K27ac, H3K4me1 and methylation QTLs, which was in agreement with the established role for monocytes in CAD aetiology (discussed above). The mean OR for the significantly enriched monocyte features, as calculated from the GARFIELD output data was 2.737. Significant neutrophil annotations showed a slightly lower average OR of 2.310, and the equivalent p values of enrichment were less significant than the equivalent p value for monocyte annotations apart from for neutrophil splicing, which was not significant in monocytes. For T cells, only gene, H3K27ac and methylation QTLs showed significant enrichment with CAD variants.

In summary, this approach has demonstrated significant enrichment of non-autoimmune disease variants in molecular QTLs, where greater enrichment was observed in all cases for at least five molecular QTL types than the negative case of T2D. For example, all five disease-associated variants showed significant enrichment in monocyte and neutrophil eQTLs. These significant enrichments suggest that these GWAS variants may result in functional changes in immune cells that could underpin mechanisms at some genetic risk loci. Therefore, using these molecular data to further study disease loci could aid the generation of biological hypotheses of downstream consequences at these loci.

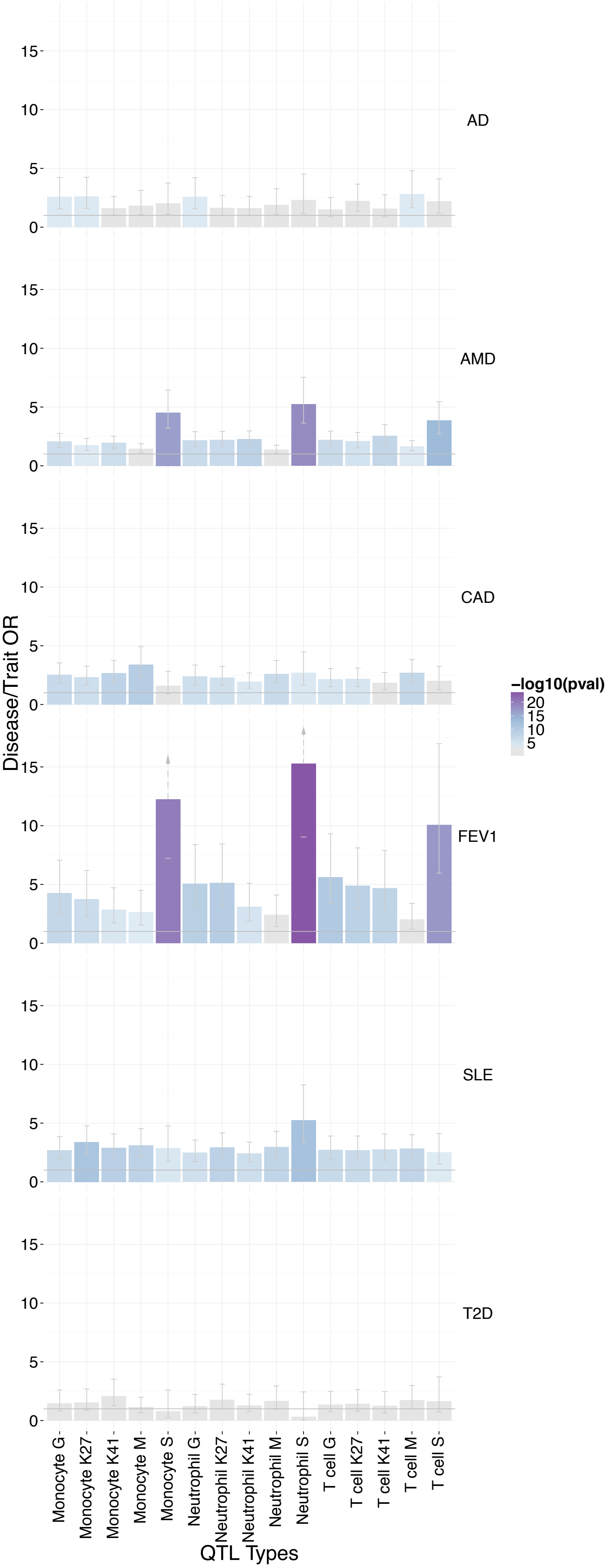

AD
AMD
CAD
FEV1
SLE
T2D
0
5
10
15
Disease/Trait OR
-log10(pval)
20
15
10
5
Monocyte G
Monocyte K27
Monocyte K41
Monocyte M
Monocyte S
Neutrophil G
Neutrophil K27
Neutrophil K41
Neutrophil M
Neutrophil S
T cell G
T cell K27
T cell K41
T cell M
T cell S
QTL Types

Figure 2.6: Enrichment of molecular QTLs in six diseases
Histogram plot summarises the enrichment of molecular QTLs in a range of diseases. Higher enrichment represents a greater overlap of disease SNPs and molecular QTLs and is indicated by an increased odds ratio (OR) on the y-axis calculated by GARFIELD. Represented here is overlap of disease variants at the suggestive GWAS p value threshold, 1 x 10^{-05}. The significance of the overlap is depicted by the colour scale of the plots (purple is more significant), and is corrected for the number of effective annotations (Methods) and for 5 diseases (where significant enrichment was expected, which excludes Type 2 diabetes). Annotations are enriched with OR ≥ 1. The OR confidence limits are reflected by the error bars, with arrows to designate values beyond the maximum OR shown in the figure. Type 2 diabetes was included as a non-immune negative control and as such no significant enrichment was observed for these SNPs. G refers to gene features i.e. eQTLs, K27 to H3K27ac hQTLs, K41 to H3K4me1 hQTLs, M to methylation QTLs and S to splicing QTLs (percent-splice in as defined in the main BLUEPRINT paper and detailed in Methods).

2.3.2 Colocalisation of molecular traits with five diseases

Formal statistical testing is required to evaluate whether the molecular trait and disease associations can be attributed to the same underlying causal genetic variant. I therefore implemented the colocalisation method, gwas-pw, which uses prior information regarding the effect sizes (expressed as a Z-score) of two traits within one locus (detailed in the Materials and Methods). I assessed colocalisation between each of the five complex diseases/traits described above; AD, SLE, CAD, AMD and FEV_1 with four molecular traits, gene expression, H3K72ac & H3K4me1 modifications and RNA splicing from three cell types monocytes, neutrophils and naïve $CD4^+$ T cells (Chen et al., 2016a). The overall aim of this approach was to identify to what extent immune molecular features could explain mechanisms at unique disease loci reported from each GWAS.

Across all five diseases, I identified that 46% (55/120) of previously reported disease loci colocalised with at least one molecular feature using a high posterior probability of colocalisation (PP ≥ 0.99) (Table 2.2). The highest percentage of colocalised loci was observed with systemic lupus erythematosus (SLE) variants (54%), which is expected as SLE is a paradigmatic autoimmune disease displaying strong immune cell involvement (Dai et al., 2014, Farh et al., 2015). SLE was included in this chapter to provide a positive control, as strong enrichment of SLE variants in lymphoid and monocyte enhancers (high H3K27ac) has been previously demonstrated using ROADMAP data (Farh et al., 2015). Neutrophil data was not assessed in this previous study, despite recent observations of the importance of this cell type to SLE pathogenesis (Weidenbusch et al., 2017). Using the BLUEPRINT neutrophil data, therefore, provided the opportunity to gain novel insight into neutrophil-mediated risk at SLE loci.

The lowest percentage of colocalised loci was observed with coronary artery disease (40%). However, of the 43 common GWAS loci excluding the MHC region that were assessed, there were still 17 CAD loci that colocalised with at least one molecular feature. Therefore, this type of analytical approach affords the potential to form molecular hypotheses of up to one-third of the known loci within clearly defined cell type populations facilitating downstream experimentation.

Disease	Disease variants (+MHC)	Disease loci (+MHC)	All coloc (%)	Gene coloc	PSI coloc	H3K4me1 coloc	H3K27ac coloc
AMD	41 (45) *	32 (33)	16 (50)	9	4	10	12
CAD	43 (44)	43 (44)	17 (40)	7	3	12	15
SLE	24 (25) *	24 (25)	13 (54)	6	2	6	9
AD	14 (15)	14 (15)	6 (43)	2	3	4	4
FEV1	7 (9) *	7 (9)	3 (43)	3	1	2	2
All	129 (138)	120 (126)	55 (46)	27	13	34	42

Table 2.2: Number of colocalised disease loci per feature type
Number of colocalised unique disease loci per feature type, per disease. The percentage of all features colocalised per disease is the ratio of colocalised loci over total common loci that were defined in the GWAS excluding the MHC region. For the AMD statistics, 7 variants and 1 locus with a MAF < 1% were also excluded and not counted in the table above. Each disease locus that colocalised with at least one feature across three cell types is counted as one colocalisation. For AMD, the percentage was calculated with respect to the disease loci (32) not the number of independent variants (41), where at some loci there were multiple independent genetic signals. Where some form of conditional analysis was performed and identified further signals this is designated by *.

Nearly half of the disease loci (27/55, 49%) colocalised with at least one eQTL in at least one cell type across all diseases. Of these 27 loci, eight loci colocalised with at least one eQTL in monocytes and no other gene effects in neutrophils or T cells (AD *MS4A6A,* AMD *TBC1D23*, AMD *CETP*, AMD *TNFRSF10A*, CAD *REST*, CAD *PPAP2B*, FEV_1 *TSEN54*, SLE *BANK1*) (Figure 2.8). Two loci, CAD *NT5C2* and FEV_1 *RP11-186N15.3*, colocalised with eQTL effects in neutrophils only i.e. no other colocalised eQTL was detected in either monocytes or T cells and similarly only two loci colocalised with T cell specific eQTLs (AMD *HIGD1AP14* and SLE *BLK*). Where gene effects were not unique to one cell type, in the majority of cases at least one colocalised gene was shared between two or three cell types (80%, 12/15).

In total, there were 13 disease loci that colocalised with at least one percent-splice-in (PSI) effect, which could include exon skipping or alternative donor or acceptor usages. A subset of loci colocalised with eQTLs and splicing effects in the same gene such as AD *MS4A6A*, AMD *PILRB*, CAD *NT5C2* and SLE *IRF7*, which highlighted cases where a disruption of alternative splicing could lead to a change in the overall level of gene expression. Alternatively, five loci colocalised with a splicing QTL but not an eQTL across all cell types. These included AD *CR1*, AMD *NPLOC4*, CAD *MIA3* and CAD *IL6R*, which all colocalised with exon skipping events in the corresponding genes as well as the SLE *FCGR2A* locus, which colocalised with exon skipping events in the *FCGR2A* gene in neutrophils and the *FCGR3A* gene in monocytes. Genetic control of splicing has previously been shown to be relevant to disease risk and also often independent of both gene expression and histone activity (Li et al., 2016c). Therefore, investigating splicing events as well as gene expression can highlight additional molecular mechanisms (Odhams et al., 2017, Li et al., 2016c, Chen et al., 2016a).

Next, I evaluated whether there existed any specific cell type patterns across the five diseases. In addition to assessing the number of disease loci that colocalised with *at least* one feature, I counted *all* colocalised gene and splicing QTL effects including where there were multiple features for one disease locus (Figure 2.7). For example, in this analysis, the AMD *TNFRSF10A* locus colocalised with eQTLs in monocytes for three genes: *TNFRSF10A; CHMP7* and *RP11-1149O23.3*, which were counted as three monocyte-specific features. I also limited my analysis to genes and splicing effects to avoid over-inflating the feature counts. In the BLUEPRINT cohort, overlapping histone modification signal peaks from multiple individuals were merged to create a unified peak list facilitating the identification of QTLs across all individuals (example in Figure 2.10) (Chen et al., 2016a). This could generate broad regions that could contain the signal of multiple correlated peaks. Elucidation of the exact putative enhancer region for the corresponding colocalised gene therefore required further molecular dissection (as discussed later in this chapter) (Figure 2.10-12, 2.17). To avoid over-estimating the importance of cell type by counting multiple correlated histone regions I assessed the gene and splicing effects for possible cell-type specific patterns.

18 monocyte features colocalised across all disease loci in total. Nine neutrophil- and nine T cell-derived features also colocalised across all loci (Figure 2.7). There was a higher degree of shared colocalised genes and/or splicing junctions between monocytes and neutrophils (seven) than between monocytes and T cells (three) or between neutrophils and T cells (one). This could be expected given that monocytes and neutrophils are derived from the common myeloid progenitor cells whereas CD4$^+$ T cells differentiate from the common lymphoid progenitor and deviate more in function (Figure 1.5 (Orkin and Zon, 2008)). Interestingly, there were also seven gene or splicing effects that occurred in all three cell types, suggesting these loci may have more ubiquitous effects (AD *EPHA1-AS1*, AMD *PILRB* and *MEPCE* at the same locus, AMD *RP11-644F5.10*, CAD *NT5C2*, CAD *GGCX* and

SLE *IRF7*). Indeed, I observed that the AD *EPHA1-AS1* (*EPHA1* antisense RNA 1) locus lead SNP was also a significant eQTL for this gene across multiple tissues assayed in the (G. TEx Consortium, 2015) including adipose, lung, spleen and whole blood. I observed the same broad tissue effect for the AMD *PILRB* locus, where the lead SNP was also a significant eQTL across more than 15 different tissues including coronary artery, brain, adipose and pancreas. Finally, the CAD *GGCX* locus, was also an eQTL in over ten tissues including aorta artery, stomach, pancreas and adipose. In summary, this approach demonstrated that across all five diseases, the most commonly colocalised features were monocyte-derived.

The majority of disease loci also colocalised with either H3K27ac, H3K4me1 QTLs or both (89%, 49/55). Across all three cell types, 42% (20/55) of loci that colocalised with gene or splicing QTLs also colocalised with a histone QTL from the corresponding cell type in at least one cell type (i.e. where a monocyte eQTL colocalised with either a monocyte H3K27ac or H3K4me1 QTL). In these cases, colocalised histone-bound regions may demarcate putative regulatory regions for the respective colocalised genes (Section 2.3.5 onwards for detailed examples). In addition, 43% (23/55) of disease loci colocalised with a histone QTL but not a gene or splicing QTL. These could represent poised enhancers indicating that the regulated genes are active in other cell types or that these regulatory QTLs affect processes beyond gene expression and splicing (Pai et al., 2015).

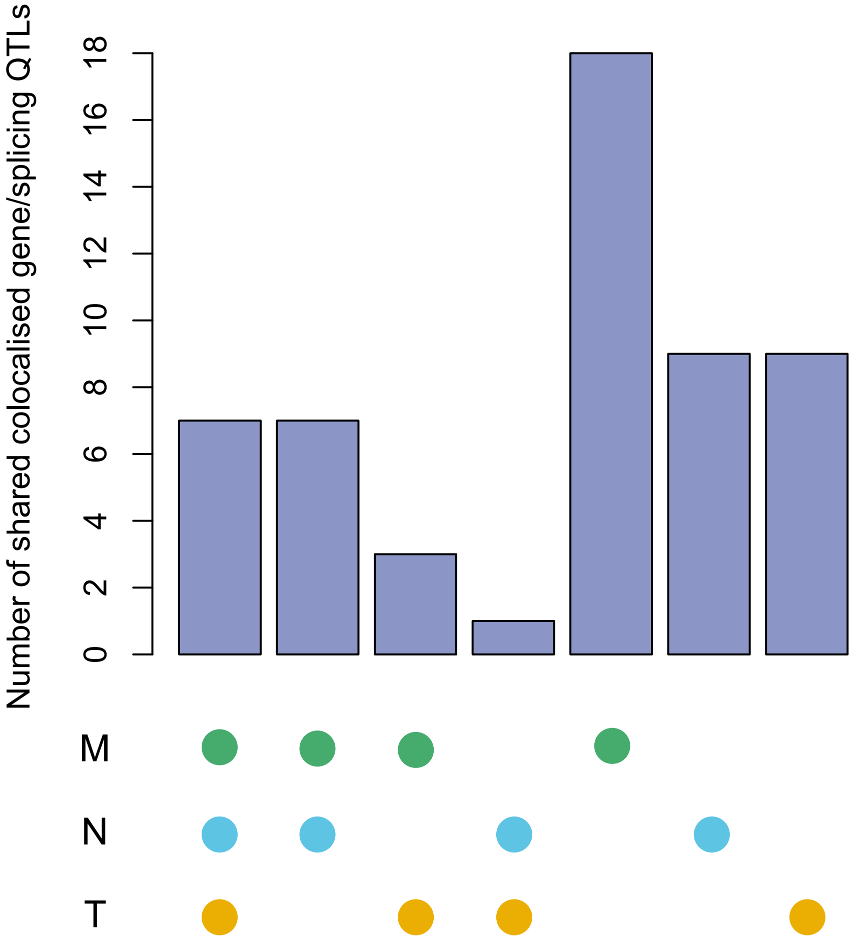


Figure 2.7: Colocalised gene and splicing QTLs per cell type and cell type combination across all diseases
All gene or splicing QTL features are counted for each cell type combination, for example the CAD *GGCX* locus colocalised with *GGCX* eQTLs across all three cell types so was counted as a shared MNT signal, but in the same locus, *VAMP8* eQTL colocalised in T cells only, so this was counted as a T cell (T) specific signal. The highest number of genes and splicing QTL features were in monocytes, demonstrating the potential importance of these cells across all diseases.

2.3.3 Colocalisation of molecular traits reveals potential molecular mechanisms at disease risk loci

I observed a high proportion of disease loci colocalised with molecular features but to form detailed mechanistic hypotheses for specific disease loci, integration of all colocalised features across cell types is required. Below, I discuss specific insights into the potential function of genetic risk loci.

Colocalisation with eQTLs offers the most intuitive interpretation of the molecular consequences at disease risk loci by identifying genes with altered expression levels. Figure 2.8 summarises the multiple molecular features I identified as colocalised with each disease locus. Figure 2.8 also describes the status of the previously predicted gene for the GWAS locus. There were several instances, 18 in total, where the colocalised gene matched previous target predictions for that locus, including AD *MS4A6A*, AMD *SRPK2*, *RDH5*, *CETP*, *CNN2*, *PILRB/A*, *TNFRSF10A*, CAD *NT5C2*, *LIPA*, *REST*, *GGCX*, *PPAP2B*, FEV_1 *TSEN54* and SLE *BANK1*, *IRF7*, *BLK*, *ITGAM*, *UBE2L3*.

All colocalised genes matching previous predictions at SLE loci had direct immune roles, which is expected for this prototypic autoimmune disease, but also in this case likely due to the use of publicly available eQTL data in the initial assignment of GWAS locus by the authors (Bentham et al., 2015). From my analysis, I identified colocalised eQTL genes, *BANK1* and *BLK,* which act in related B cell signalling pathways and are both regulated by type 1 interferons (Delgado-Vega et al., 2010). Here, these gene effects were T cell-mediated; a decrease in T cell *BLK* expression correlated with an increased disease risk, which was also observed in the original GWAS study. I also identified the gene *ITGAM*, where a neutrophil eQTL colocalised at this locus and decreased gene expression corresponded to decreased disease risk (rs9673398, EA = G, eQTL beta = -0.4896, SLE OR = 0.81). *ITGAM* encodes integrin alpha M chain, which forms the leukocyte-specific Mac-1 receptor shown to be important for neutrophil and monocyte-endothelial adherence and phagocytosis (Rebhan et al., 1998). I also identified other well-known immune genes such as *UBE2L3*, which encodes a ubiquitin conjugating enzyme E2 L3 involved in targeting proteins for degradation (Rebhan et al., 1998). Variants within the gene have also been associated with risk of Crohn’s disease, coeliac disease and rheumatoid arthritis (Fransen et al., 2010, Zhernakova et al., 2011). Here gene expression was positively correlated with SLE risk, confirming observations from the original GWAS (Bentham et al., 2015) Interestingly, the eQTL was more significant in neutrophils (eQTL lead, rs2298429, p value = 1.188×10^{-30}) than monocytes (eQTL lead rs5749485, p value = 2.901×10^{-04}) and the effect greater in magnitude (neutrophil, EA = G, beta = 1.267, SE = 0.110, monocyte, EA = C, beta = 0.448, SE = 0.124). Neutrophil eQTLs were not assessed as part of locus assignment from the Bentham *et al.* (2015) SLE GWAS, instead stimulated monocytes, B cells, CD14$^+$ monocytes, NK cells and CD4$^+$ T cells were included (Bentham et al., 2015). This suggests that neutrophils may be the effector cell for this disease risk locus, demonstrating the importance of assessing multiple cell types, particularly in this context as neutrophils have been shown to be important in the aetiology of lupus (Weidenbusch et al., 2017).

I also identified colocalised genes with immunological roles for the other traits I investigated. AD risk variants have been identified with within the *MS4A* cluster containing multiple genes encoding accessory proteins that amplify receptor function and regulate immune cell activation and survival (Proitsi et al., 2014). Significantly increased blood *MS4A6A* transcript levels have been associated with common coding SNPs in a cohort of approximately 300 AD patients, suggesting higher protein levels could contribute to the pathogenic pro-inflammatory AD phenotype (Proitsi et al., 2014). The authors found that *MS4A6A* was the only gene significantly differentially expressed, with higher expression in the patient cohort compared to the normal elderly controls. The *MS4A6A* expression effect was associated with SNP genotype but this effect was only significant in the patient group. By contrast, I identified that the AD *MS4A6A* locus colocalised with multiple genes in this locus including *MS4A6A, MS4A4A* and *MS4A4E* in monocytes and a further monocyte splicing QTL for *MS4A6A*. I observed the same positive correlation between gene expression and disease risk, where higher expression of *MS4A6A, MS4A4A* and *MS4A4E* corresponded to increased AD risk. My analysis identified *MS4A* effects in healthy individuals, which suggested in contrast to previous observations that the expression effect could contribute to the risk of AD, rather than reflect an expression effect that is perturbed in the disease state. There is evidence that the effect I identified and the previous effect could represent the same genetic signal, the LD between tested variants was moderate (r^2 0.54-0.61). Higher expression of these genes could, therefore, aid prediction of risk and also provide potential targets to investigate for therapeutic lowering of the expression levels.

For CAD, I confirmed the well-known *IL6R* locus, discussed in Section 2.1, colocalised with an exon skipping event in the *IL6R* gene generating higher levels of the transcript encoding soluble *IL6R*, thus providing clear validation of my analytical approach. Other CAD loci also highlighted the importance of immune activity, for example, I identified that the CAD *PPAP2B* colocalised with monocyte expression of this gene (rs56186267 EA = A, p value = 3.098 x 10^{-10}, beta = 0.885, SE = 0.141) as well as a H3K27ac marked region (p value = 1.884 x 10^{-07}, beta = 0.780, SE = 0.150) towards the 3' end of this gene. The *PPAP2B* intronic SNP, rs72664324, which is in high LD with the lead BLUEPRINT SNP (r^2 = 1, 1000G) has previously been identified as disrupting the binding of a C/EBPβ transcription factor (TF) within a macrophage LDL-induced dynamic open chromatin region (Reschen et al., 2015). Monocytes recruited *in vivo* to atherosclerotic plaques differentiate into macrophages, which are in turn stimulated to form foam cells by uptake of environmental lipids (Reschen et al., 2015). *PPAP2B* encodes the enzyme, LPP3, which deactivates pro-inflammatory mediators, such as those released from foam cells. Increased gene expression confers a protective CAD effect. Similarly, in the colocalised monocyte effects I identified, increased *PPAP2B* gene expression as well as increased signal of the H3K27ac modification (and therefore enhancer activity) corresponds to a decrease in CAD risk. Evidence also suggested that this was a monocyte-specific effect as *PPAP2B* expression was not tested in T cells due to low expression and in neutrophils, the association was not significant after correcting for multiple testing (rs72664324/rs56186267, p value = 1.975 x 10^{-01}). The classical $CD14^{+}CD16^{-}$ monocytes studied in the BLUEPRINT is known to differentiate into macrophages in inflamed tissues (Ohradanova-Repic et al., 2016). Therefore, the identified monocyte colocalisation at this locus suggests that enhancer activity and *PPAP2B* gene effects may be present in monocytes before stimulation and differentiation to macrophages, which in turn further supported the relevance of these tissues to CAD risk mechanisms.

A subset of AMD loci also colocalised with immune-related genes including the *TNFRSF10A* gene encoding the TRAIL 1 receptor and the *PILRB/A* genes encoding the paired immunoglobulin-like type 2 receptor beta and receptor alpha that regulate immune responses (www.genecards.org), (Rebhan et al., 1998). Similarly, I identified evidence for a possible immunological role for the AMD *CNN2* locus, for which colocalised genes were disparate across all three cell types in this study (Figure 2.8). The locus colocalised with the eQTL for *CNN2* in monocytes, for *CTB-31O20.2* in neutrophils and for *ABCA7* in T cells, a gene located approximately 1 kb downstream of *CNN2* on chr19. The transporter encoded by *ABCA7* is involved in pathogen-mediated phagocytosis, a process which requires actin skeleton reorganisation (Humphries et al., 2015). *CNN2* encodes calponin 2, a protein involved in the structural organisation of actin filaments, suggesting these two genes may have coordinated functions (Humphries et al., 2015). Genes in the *ABCA7* locus, including *CNN2*, have been shown to be involved late-onset Alzheimer’s disease aetiology (Humphries et al., 2015), although in this chapter the locus was evaluated in the context of AMD, this suggests a role for this cluster in age-related disorders. Therefore, although divergent colocalised genes across cell types may suggest differential functions in each tissue, a thorough assessment at each locus is required to evaluate whether the combined activity of genes may highlight important disease-relevant pathways.

In addition to immune function, I provided further support for the importance of lipid pathways in CAD and AMD. Important loci included the CAD *LIPA* locus that colocalised with monocyte and T cell expression of the *LIPA* gene and the AMD *CETP* locus colocalised with monocyte expression of the *CETP* gene, which encodes cholesteryl ester transfer protein. *LIPA* encodes lysosomal acid lipase A and the association of the intronic SNP, rs1412444 with increased *LIPA* gene expression levels in blood cell types and increased CAD risk is relatively well established (Zeller et al., 2010, G. TEx Consortium, 2015, Wild et al., 2011). I confirmed this direction of effect in monocytes but observed that in T cells, the effect was less significant and a decreased in *LIPA* expression corresponded to increased CAD risk (rs1412444 monocyte p value=2.044 x 10^{-49}, T cell p value= 4.588 x 10^{-06}). In neutrophils, rs1412444 was not significantly associated with *LIPA* expression (p value = 7.031 x 10^{-02}), suggesting that within the limits of the power of this cohort, the *LIPA* monocyte expression effect was most functionally relevant to CAD risk. Further supporting this is the colocalisation of this locus with a H3K27ac peak (10:90993615:91006217) and a H3K4me1 peak (10:90987967:91024823), which both directly overlap rs1412444. Together the activity of these histones could function as a putative enhancer for the *LIPA* gene as an increase in H3K27ac and H3K4me1 signal corresponds to an increased in *LIPA* expression and increased CAD risk.

I also identified colocalised genes with more general functions. The FEV_1 *TSEN54* locus colocalised with a monocyte eQTL for the *TSEN54* gene encoding a subunit of the tRNA splicing endonuclease complex (Figure 2.8). Despite a more general function, the colocalisation with monocyte expression appeared to reflect a relatively selective cell-type effect within this dataset. *TSEN54* was not tested in neutrophils due to low expression (median $\log_2$FPKM = 2.719, BLUEPRINT cohort). *TSEN54* gene expression was high in T cells and monocytes (median $\log_2$FPKM T cells = 9.433, median $\log_2$FPKM monocytes = 7.067) but the lead T cell eQTL (rs7225469 EA = C, beta = -1.197, SE = 0.176, p value = 9.348 x 10^{-12}) was not highly correlated with the lead FEV_1 variant, rs7218675 (r^2 = 0.26 1000G). These, therefore, may represent independent effects, providing both a hypothesis and a cellular model where decreased *TSEN54* expression in monocytes corresponds to an increase in FEV_1 ratio. The cell-type specific observation is an improvement on the initial observation that the GWAS locus was a significant *TSEN54* eQTL in the heterogeneous mix of whole blood (Wain et al., 2015).

I also identified a monocyte and T cell gene target, *RDH5*, that colocalised with the AMD locus where the encoded protein has a specialised role seemingly localised to the disease tissue (retina). *RDH5* encodes the 11 cis-retinol dehydrogenase enzyme, which catalyses the final step in the synthesis of the mammalian pigment chromophore, 11 cis-retinaldehyde (Liden et al., 2001). Multiple lines of evidence link *RDH5* disruption to impaired eye function; for example, the rare night blindness disorder fundus albipunctatus is caused by *RDH5* mutations and in AMD-like pathology and impaired dark adaptation is observed in the *RDH5* mouse model (MGI:1201412) (Blake et al., 2014, Fritsche et al., 2016).

Here, a decrease in *RDH5* expression in both monocytes (rs3138141 EA = A, p value = 5.512×10^{-12}, beta = -0.780, SE = 0.113) and T cells (rs3138141, p value= 1.931×10^{-10}, beta = -0.760, SE = 0.119) corresponds to an increase in AMD risk (rs3138141, beta = 0.15). *RDH5* expression is widely expressed beyond ocular tissues and the specialised retinal RPE-tissue (Wang et al., 1999). Interestingly, rs3138141 has shown to be a significant eQTL in over 10 different tissues from the (G. TEx Consortium, 2015) but I did not identify colocalisation of this locus in neutrophils. Instead a seemingly independent QTL was associated with neutrophil *RDH5* expression but not with AMD risk (rs142106092, *RDH5* p=4.536×10^{-06}, AMD p=0.147, rs3138141 $r^2 < 0.2$ 1000G). The functional significance of a specialised gene with a fairly ubiquitous effect remains unclear, but I did observe significant enrichment of AMD variants in monocyte eQTLs (Figure 2.6), providing evidence that among the ubiquitous expression, the monocyte-derived expression effect may be more likely to be disease-relevant. Certainly, there is a well-known role for Vitamin A, of which 11-cis-retinal is a derivative, in regulating the immune system and a further immune-link was demonstrated by reduced RPE *RDH5* expression in an *in vitro* system as a result of TNFα secretion from activated pro-inflammatory CD14$^+$ monocytes (Mora et al., 2008, Mathis et al., 2017). Further complicating mechanistic interpretation, this locus colocalised with a more significant gene expression effect in monocyte (p=9.234×10^{-34}), T cell (p=6.880×10^{-18}) and neutrophil (1.427×10^{-24}) of a second gene, *RP11-644F5.10*, for which there is no current characterised function. In addition, *RP11-644F5.10* and *RDH5* are directly overlapping, and rs3138141 is located within both genes.

The disease loci that colocalised with well-studied genes confirmed the validity of my analytical approach. I have also provided functional evidence for loci that may have previously been suggested without eQTL evidence and often based on genomic proximity to the lead SNP. Where eQTL data was used, utilising specific cell populations provides tractable and specific cellular models for functional follow up. Supplementary Figure 2.3 shows the regional association plots for all features colocalised at the loci described here and Supplementary Table 2.2 lists all of the colocalised features for all marks (the most significant colocalised features per disease locus is summarised in Figure 2.8).

	SNP	Locus	Mark	Feature	M	N	T
AD	rs10792832	*PICALM*	gene	*PICALM*	□	▨	□
			H3K27ac	11:85865916:85876777	■	■	□
			H3K27ac	11:85843685:85857459	▨	▨	■
			H3K4me1	11:85823185:85935960	□	■	▨
AD	rs10948363	*CD2AP*	gene	*CD2AP*	▨	▨	□
			H3K4me1	6:47512736:47517132	–	–	■
AD	rs11771145	*EPHA1*	gene	*EPHA1*	▨	□	▨
			gene	*EPHA1-AS1*	■	■	▨
			gene	*TAS2R41*	■	■	▨
			gene	*TAS2R62P*	–	■	▨
			gene	*TAS2R60*	■	■	▨
			H3K27ac	7:143133149:143136359	■	□	▨
			H3K4me1	7:143052447:143144656	■	■	▨
			psi	*EPHA1-AS1*	–	–	■
AD	rs6656401	*CR1*	gene	*CR1*	▨	□	□
			psi	*CR1*	■	■	□
AD	rs6733839	*BIN1*	gene	*BIN1*	▨	–	▨
			H3K27ac	2:127805979:127846262	▨	▨	■
AD	rs983392	*MS4A6A*	gene	*MS4A6A*	■	▨	▨
			gene	*MS4A4E*	■	–	–
			gene	*MS4A4A*	■	–	–
			H3K27ac	11:60072491:60079697	■	□	□
			H3K27ac	11:59932786:59958180	□	■	□
			H3K4me1	11:59867337:59870007	▨	■	–
			H3K4me1	11:60097581:60108825	■	□	□
			psi	*MS4A6A*	■	–	–
AMD	rs10033900	*CFI*	gene	*CFI*	–	–	–
			gene	*HIGD1AP14*	□	–	■
AMD	rs11080055	*VTN* *TMEM97*	gene	*VTN*	–	–	–
			gene	*TMEM97*	▨	–	▨
			gene	*SARM1*	□	■	▨
			gene	*TMEM199*	■	■	□
			H3K27ac	17:27592619:27623928	□	□	■
AMD	rs1142	*KMT2E* *SRPK2*	gene	*KMT2E*	□	□	□
			gene	*SRPK2*	■	■	▨
			H3K27ac	7:104840590:104849222	■	□	□
			H3K27ac	7:104982791:105001752	□	■	□
			H3K4me1	7:104817678:105045953	■	□	□
AMD	rs140647181	*COL8A1*	gene	*COL8A1*	–	–	–
			gene	*TBC1D23*	■	□	▨
			H3K27ac	3:99927877:99930243	–	–	■
			psi	*TBC1D23*	▨	■	▨
AMD	rs142450006	*MMP9*	gene	*MMP9*	▨	□	–
			H3K4me1	20:43711858:43736607	□	■	▨
AMD	rs1626340	*TGFBR1*	gene	*TGFBR1*	▨	▨	▨
			H3K27ac	9:101863436:101906508	□	■	□
			H3K4me1	9:102394208:102401582	□	□	■
			H3K4me1	9:101924908:101933600	□	■	▨
AMD	rs201459901	*C20orf85*	gene	*C20orf85*	–	–	–
			H3K27ac	20:55937467:55942133	■	□	□
AMD	rs3138141	*RDH5* *CD63*	gene	*RDH5*	■	▨	■
			gene	*CD63*	▨	□	□
			gene	*RP11-644F5.10*	■	■	■
			psi	unknown_5611300	–	■	–
			psi	unknown_5611300	–	■	–
AMD	rs5817082	*CETP*	gene	*CETP*	■	▨	□
			H3K27ac	16:56996497:57122386	■	▨	▨
			H3K4me1	16:56989796:57234898	■	▨	□
AMD	rs61985136	*RAD51B*	gene	*RAD51B*	□	▨	□
			H3K27ac	14:68744958:68766276	□	■	■
			H3K27ac	14:68807602:68809821	■	–	□
			H3K4me1	14:68786646:68813506	■	□	□
			H3K4me1	14:68705681:68768652	□	■	□
AMD	rs62247658	*ADAMTS9-AS2*	gene	*ADAMTS9-AS2*	–	–	–
			H3K27ac	3:64800574:64803236	–	■	–
			H3K4me1	3:64807275:64815846	■	▨	▨
AMD	rs6565597	*NPLOC4* *TSPAN10*	gene	*NPLOC4*	▨	▨	▨
			gene	*TSPAN10*	–	–	▨
			H3K27ac	17:79578768:79583302	▨	□	■
			H3K27ac	17:79585940:79590489	■	■	□
			H3K4me1	17:79594768:79608902	■	□	■
			psi	*NPLOC4*	■	–	■

M	N	T	
■	■	■	Colocalised
▨	▨	▨	Significant QTL
□	□	□	Non-significant QTL
–	–	–	Not tested

AMD	rs67538026	*CNN2*	gene	*CNN2*	●	◐	◐
			gene	*ABCA7*	◐	◐	●
			gene	*CTB-31O20.2*	○	●	○
			H3K27ac	19:1024681:1033920	●	○	●
AMD	rs72802342	*CTRB2* *CTRB1*	gene	*CTRB2*	▨	▨	▨
			gene	*CTRB1*	▨	▨	▨
			H3K27ac	16:75298651:75302321	▨	●	○
			H3K4me1	16:75230954:75238400	○	●	◐
			H3K4me1	16:75294112:75310094	●	◐	○
AMD	rs7803454	*PILRB* *PILRA*	gene	*PILRB*	●	◐	●
			gene	*PILRA*	○	○	●
			gene	*AC005071.2*	○	●	▨
			gene	*RP11-758P17.2*	○	◐	●
			gene	*ZCWPW1*	○	○	●
			gene	*MEPCE*	●	●	●
			gene	*STAG3*	●	●	◐
			H3K4me1	7:99906073:99912427	●	●	▨
			psi	unknown_9993580	●	●	●
			psi	*PILRB*	◐	◐	●
			psi	*PILRB*	●	●	▨
AMD	rs79037040	*TNFRSF10A*	gene	*TNFRSF10A*	●	◐	◐
			gene	*RP11-1149O23.3*	●	▨	○
			gene	*CHMP7*	●	◐	◐
			H3K27ac	8:23048166:23092260	●	◐	◐
			H3K4me1	8:22998146:23133613	●	◐	○
CAD	chr2:203828796:I	*WDR12*	gene	*WDR12*	○	◐	◐
			gene	*AC073410.1*	○	▨	●
			gene	*NBEAL1*	●	◐	◐
			gene	*ALS2CR8*	●	○	◐
			gene	*ICA1L*	●	▨	◐
			H3K27ac	2:204364327:204367436	◐	●	○
			H3K4me1	2:204391511:204403180	●	○	○
CAD	rs11065979	*SH2B3*	gene	*SH2B3*	○	◐	○
			H3K27ac	12:112386996:112391985	○	▨	●
CAD	rs11191416	*NT5C2* *CYP17A1* *CNNM2*	gene	*NT5C2*	○	●	◐
			gene	*CYP17A1*	▨	▨	▨
			gene	*CNNM2*	◐	○	○
			gene	*RP11-332O19.2*	○	●	○
			H3K27ac	10:104811999:104815290	○	○	●
			psi	*NT5C2*	●	▨	●
CAD	rs1412444	*LIPA*	gene	*LIPA*	●	◐	●
			H3K27ac	10:90248309:90252291	○	◐	●
			H3K27ac	10:90993615:91006217	●	●	○
			H3K4me1	10:90987967:91024823	●	●	○
CAD	rs17087335	*REST* *NOA1*	gene	*REST*	●	◐	◐
			gene	*NOA1*	◐	◐	◐
			H3K27ac	4:57823529:57826313	●	●	○
			H3K4me1	4:57820927:57828891	●	●	○
CAD	rs1870634	*CXCL12*	gene	*CXCL12*	▨	▨	▨
			H3K27ac	10:44339141:44344636	●	○	○
			H3K27ac	10:44499917:44501820	▨	●	▨
			H3K27ac	10:44468665:44477554	◐	◐	●
CAD	rs2487928	*KIAA1462*	gene	*KIAA1462*	▨	▨	▨
			H3K27ac	10:30314435:30318729	○	○	●
			H3K4me1	10:30286485:30293313	○	○	●
CAD	rs28451064	*KCNE2* *(gene desert)*	gene	*KCNE2*	▨	▨	▨
			H3K27ac	21:35594186:35597126	●	○	▨
			H3K27ac	21:35444064:35452944	◐	●	◐
			H3K27ac	21:35389093:35398453	○	◐	●
			H3K4me1	21:35592772:35599590	●	○	○
CAD	rs4468572	*ADAMTS7*	gene	*ADAMTS7*	▨	▨	▨
			H3K27ac	15:79049034:79056595	○	◐	●
			H3K4me1	15:79121511:79127959	●	●	○
			H3K4me1	15:79029778:79035218	◐	◐	●
CAD	rs56289821	*LDLR*	gene	*LDLR*	○	○	○
			H3K4me1	19:11105519:11214483	◐	●	○
CAD	rs6689306	*IL6R*	gene	*IL6R*	○	○	◐
			H3K27ac	1:154372031:154419908	◐	●	◐
			H3K4me1	1:154342399:154479953	○	●	○
			psi	*IL6R*	◐	◐	●

M	N	T	
●	●	●	Colocalised
◐	◐	◐	Significant QTL
○	○	○	Non-significant QTL
▨	▨	▨	Not tested

CAD	rs67180937	*MIA3*	gene	*MIA3*
			H3K4me1	1:222943024:222949002
			psi	*MIA3*
CAD	rs7212798	*BCAS3*	gene	*BCAS3*
			H3K27ac	17:58166053:58170841
CAD	rs7528419	*SORT1*	gene	*SORT1*
			gene	*PSRC1*
			H3K27ac	1:109109862:109115257
			H3K27ac	1:109812607:109818851
			H3K4me1	1:109779241:109861456
CAD	rs7568458	*VAMP8*	gene	*VAMP8*
		GGCX	gene	*GGCX*
		VAMP5	gene	*VAMP5*
			H3K27ac	2:85760296:85771243
			H3K4me1	2:85523177:85561159
CAD	rs9349379	*PHACTR1*	gene	*PHACTR1*
			H3K27ac	6:12961893:12964068
			H3K4me1	6:12953822:12975623
			H3K4me1	6:13023419:13036119
CAD	rs9970807	*PPAP2B*	gene	*PPAP2B*
			H3K27ac	1:56969801:56978117
			H3K27ac	1:56931078:56933449
FEV1	rs7218675	*TSEN54*	gene	*TSEN54*
FEV1	rs78420228; rs67863175	*CDC123*	gene	*CDC123*
			gene	RP11-186N15.3
			H3K27ac	10:12310277:12315701
			H3K4me1	10:12273289:12320006
SLE	rs10028805	*BANK1*	gene	*BANK1*
			H3K27ac	4:102711289:102714021
			H3K4me1	4:102752891:102758975
			H3K4me1	4:102739046:102740746
SLE	rs10488631	*IRF5*	gene	*IRF5*
			H3K27ac	7:128720370:128724585
			H3K27ac	7:128733525:128737997
SLE	rs10774625	*SH2B3*	gene	*SH2B3*
			H3K27ac	12:112386996:112391985
SLE	rs11644034	*IRF8*	gene	*IRF8*
			H3K4me1	16:85911735:86006346
SLE	rs11889341	*STAT4*	gene	*STAT4*
			H3K27ac	2:190816300:190818184
SLE	rs12802200	*IRF7*	gene	*IRF7*
			gene	*PHRF1*
			gene	*LRRC56*
			gene	*C11orf35*
			H3K27ac	11:600961:621989
			H3K4me1	11:601613:633623
			psi	*IRF7*
SLE	rs1801274	*FCGR2A*	gene	*FCGR2A*
			psi	*FCGR3A*
			psi	*FCGR2A*
SLE	rs2304256	*TYK2*	gene	*TYK2*
SLE	rs2663052	*WDFY4*	gene	*WDFY4*
			H3K27ac	10:49965615:49980674
SLE	rs2732549	*CD44*	gene	*CD44*
			H3K27ac	11:35087097:35089761
SLE	rs2736340	*BLK*	gene	*BLK*
			H3K27ac	8:11348864:11353299
			H3K4me1	8:11345910:11367069
			H3K4me1	8:11336396:11344630
SLE	rs34572943	*ITGAM*	gene	*ITGAM*
			gene	*C16orf58*
			gene	*RP11-388M20.2*
			gene	*RP11-347C12.10*
			H3K4me1	16:31355247:31421179
SLE	rs7444	*UBE2L3*	gene	*UBE2L3*
			H3K27ac	22:21938482:21985305
			H3K4me1	22:21917050:22033004
			H3K4me1	22:22399916:22404237

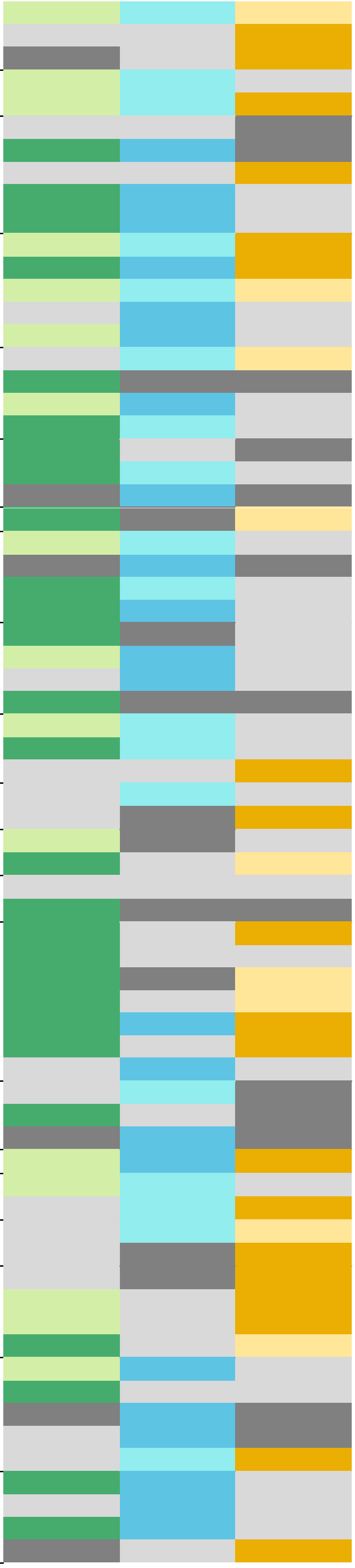


Figure 2.8 Disease loci colocalised with multiple molecular features
Summary of 54 disease loci colocalised with at least one feature. Where multiple features of the same type colocalised, the most significant feature per cell type is given. For each locus, the status of the previously reported locus is given. The most significant SNP per locus as denoted by the study is given. Colocalisation is the dark colour, non-colocalised significant QTLs, lighter colour, non-significant QTL by light grey and not tested, dark grey. The FEV$_1$ *KANSL1* inversion locus is excluded.

2.3.4 Therapeutic utility of colocalised gene targets

I have discussed how genetic studies of human cohorts provide the design of a randomised clinical trial without complications of intervention or reverse causation as genotype is randomly determined at birth (Finan et al., 2017). However, case-control GWAS are still limited in the identification of the exact mechanistic targets. Here, I have leveraged the advantage of GWAS data to identify risk-associated regions together with molecular traits to precisely detect potential putative target genes. I next evaluated whether any of the genes I identified through expression or splicing effects for all colocalised loci (excluding the *KANSL1* inversion locus) were known drug targets for the disease of interest or with other disorders highlighting the potential for drug repurposing (Finan et al., 2017). I collated information from both DrugBank (DrugBank, 2017) and the Open Targets (Open Targets, 2017) platform and identified 11 genes, which are targets for compounds or drugs that are currently under investigation, approved or experimental (Table 2.3) (Law et al., 2014, Koscielny et al., 2017). These included three CAD risk genes; *IL6R, GGCX, NT5C2,* three AMD risk genes; *RDH5*, *TNFRSF10A*, *CETP*, *SRPK2* and four SLE risk genes; *BLK*, *TYK* and the closely related and located *FCGR3A/FCGR2A* (Table 2.3).

I also queried my colocalised genes with a recently updated curation of the "druggable genome" resource from the Finan et al. (2017) study. This resource is composed of three tiers of genes predicted to encode druggable proteins including recent first-in-class drugs, biotherapeutics, drugs in late-phase development at the time of publication, preclinical small molecules with potential to bind proteins as reported in ChEMBL, secreted or plasma membrane-bound proteins that represent ideal targets for monoclonal antibodies and proteins that have greater than 50% identity with approved drug targets (Finan et al., 2017). Using this resource, I identified a further six druggable genes including the AMD *ABCA7* (small molecular or biotherapeutic), CAD *LIPA* (small molecule), CAD *VAMP8* (biotherapeutic), AMD *SRPK2* (small molecule), SLE *ITGAM* (biotherapeutic) and AD *CR1* (biotherapeutic).

In total, there was evidence of potential therapeutic utility for 17 genes identified in my analysis. In conclusion, such colocalisation approaches using molecular QTLs provides an additional layer of genetic and functional evidence for the selection of pre-clinical drug targets while also highlighting potentially affected cell types for further testing (Glinos et al., 2017, Okada et al., 2014).

Gene/Target	Disease	Drug	Target Mechanism	Status	Treated Disease	Accession /ChEMBL
NT5C2- Cytosolic purine 5'-nucleotidase (HGNC Acc:8022)	CAD	ATP, small molecule.	Unknown	-	Nutraceutical	DB00171
		Ribavirin, small molecule, guanosine nucleoside interferes with synthesis of viral mRNA,	Inducer	A	Hepatitis C, viral haemorrhagic fevers	DB00811
IL6R- Interleukin-6 receptor subunit alpha (HGNC Acc:6019)	CAD	Tocilizumab, antagonist, inhibits IL6R alpha subunit	Antibody	A	RA, SJIA, schizophrenia, temporal arteritis, AML, HIV, immune system disease	DB06273
		Sarilumab, antagonist, inhibits IL6R alpha subunit	Antibody	PIV	RA, ankylosing spondylitis, uveitis, immune system disease	DB11767
		SA237, antagonist, IL6Ralpha/GP130	Antibody	PIII	Neuromyelitis optica,	CHEMBL3833307
GGCX- Vitamin K dependent gamma-carboxylase (HGNC Acc:4247)	CAD	Phylloquinone/Vitamin K1 small molecule	Inducer	A	Haemorrhagic conditions	DB01022
		Anisindione, small molecule anticoagulant	Inhibitor	A	Venous thrombosis, embolism	DB01125
		Menadione/Vitamin K3, small molecule	Cofactor	A	Nutraceutical	DB00170
		Coagulation factor VIIa Recombination human promoting hemostasis	Unknown	A	Haemorrhagic complications	DB00036
		Drotrecogin alfa, recombinant activated human protein C	Unknown	A, I, W	Sepsis (withdrawn)	DB00055
		Coagulation Factor IX (Recombinant)	Unknown	A	Factor IX deficiency	DB00100
		Glutamic Acid	Unknown	A	Nutraceutical	DB00142
		Coagulation Factor IX Human, serine protease	Unknown	A	Factor IX deficiency	DB13152
TYK2- Non-receptor tyrosine-protein kinase (HGNC Acc:12440)	SLE	Tofacitinib, small molecule antagonist, inhibits janus kinases	Inhibitor	A, I	RA, immune system disease, UC, psoriasis, CD,	DB08895
		2-(1,1-DIMETHYLETHYL)9-FLUORO-3,6-DIHYDRO-7H-BENZ[H]-IMIDAZ[4,5-F]ISOQUINOLIN-7-ONE, small molecule	Unknown	E	-	DB04716

		Cerdulatinib, small molecule antagonist, tyrosine kinase inhibitor	Inhibitor	PI	Non-hodgkins lymphoma, chronic lymphocytic leukemia	CHEMBL3545284
		Peficitinib, small molecule antagonist, tyrosine kinase inhibitor	Inhibitor	PIII	RA, psoriasis, liver disease	CHEMBL3137308
FCGR3A- Fc fragment of IgG, low affinity IIIa, receptor/CD16a (HGNC Acc:3619) *FCGR2A*- Fc fragment of IgG, low affinity IIa, receptor/CD32 (HGNC Acc:3616)	SLE	Cetuximab, antibody binds EGFr, HER1, c-ErbB-1 and competitively inhibits binding of EGF	Unknown	A	EGFR-expressing metastatic colorectal carcinoma	DB00002
		Etanercept, protein binds to TNF	Unknown	A, I	RA, psoriasis	DB00005
		Immune Globulin Human, antibody mix binds and kills bacteria and viral particles	Antagonist	A, I	Immunodeficiencies	DB00028
		Adalimumab, human monoclonal binds and blocks TNF-alpha reducing inflammation	Unknown	A	RA, CD, psoriatic arthritis, ankylosing spondylitis	DB00051
		Abciximab, antibody binds to glycoprotein IIb/IIIa receptor and inhibits platelet aggregation	Unknown	A	Coronary intervention	DB00054
		Gemtuzumab oxogamicin, antibody binds and kills CD33 leukemic cells	Unknown	A, I, W	AML	DB00056
		Trastuzumba, antibody binds human epidermal GF receptor inhibits proliferation of tumour cells	Unknown	A, I	HER2 Breast cancer	DB00072
		Rituximab, antibody binds CD20 and kills B cells	Unknown	A	CD20^{+}non-hodgkins lymphoma, chronic lymphocytic leukaemia, RA	DB00073
		Basiliximab, immunosuppressive binds IL-2R alpha	Unknown	A, I	Prevent kidney transplant rejection	DB00074
		Muromonab, binds to and kills CD3^{+} cells	Unknown	A, I	Prevent organ rejection	DB00075
BLK- B lymphoid tyrosine kinase (HGNC Acc:1057)	SLE	Dasatinib, small molecule antagonist, SRC inhibitor	Inhibitor	A	Neoplasm, leukaemia, lymphoma	CHEMBL1421
		Ilorasertib, small molecule antagonist, SRC kinase inhibitor	Inhibitor	PII	Neoplasms	CHEMBL1980297
		ENMD-981693, small molecule antagonist, SRC inhibitor	Inhibitor	PII	Pancreatic carcinoma, breast cancer	CHEMBL52885
		XL-228, small molecular, SRC inhibitor	Inhibitor	PI	Lymphoma, leukemia	CHEMBL3545085

TNFRSF10A- Tumor necrosis factor receptor superfamily, member 10a (HGNC Acc:11904)	AMD	Dulanermin, protein, TNFRSF10A/B agonist	Agonist	PII	Non-Hodgkins lymphoma, lung/colorectal carcinoma,	CHEMBL2107846
		Mapatumumab, TNFRSF10A agonist	Agonist	PII	Non-Hodgkins lymphoma,myeloma, various carcinoma	CHEMBL2108621
CETP Cholesteryl ester transfer protein, plasma (HGNC Acc:1869)	AMD	Evacetrapib, small molecule	Inhibitor	PIII	CVD, lipid, hypercholesterolemia	CHEMBL2017179
		Anacetrapib, small molecule	Inhibitor	PIII	CHD, CVD, lipid, hypercholesterolemia	CHEMBL1800807
		Dalcetrapib, small molecule	Inhibitor	PIII	Acute coronary syndrome, CHD, CVD	CHEMBL313006
SRPK2- SRSF protein kinase 2	AMD	Adenine, small molecule	Unknown	A	Nutraceutical	DB00173
		Purvalanol, small molecule	Unknown	E	-	DB02733
		Phosphoaminophosphonic acid-adenylate ester, small molecule inhibits ATPase	Unknown	E	-	DB04395
RDH5- 11-cis retinol dehydrogenase (HGNC Acc:9940)	AMD	NADH, small molecule	-	-	Nutraceutical, possible PD, AD, CVD	DB00157
		Vitamin A, small molecule.	Unknown	A	Nutraceutical	DB00162

Table 2.3: Colocalised genes that are also known drug targets

Table summarises colocalised genes from expression and/or splicing effects that are known drug targets using DrugBank version 5.0.9, accessed August 2017 and Open Targets Platform, accessed October 2017 (Koscielny et al., 2017). The gene name was used to search the platforms for known targets. The drug, status, mechanism of action on target, diseases for which the drug is used, accession number (DrugBank, 2017, Law et al., 2014) or CHEMBL reference (Bento et al., 2014) are listed. Nutraceutical is a food source that provides health benefit. Drug status is listed as A = approved, W = withdrawn, I = investigational, E = experimental or maximum clinical trial either completed, ongoing or terminated (PI/II/III). The diseases for which the drug is currently used or investigated are also given or some example diseases where multiple conditions have been investigated. Disease abbreviations: PD = Parkinson’s disease, AD = Alzheimer’s disease, CVD = cardiovascular disease, RA = rheumatoid arthritis, SJIA = systemic juvenile idiopathic arthritis, CD = Crohn’s disease. SRC inhibitor = src family kinase inhibitor, CD = Corhn’s disease, UC = ulcerative colitis, AML = acute myeloid leukemia.

2.3.5 Complex regulatory mechanisms highlighted through integration of multiple molecular evidence

The combination of gene expression, splicing and histone QTLs enables not only a deeper description of disease risk mechanisms but is also a valuable tool in understanding how genes are regulated under homeostatic conditions. Above I summarised broad themes from my analysis, but a definitive description of the regulatory mechanism at each locus requires in-depth analysis and integration of multiple data sources. Here, I discuss approaches to distinguish likely putative regulatory mechanisms from multiple colocalised phenotypes involving complex histone activity, transcription factor binding and non-coding RNA function, demonstrating the importance of in-depth investigation at every disease locus.

2.3.5.1 Cell-type specific regulatory activity at the CAD *SORT1* locus

Not all loci colocalised with genes that matched previous predictions. One clear example was the CAD *SORT*1 locus, where I identified colocalisation with monocyte and neutrophil *PSRC1* gene expression as well as H3K27ac and H3K4me1 signal in these cell types (Figure 2.9). *SORT1*, a gene which encodes the multi-ligand sortilin receptor protein, has been previously identified as the liver target gene of the causal SNP, rs12740374 (Musunuru et al., 2010). There was also a significant *PSRC1* expression effect in human liver, but the largest observed effect was with the expression of *SORT1* (Musunuru et al., 2010). Increased hepatic *Sort1* expression in mice was also shown to modulate hepatic very-low density lipoprotein (VLDL) secretion resulting in reduced secretion of LDL-C therefore lowering LDL levels, which is known to decrease CAD risk (Musunuru et al., 2010). These effects were recently reproduced in iPSC-differentiated hepatocyte-like cells (HLCs) from 68 lines (Warren et al., 2017). Intracellular metabolites were extracted from these HLCs, and it was demonstrated that in minor allele-carrying individuals (rs12740374, T), there was a significant decrease in lipid metabolites such as triacylglycerol, diacylglycerol and aminoadipic acid, which has been associated with CAD (Warren et al., 2017). There is, therefore, clear evidence that in liver cells, *SORT1* hepatic expression and Sortilin protein levels were significantly associated with rs12740374 genotype and that this protein has a causal role in lipid regulation conferring protection (minor allele, T) to CAD risk. How sortilin exactly modifies lipid phenotypes is not yet clear and will require further experimental investigation (Kjolby et al., 2015).

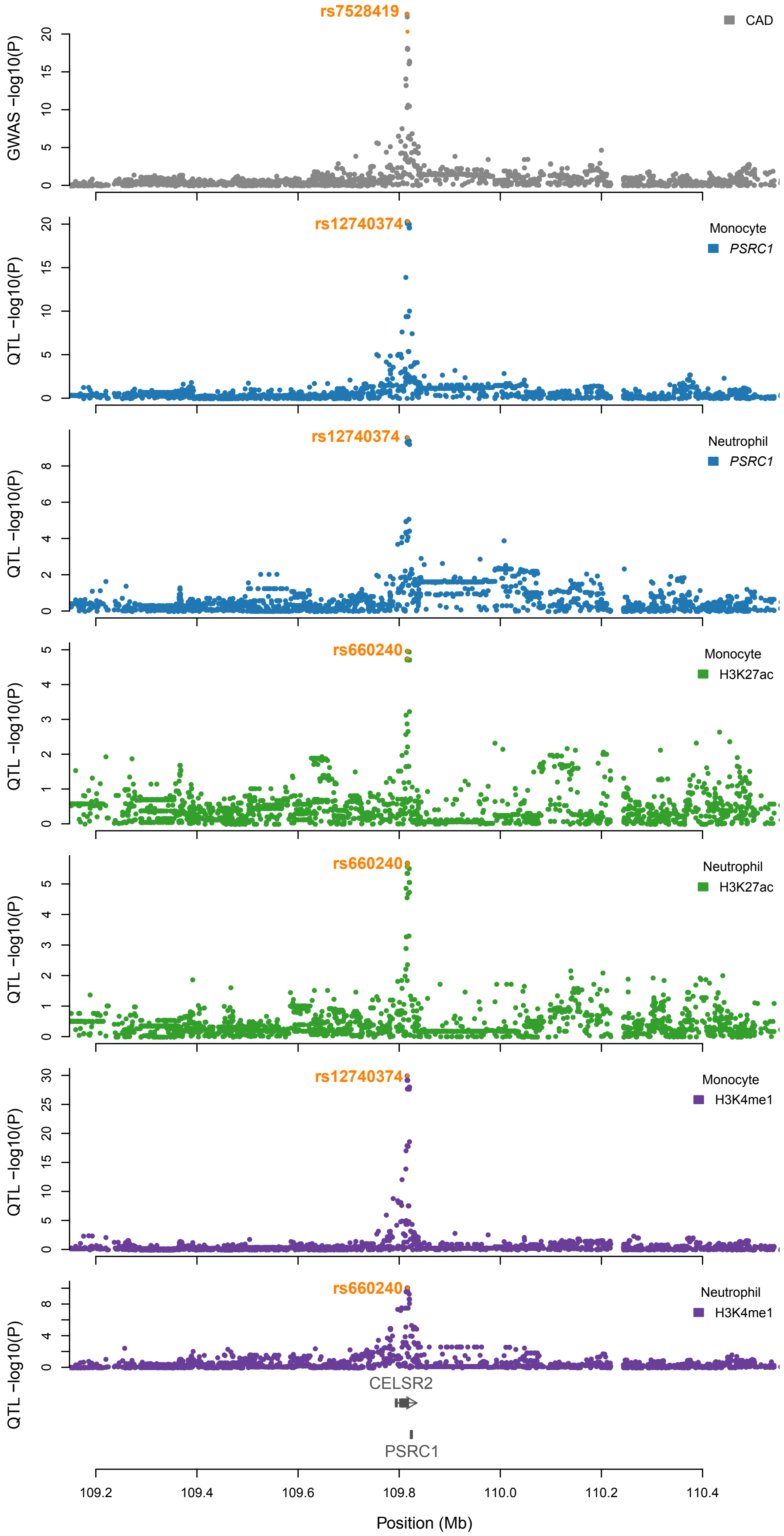


Figure 2.9: Regional association plot of the CAD *SORT1* locus
Locus zoom plots show association strength in -log10(p value) for variants that were shared between studies testing CAD (grey), monocyte gene expression (blue), monocyte H3K27ac (green) and monocyte H3K4me1, (purple). The respective lead SNPs for each feature are highlighted in orange.

In my analysis with the BLUEPRINT cohort, I identified that increased *PSRC1* expression (rs12740374 EA = T, monocyte p value = 4.364 x 10^{-21}, beta = 1.016 SE = 0.108, neutrophil p value = 2.572 x 10^{-10}, beta = 0.723, SE = 0.114) corresponds to a decreased risk of CAD (rs12740374 EA = T, p value = 4.63 x 10^{-23}, beta = -0.114, SE = 0.011). The predicted causal SNP, rs12740374 is located upstream of the *PSRC1* gene in the 3'UTR of the *CELSR2* gene (Figure 2.10). Compared to *SORT1*, very little is known regarding the function of *PSRC1*, which encodes a proline/serine-rich coiled coil protein 1 (Kjolby et al., 2015).

In monocytes and neutrophils, I identified that *SORT1* was expressed, but was not significantly associated with an eQTL. Both *PSRC1* and *SORT1* expression were low in T cells and below the threshold for association testing. The locus also colocalised with monocyte and neutrophil H3K4me1 and H3K27ac QTLs for peaks located just upstream of the *PSRC1* gene (Figure 2.10). The difference in modification level between individuals of discordant genotype was greater for H3K4me1 than H3K27ac (Figure 2.10) and the association more significant in both monocytes and neutrophils (Figure 2.9). Combined with the upstream location relative to the gene of these histone modifications, I postulated that this region acted as an enhancer for both *PSRC1* and *SORT1.* Genetic disruption of H3K4me1 could, in turn, alter downstream *PSRC1* gene expression in monocytes and neutrophils.

I investigated whether differences in regulatory function may explain the difference in the primary gene targets between haematopoietic and hepatic cell types, i.e. why the strongest effect was *PSRC1* in monocytes and neutrophils but as previously established, *SORT1* in liver (Musunuru et al., 2010). First, I used ENCODE ChIP-seq data from the hepatocyte cell line HepG2 to show that there was equivalent histone signal in the region overlapping rs12740374 as well as proximal to the *SORT1* promoter compared to both monocytes and neutrophils (Figure 2.10). Therefore, differences in enhancer activity seemed not explain the observation that *PSRC1* was significantly associated with rs12740374 and not *SORT1*. Instead, additional regulatory factors may generate cell-type genetic regulation of gene expression.

Binding of the liver-enriched transcription factor (TF), C/EBPα, to the motif containing the rs12740374 SNP was previously demonstrated in cultured human hepatocellular carcinoma cells (Hep3B) (Musunuru et al., 2010). Further, the major allele of rs12740374, G, disrupts a nucleotide of the C/EBP motif, G***TT***GCTC*AA*T, where *TT* and *AA* are the consensus nucleotides (Musunuru et al., 2010). The liver-enriched TF, C/EBPα, bound to the minor allele and directly affected *SORT1* expression levels in Hep3B cells (Musunuru et al., 2010). In addition to regulating many metabolic liver genes, C/EBPα is essential for granulopoiesis, regulating important genes such as the granulocyte-stimulating factor receptor (G-CSFR), but its expression decreases as maturation advances (Bardoel et al., 2014, Jakobsen et al., 2013). I hypothesised that binding of C/EBPα may be divergent between hepatocyte and haematopoietic cells as differential TF binding at regulatory regions is important for generating cell-type specific gene expression (Hardison and Taylor, 2012, Heinz et al., 2015). First, I used ChIP-seq data from the Soranzo team of C/EBPα in the monocytic cell line, U937 (unpublished data) to investigate regions of binding. I used two repeats of C/EBPα U937 ChIP-seq data and demonstrated that there was no equivalent C/EBPα binding directly over rs12740374 in U937 cell lines (peaks are represented by blocks as designated by a ChIP-seq peak caller) (Figure 2.10).

I further investigated whether different TFs may bind at the rs12740374 locus, instead of C/EBPα. C/EBPβ is a member of the same basic region leucine zipper-family (bZIP) as C/EBPα and as well as playing an important role in regulating chromatin dynamics and regrowth in liver, it is known to have an important role in haematopoietic cell differentiation and function (Jakobsen et al., 2013, Grontved et al., 2013, Bardoel et al., 2014). C/EBPα and C/EBPβ also bind to the same motif (Jakobsen et al., 2013). In addition, the TF, PU.1, is a crucial factor in promoting lymphomyeloid differentiation and acts as a pioneer factor binding to nucleosomes and preceding deposition of H3K4me1 (Heinz et al., 2010). Based on these observations, I postulated that this highly functional region could be bound by different combinations of lineage-specific master regulators in alternative cell types.

In order to establish whether these factors were bound, I generated C/EBPβ and PU.1 binding data using ChIP-seq in a differentiated HL60 cell line (Materials and Methods). HL60 is an immortalised cancer cell line established from a patient with acute myeloid leukaemia and is thought to resemble the granulocyte precursors, promyelocytes (Birnie, 1988). Using a well-established method, I differentiated HL60 to a more mature neutrophil-like stage by addition of all trans-retinoic acid (ATRA) or dimethyl sulfoxide (DMSO) (Materials and Methods). Using ChIP-seq, I then generated genome-wide binding data for transcription factors PU.1 and C/EBPβ as well as the histone modifications, H3K27ac and H3K4me1.

Figure 2.11 shows that the neutrophil H3K4me1 modification activity is recapitulated in differentiated HL60, confirming that enhancer activity is likely in this region. Further, there are strong binding peaks shown for C/EBPβ and PU.1 in both differentiated HL60 models. This confirms that master haematopoietic regulators are bound in this region in myeloid cells.

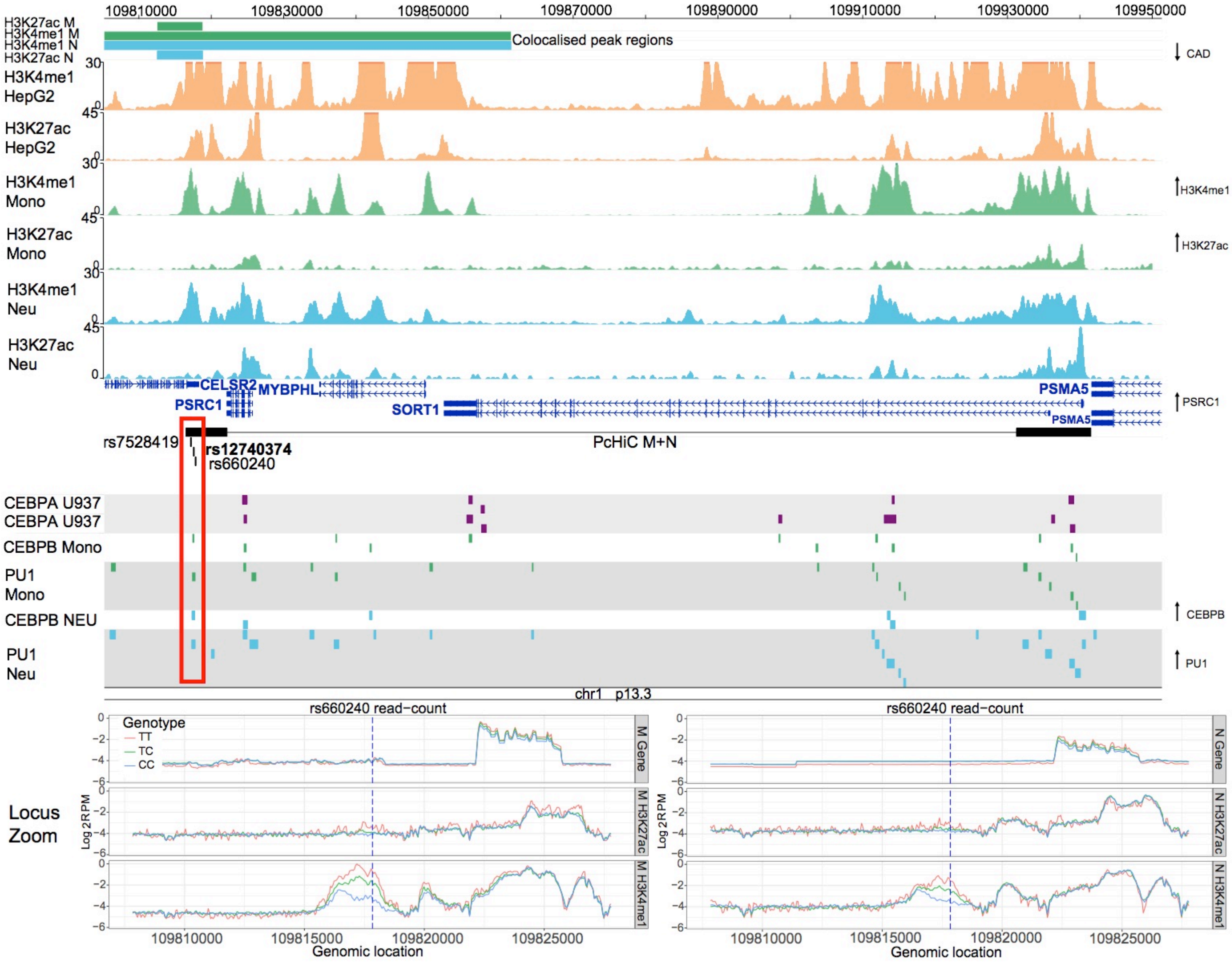


Figure 2.10: Genomic context of the CAD *SORT1* locus

Genome browser plot of the rs12740374 CAD locus with H3K4me1 and H3K27ac colocalised features shown for HepG2 (liver) cells, primary monocytes and neutrophils. Signal peaks are shown for representative individuals from the BLUEPRINT cohort and the x-axis for peaks is given in reads per million. The exact location of the predicted causal SNP, rs12740374, is highlighted in a red box which shows that the SNP intersects with histone marks as well as C/EBPβ and PU.1 in monocytes and neutrophils. C/EBPα binding in the monocyte-like cell line, U937, shows no binding in this region, suggesting differential TF binding at this cis-regulatory element underpins cell-type specific regulation of gene targets. The lower panel shows a zoom in on the region around the SNP upstream of the monocyte and neutrophil colocalised eQTL gene, *PSRC1*. Gene expression, H3K27ac and H3K4me1 signal is shown stratified by genotype. The signal is shown for the lead SNP for the H3K27ac, in this case, rs660240, which is in high LD with rs12740374. The genotype-associated difference in H3K4me1 signal is greater than the signal for H3K27ac, suggesting this is predominantly an enhancer effect due to changes in H3K4me1 activity. The directions of all features at this locus with respect to CAD risk are shown on the right-hand side.

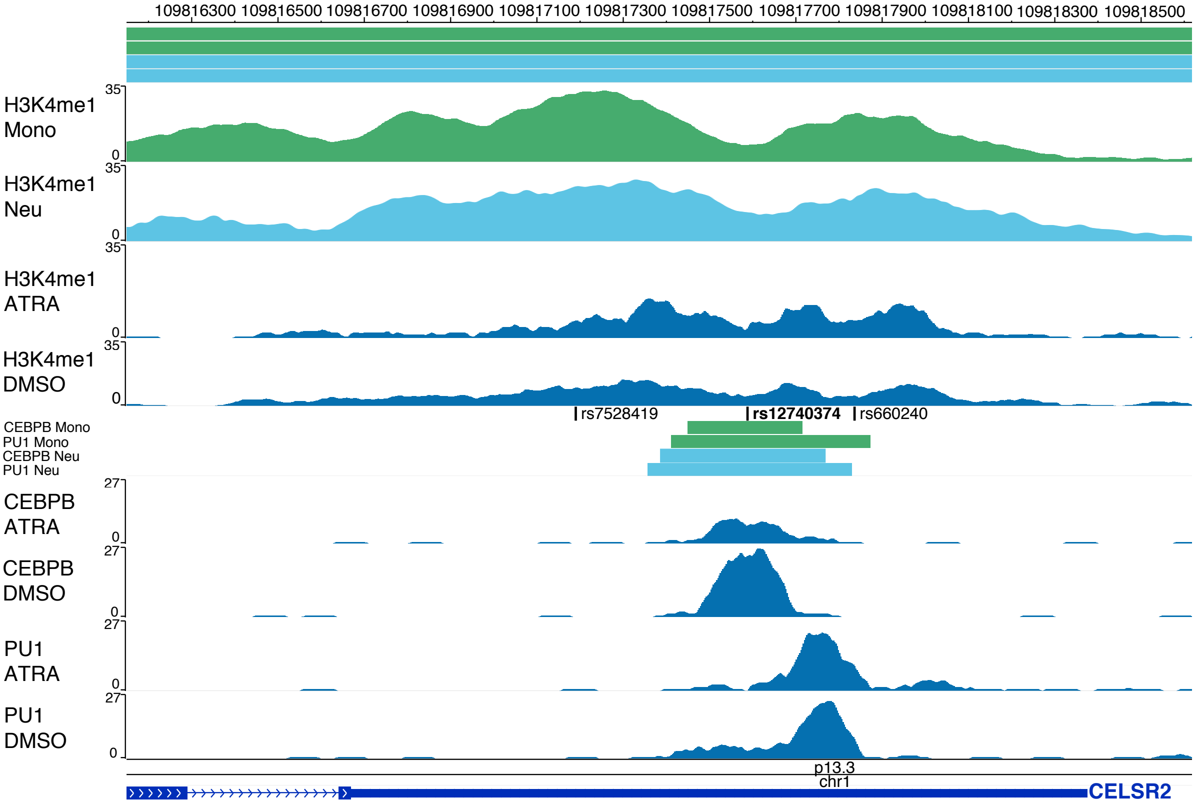


Figure 2.11 C/EBPβ and PU.1 are bound directly over rs12740374 in differentiated HL60 cells

Zoom in of the genomic region of rs12740374, which is located directly within C/EBPβ and PU.1 peaks from differentiated HL60 cells. H3K4me1 peaks for monocyte neutrophil and differentiated HL60 are shown to confirm that, at this locus, this model cell line recapitulates primary human cells. Neighbouring SNPs, rs7528419 (CAD GWAS lead) and rs660240 (certain histone peak lead) are located just outside or on the edge of the peak, showing how using TF data can aid in identifying the putative causal variant at disease loci.

Having demonstrated clear binding signal, I wanted to establish whether rs12740374 disrupts the haematopoietic binding of C/EBPβ and/or PU.1 as was previously shown for C/EBPα in Hep3B cells. In order to assess this, I accessed an unpublished dataset from the Soranzo team of C/EBPβ and PU.1 binding in primary human neutrophils (22 and 93 individuals respectively) and monocytes (nine and ten individuals respectively) (Stephen Watt, manuscript in preparation). Figure 2.12 highlights that in C/EBPβ and PU.1 are also bound directly over rs12740374 in monocytes. The binding of these TFs appeared lower in primary neutrophils than in monocytes and the differentiated HL60 model. This was likely due to the increased technical difficulties associated with applying these approaches in primary neutrophils, as we have observed within our team. This further demonstrates the importance of confirming TF binding using the more tractable, differentiated HL60 model.

Despite this lower level of binding, the higher number of individuals discordant at the rs12740374 genotype in the primary neutrophil cohort enabled me to investigate whether binding was associated with SNP genotype. Figure 2.12 shows primary monocyte and neutrophil binding of C/EBPβ and PU1 around the *PSRC1* locus and also binding of specific peaks stratified by the genotype of this SNP. Both C/EBPβ and PU.1 peaks directly overlapping rs12740374 are significantly associated with genotype as evaluated using linear regression (C/EBPβ, p value = 7.351 x 10^{-04}, PU.1 p value = 1.584 x 10^{-06}). I confirmed that no other immediate surrounding peaks are significantly associated with rs12740374 (Figure 2.12 and data not shown). This evidence suggested that in neutrophils, the major allele of rs12740374 may disrupt binding of PU.1 and C/EBPβ as well as H3K4me1 activity, which could result in disruption of *PSRC1* expression. I also demonstrated that binding of C/EBPβ and PU.1 occurs over rs12740374 in monocytes by showing representative binding of an individual heterozygous for rs12740374 (Figure 2.12). Although there was a limited number of individuals for which monocyte data was available, the concordance of the other molecular effects between monocytes and neutrophils suggests that C/EBPβ and PU.1 binding could also be disrupted in monocytes. However, more individuals would be required to fully validate this effect.

Interestingly, using publicly available promoter-capture HiC (Schofield et al., 2016, Javierre et al., 2016) data, I observed that a significant neutrophil and monocyte chromatin interaction fragment links rs12740374 to the promoter of the *SORT1* gene (Figure 2.10). Despite this physical connection, *SORT1* expression is not significantly associated with this SNP in monocytes or neutrophils. Therefore, this demonstrates that at some loci, the combination of TF bound is an important driving factor over chromatin interactions and enhancer activity in generating cell-type specific gene regulatory mechanisms, although further functional experiments would be required to fully ascertain this potential hierarchical regulation.

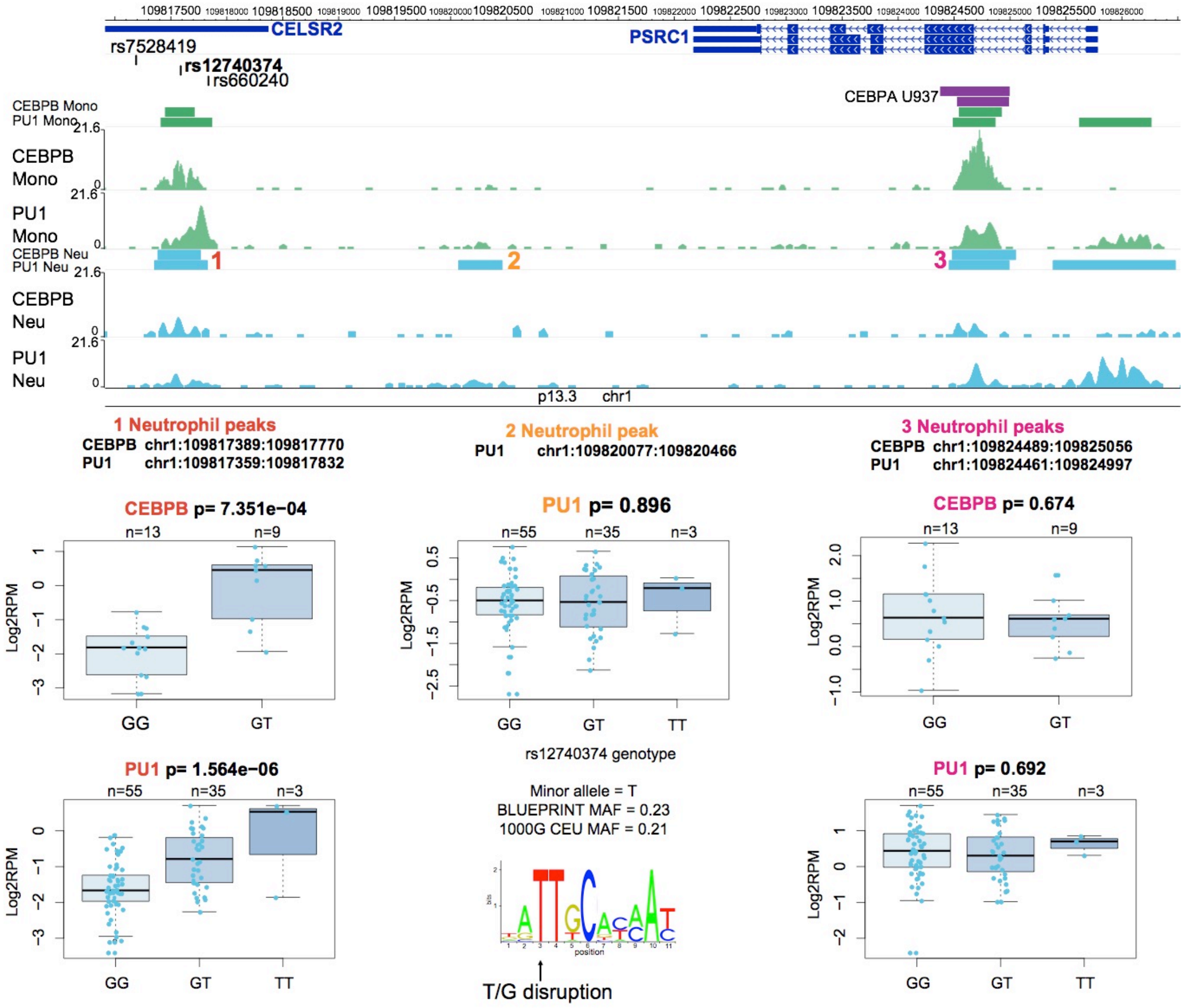


Figure 2.12: Transcription factor binding at the CAD *SORT1/PSRC1* locus in monocytes and neutrophils

Genomic region of *PSRC1* and predicted causal SNP, rs12740374 showing binding of C/EBPβ and PU.1 in monocytes and neutrophils (upper panel). In the lower panel, boxplots show the binding signal in $\log_2$RPM of neutrophil TFs at three peaks in the locus stratified by rs12740374 genotype. The p value is shown for the association with genotype as calculated using linear regression on standardised inverse normalised binding values in $\log_2$RPM. The only peaks significantly associated with rs12740374 are the C/EBPβ and PU.1 that are bound directly over the SNP. The consensus C/EBPβ motif is also shown with the position of the nucleotide disrupted by rs12740374, where the minor allele T creates the binding site and the major allele, G disrupts the binding site.

2.3.5.2 In-depth dissection of the molecular mechanisms at the AMD *TNFRSF10A* disease locus

I identified colocalisation between the *TNFRSF10A* advanced age-related macular degeneration locus and three monocyte eQTLs; *TNFRSF10A, RP11-1149O23.3* and *CHMP7.* In addition, this locus colocalised with three histone peaks; H3K27ac (8:23048166:23092260), H3K27ac (8.23092704.23132254) and H3K4me1 (8:22998146:23133613) (Figure 2.14). I discuss here approaches to resolve this complex locus and provide paradigms for future efforts to identify mechanisms that influence disease risk in a cell type-specific manner.

TNRFSF10A encodes the TRAILR1 receptor that binds the tumour necrosis factor-related apoptosis-inducing ligand (TRAIL) (Diehl et al., 2004). TRAIL can bind four possible receptors: TNFRSF10A, TNFRSF10B (TRAIL-R2), TNFRSF10C (TRAIL-R3) and TNFRSF10D (TRAIL-R4) (Diehl et al., 2004). TRAIL-R1 and TRAIL-R2 are functional proteins that include an intracellular tail containing the death domain (Figure 2.14) (Diehl et al., 2004). TRAIL-R3, a GPI-linked protein and TRAIL-R4, a truncated protein that misses the death domain in the cytoplasmic tail, are both decoy receptors that do not activate TRAIL-mediated apoptosis and can antagonise TRAILR1-2 signalling (Diehl et al., 2004, Guicciardi and Gores, 2009). Functional TRAILRs activate apoptosis in tumour cells, and were originally not thought to induce cell death in non-transformed cells (Diehl et al., 2004, Liguori et al., 2016). Recently, however, TRAIL susceptibility leading to caspase-8-dependent apoptosis was observed in primary mononuclear phagocytes, where the expression of functional *TNFRSF10A/TRAILR1* was highest compared to the expression on neutrophils and T lymphocytes (Liguori et al., 2016). No caspase-8 activation was observed in neutrophils or lymphocytes (Liguori et al., 2016). Macrophages may be more resistant to death signals as they represent a more activated immune cell than monocytes (Liguori et al., 2016). TRAILR2 seems to have a more important role in stimulating apoptosis than TRAILR1 (Guicciardi and Gores, 2009).

Up-regulation of the *TNFRSF10A/TRAILR1* receptor has been shown to be associated with anti-inflammatory signals such as stimulation by the cytokine IL-10 (Liguori et al., 2016). In *TRAIL-/-* mice, cytokine production from macrophages and dendritic cells was increased and these mice had increased susceptibility to certain immune disorders such as autoimmune arthritis and diabetes (Diehl et al., 2004, Falschlehner et al., 2009). In the MS mouse model, EAE, blocking TRAIL resulted in increased CNS inflammation (Falschlehner et al., 2009). Therefore, in normal immune cells, the TRAIL system seems to exert a regulatory and suppressive role in the functioning of the immune response.

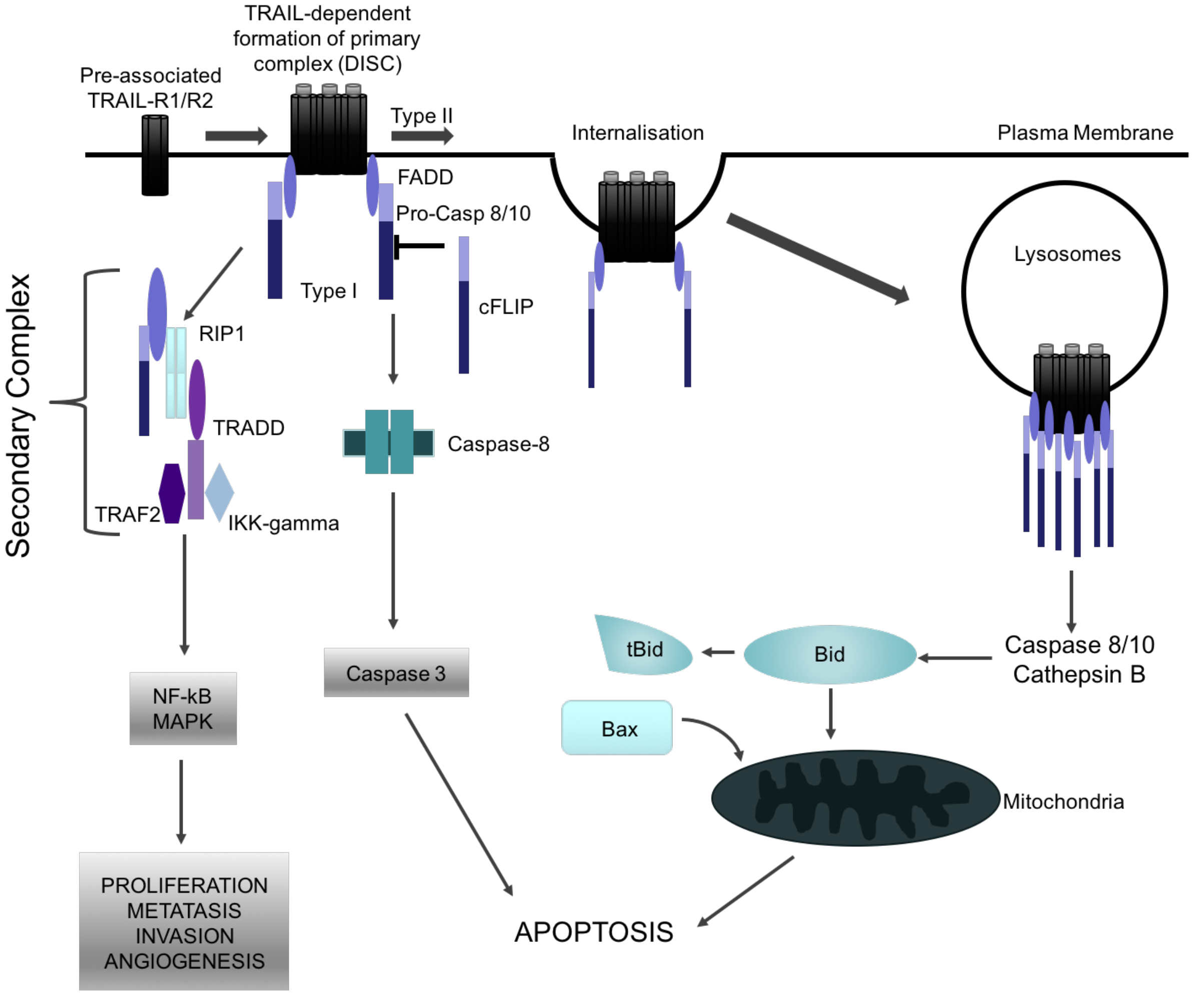


Figure 2.13: TRAILR1/2 signalling pathways
The different pathways stimulated by TRAIL binding to TRAILR1/2, which are often both expressed on the same cell. TRAILR1/2 do not require internalisation for stimulation of apoptosis in type 1 cells, but is essential in hepatocytes, as an example of type 2 cells. In addition to inducing cell death, TRAIL promotes activation of pro-survival mediators such as NF-kB and MAP kinases through a distinct pathway as shown above left. Activation of NF-kB cannot overcome TRAIL-mediated apoptosis in all cell types. Slight differences occur between TRAILR1 and TRAILR2 signalling at the TRAF2 level. TRAILR1 instead activates JNK/SAPK (stress-activated protein kinase) via a TRAF2-MKK4 (mitogen-activated protein/ERK kinase 4)-dependent pathway. Caspases are apoptosis activators. Bid, the truncated Bid (tBid) and Bax are all apoptotic proteins. cFLIP is a caspase 8-like inhibitory protein. RIP1 is the receptor-interacting protein 1, which is a death domain-containing serine/threonine kinase crucial in the balance between death and survival signalling, binds to all death receptors and can stimulate either a death cascade or a survival signal, in this case NF-kB activation by RIP1 activates survival pathways. TRADD is the TNF receptor-associated protein with death domain and acts as an adaptor protein. This figure and associated details described here were adapted from (Guicciardi and Gores, 2009).

I identified that decreased histone modification signal corresponds to decreased expression of both *TNFRSF10A* and *RP11-1149O23.3* genes, which in turn corresponds to an increased AMD risk (Table 2.4 and 2.5). Given the evidence that *TNFRSF10A* functions in monocytes to negatively regulate immune responses, a decrease in expression could result in an increased inflammatory state that over a prolonged period could add to increased risk of AMD.

There were two H3K27ac peaks that colocalised with this disease locus: 8:23048166:23092260 that directly overlapped the SNP and 8:23092704:23132254, located downstream. The peak directly overlapping the SNP was more significantly associated (p value = 3.941 x 10^{-45}, beta = -1.200, SE = 0.085, Table 2.5) than the downstream peak (p value = 1.647 x 10^{-09}, beta = -0.638, SE = 0.106). Therefore, based on location and strength of association, I postulated that the overlapping peak contained a putative regulatory element. Indeed, in previous molecular QTL studies, for example with DNase I hypersensitive (open chromatin) regions, it has been observed that most significant QTLs lie close to the DHS peak and proximal region (target window), specifically 56% of dsQTLs are located within the associated DHS and 67% are within a window of 1 kb around the feature (Degner et al., 2012). In addition, molecular strength of association is known to decay with increasing distance from the SNP (Waszak et al., 2015). I also excluded the colocalisation with the eQTL of the *CHMP7* gene from further analysis as the p values of association was also much less significant than the others (*CHMP7* beta = -0.473, SE = 0.099, value = 1.880 x 10^{-06}, H3K27ac beta = -0.638, SE = 0.1058, p value = 1.647 x 10^{-09}). Figure 2.14 and Table 2.4 and Table 2.5 summarise the association statistics of the four AMD-colocalised features that I investigated further; the two genes *TNFRSF10A* and *RP11-1149O23.3* as well as the H3K4me1 peak (8:22998146:23133613) and the single H3K27ac peak (8:23048166:23092260).

For all monocyte colocalised molecular features, the lead SNP was rs13255394, a common SNP (EAF = 0.575) located just downstream of the *TNFRSF10A* gene start site (Figure 2.17, Table 2.4). I next evaluated whether the lack of colocalisation with either neutrophil or T cell features represented a true cell-type specific disease effect, or whether neutrophil or T cells effects are missed due to a limitation in power.

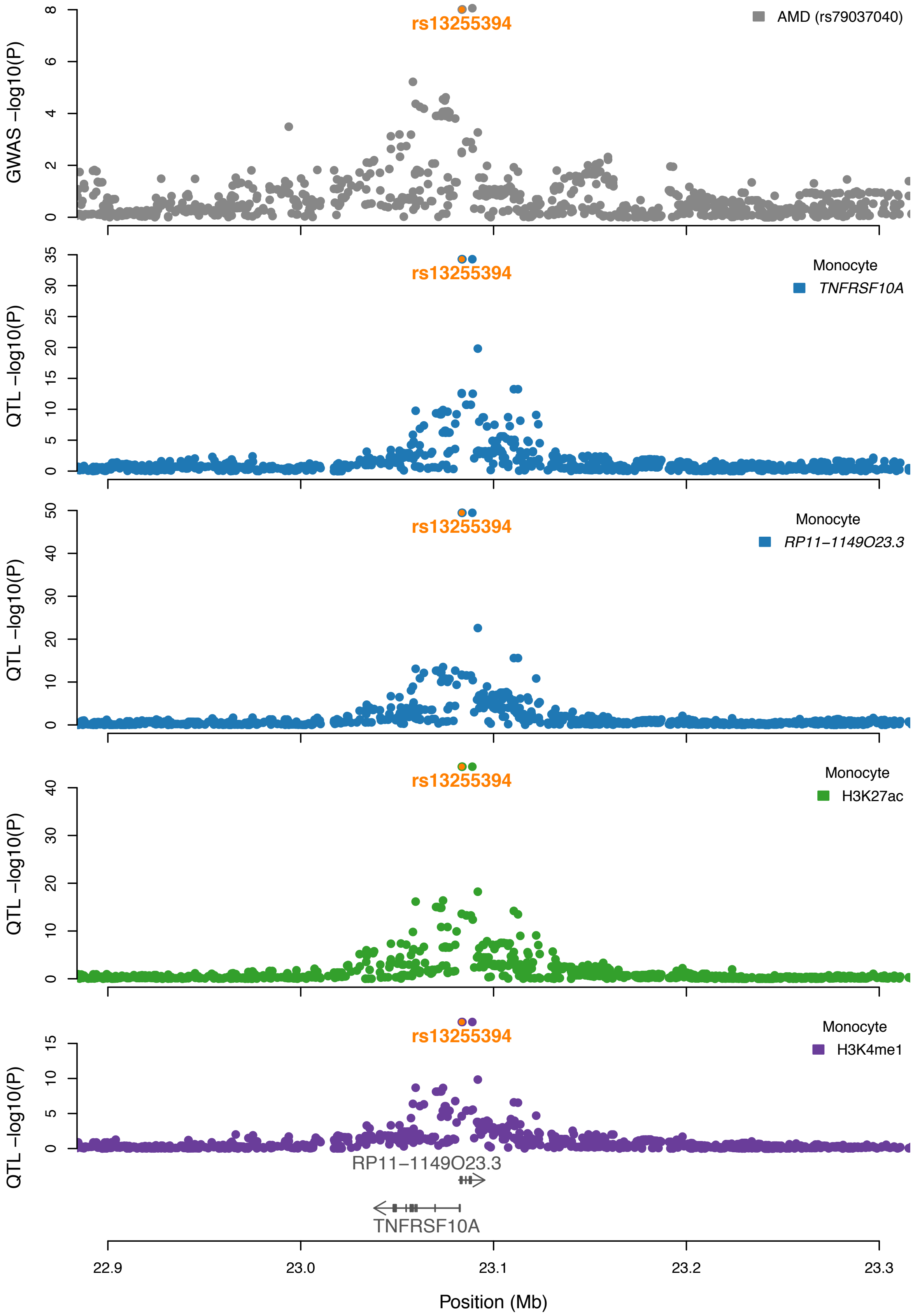


Figure 2.14 Regional association plots for the *TNFRSF10A* locus

Locus zoom plots show association strength in -log10 p value for variants that were shared between studies testing AMD (grey), monocyte gene expression (blue), monocyte H3K27ac, 8:23048166:23092260 (green) and monocyte H3K4me1, 8:22998146:23133613) (purple). The index disease SNP defined for the locus by Fritsche *et al* is rs79037040, but was not tested as part of the Blueprint study. The index molecular SNP, rs13255394 is labelled in orange and shown with respect to the genomic location, within exon 1 of *RP11-1149O23.*

Trait	SNP	R^2	EA/OA	EAF	AMD beta	AMD P	Cell type	Beta (SE)	P
TNFRSF10A	rs13255394 (M lead)	-	C/T	0.575	+	9.92 x 10^{-09}	Monocyte	-1.047 (0.085)	**5.249 x 10^{-35}**
							T cell	0.299 (0.110)	6.493 x 10^{-03}
							Neutrophil	-0.376 (0.096)	1.356 x 10^{-04}
	rs7820465 (T lead)	0.141	A/G	0.23	-	8.64 x 10^{-05}	Monocyte	0.585 (0.117)	5.892 x 10^{-07}
							T cell	-1.164 (0.108)	**3.279 x 10^{-27}**
							Neutrophil	0.536 (0.117)	4.999 x 10^{-06}
	rs4872078 (N lead)	0.005	T/G	0.47	-	1.950 x 10^{-02}	Monocyte	0.132 (0.098)	1.761 x 10^{-01}
							T cell	-0.813 (0.094)	4.833 x 10^{-18}
							Neutrophil	0.841 (0.087)	**6.145 x 10^{-22}**
RP11-1149O23.3	rs13255394 (M lead)	-	C/T	0.575	+	9.92 x 10^{-09}	Monocyte	-1.171 (0.079)	**3.477 x 10^{-50}**
							T cell	-0.295 (0.109)	6.734 x 10^{-03}
							Neutrophil	Not tested	Not tested

Table 2.4: Summary statistics of lead SNPs with gene expression of *TNFRSF10A* and *RP11-1149O23.3* expression, H3K27ac and H3K4me1 modification phenotypes in monocytes, neutrophils and T cells
Association statistics for the cell-specific lead SNPs for BLUEPRINT traits (Chen et al., 2016a). In bold are highlighted the lead associations in that cell type. AMD beta values and standard error estimates can be obtained by application to the IAMDGC consortium.

Trait	SNP	R^2	EA/OA	EAF	AMD beta	AMD P	Cell type	Beta (SE)	P
H3K27ac	rs13255394 (M lead)	-	C/T	0.575	+	9.92×10^{-09}	Monocyte	-1.200 (0.085)	**3.941×10^{-45}**
							T cell	-0.027 (0.120)	8.201×10^{-01}
							Neutrophil	-0.873 (0.096)	6.868×10^{-20}
	rs13255997 (T lead)	NT	G/A	0.510	+	8.540×10^{-03}	Monocyte	-0.261 (0.105)	1.301×10^{-02}
							T cell	0.508 (0.111)	4.669×10^{-06}
							Neutrophil	-0.088 (0.103)	3.944×10^{-01}
	rs4872090 (N lead)	0.402	A/T	0.763	+	1.240×10^{-03}	Monocyte	-0.877 (0.117)	5.407×10^{-14}
							T cell	-0.239 (0.130)	6.472×10^{-02}
							Neutrophil	-1.107 (0.105)	**4.599×10^{-26}**
H3K4me1	rs13255394 (M lead)	-	C/T	0.575	+	9.92×10^{-09}	Monocyte	-0.858 (0.097)	8.652×10^{-19}
							T cell	-0.054 (0.140)	6.984×10^{-01}
							Neutrophil	-0.603 (0.102)	2.944×10^{-09}
	rs8192332 (T lead)	NT	T/C	0.288	+	6.930×10^{-02}	Monocyte	-0.107 (0.125)	3.936×10^{-01}
							T cell	0.651 (0.172)	1.592×10^{-04}
							Neutrophil	-0.064 (0.125)	6.091×10^{-01}
	rs4872090 (N lead)	0.402	A/T	0.763	+	1.240×10^{-03}	Monocyte	-0.546 (0.119)	3.833×10^{-06}
							T cell	-0.188 (0.151)	2.142×10^{-01}
							Neutrophil	-1.083 (0.108)	**1.583×10^{-23}**

Table 2.5: Summary statistics of lead SNPs for H3K27ac with H3K4me1 modification phenotypes in monocytes, neutrophils and T cells

Figures 2.8 and 2.15 show there was a significant eQTL for *TNFRSF10A* in both T cells and neutrophils, but the lead SNPs associated with these signals are different and did not colocalise with AMD at this locus (summarised in Table 2.4). In T cells, the lead SNP was rs7820465 (EA = A, EAF = 0.23, beta = -1.164, SE = 0.108, p value = 3.279 x 10^{-27}, number of individuals = 169). In neutrophils, the lead *TNFRSF10A* eQTL was rs4872078 (EA = T, EAF = 0.47, beta = 0.841, SE = 0.087, p value = 6.145 x 10^{-22}, number of individuals = 196). Neither of these SNPs were significantly associated with AMD (rs7820465 p value = 8.64 x 10^{-05}, rs4872078 p value = 0.020).

I performed conditional analysis using GCTA and the eQTL summary statistics in each cell type (Materials and Methods). I tested for association of remaining SNPs after conditioning on the corresponding lead SNP for each cell type (Table 2.4 and Figure 2.15). In order to evaluate if any residual significant signals remained, I corrected the p value for the number of variants tested in the cis-region using the qvalue R package (Bass JDSwcfAJ, 2015) (Materials and Methods). There were no significant associations after conditioning on the respective lead SNP in monocytes or neutrophils (qvalue < 5%), which was evidence that within the power limitations of the cohort, there was one independent genetic signal in this region driven by the respective lead SNPs. In T cells, after conditioning on the lead SNP, rs7820465, there remained a marginally significant signal driven by the neutrophil lead SNP, rs4872078 (conditional beta = -0.394, conditional SE = 0.098, conditional p value = 5.946 x 10^{-05}, conditional q value = 0.046). I performed an iterative second stage of conditional analysis, using the output summary statistics generated by conditioning on rs7820465. In this second stage, I conditioned on rs4872078 and found no significant associations remained. None of these cell type lead SNPs were highly correlated; I calculated the LD r^2 estimates using the UKBB cohort of nearly 175,000 individuals and found an r^2 of less than 0.2 for each pairwise comparison (Table 2.4). Therefore, the association evidence suggests that expression of *TNFRSF10A* is regulated by varying independent signals across cell types and only the monocyte signal colocalised with AMD risk.

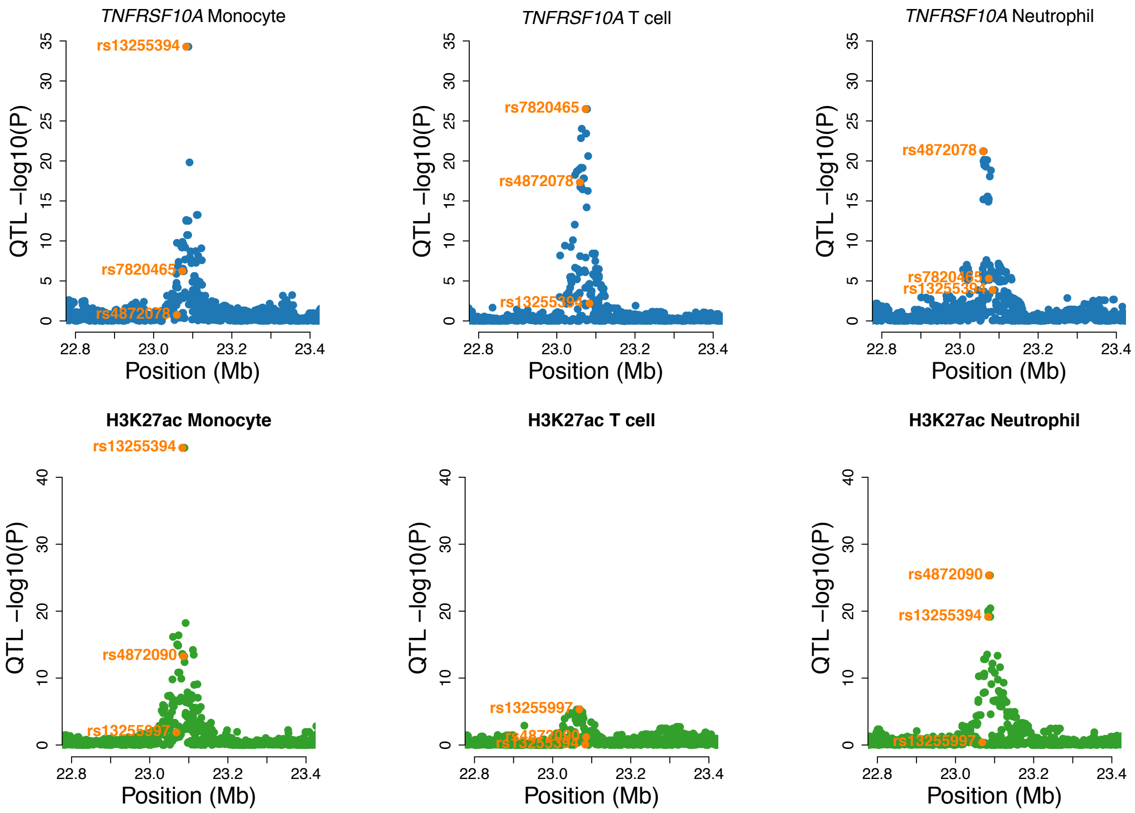


Figure 2.15: Different genetic signals across cell types at the *TNFRSF10A* locus
Regional association plots show the association signals for two colocalised features, *TNFRSF10A* gene expression and H3K27ac signal (8:23048166:23092260) in each cell type. The respective lead SNPs are highlighted in orange in each association plot. For the *TNFRSF10A* gene expression effect, there were three lead SNPs with evidence from LD and conditional analysis suggesting these represented three independent genetic signals explained by rs13255394 (monocytes), rs7820465 (T cells) and rs4872078 (neutrophils and secondary T cells). For the H3K27ac signal effect, evidence of variant LD and conditional analysis suggested that there were two genetic signals (rs13255394/rs4872090 monocytes and neutrophils) and a marginal signal in T cells (rs13255997).

I next investigated the histone modified region, which I postulated, was a regulatory control region for this locus. Monocyte H3K27ac signal across all individuals within the cohort showed greater correlation with monocyte *TNFRSF10A* signal than monocyte H3K4me1 (Figure 2.16, H3K27ac and *TNFRSF10A* Pearson's *r* = 0.567, p value = 1.254 x 10^{-16}, H3K4me1 and *TNFRSF10A* Pearson's *r* = 0.177, p value = 0.03). The higher correlation with H3K27ac than H3K4me1 could reflect the different roles of these histone marks. H3K4me1 is known to demarcate poised enhancers, that may not be active in the current cellular context (Creyghton et al., 2010). H3K27ac marks promoters but also active enhancers when modified in combination with H3K4me1, and therefore is required for active gene expression in specific cellular contexts (Creyghton et al., 2010, Heintzman et al., 2009). Using the Blueprint consortium cohort, we also observed a strong positive correlation between per-allele effect sizes of eQTLs and hQTLs (H3K27ac and H3K4me1) (Chen et al., 2016a). It is highly possible that at this locus, genetic disruption of H3K27ac is functionally more directly linked to gene expression.

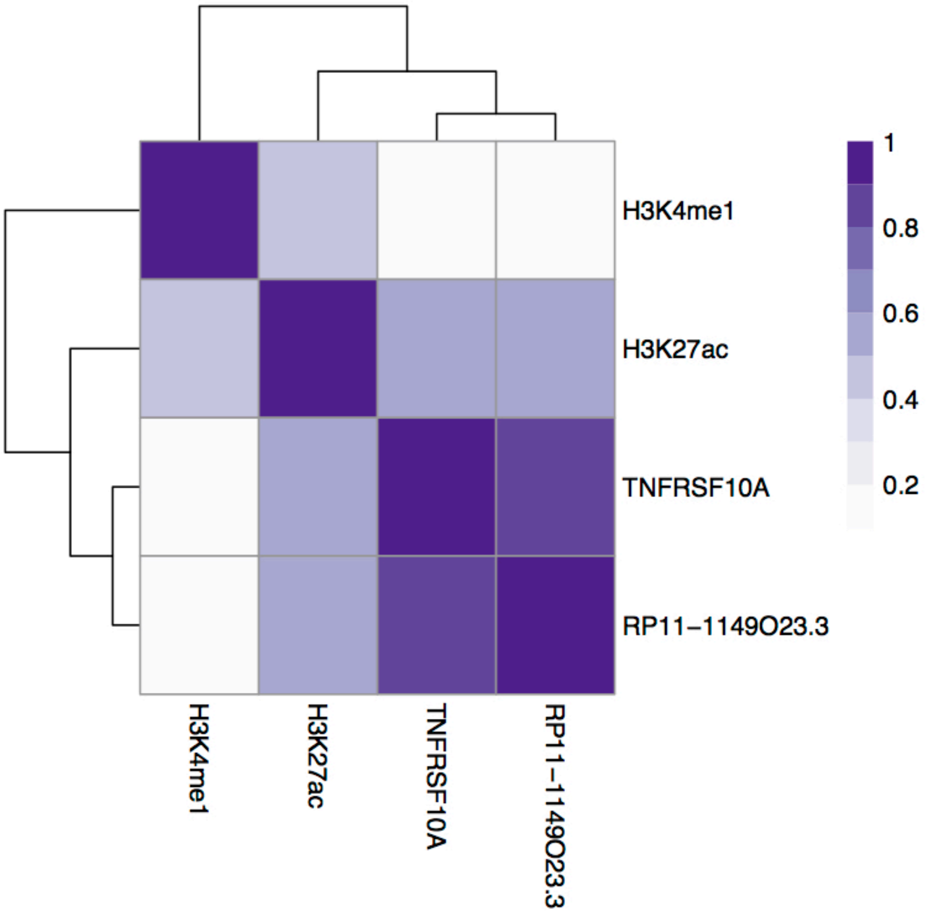


Figure 2.16: Correlation of molecular features at the *TNFRSF10A* locus
Heatmap shows unsupervised hierarchical clustering based on Pearson correlations between the molecular features. The Pearson correlation estimate is plotted between monocyte gene expression values or monocyte histone signal across the 158 individuals for which all of the molecular feature data was available.

In T cells, the H3K27ac signal was both weaker and explained by a different SNP (H3K27ac rs13255997, EA = G, EAF= 0.51, beta = 0.508, SE = 0.111, p value = 4.669 x 10^{-06}, Table 2.5, Figure 2.15) than in monocytes and there were no variants that reached the significant threshold for the H3K4me1 peak. There was evidence of regional histone signal in T cells as shown by the H3K27ac median $\log_2$RPM of 6.333 in T cells and 4.746 in monocytes and H3K4me1 median $\log_2$RPM of 7.246 in T cells and 7.146 in monocytes. This suggests that histone activity is present in both cell types, but the genetic regulatory mechanisms are divergent.

In neutrophils, both peaks were strongly associated with the lead neutrophil eQTL, rs4872090 (Table 2.5, Figure 2.15). The r^2 between the neutrophil histone lead SNP, rs4872090 and the monocyte histone lead SNP, rs13255394 was 0.402, and these SNPs are in close proximity approximately 2.29 kb apart (Figure 2.17). Conditional analysis using GCTA on hQTL summary statistics demonstrated that no significant signal remained after conditioning on the respective lead SNPs in either monocytes or neutrophils (qvalue < 5%). Interestingly, this was evidence that the histone signals can be explained by the same genetic signal in monocytes and neutrophils, despite the observation that the *TNFRSF10A* expression was controlled by independent SNPs across all cell types. Only in monocytes was there evidence for coordinated gene expression, histone activity and disease risk, which were all regulated by same SNP, rs13255394. Given that there was evidence of histone activity in other cell types, similarly, to the CAD *PSRC1* locus explained above, this suggests that there are additional regulatory features that coordinate in generating cell type-specific regulatory mechanisms.

I evaluated whether the additional colocalised gene, *RP11-1149O23.3*, could represent such an additional regulatory mechanism. Figure 2.17 shows that the lead SNP, rs13255394, is located within exon 1 of the non-coding RNA gene, *RP11-1149O23.3*. The p value for association with this RNA was more significant (p value=3.477 x 10^{-50}) and the effect size larger (beta= -1.171, SE = 0.079) than with *TNFRSF10A* expression (Table 2.4), suggesting this is an important functional effect at this locus.

The lead SNP for this locus identified in the original GWAS discovery (Fritsche et al., 2016) was rs79037040 which is also located in exon 1 of *RP11-1149O23.3,* but closer to the TSS of both *RP11-1149O23.3* and *TNFRSF10A* (Figure 2.17). Figure 2.17 shows the raw histone signal (in $\log_2$RPM) stratified by genotype (bottom panel), which enabled identification of the location of disrupted histone activity with a greater resolution (50bp across the genome). It was clear that the position of rs79037040 was directly in the centre of the monocyte and neutrophil H3K27ac peaks. This SNP is in high LD with rs13255394 (r^2 = 0.837, 1000G) suggesting that rs13255394 and rs79037040 likely explain the same genetic signal. rs79037040 was filtered from the first phase of BLUEPRINT cohort analysis due to stringent quality control thresholds. However, efforts within our group have been undertaken to reanalyse this cohort with improved imputation procedures generating a denser SNP panel, which included rs79037040. This analysis, referred to as phase 2, was performed by Kousik Kundu in the Soranzo team. I confirmed that in the phase 2 association testing, rs79037040 was now the lead SNP for all of the colocalised molecular features described thus far (Table 2.6). In addition, the colocalisation method used in this chapter, gwas-pw, calculates the posterior probability (PP) of all tested variants being causal for the two colocalised traits. Colocalisation with the new phase 2 summary statistics, calculated a PP for rs79037040 of 1 for *TNFRSF10A, RP11-1149O23.3,* H3K27ac and H3K4me1 traits. Therefore, all evidence supports that rs79037040 is the single causal variant for AMD risk and all monocyte molecular features.

Feature	SNP (EA/OA)	Beta	SE	P
Mono *TNFRSF10A*	rs79037040 (T/G)	-1.200	0.079	1.540×10^{-51}
Mono *RP11-114O23.3*	rs79037040 (T/G)	-1.290	0.073	8.497×10^{-70}
Mono H3K27ac (8:23048166:23092260)	rs79037040 (T/G)	-1.291	0.082	7.830×10^{-56}
Mono H3K4me1 (8:22998146:23133613)	rs79037040 (T/G)	-0.9856	0.095	2.057×10^{-25}
AMD	rs79037040 (T/G)	0.109	0.016	4.5×10^{-11}

Table 2.6: Association summary statistics of the lead SNP from Blueprint phase 2 genetic analysis
The AMD disease lead, rs79037040 (later merged into the ID rs13278062) was tested only as part of the latest Blueprint genetic analyses (phase 2). Beta, standard error (SE) and p value for association are listed here. The effect allele, T is associated with a decrease in gene expression and histone signal is also associated with an increase in AMD-risk. The effect allele (T) frequency is 0.556.

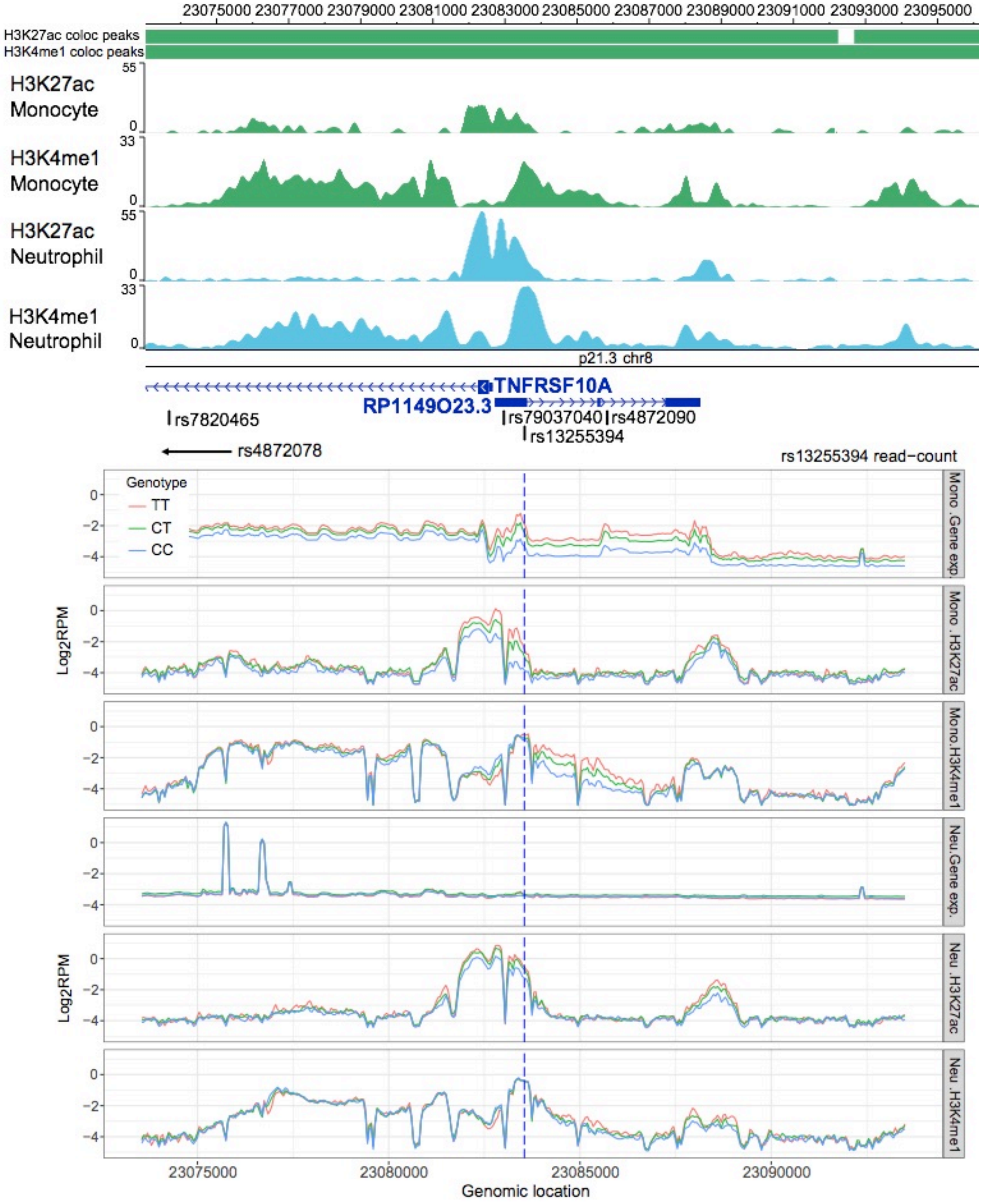


Figure 2.17: Epigenome characteristics of the *TNFRSF10A* locus
Genome browser figure of the genomic region around the *TNFRSF10A* gene and the proximal regulatory RNA, *RP11-1149O23.3* located upstream on the opposite strand (top panel). Peaks shown were generated from a representative individual from the BLUEPRINT cohort predicted to carry the allele associated with the highest signal (Chen et al., 2016a). Locations of the lead QTL variants and their genomic locations are shown. The lead monocyte QTLs, rs13255394 (phase 1) and rs79037040 (phase 2) are located within exon 1 of the RP11-1149O23.3-002/003 transcript. Solid blocks represent the colocalised peak, the second H3K27ac and H3K4me1 peaks extend further downstream than is shown. The bottom panel shows the raw histone signal, in monocytes and neutrophils, in Log_2RPM (reads per million) calculated in windows of 50 bp across the genome. The gene expression signal is plotted from the raw signal expressed in Log_2RPM. Each genotype of the lead monocyte SNP, rs13255394 is plotted (red for homozygous reference, TT). The genome browser plot (top panel) was generated with custom tracks using the Washu Epigenome browser (Zhou et al., 2011). The raw signal plots were generated using Blueprint data by Kousik Kundu (Chen et al., 2016a).

Less is known about the function of the non-coding RNA, *RP11-114O23.3* compared to the *TNFRSF10A* encoded receptor. However, previously a regulatory relationship between *RP11-114O23.3* and *TNFRSF10A*, two genes located on opposite strands, has been suggested (Zheng et al., 2016). Using microarrays, it was demonstrated that the expression of *RP11-114O23.3* (also known as *LOC389641*) is increased in pancreatic ductal adenocarcinoma (PDAC) tissues and correlated with patient prognosis and high expression levels reduced overall survival (Zheng et al., 2016). *RP11-114O23.3* expression increased proliferation and decreased apoptosis of cancer cell lines (Zheng et al., 2016). The authors identified a 378bp region that contained a highly conserved sequence directly upstream of 5' end of *RP11-114O23.3* that is the reverse complement of the *TNFRSF10A* promoter. In the same study upregulation of *TNFRSF10A* expression was observed in PDAC patient tissue compared to non-tumorous tissues. Crucially, siRNA mediated knock-down of *RP11-114O23.3* in SW1990 cells, significantly decreased *TNFRSF10A* expression but knock-down of *TNFRSF10A* had no effect on the expression of *RP11-114O23.3*. This suggests that *RP11-114O23.3* regulates expression of *TNFRSF10A* through complementary sequence-mediated binding.

These analyses were performed either in cancer cell lines or in PDAC patient tissue. It could be possible that this relationship is not observed in healthy primary immune cells. In addition, to the best of my knowledge, possible genetic regulation of the relationship between these genes has not been previously explored. I therefore sought to investigate the relationship between *RP11-114O23.3* and *TNFRSF10A* in monocytes, neutrophils and T cells.

First, similar to the high significant correlation observed between the *TNFRSF10A* and *RP11-114O23.3* expression in PDAC tissues (r^2=0.606, p<0.001, N = 106 patients), I also identified high correlation between monocyte expression values of the two genes across all BLUEPRINT individuals tested for these monocyte features (r= 0.804, p value = 3.962 x 10^{-45}, Figure 2.18, N = 194). However, in T cells, the correlation between *RP11-1149O23.3* and *TNFRSF10A* expression was lower and less significant than that observed with monocytes (r= 0.172, p value = 0.025, N = 169) (Figure 2.18). Expression of *RP11-1149O23.3* was not significantly associated with the lead T cell *TNFRSF10A* eQTL, when evaluating local FDR (Figure 2.18). In neutrophils, *RP11-1149O23.3* was not tested due to low expression, where the median log2FPKM was 2.672 (N = 196), compared to 5.852 in monocytes and 6.644 in T cells. Combined, this is evidence that the relationship between expression of the two genes and the genetic regulation of each gene was monocyte specific.

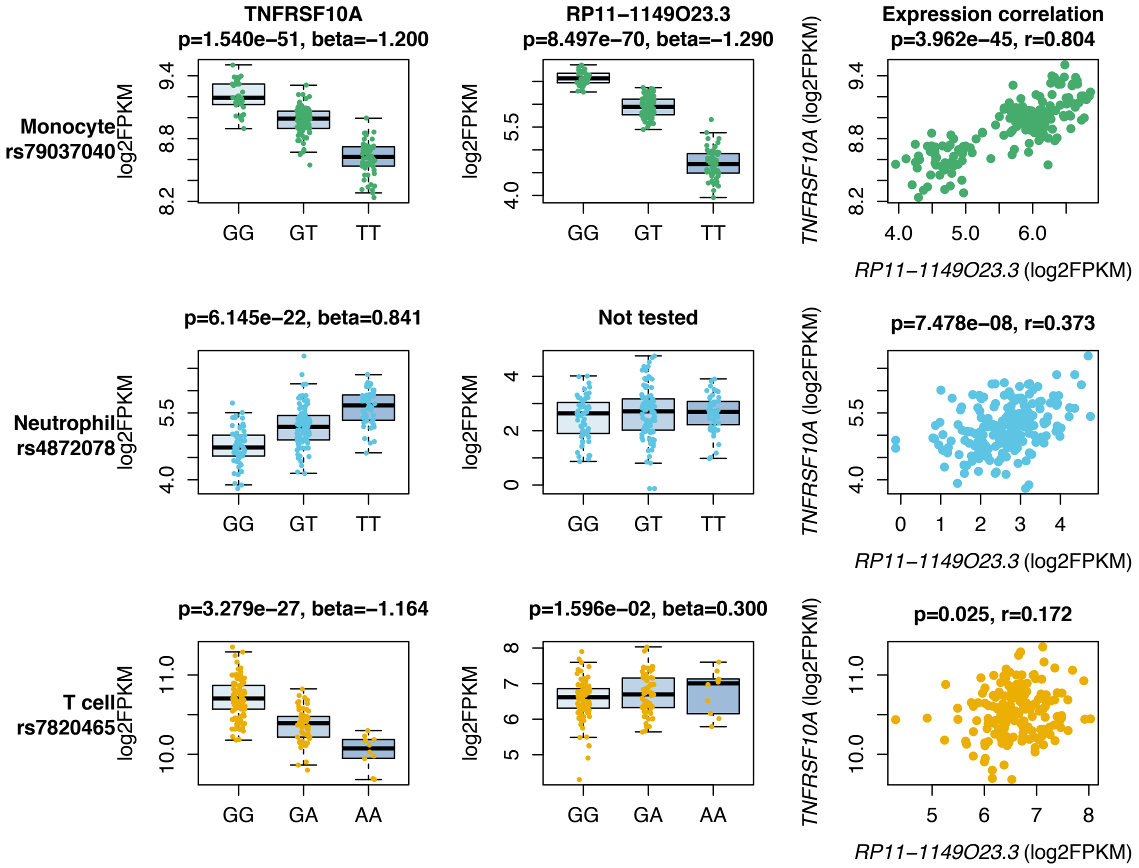


Figure 2.18: Genetic control of *TNFRSF10A* and *RP11-1149O23.3* is not shared across monocytes, T cells and neutrophils
Gene expression of the two genes, in log2FPKM, is stratified by genotype of the respective lead *TNFRSF10A* SNPs in each cell type. The beta and p values of association are shown for each case, as calculated by the BLUEPRINT study (Chen et al., 2016a). The correlation between the gene expression raw signals is shown for each cell type and is strongest in monocytes. Here the phase 2 lead SNP was used, rs79037040, but similar correlation was observed for the phase 1 lead, rs13255394 (data not shown).

I next sought to confirm that the observed correlation in healthy cells was consistent with a hierarchical regulatory mechanism as demonstrated by previous siRNA experiments where knockdown of *RP11-114O23.3* in cancer cells affected *TNFRSF10A* expression and not vice versa (Zheng et al., 2016). To test this, I implemented a linear regression approach to model the gene expression of each gene using expression of the alternative gene as a covariate in the model. In this conditional approach, variation due to one gene was removed by correcting for expression of all individuals in the BLUEPRINT cohort (Materials and Methods). Following this, I tested for association of any variation in expression remaining in the residuals with rs79037040 genotype. I performed the analysis using only the phase 2 lead SNP, given the evidence that rs79037040 was the single causal SNP for all monocyte molecular feature associations.

I first tested the univariate associations with rs79037040 for both genes using the lm() R function and confirmed significant associations observed in the QTL testing method from the BLUEPRINT study (Table 2.7) (Chen et al., 2016a). I applied the two-stage models (above and Materials and Methods) to test if the significant association with the casual SNP remains after removing any variation in *TNFRSF10A* expression that is due to variation in *RP11-1149O23.3* expression levels. I identified the strength of the rs79037040 association with *TNFRSF10A* gene expression decreased and was no longer significant after removing the RNA effect. The p value increased from 2.334 x 10^{-40} to 0.076 (non-significant) and the effect size (beta) decreased by 10-fold (Table 2.7). This demonstrated that *RP11-1149O23.3* contributes a high degree to variation in *TNFRSF10A* expression. I also applied the reverse model: *RP11-1149023.3 ~ TNFRSF10A* expression followed by testing the residuals for association with rs79037040. The results demonstrated that the p value increased from 8.349 x 10^{-66} to 1.124 x 10^{-08}, remaining significant and beta decreased only by 3.5 times. This reduction suggested that there may also be a smaller effect of *TNFRSF10A* expression on that of *RP11-1149023.3,* but the dominant effect is regulation of *RP11-1149023.3* on *TNFRSF10A* expression.

I extended the causality analysis of the different monocyte features to assess whether this approach may indicate that H3K27ac regulation also exerted an effect on gene expression. I applied the following model: *TNFRSF10A* gene expression ~ *RP11-1149O23.3* expression + H3K27ac signal and then tested the residuals for association with rs79037040 (Table 2.6). I used R^2 estimates, adjusted for the number of covariates in the model, to evaluate whether the model fit improves with the addition of the histone modification effect. The adjusted R^2 slightly increased when adding H3K27ac signal as a covariate, from 0.643 to 0.669, and a significant difference was confirmed using the ANOVA significance test for nested models (p value = 2.883 x 10^{-04}, N = 158). Additionally, the inclusion of the colocalised H3K4me1 peak

(8:22998146:23133613) as a covariate in the model, decreased the R^2 estimate to 0.667 and was not significant when compared to the *TNFRSF10A* ~ *RP11-1149O23.3* + H3K27ac model (p value = 0.812, N = 158).

All together, these results suggested that both the RNA and H3K27ac influence gene expression of *TNFRSF10A*, but H3K4me1 has limited effect and that *RP11-1149O23.3* contributes a high degree to the variation in *TNFRSF10A* expression.

Feature	SNP (EA/OA) (EAF)	Beta	SE
TNFRSF10A (rs79037040)	Univariate	-0.303 (0.017)	2.334 x 10^{-40}
	Conditional (*RP11-114O23.3* expression)	-0.031 (0.017)	0.076
	Conditional (*RP11-114O23.3* expression + H3K27ac signal)	-0.025 (0.017)	0.140
RP11-114O23.3 (rs79037040)	Univariate	-1.001 (0.034)	8.349 x 10^{-66}
	Conditional (*TNFRSF10A* expression)	-0.283 (0.047)	1.124 x 10^{-08}
	Conditional (*TNFRSF10A* expression + H3K27ac signal)	-0.264 (0.047)	6.235 x 10^{-08}

Table 2.7: Conditional causality analysis in the *TNFRSF10A* locus
Association results (beta, SE, and p value) from a simple linear regression model using non-transformed phenotype values (as this demonstrated good model fit and normally distributed residuals). Similar trends were also observed for inverse normalised phenotype values. The univariate approach tests for association of the respective gene expression with the genotype of rs79037040 (lead monocyte SNP from Blueprint release 2 and lead AMD SNP). Conditional analysis then tests for association of the gene with genotype whilst conditioning on the expression of the alternative gene. The increase in p value and decrease in significance is greatest when testing for association between the SNP and *TNFRSF10A* expression whilst conditioning on *RP11-114O23.3* expression, which suggests the RNA may be causal for variation in expression of *TNFRSF10A*. The further approach conditions on the gene expression and H3K27ac signal and then tests the resulting residuals for association with the SNP genotypes. For all models, data from 158 individuals was used.

2.4 Discussion

In this chapter, I applied enrichment and colocalisation methods to evaluate the utility of immune molecular phenotypes, specifically of monocytes, neutrophils and naïve CD4 T cells, to dissect mechanisms of disease risk for a variety of disorders. I demonstrated that a high number, 46%, of tested disease loci colocalised with at least one molecular feature in at least one cell type and highlighted many important gene targets, some of which already have therapeutic utility (Table 2.3). Following this, I performed an in-depth analysis of two example disease loci, CAD *SORT1* and AMD *TNFRSF10A* and demonstrated how the integration of multiple data sources is required to generate plausible mechanistic hypotheses.

I identified significant enrichment of GWAS variants in regions of the genome known to be associated with immune molecular traits, particularly of monocyte and neutrophil eQTLs in all of the range of five diseases I studied. The relative absence of strong cell-type specific patterns for most diseases was consistent with previous observations using similar analytical approaches for the same molecular data but with a wider range of classical autoimmune diseases (Chen et al., 2016a). Here, for disease loci associated with coeliac disease, Crohn's disease, inflammatory bowel disease, ulcerative colitis, multiple sclerosis, rheumatoid arthritis or Type 1 diabetes, colocalisation of 54% with eQTLs, 55% with splicing QTLs, 62% with H3K27ac and 54% with H3K4me1 QTLs was observed (Chen et al., 2016a). However, this analysis included the MHC region for all diseases, which may affect the estimates given the complex genetic architecture of this region. Comparison of colocalisation estimates with other studies and/or diseases is challenging, given the different methods available and approaches to evaluate colocalised loci. Using iPSC-differentiated unstimulated and stimulated macrophages, the highest number of colocalised eQTLs or chromatin accessibility (ca)QTLs were observed with inflammatory bowel disease variants (11 and five loci respectively) (Alasoo et al., 2017). The recent G. TEx eQTL analysis identified a similar percentage to my study; 52% of trait-associated variants colocalised with an eQTL in one or more tissues (G. TEx Consortium, 2017).

Many of the colocalised genes identified had well-established or suggested roles in immune function. This is in agreement with increasing insight into the pathogenic involvement of inflammation in wider range of disorders and with the early promise of therapeutically targeting these pathways (Section 2.1). Based on these observations, I concluded that functional insight can be gleaned from using peripheral immune cell types in these diseases, which were traditionally not considered prototypic immune-mediated diseases. I also provided support for the importance of lipid-pathway genes such as the CAD loci colocalised with *LIPA* and the AMD locus colocalised with *CETP*. Through the identification of well-known examples such as these I confirmed the validity of my analytical approach as well as providing further mechanistic evidence for disease loci.

This study is not the first to integrate GWAS loci with molecular features, certainly for autoimmune diseases, this is fairly widespread and as discussed in Section 2.1 has generated important insight (Farh et al., 2015, Chun et al., 2017). Using peripheral whole blood as a tissue source enabled QTL identification in large cohorts of healthy individuals, such as a study from 2013 of 5,311 individuals with replication in 2775 individuals (Westra et al., 2013). These data have been used to dissect gene expression consequences of trait-associated loci with for example inflammatory bowel disease and lung function (FEV_1) (Wain et al., 2015, Huang et al., 2017b). In comparison, an advantage of cohorts such as BLUEPRINT is in facilitating identification of the specific cell-type source of the genetic effect. In addition, BLUEPRINT enables the study of neutrophil effects, which are historically understudied despite that the important role in inflammation, immune cross-talk and certain disease aetiology. I highlighted one particular SLE locus, *UBE2L3,* where the strongest colocalised eQTL was from neutrophils compared to previous observations of correlation of this locus with *UBE2L3* expression across monocytes, CD4 T cells, B cells and NK cells at this locus in the original GWAS study (Bentham et al., 2015). Definitive confirmation of disease-relevant mechanisms requires functional validation, but if clear demarcation of cell types is possible for at least a proportion of loci, this is an important preliminary step in designing these experiments and selecting experimental cellular models. Blood is an experimentally tractable and easily accessible tissue source and function is conserved across organisms facilitating the use of animal models (Orkin and Zon, 2008, Vasquez et al., 2016). Providing colocalisations within blood for diseases where human biosamples for other relevant tissues are challenging to obtain, such as brain or ocular tissue, is a clear advantage of these findings.

However, using whole blood and purified cell cohorts to fully resolve functional genetic mechanisms can be thought of as complementary. For example, the smaller sample sizes of cohorts such as BLUEPRINT (N = 200) limits the identification of *trans* QTLs. These are variants that affect molecular features located more than 1Mb away, or even on different chromosomes. Highly powered studies are required to detect these effects due to the increase in the multiple testing burden when expanding the testing window beyond variants in *cis*. *Trans* eQTLs were identified and replicated for 103 independent loci using the large whole-blood cohort described above (Westra et al., 2013). Complex-trait associated variants showed a high number of *trans* eQTL effects. Interesting insights into CAD variants were also gleaned from both *cis*- and *trans*-eQTLs identified in another large microarray gene- and exon-based QTL dataset of whole blood from 5257 individuals (Joehanes et al., 2017).

19,000 independent lead cis-eQTLs were detected compared to just over 6000 *trans*-eQTLs. By overlapping blood eQTLs with CAD SNPs or those in LD $r^2 \geq 0.8$, Joehanes *et al.* (2017) identified genes for 21 of the 58 GWAS loci (Joehanes et al., 2017). Those in agreement with the effects identified in this chapter were eQTLs for *LIPA*, *NT5C2*, *VAMP8 and GGCX, REST* and also the *PSRC1* target at the *SORT1* locus. In my study, only the *GGCX* and *NT5C2* genes showed evidence of either expression or splicing effects across all three of the cell types studied here, enabling some assessment of cell-type specificity of the other loci among the three subsets studied in BLUEPRINT. Joehanes *et al.* (2017) identified *trans* effects at some CAD loci. For example, the evidence suggested that the *VAMP5-VAMP8-GGCX* locus (rs7568458) affected the expression of 5 genes in *trans*; *CASP5, DPEP3*, *CRISPLD2*, *SLC26A8* and *PKN2* (Joehanes et al., 2017). The expression of *CASP5* has been previously shown to be associated with blood pressure, suggesting trait-relevant effects may occur in *trans* (Joehanes et al., 2017). In total, Joehanes *et al.* (2017) identified more CAD SNPs overlapping with gene QTLs than my analysis (21 compared to 8 e/sQTLs here), which could be due to the increased study power but could also represent overlaps occurring by chance without formal assessment such as those applied in colocalisation methods (Joehanes et al., 2017). Future studies with larger cohorts and defined cell populations will combine the advantages of these two study types and provide the opportunity for identifying further regulatory pathways that also influence disease risk. Of course, blood cell types will not be the disease-relevant tissue for all loci, which may be the explanation for why I did not detect colocalisations for all loci. Part of the future work of this thesis will involve fully integrating disease loci with eQTLs from the (G. TEx Consortium, 2015) to evaluate effects across a wider range of tissue types.

A further advantage of the BLUEPRINT cohort is that the it enabled concomitant assessment of multiple regulatory features rather than being limited to gene expression effects. Consideration of the genetic effects on chromatin state enables resolution of regulatory mechanisms at disease loci. It has been demonstrated that transcriptional and local epigenetic states are highly coordinated and that genetically controlled changes in gene expression may occur through disruption of chromatin states (Grubert et al., 2015, Waszak et al., 2015). My analysis focused on the histone modifications, H3K27ac and H3K4me1 and also RNA splicing effects, which were either independent of or in addition eQTL effects depending on the locus. In describing detailed mechanisms, I showed two loci that differed in the strongest chromatin effect being either H3K27ac or H3K4me1. These observations support the use of multiple sources and types of molecular data in fully investigating regulatory function.

The *SORT1* CAD locus colocalised with *PSRC1* eQTL but not a *SORT1* QTL effect in these haematopoietic cell types. That *PSRC1* is regulated (and not *SORT1*) in blood has been observed in independent cohorts (Zeller et al., 2010, Joehanes et al., 2017). However, I demonstrated this effect was also present in neutrophils and absent in CD4$^+$ T cells. I further provided a potential molecular mechanism underlying the differences between hepatocytes and myeloid cells. Principally, this difference seemed to be explained by the binding of important haematopoietic TFs, C/EBPβ and PU.1, to a C/EBP motif disrupted by rs12740374. In liver, C/EBPα was found to be bound at this motif and *SORT1* was the strongest genetically regulated expression effect (Musunuru et al., 2010). Interestingly, using ENCODE data, I found that the liver pioneer factors, FOXA2 and FOXA2, were bound at the rs12740374 locus in HepG2 cells (data not shown) (Odom et al., 2006, Iwafuchi-Doi et al., 2016, Zaret et al., 2008). I postulated that the binding of different pioneer TFs in different cells may promote regulation of the expression of alternative genes. The importance of pioneer factors in directing cell-type specific binding and expression has previously been observed (Heinz et al., 2010, Heinz et al., 2013, Mullen et al., 2011). In this way, SNP-mediated disruption of a TF motif can result in opposing molecular consequences in different cells while the sequence effect remains the same.

The challenge remains of interpreting the role of *PSRC1* in blood cells and whether this gene is causally related to CAD. Gain-of-function studies in mice and genetic findings in human cohorts have both supported an association of the 1p13 minor haplotype (rs12740374 T allele) with increased hepatic *SORT1* expression and decreased LDL-C and VLDL (Musunuru et al., 2010). However, in a mechanism thought to be independent of lipoprotein metabolism, SORT1 mediates LDL uptake in macrophages stimulating their differentiation to foam cells and therefore *promoting* atherosclerosis (Mortensen et al., 2014, Westerterp and Tall, 2015). It is conceivable, therefore, that the differences in response to LDL between liver and myeloid cells manifest in differences in CAD risk. This is supported by the observation of Musunuru *et al.* (2010) that although *PSRC1* expression in human liver was significantly associated with rs12740374, overexpression of *Psrc1* in mouse liver was not associated with any significant changes in total cholesterol (Musunuru et al., 2010). In my analysis, increased *PSRC1* expression was associated with a protective CAD effect. Given little is known regarding the function of *PSRC1*, further work is required to evaluate this effect in haematopoietic cells and whether this effect could also be causal to CAD or whether the hepatocyte *SORT1* effect is the only causal contribution of this locus to disease risk. For example, using CRISPR-Cas9 to knock-out this gene in iPSCs and differentiation to macrophages followed by stimulation with oxidised LDL to promote foam cell formation could highlight whether *PSRC1* is involved through regulating this process. These experimental approaches are well established (Hale et al., 2015, Reschen et al., 2015). Implementation of

Mendelian randomization approaches could also help to ascertain whether the blood *PSRC1* effect of this locus represents a disease causal mechanism or purely pleiotropy.

It would also be interesting to further evaluate the mechanisms through which cell-type specific genetic regulation of gene expression is achieved and confirm that the binding of PU.1 and C/EBPβ is linked to *PSRC1* gene expression. This is important for the design of novel treatments for understanding the effect of a drug on multiple tissues. Specific siRNA knock-down of PU.1, C/EBPβ and C/EBPα in haematopoietic and hepatic cell lines, coupled with an assessment of the effect on *PSRC1* and *SORT1* gene expression, would demonstrate whether the binding of both factors is required for downstream gene expression or whether one TF acts as the putative regulatory factor. These experiments could be performed in the cell line models discussed above; differentiated HL60, the monocyte-like cell line, U937 and the hepatic cell line, HepG2.

I also investigated a monocyte-specific effect at the AMD *TNFRSF10A* locus and identified a putative regulatory element for the *TNFRSF10A* and *RP11-1149O23.3* genes that colocalised with the AMD locus. Although colocalisation was identified between both H3K4me1 and H3K27ac peaks, the effect was more significant with H3K27ac (Figure 2.14, Figure 2.17). In addition, in a linear model of *TNFRSF10A* gene expression, inclusion of H3K27ac improved the model fit, but not inclusion of H3K4me1. I also demonstrated the importance of other regulatory mechanisms at this locus, by identifying the colocalisation of an eQTL for *RP11-1149O23.3* and further that expression of this non-coding RNA explained a significant degree of variation in *TNFRSF10A* expression. This relationship is a “local trans” regulatory mechanism, where a genetic variant affects the expression of one gene, which in turn regulates a proximally located, but distinct gene. Clearly genomic regulation in this locus involved both an RNA and histone effect, but the exact linear relationship between these effects is difficult to ascertain without functional validation. Open chromatin is required for active gene expression of most genes in order to enable access of the RNA polymerase II machinery to the transcription start site (Venters and Pugh, 2009). It is conceivable that *RP11-1149O23.3* could either require established open chromatin to be expressed or to bind this region or could recruit further chromatin remodellers. Using CRISPR to knock-out the RNA and independently the histone region could allow an assessment of the downstream effect on *TNFRSF10A* expression may aid dissection of these relationships. It is thought that disruption of chromatin is proceeded by the alteration of transcription factor binding (McVicker et al., 2013, Kilpinen et al., 2013). Experiments such as ChIP-seq could also be used to identify other bound co-factors. The lead AMD and molecular feature SNP, rs79037040, located within exon 1 of *RP11-1149O23.3* has also been predicted to disrupt the motifs of TFs LXR and NERF1a (Kheradpour and Kellis, 2014). These TFs are both highly expressed in monocytes (>6 and >11 log2FPKM respectively from RNA-seq gene expression from (Chen et al., 2016a). Disruption of TFs bound, and histone activity at the RNA promoter could lead to changes in *RP11-1149O23.3* expression, which further propagate to corresponding changes in the expression of *TNFRSF10A*.

Given that TNFRSF10A is a surface expressed receptor, I postulated that downstream alterations in receptor surface expression or function would ultimately impact AMD risk. Expression of the *TNFRSF10A* gene in the AMD-relevant tissue, peripheral retinal pigment epithelium/choroid/sclera (PRCS) (FPKM PRCS = 1.91), is low compared to AMD-drug target gene, VEGFA (FPKM PRCS = 56.38) (Li et al., 2014), which potentially further provides evidence that the disease-relevant effect of this locus could be exerted in monocytes. TNFRSF10A (TRAIL-R1), TNFRSF10B (TRAIL-R2) and the decoy receptor TNFRSF10D (TRAIL-R4) are all highly expressed on the surface of primary monocytes from healthy individuals, with the highest surface expression observed for TNFRSF10B (Deligezer and Dalay, 2007, Liguori et al., 2016). The disease SNP is not associated with monocyte expression of the other TNF receptors (rs79037040 TRAILR2 p value = 0.976, TRAILR3 p value = 0.144, TRAILR4 = 0.504). Interestingly, despite the significant *TNFRSF10A* eQTLs in neutrophils and T cells, this receptor has been shown to be lowly expressed on freshly isolated primary neutrophils and T cells (Kamohara et al., 2004, Liguori et al., 2016). Instead the decoy protein TNFRSF10C (TRAIL-R3) receptor is highly expressed on the surface of neutrophils and to a lesser extent on the surface of lymphocytes (Kamohara et al., 2004, Liguori et al., 2016). TNFRSF10C expression was associated with a strong eQTL in neutrophils (rs7009522, EA = A, beta = 1.098, p value = 2.519 x 10^{-23}) but was not tested in T cells. Therefore, post-transcriptional processes could play an important role in reducing surface expression of *TNFRSF10A* in certain cell types, highlighting the importance of integrating multiple sources of functional information to interpret the mechanistic consequences at disease loci. Similar experiments to those described in Chapter 4 of this thesis, where surface receptor expression was measured using flow cytometry in a recall-by-genotype design could establish possible differences in surface expression associated with genotype.

Lower expression of *TNFRSF10A* in monocytes corresponded to an increase in AMD risk. It has been postulated that recruitment of blood cells such as macrophages to the damaged retinal tissue in AMD could contribute to a pathogenic pro-inflammatory environment (Nussenblatt and Ferris, 2007). TNFRSF10A is known to be immunosuppressive. The evidence presented here suggests that lower *TNFRSF10A* gene expression could result in a reduced inability to downregulate pro-inflammatory responses in monocytes or in macrophages if differentiated to monocytes. It would be interesting to study these effects in monocyte-derived macrophages to observe if cells with reduced *TNFRSF10A* expression show a more inflammatory profile.

Not all loci that colocalised with histone features also shared a gene effect, either expression or RNA splicing, an observation that was also made in the Chen *et al*. (2016a) study and in other independent cohorts. For example, a study identifying iPSC-differentiated macrophage gene expression and chromatin accessibility QTLs (open chromatin using ATAC-seq) also found that of the 23 caQTLs that colocalised with a GWAS variant, only two of these also colocalised with an eQTL (Alasoo et al., 2017). These regulatory QTLs might impact gene expression in different cells, contexts or affect post-transcriptional processes (Alasoo et al., 2017, Fairfax et al., 2014, Pai et al., 2015).

Colocalisation approaches provide a statistical assessment of regions of the genome that are associated with two traits, but there are still some limitations to this approach. First, the power of the method to detect true colocalisation when the lead variants of each trait are in high LD ($r^2 \geq 0.8$) is limited and the method assumes one causal variant at each locus. Definitive demonstration of causality between traits, specifically whether the shared molecular effect is causal to disease risk, is not possible (discussed in detail in Chapter 5).

Further, the colocalisation method does not distinguish between multiple independent genetic signals, which have been observed at molecular loci and in some cases colocalised with disease variants over or in addition to primary signals (Dobbyn et al., 2017, Ke, 2012). I used the pre-defined association signals from the Chen *et al.* (2016) where multiple independent signals were not investigated. Visual inspection of some colocalised loci in this analysis suggested the colocalisation may not be between the primary molecular association. Supplementary Figure 2.3 shows the colocalisation of the SLE *FCGR2A* locus with a splicing QTL for *FCGR3A* (CD16) in monocytes and *FCGR2A* (CD32) in neutrophils. Both genes are expressed in neutrophils and monocytes (≥ 8 $\log_2$FPKM), and both receptors are expressed on the surface of each cell type (Stenberg et al., 2013, Cooper et al., 2012, Ziegler-Heitbrock, 2007, Devaraj et al., 2013). The disease lead SNP and the *FCGR3A* splicing lead QTL are highly correlated (rs6671847, rs4657041 1000G r^2=0.89) but the lead splicing SNP for *FCGR2A* is not highly correlated with the disease SNP (rs12129787 1000G $r^2 < 0.2$). To assess this, future work will implement conditional analysis to identify independent genetic signals.

In conclusion, I demonstrated that applying colocalisation methods to GWAS and molecular QTL data can provide detailed mechanistic hypotheses at disease risk loci, which are invaluable for facilitating further experimental investigation.

Chapter 3

Expanding genetic studies to cellular phenotypes: analytical exploration of neutrophil function phenotypes

Collaboration Note

The neutrophil function experiments were performed by Kate Waller, Carly Kempster, Harriet McKinney and Joana Batista under the supervision of Kate Downes at the NHS Blood and Transplant Unit, Addenbrookes Hospital. The project was also coordinated by Willem Ouwehand, Department of Haematology and Nicole Soranzo, Wellcome Trust Sanger Institute.

The discovery cohort genotyping data from the Cambridge BioResource was analysed and processed by Heather Elding at the Wellcome Trust Sanger Institute. The whole-genome sequence data was analysed as part of the BLUEPRINT consortium (Chen et al., 2016a).

Analysis of these data was performed in close partnership and supervised by Klaudia Walter (Wellcome Trust Sanger Institute). Klaudia also analysed the genotype data for the Sanquin replication cohort and devised the custom scripts to calculate the parameters from the real-time data. All the other analyses described here otherwise were performed by myself.

Taco Kuijpers and Judy Geissler (Sanquin Research, The Netherlands) coordinated the Sanquin replication cohort, which was genotyped at the Wellcome Trust Sanger Institute. The replication experiments were performed by Anton Tool at Sanquin Research, The Netherlands. Anton and Taco also provided helpful discussions on the details of the neutrophil function assays and on the approaches to analyse the Cambridge discovery cohort data.

3 Expanding genetic studies to cellular phenotypes: analytical exploration of neutrophil function phenotypes

3.1 Introduction

3.1.1 Neutrophil Biology and central role in immune responses

Neutrophils, or polymorphonuclear leukocytes (PMNs), are the most abundant type of white blood cell, constituting approximately 40-60% of total white blood cells (Wright et al., 2010). Neutrophils are characterised by two distinctive morphologies; the lobulated nucleus and presence of protease-containing granules (Kaplan, 2013) (Figure 3.1). The closely related eosinophils and basophils, together with neutrophils, form the granulocytic family of leukocytes (Amulic et al., 2012). Often the first responders in an immune response, neutrophils deploy antimicrobial functions such as phagocytosis to remove pathogens and cell debris, degranulation to release granular lytic enzymes and the respiratory burst to produce reactive oxygen species (ROS) (Kaplan, 2013, Amulic et al., 2012). In healthy homeostatic conditions, the release of mature neutrophils from the bone marrow must be highly controlled to prevent inadvertent activation and possible tissue damage. During infection when the demand is increased, high numbers of neutrophils are released (Amulic et al., 2012).

Neutrophils are technically complex to study being refractory to techniques such as transfection and RNA knock-down. In addition, as terminally differentiated cells they cannot be grown in tissue culture (Amulic et al., 2012). Many insights have been gleaned from either *in vitro* assessment, cell-line models (as discussed in Chapter 2) or mouse models. There are certain differences between murine and human PMNs that can complicate findings. For example, there is a lower number of circulating neutrophils in mice compared to humans (Amulic et al., 2012).

Neutrophils were traditionally thought of as short-lived cells (6-8 hours), incapable of further expansion (terminally-differentiated) and therefore were assumed to play a more passive role in responding to activating signals (Wright et al., 2010, Amulic et al., 2012). However, it is now known that activated neutrophils possess the ability to perform most regulatory or immune-related functions possessed by macrophages, particularly when neutrophils are primed and have longer life spans (Wright et al., 2010). When stimulated, neutrophils can synthesise pro-inflammatory mediators, present antigen through MHC class II receptors to T lymphocytes as well as mediate extensive immune cell cross-talk as summarised in Figure

3.1 (Wright et al., 2010). Below, I discuss in detail the neutrophil functional responses relevant to this chapter.

Neutrophil activation occurs via two-stages. First, resting circulating neutrophils can be “primed” by bacterial compounds and host cytokines and chemokines such as TNF-α, GM-CSF or IFN-γ (Wright et al., 2010). Upon activation, primed neutrophil responses are much greater than those of non-primed activated neutrophils (Hallett and Lloyds, 1995). For example, the gram-negative bacterial outer cell membrane lipopolysaccharides (LPS) prime neutrophils by stimulating the assembly of the NADPH oxidase complex on the membrane. Subsequent stimulation by the bacterial chemoattractant N-formylmethionine-leucyl-phenylalanine (fMLP) then activates the complex (El-Benna et al., 2008). Priming can occur over minutes where pre-formed receptors contained within intracellular granules are mobilized and transported to the plasma membrane. In some cases, over longer periods of transcription further inflammatory molecules can be induced *de novo* (Wright et al., 2010). Full neutrophil activation and mobilisation of all neutrophil killing activities requires integration of multiple environmental signals and is the result of a cascade of activating signalling processes (Amulic et al., 2012).

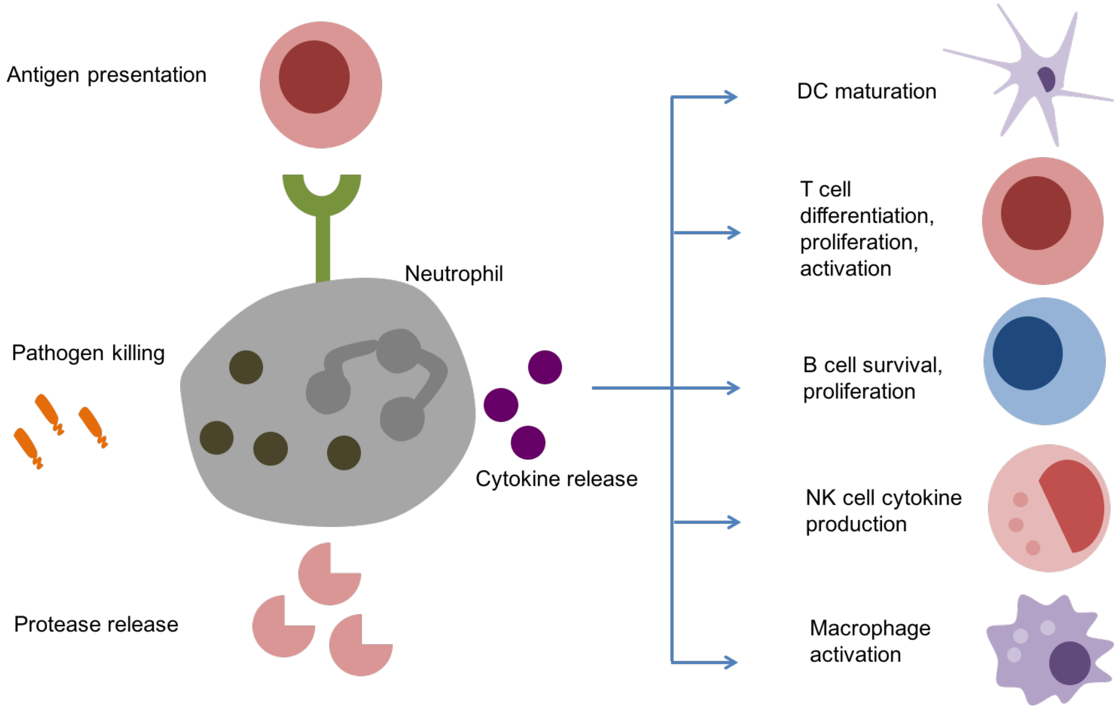


Figure 3.1: The central role of neutrophils in the immune response
Neutrophils are one of the first immune cells to respond to infection and possess a variety of anti-microbial functions. Through cytokine release neutrophils can activate many other cells of the immune system leading to a coordinated adaptive immune response as well as innate response. A relatively new functionality has been observed in the presentation of antigens through MHC class II molecules to elicit T cell activation and proliferation. Some of these functionalities have also been observed to become dysregulated in the pathology of autoimmune disorders. Adapted from (Wright et al., 2014).

Neutrophil activation requires the recruitment of these cells to inflamed tissues, which is achieved through neutrophil recognition of stimulated endothelial cells. Signals derived from either bacterial (LPS and fMLP) or host mediators (TNF-α, IL-1β, Il-17) stimulate the surrounding endothelial cells to express selectin adhesion molecules and members of the integrin super-family, intercellular-adhesion molecules (ICAMs) (Figure 3.3) (Amulic et al., 2012, Borregaard, 2010). Tethering of neutrophils to activated endothelial cells is mediated through neutrophil surface molecules, P-selectin glycoprotein ligand-1 (PSGL-1) and L-selectin, which interact with the endothelial-expressed P- and E-selectins (Figure 3.3) (McEver and Cummings, 1997, Amulic et al., 2012). Neutrophils then roll along the endothelial wall with concomitant activation of signalling kinases including Src family kinases (Syk), phosphoinositide 3-kinase (PI3K) and p38 mitogen-activated protein kinase (Mueller et al., 2010, Amulic et al., 2012). Firm adhesion and the arrest of rolling occurs through integrin contact mediated by the neutrophil-expressed LFA-1 and Mac-1 receptors. Combined with activation by cytokines and chemoattractants, sustained interactions generate changes in neutrophil morphology and a process known as cell spreading (Figure 3.3) (Sengupta et al., 2006). Cytoskeleton rearrangements enable neutrophils to move along chemotactic gradients. At this stage, the respiratory burst is initiated (Amulic et al., 2012). Firm adhesion allows neutrophils to cross the cell membrane once they reach an endothelial cell junction in a process known as transendothelial migration (Figure 3.3). Neutrophil adherence to the endothelial surface is referred to as adhesion and is a vital step in recruiting neutrophils to the site of inflammation ensuring effector functions are appropriately stimulated reducing the risk of spurious tissue damage.

At the site of inflamed tissue, further host and bacterial inflammatory signals activate the later stages of neutrophil activation. Chemoattractants signalling through GPCRs, such as the fMLP receptor, activate the MAPK/ERK signalling cascade culminating in the assembly of the respiratory burst complexes (Zarbock and Ley, 2008, Selvatici et al., 2006). The NADPH oxidase complex is a multi-protein complex that catalyses the production of powerful oxidising agents known as ROS (Figure 3.4) (Segal et al., 2000). ROS are directly antimicrobial but can also modify host molecules and responses and also influence the activity of granule proteins (Amulic et al., 2012).

Sustained activation by chemoattractants along a chemical gradient also stimulates degranulation, which is the release of antimicrobial contents from the specialised organelles known as granules (Table 3.1). Granules are formed throughout the differentiation process, and their contents vary based on the changing transcriptional programme during development (Amulic et al., 2012). Granules fuse with either the plasma membrane or phagosome, releasing the antimicrobial contents and permanently changing the composition of those membranes (Amulic et al., 2012). Granule deployment has important functional consequences. For example, the specific granules (Table 3.1) contain with flavocytochrome b558, which is a component of the NADPH and therefore, the fusion of these granules with the phagosomal or plasma membrane promotes the respiratory burst response (Amulic et al., 2012, Uriarte et al., 2011). Antimicrobial proteins can be categorised into three groups: those that bind to microbial membranes, those that possess enzymatic activity and those that deprive microbes of nutrients (Amulic et al., 2012). Some examples are given in Table 3.1.

Activated neutrophils can also release extracellular traps (NETs), which are web-like structures of granule proteins and decondensed chromatin. NETs enable the neutralisation of a wide range of pathogens (Papayannopoulos, 2017, Brinkmann et al., 2004). This particular function is not studied in this chapter, but the dysregulation of this process is known to contribute to the aetiology of inflammatory disorders (Papayannopoulos, 2017).

Granule protein	Granule Type	Function
Myeloperoxidase (MPO)	Azurophillic/primary	Can react with H_2O_2 to produce ROS including hypohalous acids
Lysozyme	Azurophillic/primary, Specific (secondary)	Degrades bacterial cell wall
Elastase, Cathepsin G	Azurophillic/primary	Cleaves bacterial virulence factors and outer membrane proteins, binds to bacterial membranes
Defensin	Azurophillic/primary	Arginine-rich cationic peptides, antimicrobial often by disrupting bacterial membranes
Laminin receptor	Specific (secondary), Gelatinase (tertiary)	Cell surface receptor, important for cell adhesion. Binds laminin, an extracellular matrix protein
Bactericidal/permeability-increasing protein (BPI)	Azurophillic/primary	Binds to LPS and increased bacterial permeability and bacterial phospholipid hydrolysis
Azurocidin	Azurophillic/primary	Binds to bacterial membranes
Lactoferrin	Specific (secondary)	Binds to and sequesters iron, which is a bacterial nutrient and inhibits bacterial growth. Binds to lipid A of LPS resulting in a release of LPS from the membrane and increased in permeability
Cytochrome b_{558}	Specific (secondary), Gelatinase (tertiary), Secretory	Component of phagocyte NADPH oxidase
fMLP receptor	Specific (secondary), Gelatinase (tertiary), Secretory	Receptor for bacterial chemoattractant fMLP
MAC-1 (CD11b/CD18)	Specific (secondary), Gelatinase (tertiary), Secretory	Complement receptor
Gelatinase	Specific (secondary)	Gp91phox/p22phox, CD11b, MMP25, arginase-1, β2-microglobulin, CRISP3
Complement receptor 1 (CR1)	Secretory	Complement receptor
LFA-1 (CD11a/CD18)	Secretory	Integrin important for adhesion
Proteinase 3 (PR3)	Azurophillic/primary	Serine protease

Table 3.1: Examples of granule proteins, which granule(s) they are contained within and the function

This table describes the four different types of granules and examples of granule protein content. The four granules include azurophilic, specific, gelatinase and secretory. The list is not exhaustive and there are other proteins contained in neutrophil granules. Exocytosis of neutrophil granules is an important process in activation in response to a stimulus and the destruction of phagocytosed pathogens. This table was adapted from (Amulic et al., 2012, Wright et al., 2010, Nelson et al., 2008).

3.1.2 Neutrophils and disease

Fully functional neutrophil responses are important for appropriate immune responses, which is clearly shown by the inability to fight infections due to defects in neutrophil activation and function in patients with certain primary immunodeficiencies (Bouma et al., 2010, Record et al., 2015). Mutations found in chronic granulomatous disease patients result in a non-functional NADPH oxidase and deficient ROS production (Gennery, 2017, Segal et al., 2000). As a result, these patients are susceptible to severe infection and autoinflammation (Gennery, 2017, Segal et al., 2000, Amulic et al., 2012).

In Chapter 2, I also highlight examples where neutrophil function has been linked to complex diseases. Indeed, dysregulated neutrophil function is a key factor in the pathogenesis of certain inflammatory diseases, highlighting the importance of regulating neutrophil activity to balance effective immune responses while limiting damage to the host (Gupta and Kaplan, 2016). Apoptosis of activated neutrophils is important in the return to homeostatic conditions after an inflammatory response (Wright et al., 2010). Failure of neutrophil apoptosis or deficient clearance of neutrophil apoptotic particles can cause chronic inflammation as observed for example, in chronic obstructive pulmonary disease (COPD) and rheumatoid arthritis (RA) (Amulic et al., 2012, Wright et al., 2010). Neutrophil products, such as MPO and PR3 (Table 3.1) are also known targets of autoantibodies, referred to as antineutrophil cytoplasmic antibodies (ANCA) and have been detected for example in the systemic autoimmune disease, antineutrophil cytoplasmic antibody (ANCA)-associated vasculitis (AAV) (Gupta and Kaplan, 2016). The interaction of ANCAs with antigens on primed neutrophils can activate neutrophil effector processes as described above (Kaplan, 2013). Observations of the presence of ANCAs, activated neutrophils in the synovial fluid and granulocyte-dependent cartilage damage in RA patients also support a role for neutrophil dysregulation in this disease (Emery et al., 1988, Mohr and Wessinghage, 1978, Kaplan, 2013). Indeed, neutrophils from RA patients in remission showed lowered adhesion and chemotactic characteristics, suggesting that migration to the synovial fluid may contribute to disease severity (Dominical et al., 2011).

Beyond autoimmune diseases, there is a well-known association between inflammation and cancer, and neutrophils are present in high numbers in tumours where their infiltration is linked a worse prognosis (Jensen et al., 2009, Amulic et al., 2012). I also discussed the potential role of neutrophils in Alzheimer’s disease in Chapter 2 and more widely the contribution of inflammation to other complex diseases including coronary artery disease and age-related macular degeneration.

3.1.3 Functional phenotypes

The observed dysregulation of neutrophil function in multiple immune disorders makes the therapeutic targeting of these functions attractive. Understanding the mechanisms of neutrophil function and how these processes can lead to host tissues damage is an important step in enabling their therapeutic manipulation (Mayadas et al., 2014). Genetic studies of these cells afford the opportunity to discover new biological pathways involved in function, which could aid the identification of potential intervention targets.

In Chapter 1, I described how this goal is helped by using molecular phenotypes to understand disease- and complex trait-associated loci. Such approaches have already been successfully applied to neutrophils, as I highlighted in Chapter 2 with the BLUEPRINT consortium (Chen et al., 2016a). Another recent study also demonstrated the value of studying stimulated neutrophils by assaying neutrophil gene expression measured using a microarray (Naranbhai et al., 2015). The authors identified that 30% of 9,147 genes tested had at least one significant cis-eQTL (Naranbhai et al., 2015). Interestingly, many of these genes were known to function in central processes in neutrophil biology including differentiation, trafficking, granule formation, cytokine secretion, respiratory burst, phagocytosis and migration (Naranbhai et al., 2015). Some differences were observed with stimulated neutrophils, for example, rs1981760 was an eQTL for *NOD2* in unstimulated neutrophils but regulated the expression of the interferon β gene, *IFNB*, in neutrophils stimulated by the NOD2 ligand muramyl dipeptide (Naranbhai et al., 2015). Interferon β is involved in response to NOD2 activation, showing rs1981760 acts at multiple stages of the NOD2 pathway in resting and activated cells. The variant is associated with the risk of the bacterial disease, leprosy (EA = T) but is protective for Crohn’s disease (Naranbhai et al., 2015).

In efforts to gain further insight into immunology and its genetic control, there has been a recent expansion in the type of phenotypes studied using genetic approaches. For example, measuring the immune cell production of cytokines in the blood, which has been shown to be highly heritable (Brodin et al., 2015). Indeed, studying protein-level intermediates provides a comprehensive picture of functional variation, particularly as the previous integration of eQTL and protein (p)QTLs showed that some gene effects are buffered at the protein level (Battle et al., 2015). In Chapter 1, I discussed some examples of genetic studies using cellular phenotypes.

Measuring functional phenotypes in stimulated conditions is particularly important for studying immune function. The observed variation in cytokine responses was higher when blood cells were stimulated by a range of physiological stimuli than when compared to the resting state (Li et al., 2016a). Combining loci associated with protein-level and molecular phenotypes allows identification of more of the functional steps involved in the pathway from sequence variation to organismal traits. For example, rs11141235 was associated with IL6 levels after *Candida* stimulation and through using gene expression data acquired in PBMCs similarly stimulated, the locus was associated with expression of the gene, *GOLM1*, encoding the Golgi membrane protein 1 (Li et al., 2016a). Using a patient cohort, the authors further demonstrated that the *GOLM1* locus was associated with candidemia, suggesting susceptibility is the result of genetically modulated *GOLM1* gene expression and altered IL6 cytokine production (Li et al., 2016a). This genetic approach, therefore, highlights potential novel genes involved in cytokine responses and infection susceptibility.

Cohorts of functional data, therefore, have demonstrated that it is possible to study natural variation in a wide range of intermediate traits. These allow the identification of variants independently of their effects (or lack of effects) on molecular phenotypes, but also provide additional datasets with which to further annotate variants and move closer to the full description of regulation from variant to disease. As yet, there have been no large-scale efforts aimed at studying neutrophil functional phenotypes, due in large part to the technical complexities associated with working with these cells. However, given the central role of these cells in the immune response, such an approach could be highly impactful in informing our knowledge of neutrophil function.

3.1.4 Aims of this chapter: Investigating neutrophil functional responses

In this chapter, I aimed to build on recent efforts to reproducibly measure immune functions and subsequently identify genetic variants associated with functional readouts in healthy individuals. Rather than using a heterogeneous mix of blood cell types, we aimed to specifically study neutrophil responses, given the importance of these cells and their limited inclusion in genetic studies to date. Further, in generating a complementary neutrophil functional genetic dataset to the already established BLUEPRINT epigenome, we hoped to provide additional information for annotating genetic loci of immune and disease interest (Chen et al., 2016a).

I focused on three neutrophil functional responses that represented key stages in activation of these cells; adhesion, degranulation and respiratory burst (Figures 3.3-3.5). I summarise the study design in Figure 3.2 below. Experimental measurement of these responses was carried out by our collaborators Kate Downes and team at the NHS Blood and Transplant Department. Here, I implemented the analytical exploration of what represented the first application of these traits to larger healthy cohorts, having previously been used to study neutrophils from patients with rare disorders (Anton Tool and Taco Kuijpers, Sanquin

Research, the Netherlands). First, I selected parameters representing biologically meaningful estimates of functional response across the whole cohort. Next, I investigated the technical reproducibility of these assays and the effect of known covariates. I then explored possible biological relationships between these functional phenotypes and last assessed whether any observed variation in the responses can be explained by identified genetic variants.

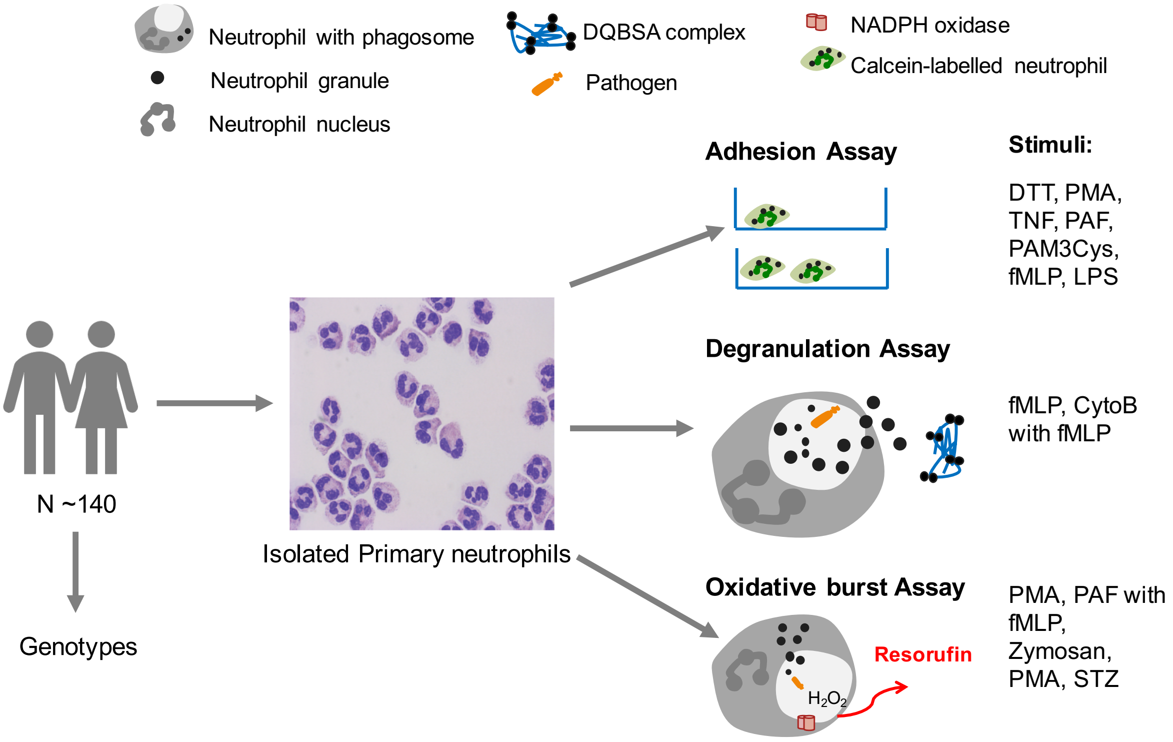


Figure 3.2: Neutrophil function study design

Approximately 200 healthy blood donors were recruited and primary neutrophils were isolated to a high purity (Materials and methods). Three assays were then performed per donor with a range of stimuli to activate neutrophils. DNA was extracted and either processed by whole genome sequencing (part of BLUEPRINT) or genotyped as part of a larger Cambridge BioResource cohort.

3.2 Materials and Methods

3.2.1 Neutrophil function data collection and experimental assays

Neutrophil adhesion, degranulation and respiratory burst were experimentally measured by our collaborators at the NHSBT at Addenbrooke's hospital (Kate Downes and Team). Here, I discuss the technical details of these assays as they are relevant to my quality control analysis and data exploration. To briefly summarise, for each assay fluorescence emitted from different fluorophores was measured for each individual using a plate-reader (Tecan, Infinite F200 PRO), and this represented the strength of neutrophil response.

Sample collection and cell isolation: All sample collection and neutrophil purification was performed at the NHSBT and details of this process are detailed in (Chen et al., 2016a). Briefly, neutrophils were purified from whole blood using a series of Percoll gradients. The resulting cells (neutrophils and eosinophils) were washed, and neutrophils were positively selected using CD16 microbeads (Miltenyi) according to the manufacturer's instructions. An average purity of CD66+CD16+ neutrophils was 98% as assessed by multicolour flow cytometry. Donors were obtained as part of the NIHR Cambridge BioResource (http://www.cambridgebioresource.org.uk/). Peripheral blood samples were collected from healthy donors with informed consent (A Blueprint of blood cells, REC 12/EE/0040, East of England-Hertfordshire Research Ethics committee). After broad outlier exclusion (see below), the cohort consisted of 138 donors where genetic data was available. There were 82 females and 56 males within the cohort. The age of donors ranged from 22 to 75 with 75% of the cohort falling between the ages of 55 and 75. For 87 donors (59%) whole genome sequence data was available as part of the Blueprint consortium cohort and for 102 donors (73%), genotype data from genotyping arrays was available (see below). The data were collected over a period of one year, with between one and four donors measured per day. In 56% of cases, a single donor was measured per day, and in 7% of cases, four donors were measured on the same day. Other exclusions, for example, outliers of the specific assay responses, are explained below.

Adhesion assay: The adhesion assay measures activated neutrophils adhering to a plate as a model system for circulating neutrophils attaching to endothelial cells, a process which is essential for neutrophils to access infected or damaged tissues (Figure 3.3). Neutrophils were first labelled with calcein-AM, supplied in 50 μg/vial. 12.5 μL of DMSO was added to one vial 50 μg vial of calcein-AM. Cells were resuspended in HEPES buffer at 5.0 x 10^6 per ml. 1 μL of the calcein-AM mix was added per ml of cell suspension. The suspension was incubated in a shaking water bath for 30 minutes at 37°C and after 15 minutes tubes were shaken by hand and replaced. 12 mL of PBS was then added and the suspension was centrifuged at 1500 rpm for five minutes, resuspended again in 12 mL of PBS and centrifugation was repeated as before. Finally, cells were resuspended in HEPES buffer and the concentration determined. For the adhesion assay, a calcein-labelled cell concentration of 2 x 10^6 per mL was used, and 80 µL of cell suspension was added to each well of a 96-well maxisorp plate (Fisher, DIS-971-090X). Eight different stimuli were used in this assay. Final concentrations used to stimulate the neutrophils on the plate were as follows: 1 µM PAF (Sigma, P4904), 10 ng/mL TNF (PeptroTech, 300-01A), 20 µg/mL Pam3Cys (EMC microcollections, L2000) 10mM DTT (Sigma, D0632) 10 µM fMLP (Sigma, F3506), 1 µg/mL PMA (Sigma, P8139), 20 ng/mL LPS (Sigma, L2880) and 50 ng/mL LBP (R&D Systems, 970-LP-025). After addition of the stimulus, the plate was covered with a sealer and incubated in a 37^0C CO_2 incubator for 30 minutes to allow activated neutrophils to adhere to the plastic surface of the plate. Neutrophil adherence to the plate is a known non-specific interaction mediated through the neutrophil integrin receptor, CD11b/CD18, that is blocked by antibodies against CD11b or CD18 (Anton Tool, personal communication). After 30 minutes, non-adherent neutrophils were washed from the plate using 100 µL of PBS at room temperature (RT). 100 µL of 0.5% Triton X-100 (Sigma, T8787) was then added to each well that contained cells, and the plate was incubated at RT for ten minutes. For the 100% input control, 20 µL of 2.5% of Trion X-100 was added to separate wells containing 80 µL of cell suspension. After incubation with Triton X-100, the plate was loaded onto the plate reader and one final fluorescence measurements was recorded. In all three assays, an unstimulated condition with the addition of only HEPES buffer was also measured on the plate reader.

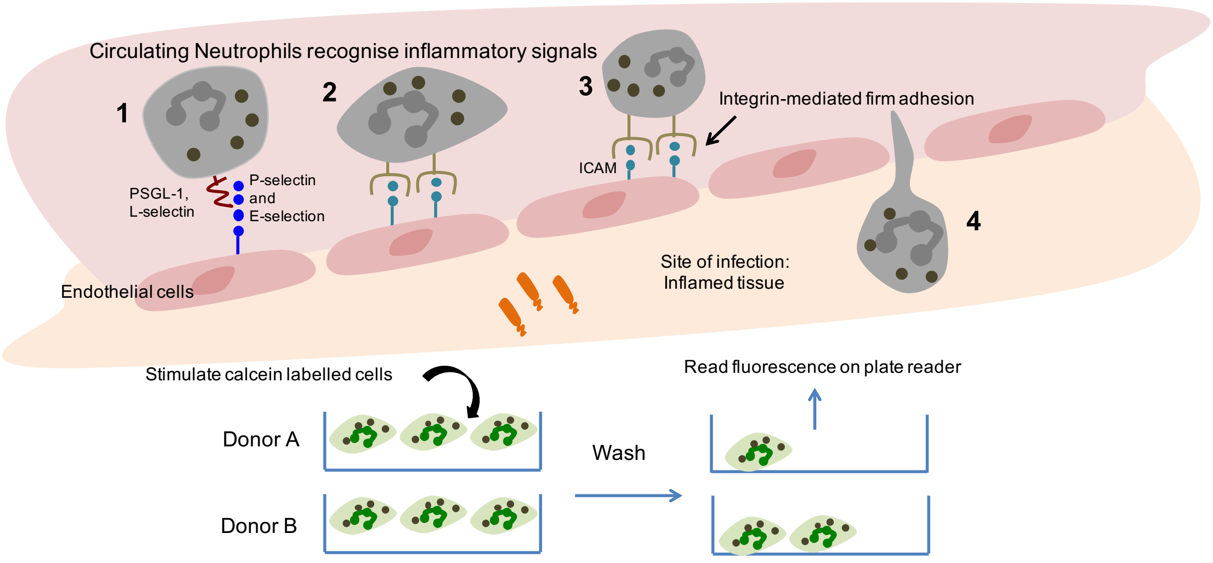

Circulating Neutrophils recognise inflammatory signals
1
2
3
4
PSGL-1,
L-selectin
P-selectin
and
E-selection
ICAM
Integrin-mediated firm adhesion
Endothelial cells
Site of infection:
Inflamed tissue
Stimulate calcein labelled cells
Read fluorescence on plate reader
Donor A
Donor B
Wash

Fig 3.3: Neutrophil adhesion
This schematic shows the biological process of adhesion and experimental measurement (lower panel). Neutrophils are recruited to inflamed tissues described in four-stage process. Initial contact is mediated through selectins (1), initial tethering mediated by the constitutively-expressed neutrophil molecules, PSGL-1 and L-selectin. P-selectin and E-selectin are expressed by endothelial cells under infection or inflammatory conditions. (2) neutrophils then roll along the endothelial wall leading to strong adherence mediated by integrins (3) (Zarbock and Ley, 2008). During the rolling process, other receptor interactions activate further signalling processes in neutrophils to initiate neutrophil extravasation, cytoskeletal rearrangement leading to the release of neutrophil cytotoxic granules and production of reactive oxygen species. The neutrophil crosses the endothelial cell wall in a process known as transendothelial migration (4). To assay this response, neutrophils are labelled with the fluorescent molecule, calcein-AM and activated with the stimuli in the plate well. Fluorescence of adherent neutrophils is measured by a plate reader. Figure adapted from (Amulic et al., 2012).

Respiratory burst assay: This assay measured the NADPH-oxidase activity of neutrophils known as the respiratory burst response. This is the production of reactive oxygen species (ROS) by an activated neutrophil (Figure 3.4). Hydrogen peroxide (H_2O_2) is produced by the neutrophil respiratory burst and reacts with the Amplex® Red reagent in the presence of horseradish peroxidase (HRP) (ThermoFisher Scientific, 2017). The 1:1 reaction of Amplex® Red and H_2O_2 produces the red fluorescent molecule, resorufin (Figure 3.4). Resorufin is excited by 571 nm and emits at 585nm enabling the measurement of the cellular production of H_2O_2 using a plate reader. For this assay, the responses are measured by the plate reader in real-time, rather than an end-point measurement. The fluorescence measured was produced not from labelled cells (as with adhesion) but from the production of fluorescent resorufin as a by-product of the stimulated functional response.

Unlabelled cell concentration used for the respiratory burst assay was 1 x 10^6 cells per mL in HEPES medium. The neutrophil cell suspension was pipetted into a black opaque 96 well plate (Fisher, DIS-210-190W). 100 µL of the 2x reaction mix containing a final concentration of 25 µM Amplex® Red (10-acetyl-3,7-dihydroxyphenozazine, Molecular Probes, A-12212) and 0.5 unit/mL HRP (Sigma, P-8250) was added to the plate along with 50 µL of cell suspension. The plate was then incubated at 37˚C on the plate reader for five minutes. 50 µL of each stimulus was then added to the relevant wells, and the reaction is then measured in the plate reader immediately after. There were four stimuli used in the respiratory burst assay and the final concentrations used in the assay are as follows: 4 mg/mL Zymosan (Sigma, Z4250), 1 mg/mL STZ (Sigma, P8139), 1 µg/mL PMA (sigma, P8139), 2.5 µM PAF (Sigma, P4904) and 25 µM fMLP (Sigma, F3506) where PAF and fMLP were added to the same well as one condition. The reaction was measured on the plate reader for 60 cycles (30 minutes). The range of different stimuli, both biological and small chemical molecules are described in Table 3.2 with a description of how these stimuli activate the respective neutrophil functions.

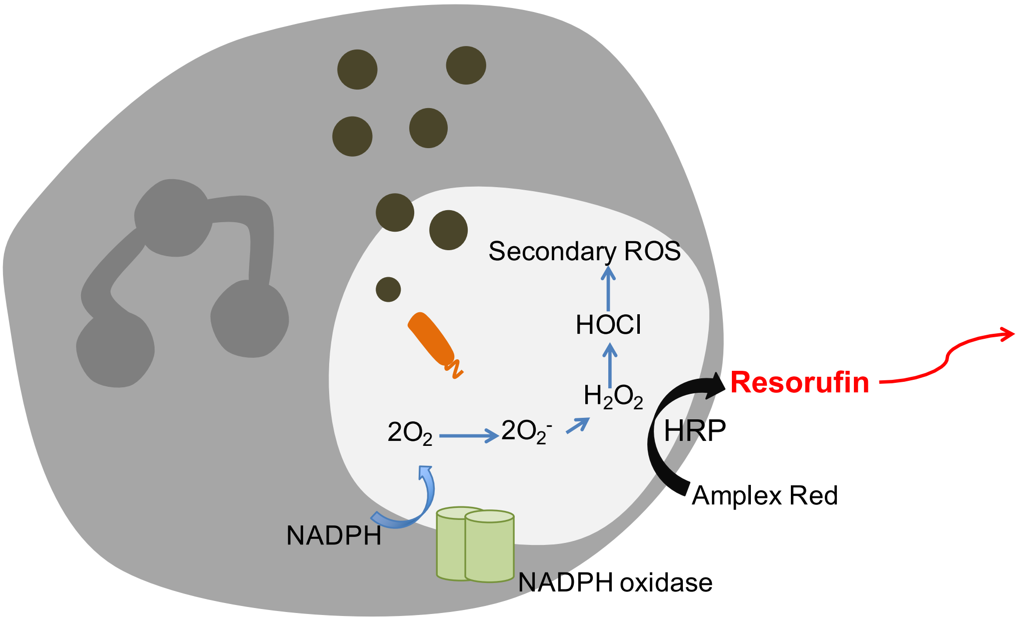


Fig 3.4: The neutrophil respiratory burst response
This schematic summarises the molecular reaction and experimental measurement of the respiratory burst response. Reaction of hydrogen peroxide (H_2O_2) with the Amplex® Red reagent in the presence of horseradish peroxidase (HRP) produces the red fluorescent molecule, resorufin. Resorufin can be excited by 571 nm and emits at 585nm enabling the measurement using a plate reader. Adapted from (Kobayashi et al., 2005).

Degranulation assay: This assay measured the release of granule contents that occurs when a neutrophil is activated (Figure 3.5). Table 3.1 lists the different neutrophil granules and their contents. Here, neutrophil degranulation was measured by using a complex of a DQ^{TM}Green BSA (DQBSA) (Molecular probes, D12050). In this form fluorescence from the green-fluorescent BODIPY® FL dye is quenched (ThermoFisher Scientific, 2017). Proteases released from the neutrophil digest the DQBSA molecule so that the fluorescence is no longer quenched and can be measured by the plate reader. The experimental process mirrors the recognition and internalisation of pathogens by neutrophils that activates a series of molecular processing leading to the release of neutrophil granular contents into the phagosome and also the surrounding cellular environment. The release of antimicrobial peptides and other immune-related molecules leads to the destruction of the pathogen and recruitment of further immune cell types.

HEPES medium was added to the relevant wells in the plate (black opaque 96-well plate as above for respiratory burst assay). 50 µL of unlabelled cells at 5 x 10^6 cells per mL were added to the wells containing HEPES. A final concentration of 10 µg/mL of DQBSA was added to wells containing cells. For the stimuli, a final concentration of 5 µg/mL Cytochalasin B (CytoB, Sigma, C6762) was used, and 10 µL of this solution was added to the relevant wells containing neutrophil cell suspension and DQBSA. The plate was then incubated for five minutes at 37˚C in the plate reader. The plate was then removed and 10 µM fMLP added to the relevant wells, where neutrophils were stimulated with a combination of CytoB and fMLP. After activation, the released DQ^{TM} Green fluorophores released are excited at 505nm and emit at 515nm. The reaction was measured in the plate reader for 120 cycles (60 minutes).

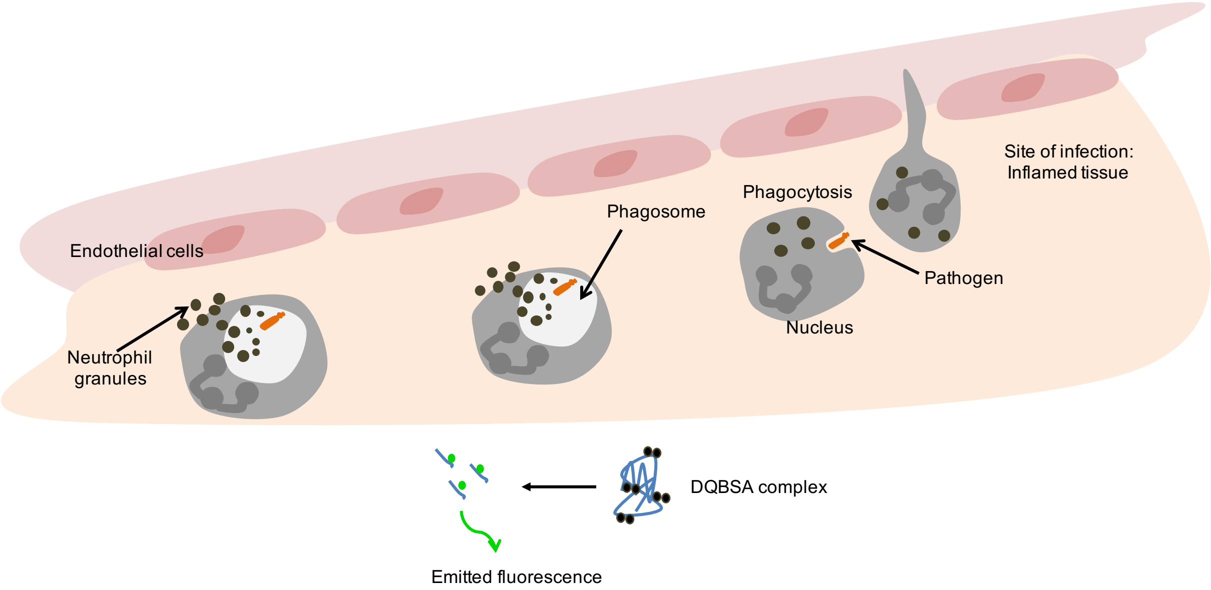


Fig 3.5: Neutrophil degranulation
This schematic shows the biological process and experimental measurement of degranulation, which is the release of granule contents. The response from purified neutrophils in solution is measured using a fluorescent marker. Released neutrophil proteolytic enzymes break up the DQBSA molecule so that DQ is no longer quenched and the fluorescence can be measured using a plate reader. Adapted from (Amulic et al., 2012).

3.2.2 Data interpretation and exploration

Given the novelty of the traits, we carried out extensive exploration of the data to identify potential artefacts. These next sections describe the process used to evaluate the utility of these traits in a genetic context and was carried out with close collaboration and supervision by Klaudia Walter.

Initial exploration of data over time: To assess the variability of the measures over time, I plotted each trait value by date of acquisition across the 12 months. As a first diagnosis, we assessed traits as calculated by and provided by the plate reader software. These were relative fluorescence units measured at cycle 20 and 40 for respiratory burst and degranulation respectively. For adhesion RFU measurement was an end-point value after a 30 minutes incubation (see above). Figure 3.6 also shows the two replicates of the 100% input control used in the adhesion assay and measured for each donor. The 100% input control is the measured calcein fluorescence (in RFU) from labelled cells after addition of a high concentration of Triton-X100 (Materials and Methods), which should release calcein from the cells and reflect a high adhesion response. This value is used to normalise the adhesion response for all other stimuli by dividing the stimulated RFU by the 100% input control (four of the total eight stimuli are shown in Figure 3.6 and listed in Table 3.2).

We observed a substantial shift in trait values and distribution between the first 68 samples and those acquired after this point (Figure 3.6). In particular, for the adhesion response, the HEPES values were inflated compared to the samples acquired later (Figure 3.6). This elevated HEPES response level could reflect possible bacterial contamination in the original HEPES batch, resulting in neutrophil activation. The HEPES buffer was used in all parts of the experimental process including stimulus dilution. Therefore, the decision was taken to remove the first 68 samples.

Selection of real-time assay parameters: One major challenge for the respiratory burst and degranulation real-time assays was the selection of a comparable measurement that can be used as a phenotypic trait for subsequent genetic studies. We explored the raw response distributions to select parameters that would best capture the highest dynamic range in response across the cohort yielding the largest resolution (Figure 3.7).

The shape of the response distributions often varied considerably between stimuli (Figure 3.7). For example, the respiratory burst response stimulated by Zymosan, a yeast particle, shows a slower activation profile than with serum treated Zymosan (STZ). STZ is Zymosan opsonised with immunoglobulin (Ig) and complement receptors that together stimulate a faster and higher neutrophil response by activating the integrin receptor CD11b/CD18.

Addition of Cytochalasin (CytoB) with fMLP also elicits a higher response (Figure 3.7). fMLP stimulates the production of diacylglycerol (DG) by phospholipase C (PLC). DG activates protein kinase C (PKC), a kinase that has been suggested to be involved in both ROS production and degranulation (Sato et al., 2013). CytoB has been shown to increase the diacylglycerol-mediated response stimulated by fMLP in neutrophils and therefore augments the fMLP response (Honeycutt and Niedel, 1986).

We combined the observed reaction distributions with prior knowledge regarding the biological relevance of reaction ranges (Anton Tool, personal communication) to calculate a range of parameters directly from the raw responses for both the respiratory burst and degranulation assays. For example, we were advised that the respiratory burst reaction stimulated by PAF combined with fMLP, is an extremely rapid reaction reaching saturation within minutes. Therefore, we calculated parameters for the PAF & fMLP response within the first ten cycles (five minutes) of the reaction to avoid missing the oxidative burst peak.

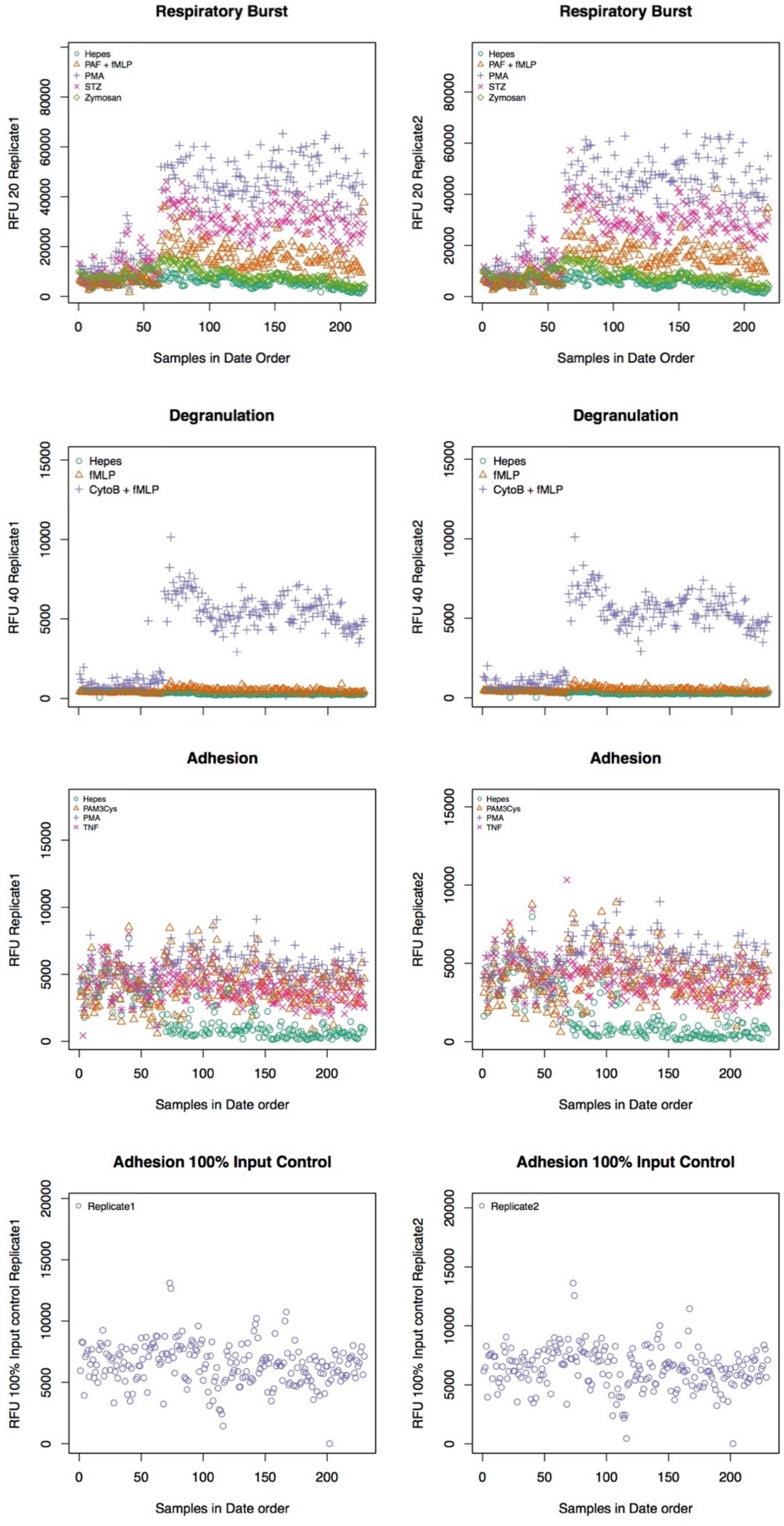
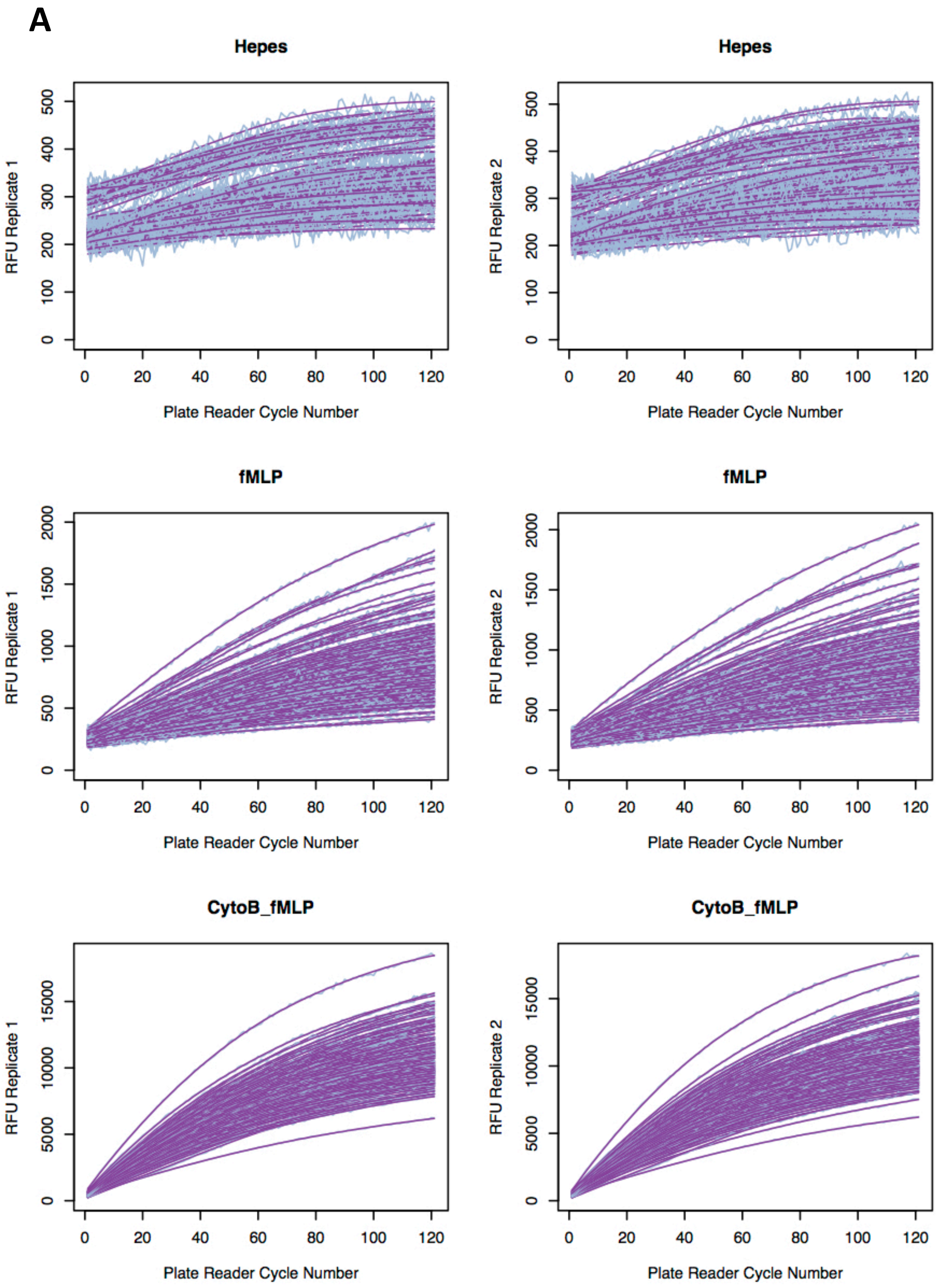
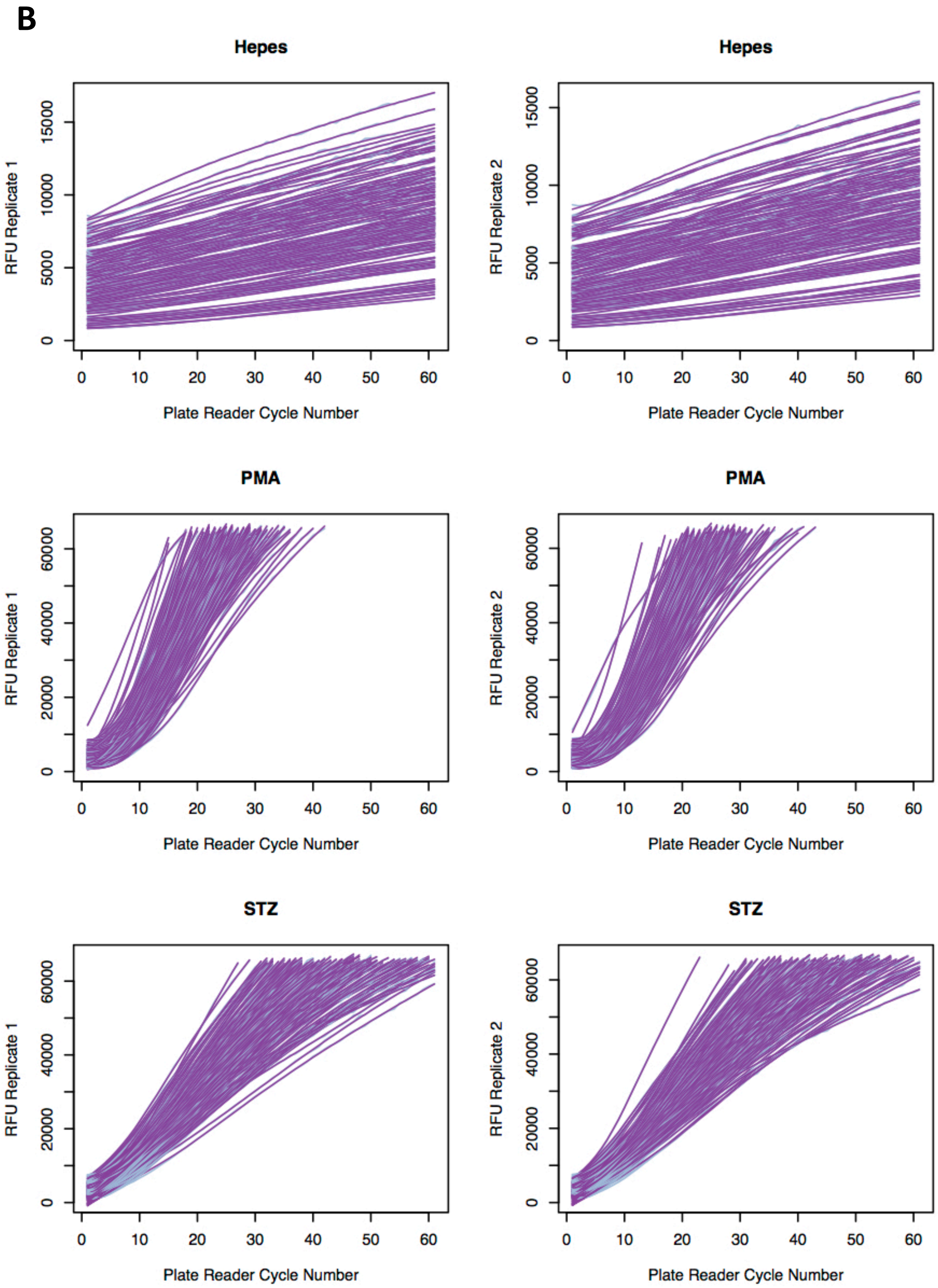

Respiratory Burst
Hepes
PAF + fMLP
PMA
STZ
Zymosan
RFU 20 Replicate1
0
20000
40000
60000
80000
0
50
100
150
200
Samples in Date Order
Respiratory Burst
Hepes
PAF + fMLP
PMA
STZ
Zymosan
RFU 20 Replicate2
0
20000
40000
60000
80000
0
50
100
150
200
Samples in Date Order
Degranulation
Hepes
fMLP
CytoB + fMLP
RFU 40 Replicate1
0
5000
10000
15000
Degranulation
Hepes
fMLP
CytoB + fMLP
RFU 40 Replicate2
0
5000
10000
15000
Adhesion
Hepes
PAM3Cys
PMA
TNF
RFU Replicate1
0
5000
10000
15000
0
50
100
150
200
Samples in Date order
Adhesion
Hepes
PAM3Cys
PMA
TNF
RFU Replicate2
0
5000
10000
15000
0
50
100
150
200
Samples in Date order
Adhesion 100% Input Control
Replicate1
RFU 100% Input control Replicate1
0
5000
10000
15000
20000
0
50
100
150
200
Samples in Date order
Adhesion 100% Input Control
Replicate2
RFU 100% Input control Replicate2
0
5000
10000
15000
20000
0
50
100
150
200
Samples in Date order

Fig 3.6: Exploratory data analysis: measurements by time
These plots depict the raw data distributions for example RFU values as collected directly by the plate reader software. All donor values for each assay are plotted against the date of acquisition of that sample. The top panel shows an example measurement of the RFU at 20 cycles for all respiratory burst conditions and the second panel shows RFU at 40 cycles for all degranulation conditions. The third panel depicts some example stimuli for the end-point RFU reflecting the degree of neutrophil adhesion on the plate. The fourth panel shows the adhesion 100% input control against date of acquisition. In panels 1-3 it can be observed how the first 68 samples show a very different relative response distribution to the remaining samples.

Figure 3.7 shows that for an individual's response distribution and within each well, there were small fluctuations between successive cycles, particularly apparent for the HEPES degranulation response. These small changes were likely due to plate reader resolution but could potentially have introduced errors in calculating our maximum and mean gradients from the raw distribution. These gradients were used as a comparative measure of the response of each donor in the real-time assays. Mean and maximum gradients were calculated using two successive cycles forming a tangent against the curve for which the gradient was calculated. For mean slope traits within the defined interval, gradients were calculated for each of successive cycles, and the mean was calculated of all these values (all summarised in Table 3.2). To remove these cycle-to-cycle deviations within each individual distribution, we applied the LOESS curve-fitting function to smooth each distribution (Figure 3.7). LOESS is a non-parametric, local regression method, where the fitted points and standard errors are calculated with respect to the whole reaction distribution. An estimated curve is fitted to the reaction distribution curve. This was implemented using the loess() R function. Parameters were then calculated for each replicate separately and then averaged across trait replicates. The function did not effectively fit the PAF fMLP respiratory burst response before cycle 10 (data not shown) due to the rapid reaction in the early cycle-stages resulting in steep response gradients. This rapid reaction also meant that in the early part of the reaction, the response distribution was already smooth and therefore, the raw data without LOESS fitting was used to calculate the parameters for this response. The unsmoothed and smoothed distributions are shown in grey and purple respectively in Figure 3.7.

For adhesion data, there was an extra normalisation step with division of RFU by the 100% input control from the stimuli RFU. Following this, I calculated the averaged across the two technical replicates as described for the other two assays. In total, there were 29 traits across all three assays in the final list of parameters (Table 3.2).


A
Hepes
RFU Replicate 1
0
100
200
300
400
500
0
20
40
60
80
100
120
Plate Reader Cycle Number
Hepes
RFU Replicate 2
0
100
200
300
400
500
0
20
40
60
80
100
120
Plate Reader Cycle Number
fMLP
RFU Replicate 1
0
500
1000
1500
2000
0
20
40
60
80
100
120
Plate Reader Cycle Number
fMLP
RFU Replicate 2
0
500
1000
1500
2000
0
20
40
60
80
100
120
Plate Reader Cycle Number
CytoB_fMLP
RFU Replicate 1
0
5000
10000
15000
0
20
40
60
80
100
120
Plate Reader Cycle Number
CytoB_fMLP
RFU Replicate 2
0
5000
10000
15000
0
20
40
60
80
100
120
Plate Reader Cycle Number



Hepes

Hepes

PMA

PMA

STZ

STZ

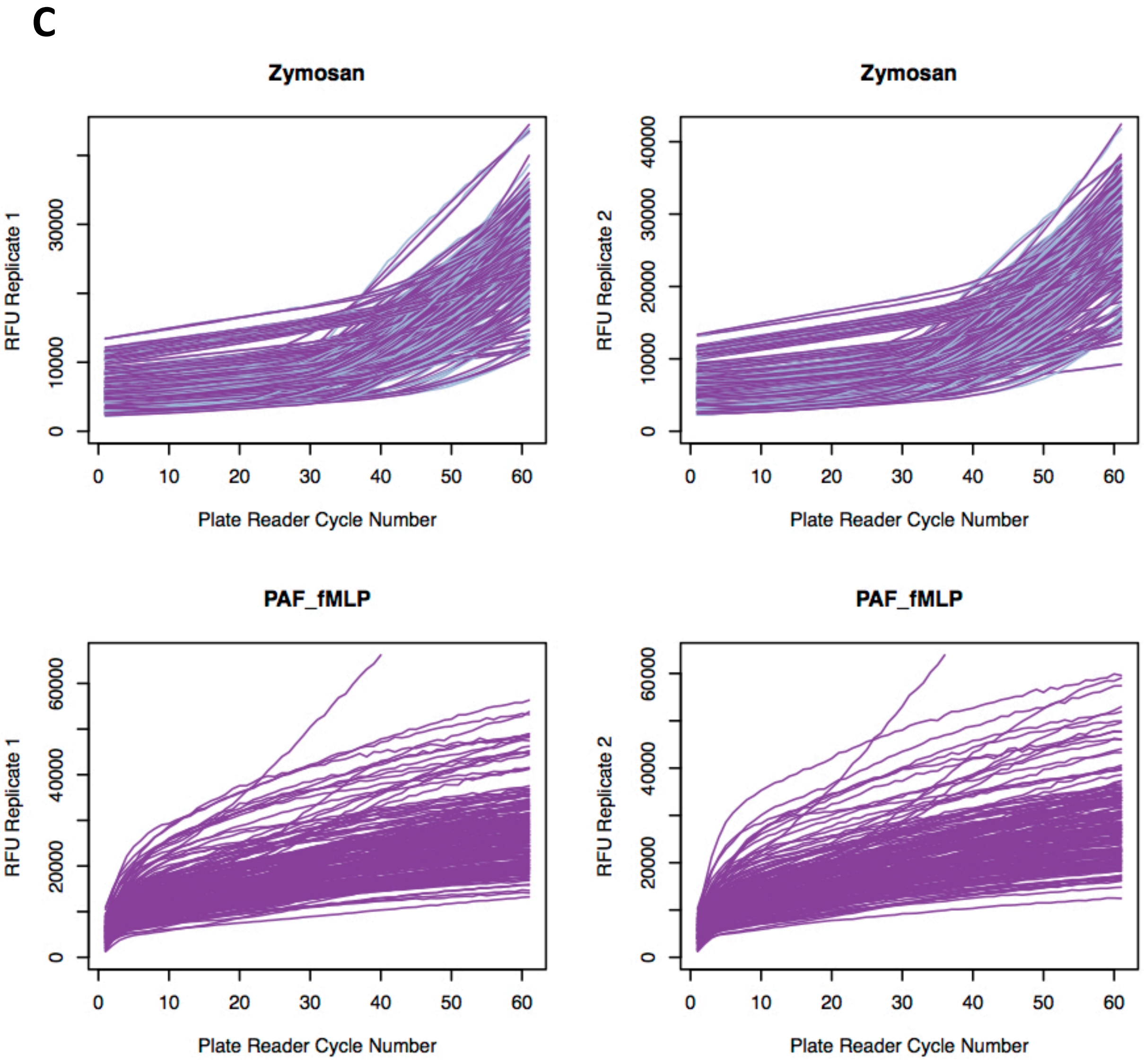


Fig 3.7: Raw data response distributions for respiratory burst and degranulation
These plots show the raw RFU at each cycle for replicate 1 and replicate 2 for every stimulus starting from the first measurement at cycle 1 up to the final number of cycles (60 or 120). Measurements are recorded every 30 seconds (one cycle) plotted on the x-axis. Fluorescence units are measured cumulatively. HEPES buffer is used in each assay as a measure of unstimulated functional response. Each line represents the whole RFU response distribution for one donor. Above is shown the response distributions for all donors in the cohort excluding the first 68 samples. A) Cumulative fluorescence recorded by a plate reader for the degranulation assay for each stimulus used. For the second HEPES replicate there was one donor which recorded a very low HEPES response, near to zero. It was suggested that this could be due to a technical artefact, therefore for this donor, the value for replicate 1 was used in place. B-C) Fluorescence measurements as recorded by a plate reader for the respiratory burst assay for each stimulus used. The raw relative fluorescence values (RFU) from the plate reader are plotted in grey. The values were fitted using the LOESS smoothing function and are shown in purple on the same plot. For respiratory burst responses stimulated by PAF fMLP the plot is shown with raw data. The LOESS function did not fit the raw data well below 10 cycles and so was not used to calculated the trait values. The individual response showing a very different reaction distribution, rising very quickly outside of the other donor responses, in the PAF & fMLP reaction (C bottom panel), was removed (discussed below).

Assay	Stimulus	Trait	N
Respiratory burst	PMA: phorbol myristate acetate, induces NADPH oxidase by direct stimulation of protein kinase C (PKC)	RFU at 10 cycles	130
		RFU at 15 cycles	130
		Time to RFU at 40000 cycles	130
		Mean Slope: 10 to 15 cycles	130
		Maximum Slope	130
	STZ: serum-treated Zymosan, which is opsonised with Ig and complement receptors, involves the neutrophil CD11b/CD18 and Fcγ receptors	RFU at 15 cycles	129
		RFU at 20 cycles	129
		Time to RFU at 30000	129
		Mean Slope: 10 to 20 cycles	129
		Maximum slope	129
	Zymosan: cell wall preparation from *S.cerevisiae,* induces NADPH oxidase via CD11b/CD18	RFU at 60 cycles	130
		Difference in RFU: 60 and 1 cycles	130
		Maximum Slope	130
	PAF + fMLP: PAF amplifies the RB response. fMLP stimulates the fMLP receptor, which activates NADPH oxidase via p47phox	RFU at 6 cycles	127
		Maximum Slope	128
Degranulation	fMLP: N-formyl peptide released by bacteria, bacterial degradation or mitochondrial protein. Stimulates release of gelatinase granules	RFU at 20 cycles	135
		Mean Slope: 20 to 40 cycles	135
		Maximum Slope	135
	Cytochalasin B + fMLP: CytoB, a fungal mycotoxin, inhibits actin polymerisation in cells and amplifies azurophilic and specific granule-mediated degranulation. Generates an increased degranulation response compared to fMLP alone	RFU at 20 cycles	135
		Mean Slope: 20 to 40 cycles	135
		Maximum Slope	137
Adhesion	PMA: activates partially by the stimulation of PKC	Final RFU	131
	PAF: platelet activating factor, binds to the GPCR, PAF receptor	Final RFU	131
	fMLP: activates adhesion through the fMLP receptor	Final RFU	131
	DTT: dithiothreitol reducing agent. Breaks disulphide bridges and activates integrin receptors	Final RFU	129
	TNF: tumour necrosis factor, activates adhesion via the TNF receptor	Final RFU	132
	LBP + LPS: LPS is a bacterial molecule that binds TLR4, LBP is LPS binding protein, an acute phase protein that binds to bacterial LPS to stimulate an immune response	Final RFU	131
	Pam3Cys: TLR1/2 agonist	Final RFU	132
	Hepes: control buffer, may reflect pre-stimulated adhesion	Final RFU	128

Table 3.2: Functional traits and number of individuals for each trait across all three assays and stimuli used
Table describes the different stimuli and their mechanism of action in each assay. The final traits used for each assay is also listed and the number of individuals (N) after final outlier removal within each trait is listed. A full understanding of the signalling pathways involved is not yet established, presented here is a mix of published and unpublished observations from our collaborators (Anton Tool, Taco Kuijpers).

Removal of outliers and technical reproducibility: For each individual, two repeat measurements were acquired where cell suspension from the same donor was added to two wells of the plate (referred to as technical replicates). I used the R function, cor.test to calculate the Spearman correlation between technical replicates across all individuals, for each trait. Before outlier removal the correlation averaged across all traits was high (respiratory burst rho= 0.906, degranulation rho= 0.972, adhesion rho= 0.956), suggesting good technical reproducibility.

I next used these technical replicates to remove outlier measurements or extreme measurements using two thresholds. First, outliers beyond a threshold of 5 x standard deviation (5SD) between technical replicates were excluded. Second, I calculated a 3SD mean *distribution* threshold, to assess potential outliers with very high or low trait values, which could reflect extreme responders or technical artefacts. To differentiate between true high and low responders and technical outliers, I kept measurements that were outside of the mean distribution thresholds but well replicated (inside the 3SD replicate thresholds).

For the adhesion assay, outliers were first excluded using the above thresholds from the un-normalised stimulated RFUs if outside the two thresholds and also from the for raw 100% input RFUs. Finally, outliers were also excluded after normalisation with the 100% input control if this generated additional outliers by normalising with very high or low input control values.

We further investigated four donor measurements that generated very extreme responses that lay far beyond the 3SD mean distribution threshold in most adhesion conditions (data not shown). We identified that these four donors were measured on the same day, the 22nd April 2014 and were the only donors processed using a single batch of a reagent Buffer 3, which is used in the neutrophil purification process. We excluded these donors from further analysis. The extensive investigation into batch effects is discussed below.

I also excluded donors for which there was no available genetic data and then inverse normalised the trait values across the whole cohort to generate a normalised trait distribution.

Covariate investigation: The identification of the four adhesion outliers described above suggested that these functional data may be subject to variation introduced by experimental covariates. We were able to investigate known experimental covariates such as reagent batch as changes were recorded during data acquisition. I also investigated the effect of environmental factors such as age, gender and season. Batch effects, which are sub-groups of trait values that exhibit different behaviour that is not related to genotype, must be removed to reduce noise, improve power and avoid systematic stratification that can cause bias in association testing (Leek et al., 2010).

In order to investigate the effect of season, we assigned the trait values to the different seasons based on when they were experimentally measured. These included: Winter (Dec-Feb), Spring (Mar-May), Summer (Jun-Aug) and Autumn (Sep-Nov). In addition, if multiple donors were processed on one day (in some cases up to 4), all samples were read on one plate-reader plate and positioned from left to right. The fluorescent signal from each well is also read by the plate reader from the left to right, but the reaction was started once the stimulus was added to the plate prior to placing in the plate reader. We annotated each measurement with plate position to investigate whether machine reading could be a source of co-variation.

I stratified trait values by these potential covariates to visualise the potential effect and Figure 3.8 shows these effects for the adhesion traits as an example. I used inverse-normalised trait values after removal of donor outliers and donors with missing genotypes, to allow better comparison between traits. Patterns for inverse normalised and raw data were similar (data not shown). I calculated the significance of the effect of the covariate by using the one-way ANOVA, using the aov R function, to test if the means of each group were significantly different (shown in Figure 3.8). Certain covariates, such as buffer 3 significantly affected all traits from all three assays, except adhesion PAF (p value = 0.091) and respiratory burst PMA Time to RFU 40000 (p value = 0.11) (Figure 3.8). As described above, buffer 3, was used in the purification of neutrophils from whole blood and also had multiple batches.

I observed that some covariates, such as season, showed variable significance across the different traits. In the case of adhesion, season has a significant effect on the response stimulated by LBP and LPS (p value = 1.6 x 10^{-07}), HEPES (p value = 7.6 x 10^{-07}), TNF (p value = 8.6 x 10^{-03}) and DTT (p value = 2.2 x 10^{-03}) but not Pam3Cys (p value = 0.18), PAF (p value = 0.25), PMA (p value = 0.15) or fMLP (p value = 0.65). The LBP/LPS response, for example, was lower in winter than it was in spring. LPS is the major component of gram-negative bacterial outer membranes, eliciting an anti-bacterial neutrophil response. In addition, the HEPES response, which may represent the pre-stimulated neutrophil adhesion response was also lower in winter. For the respiratory burst, season as a significant effect on PMA RFU.10 (p = 6.3 x 10^{-05}), PMA RFU 15 (p = 8.2 x 10^{-03}), STZ RFU 15 (p = 4.4x 10^{-06}), STZ RFU 20 (p = 2.2 x 10^{-04}), STZ Time to RFU 30000 (p = 6.9 x 10^{-04}), PAF + fMLP (p = 6.5 x 10^{-06}), Zymosan RFU 60 (p = 1.0 x 10^{-11}), Zymosan Diff RFU 1.60 (p value = 2.2 x 10^{-19}) and Zymosan max slope (p = 8.2 x 10^{-17}). For these traits, the pattern was similar to adhesion, highest levels in summer, followed by spring and lowest in winter. For the degranulation response, season was a significant effect on all traits, with the peak response in spring and decreasing in winter (data not shown).

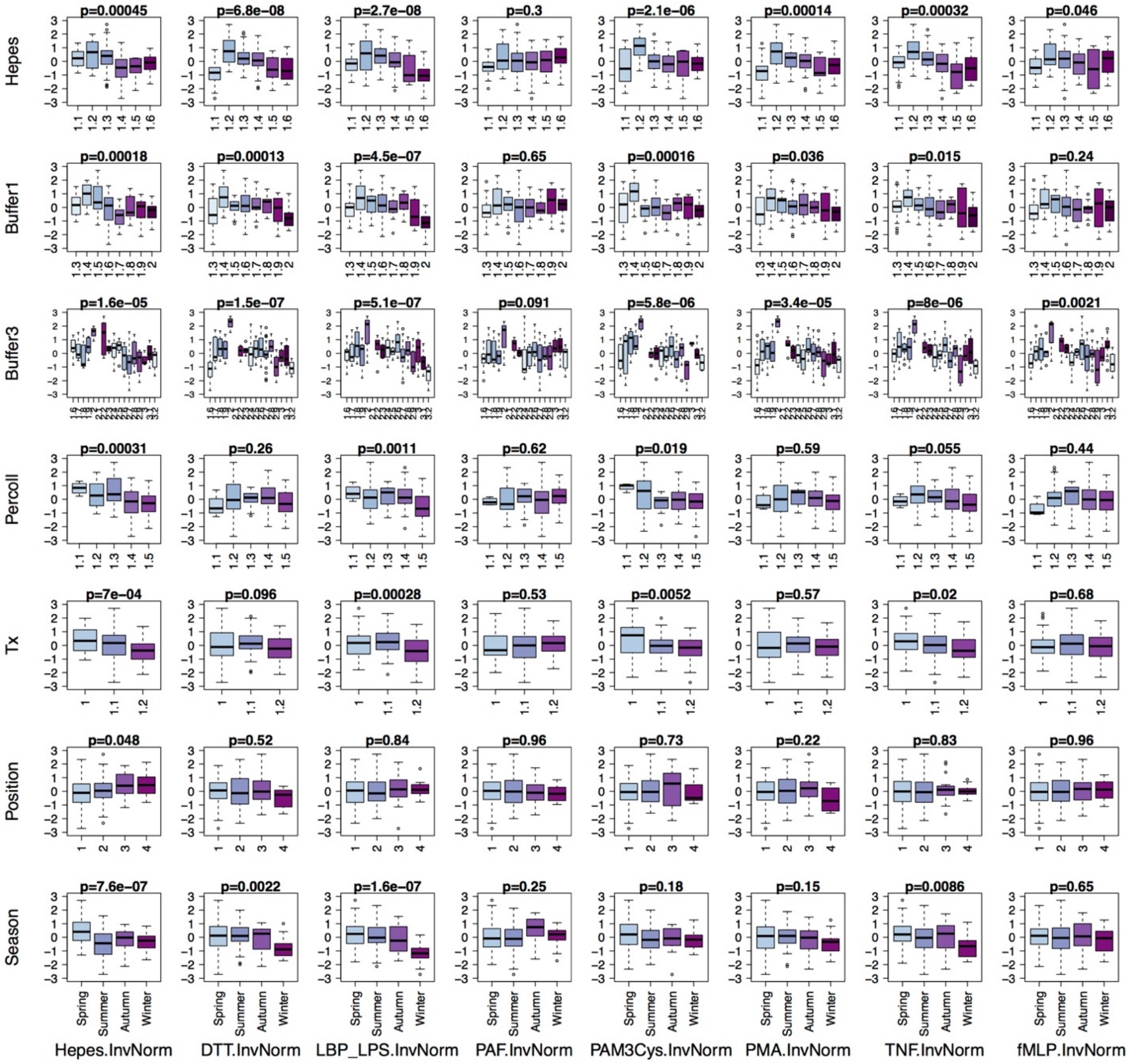


Figure 3.8: Reagent batch effect has a substantial effect on adhesion data
Inverse normalised trait values for neutrophil adhesion stimulated by a variety of stimuli (columns) are shown stratified by covariate levels (rows). HEPES, buffer1, buffer 3 and Percoll are all used in the purification of neutrophils. Triton (Tx) is used in the input control and in the main assay processing. Position explains the position of the donor cells in the well depending on the number of donors assayed. Season is the time of year at which the measurement was acquired.

To further evaluate the effect of each covariate on trait values, I calculated the R^2 value, which determines the percentage of the variation in response that is explained by the model. It is calculated by dividing the variation explained by the model by the total variation. The higher the R^2, the more trait variation is explained by the model. For each trait, I calculated the R^2 value for a linear model where the trait was the response variable and each covariate was fitted independently as the predictor variable. Here, R^2 is expressed as a percentage of trait variation where 0.1 is 10%. Figure 3.9 shows an example of R^2 of covariation for adhesion responses and the remaining traits in Supplementary Figure 3.9.

Reagent batch, particularly that of the neutrophil purification buffers 1 and 3 explain a high proportion of the total variation in each neutrophil trait value. The R^2 averaged across all traits from all three assays was 0.385 and 0.306 for buffer 3 and buffer 1 respectively. These high values underline the considerable effect that reagent batch had on these functional readouts. Interestingly, for all traits, season (mean $R^2 = 0.163$) demonstrated a greater R^2 than age (mean $R^2 = 0.029$) and gender (mean $R^2 = 0.036$) of the donor, suggesting that season contributes a reasonable degree of trait variation. In addition, age was not a significant covariate for any of the traits tested. The observation of the effect of reagent batch and to a certain extent, season, has important implications for the experimental design of future studies involving these cellular functional responses.

Covariate correction: Various approaches for correcting for these covariates were explored, but were often complicated by limited sample size (Klaudia Walter, data not shown). We decided to implement a conservative approach in correcting for covariates given the small sample size of this study, the large observed reagent batch effects and the lack of a complete, genome-wide replication cohort. We included all of the covariates in the trait correction process, termed the “full model”. This approach should lead to the lowest number of false positive associations in the downstream genetic analysis, as opposed to correcting only for those covariates with a significant ANOVA p-value or high R^2 value.

I used linear regression to correct for the technical variation related to these factors. Covariates were input as fixed effects and regressed onto the inverse normalised trait values using the R lm() function. The model fit was assessed using QQ plots (data not shown) and visualisation of the residuals both against time and stratified by covariates to assess whether the previous patterns and waves of variation were still present in the data. We concluded that the residuals generated from the linear regression demonstrated that a considerable proportion of the variation in the response had been removed (Figure 3.10). The mean of the trait values of each covariate batch was now similar, and the varying distribution of values has been removed (Figure 3.10).

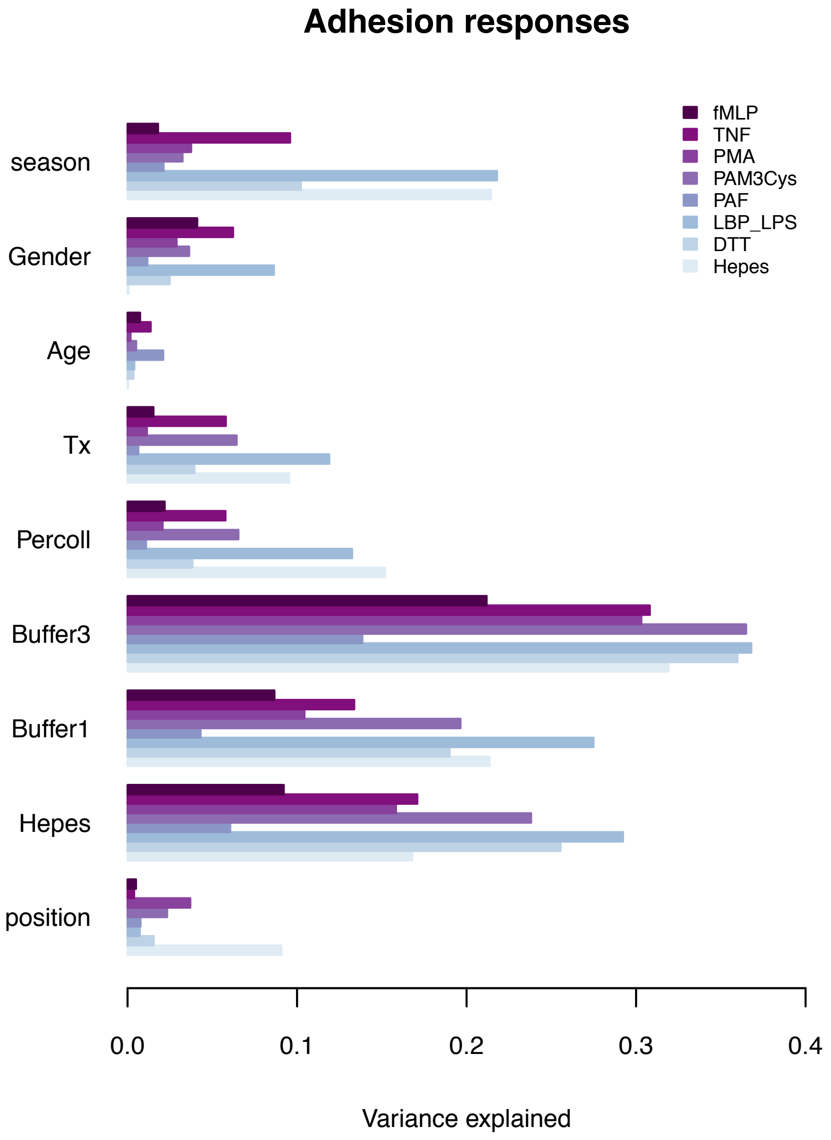


Figure 3.9: Reagent batch explains a high proportion of variation in neutrophil adhesion responses
Barplot shows the R^2 estimates calculated from fitting a linear model for each trait (coloured) independently with each covariate (y axis). This shows that reagent batch explains a high proportion of variance in the trait values.

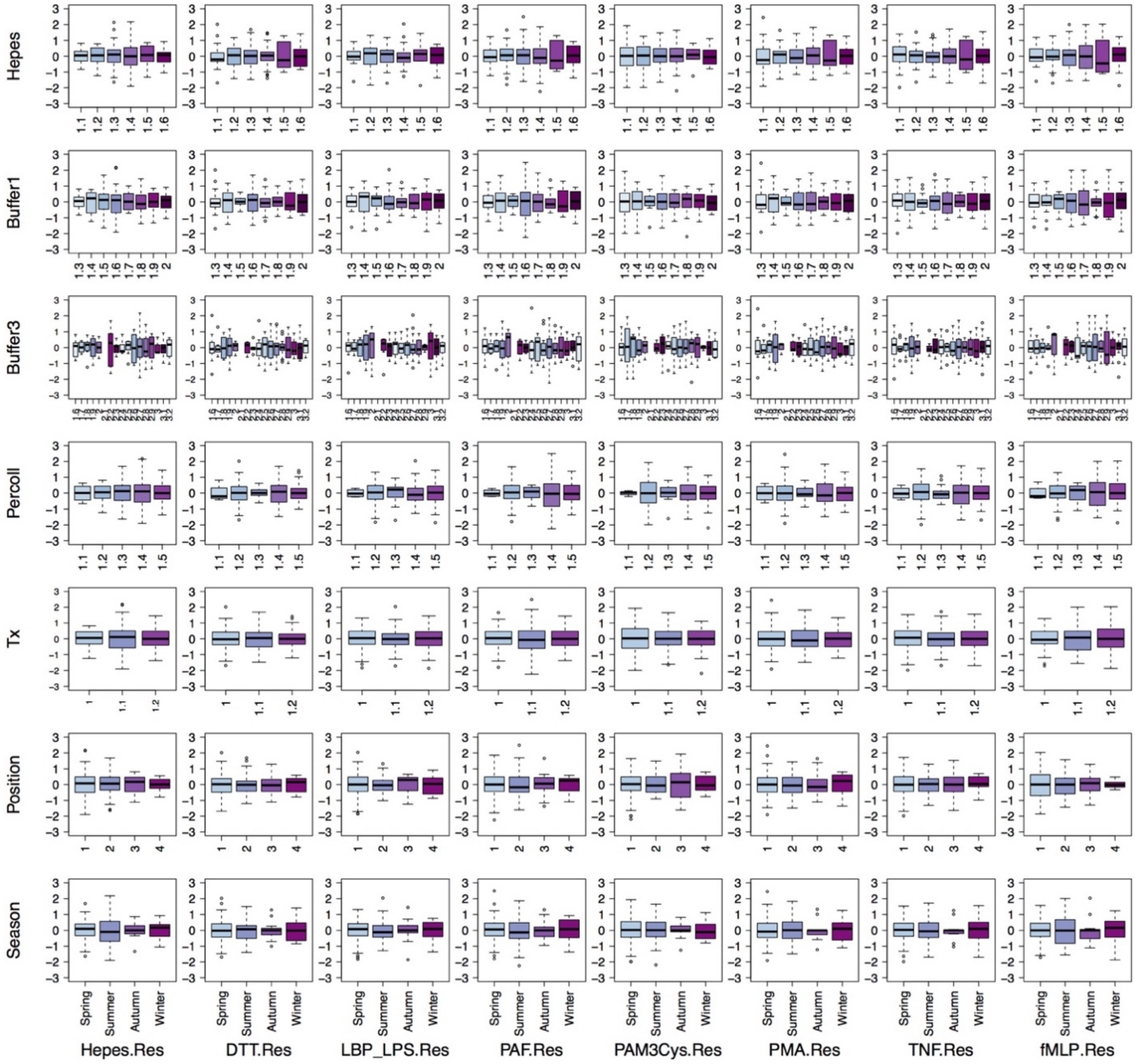


Figure 3.10: Correction of known batch effects
Boxplots show the residuals stratified by covariate levels. The residuals are shown from linear regression models with all of the listed covariates were applied to each trait to remove variation due to changes in these factors. season. This demonstrated that the means of each level were equalised even for Buffer 3 for which there were 17 batches.

Phenotype correlation: In order to investigate whether there was evidence of any similarity between phenotype traits, I used unsupervised clustering analysis by calculating Pearson correlations between all pairwise comparisons between residualised trait values. The R package, pheatmap, was used to plot the heatmap demonstrating clusters between particular traits.

3.2.3 Genetic analysis and integration with epigenomic datasets

Genetic data: This cohort consisted of a mix of individuals either from the Blueprint consortium or the Cambridge BioResource (CBR) (http://www.cambridgebioresource.org.uk/). If individuals were part of the Blueprint consortium, whole-genome sequence (WGS) data was available and was analysed as part of this project (Chen et al., 2016a). For CBR individuals, genotypes were available from the HumanOmniExpress-12v1 chip, the HumanCoreExome-12v1-0 and HumanCoreExome-12v1-1 genotyping chips, imputed using the combined reference panel of UK10K and 1000G Phase 1. Individuals were analysed as part of the large cohort (more than 3000 individuals) by Heather Elding, following all of the standard genotype QC procedures (Anderson et al., 2010). A logistic regression of SNPs with genotyping batch was performed to remove variants with discordant allele frequencies across batches. For this study, only variants that were shared between the WGS and genotyping datasets were used in downstream association analysis. All alignments and analyses in the Blueprint EpiVar project were carried out using GRCh37/hg19 and GENCODE v15 (Harrow et al., 2012). The analysis of the WGS data was performed as part of the Chen *et al.* (2016) study (Chen et al., 2016a).

Genetic association: The residuals of each trait were standardised and used as phenotypic traits in downstream genetic association analyses. Single variant association tests were performed using SNPTEST v2.4.0, and tested the association for each variant with each trait using an additive model (-frequentist 1). For each residualised trait, y_i and variant genotypes, x_i, a linear model $y_i = \beta_0 + \beta_1 x_i$ was fitted for $i = 1,2,3,\dots.,n$, where n is the number of individuals in the cohort. Genotype dosages were used (-method expected) to account for genotype uncertainty and expressed as the probability of each SNP genotype (AA, AB, and BB) per individual with 1 being most certain and NA for missing. Default quantile normalisation was disabled using the option -use_raw_phenotypes as the input phenotypes values were standardised at the stage of preparing the sample file. Variants with a MAF of less than 1% were excluded given the lack of statistical power to detect rare variant associations in this particular cohort.

Visualisation: Manhattan and QQplots were produced either using custom in-house scripts or using the QQMAN R package (Turner, 2014). Locus zoom plots were generated using the online tool (Pruim et al., 2010). Promoter-capture HiC plots were visualised using the Capture HiC plotter (CHiCP) (https://www.chicp.org) (Schofield et al., 2016).

Significance threshold: All variants meeting the standard-genome wide association threshold of 5 x 10^{-08} were identified. 5 x 10^{-08} is an appropriate testing thresholds for common variants in a European population, as tested here. In addition, neutrophil biological and genomic data were used to annotate these associations (see below) and therefore for suggestive associations that could be evaluated in a biological context, a threshold of 1 x 10^{-07} was applied.

Investigation of biological mechanism: I assessed whether significant or suggestive variants overlapped with any epigenomic or similar biological data in order to make predictions of possible variant functionality. I used ChIP-seq data from undifferentiated and differentiated HL60 cells, data from primary neutrophils from the BLUEPRINT consortium and epigenomic data including CTCF, PU.1 C/EBPβ, H3K27me3 (repressive) and H3K4me3 (active transcription) from primary neutrophils as part of an unpublished dataset (Stephen Watt, manuscript in preparation) (all described in Chapter 2). Further, I used binding data from undifferentiated HL60 cells including P300 (enhancer co-factor), C/EBPε (TF) and cohesin subunits SA1 and Rad21 (cis-regulatory module protein) that was collected within the Soranzo team by myself and Stephen Watt. I used the bedtools analysis suite (bedtools version 2.23.0) with the -intersect option to assess intersection of genetic variants and molecular features. Promoter-capture HiC data from primary neutrophils, which describes long-range interactions between genomic locations, was used to identify potential target genes of significant variants (Javierre et al., 2016). In addition, the potential effects on neutrophil gene expression of function-associated variants were assessed using RNA-seq data as part of the Blueprint consortium (Chen et al., 2016a). Rare or lower-in-frequency variants were not tested within the BLUEPRINT study, therefore to evaluate the function of rare/low-frequency variants, associations were tested using the RNA-seq gene expression data in FPKM.

Replication cohort: To confirm neutrophil function effects identified in the Cambridge discovery cohort, together with our collaborators, we established a cohort of healthy individuals in the Netherlands at Sanquin Research, University of Amsterdam (Sanquin cohort). Samples were genotyped at the Wellcome Trust Sanger Institute using the Illumina Human CoreExome Beadchip (coreex24) array. All genetic analysis for this cohort was carried out by Klaudia Walter (WTSI), but is summarised briefly here. A standard quality control protocol was implemented that included identity (≥ 0.9), duplicate (≤ 0.98) and gender checks (males ≤ 0.005, females ≥ 0.174), as well as a minimal call rate (≥ 0.95) and no excess autosomal heterozygosity (three standard deviations). 157 donors were genotyped in two batches (83 donors in the first and 74 donors in the second batch). In total eight donors failed the heterozygosity threshold and three donors failed the duplicate threshold. Additionally, a principal component analysis (PCA) was carried out together with HapMap3 samples which identified 14 population outliers, of which eight samples had also failed the heterozygosity threshold. In total 17 samples were excluded from further analysis. Samples were imputed using the Haplotype Reference Consortium (r1.1) using the EAGLE2+PBWT pipeline. The samples were processed with ethical consent approved by the WTSI Human Materials and Data Management Committee, reference 16/042 and titled: “Genetic variation in neutrophil cellular function- Biobank sample donors of Sanquin Research”.

Genotype validation: We identified a genome-wide significant association for a low-frequency locus (rs116811177/rs115109232, MAF = 2%, Figure 3.14). Given the low-frequency of these SNPs, we used Sanger sequencing with probes designed for rs116811177 and rs115109232 to confirm the heterozygous genotypes for the five individuals in the discovery cohort association (Figure 3.15). The genotyping assay was designed by Agena Bioscience using the MassARRAY® System with the iPLEX® chemistry.

3.3 Results

3.3.1 Phenotype Correlation

This study describes the first large-scale exploration of neutrophil adhesion, degranulation and respiratory burst from a predominantly healthy cohort, which enables a direct comparison of the relationship between these responses. In total, 29 traits were assessed across three assays and 12 different stimuli (Table 3.2). For adhesion, eight different stimuli were used and the final response measured as a single relative fluorescence unit (RFU). For respiratory burst and degranulation, four and two different stimuli were used respectively. Traits were calculated from the response distributions for all donors measured for either 120 or 60 cycles for each individual (Table 3.2).

I implemented unsupervised clustering analysis using the Pearson correlation between each pairwise residualised trait comparison to investigate similarities between the different neutrophil functions (Figure 3.11). Interestingly, two higher order clusters were observed, one between respiratory burst and degranulation responses and the other of adhesion responses. Mean correlations within these clusters revealed that the correlation between degranulation and adhesion responses was particularly low (mean *r* = 0.068, SD = 0.052). By contrast, the average correlation between respiratory burst traits and degranulation traits was 0.148 (SD = 0.093). However, the correlation between adhesion traits and respiratory burst traits was higher than that between degranulation and adhesion (*r* = 0.144, SD = 0.100). The correlations within responses of the same assay were all above 0.48 with the highest between degranulation traits (*r* =0.732, SD = 0.187).

There was a negative correlation between both time to particular RFU respiratory burst traits (PMA time to reach an RFU at 40000 cycles and STZ time to reach an RFU of 30000 cycles) and the rest of the respiratory burst responses. A high responder will reach a high maximum slope or RFU in a shorter time period due to the fast reaction response. These traits also clustered with the rest of the adhesion responses. The biological reason underlying the slight correlation of these time traits with adhesion responses is unclear.

The adhesion HEPES response was clustered with the rest of the adhesion responses, which could suggest that this condition demonstrates a degree of pre-stimulated activity in this particular function, which has been observed primarily in the adhesion assay over respiratory burst and degranulation by our collaborators (personal communication, Anton Tool).

The higher observed correlation and similarity between degranulation and respiratory burst responses may reflect activation of shared components of biological pathways or that certain stimuli lead to a concomitant activation of multiple biological pathways culminating in the activation of multiple neutrophil functions.

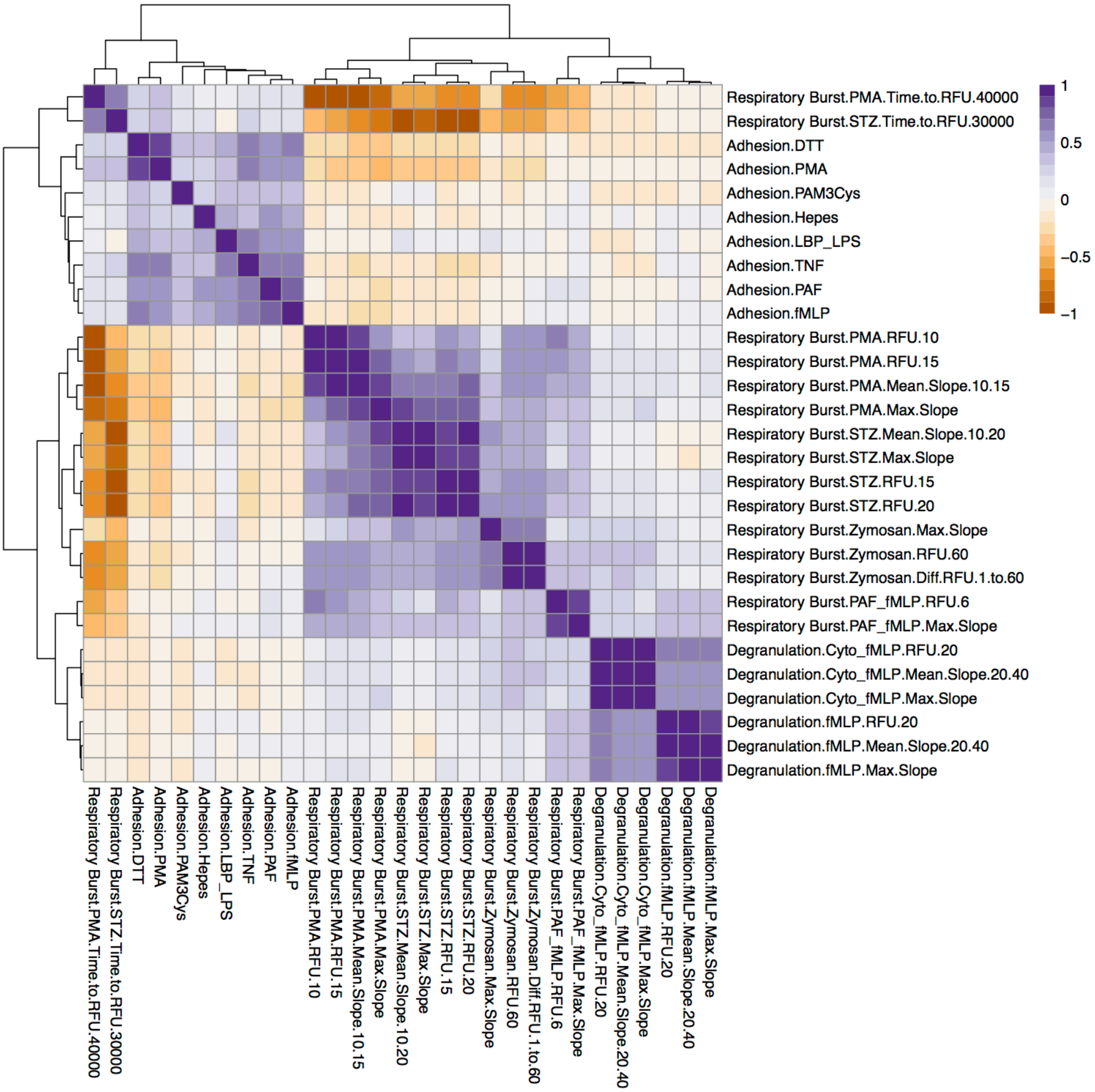


Figure 3.11: Correlation of neutrophil function phenotypes
Heatmap shows the correlation between standardised residuals from all traits and stimuli used in genetic association analyses. The Pearson correlation between all traits was calculated and used in unsupervised clustering analysis to assess the relationship between neutrophil functional responses. Respiratory burst and degranulation responses correlated more closely than to adhesion response with the exception of the response to HEPES (which should represent unstimulated responses) and the time to a specific RFU response.

3.3.2 Genetic variants associated with inter-individual variation in neutrophil function

I used the standardised residualised values for the 29 traits described above in independent single variant genome-wide association tests. Just over six million SNPs after filtering for variants below 1% MAF were tested for association with each trait. We estimated that this study was powered to detect common variants of moderate to high effect size and low-frequency variants with standard beta estimates of approximately 3 SD (Supplementary Figure 3.2).

I identified two SNPs reaching genome wide significance that were associated with respiratory burst of neutrophils stimulated with PAF and fMLP (p value < 5 x 10^{-08}, Table 3.3, Figures 3.12-3.14). The two low-frequency SNPs, rs116811177 (EA = G, EAF = 0.02, beta = 2.92, SE = 0.398, p value = 2.39 x 10^{-11}) and rs115109232 (EA = A, EAF = 0.02, beta = 2.92, SE = 0.398, p value = 2.39 x 10^{-11}), are perfectly correlated with r^2 of 1 (1000G). There were five individuals in the heterozygote state that demonstrated an increase in the respiratory burst response (Figure 3.15). No other genetic variants reached genome-wide levels of significance with any of the other functional traits. There were six variants that were associated at the suggestive p value threshold (p value < 1 x 10^{-07}, summarised in Table 3.3). All of these SNPs were low frequency (MAF < 5%) except one that was associated with adhesion response stimulated by PMA (rs57784565, p value = 8.59 x 10^{-08}, MAF = 7.6%). The remaining variants were all associated with respiratory burst of neutrophils stimulated with PAF and fMLP.

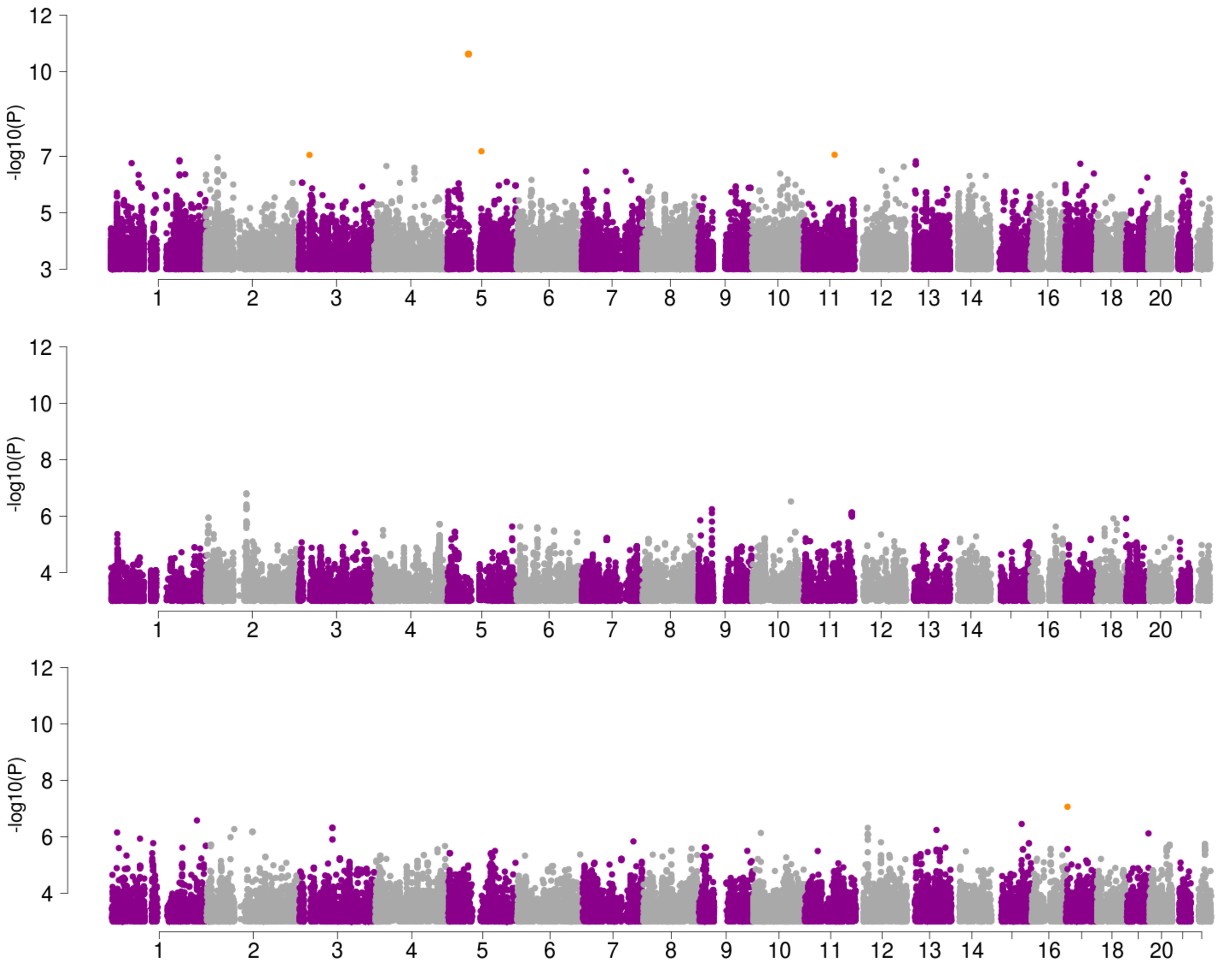


Figure 3.12: Association results for neutrophil function responses
Manhattan plot showing all variants associated with all traits at a p value threshold of 1 x 10^{-03} or less. A) Respiratory burst-associated variants, with a genome wide significant locus signal on chromosome 5. B) Degranulation-associated variants showed a lower number of variants associated at the threshold than respiratory burst but possible suggestive signals on chromosome 2 and 9. C) Adhesion-associated variants with a signal that just missed the significance threshold (approximately 9.0 x 10^{-08}) on chromosome 17. Responses for all the conditions and traits were combined into one Manhattan plot per assay. Variants associated with a p-value threshold of less than 1.0 x 10^{-07} are highlighted in orange.

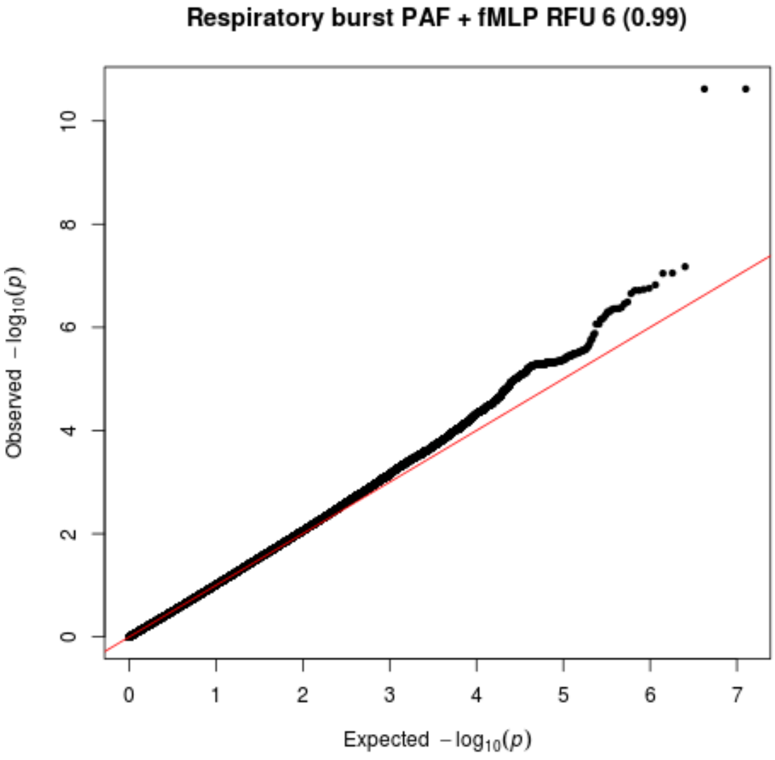


Figure 3.13: QQ plot of the respiratory burst PAF and fMLP RFU at 6 cycles response
Expected -log10(p value) calculated using a uniform distribution against the observed -log10(p value) from the function single variant association test. The deviation from the expected line demonstrates that genome-wide significant variants were more associated with the trait than expected by chance. The genomic control factor, lambda, was 0.99, suggesting there was no evidence for population stratification in this sample cohort.

rsID	Trait	chr:pos	EA:OA	P value	Beta	SE	MAF	Annotation	PcHiC	Protein Expression	Epigenome
rs116811177	Respiratory burst PAF + fMLP RFU 6	5:55597870	G:T	2.392 x 10^{-11}	2.92	0.398	0.020	Upstream, RNU6ATAC2P	MAP3K1	Expressed SV neu	
rs115109232	Respiratory burst PAF + fMLP RFU 6	5:55599891	A:G	2.392 x 10^{-11}	2.92	0.398	0.020	Intergenic	MAP3K1	Expressed SV neu	
rs147669752	Respiratory burst PAF + fMLP RFU 6	5:89571908	G:A	6.649 x 10^{-08}	2.82	0.491	0.016	Intron RP11-61G23.1	LINC00461 CTC-467M3.1		
rs117183808	Respiratory burst PAF + fMLP RFU 6	11:80312077	T:G	8.856 x 10^{-08}	2.97	0.523	0.012	Intergenic			
rs7623696	Respiratory burst PAF + fMLP RFU 6	3:28332261	T:A	8.956 x 10^{-08}	2.12	0.373	0.03	Intronic CMC1			H3K4me1
rs57784565	Adhesion PMA	17:3635489	A:G	8.596 x 10^{-08}	1.31	0.231	0.076	Intronic ITGAE	ZZEF1	Expressed PM neu	H3K4me1

Table 3.3: Function-associated variants across all assays
Table summarising statistics and annotations of variants reaching a suggestive p value threshold of 1.0 x 10^{-07}. Genome-wide significant variants, rs116811177 and rs115109232 are low-frequency and associated with respiratory burst. For each variant, association statistics including p-value, beta, standard error of beta (SE) and minor allele-frequency (MAF) (in this cohort) are listed. Variant effect predictor (VEP) was used to annotation variants (McLaren et al., 2016). A range of epigenomic data was used to annotate potential function to these variants (Materials and Methods). PcHiC data describes neutrophil-specific promoter-capture HiC data. If a variant intersects with either fragment end of a long-range interaction the corresponding gene of the other end is listed. Intersection with histone modified regions or transcription factor binding regions as assayed by ChIP-seq is also listed. The protein expression column refers to a study that compared proteins expressed on the plasma membrane (PM) or secretary vesicles of human neutrophils (Uriarte et al., 2008).

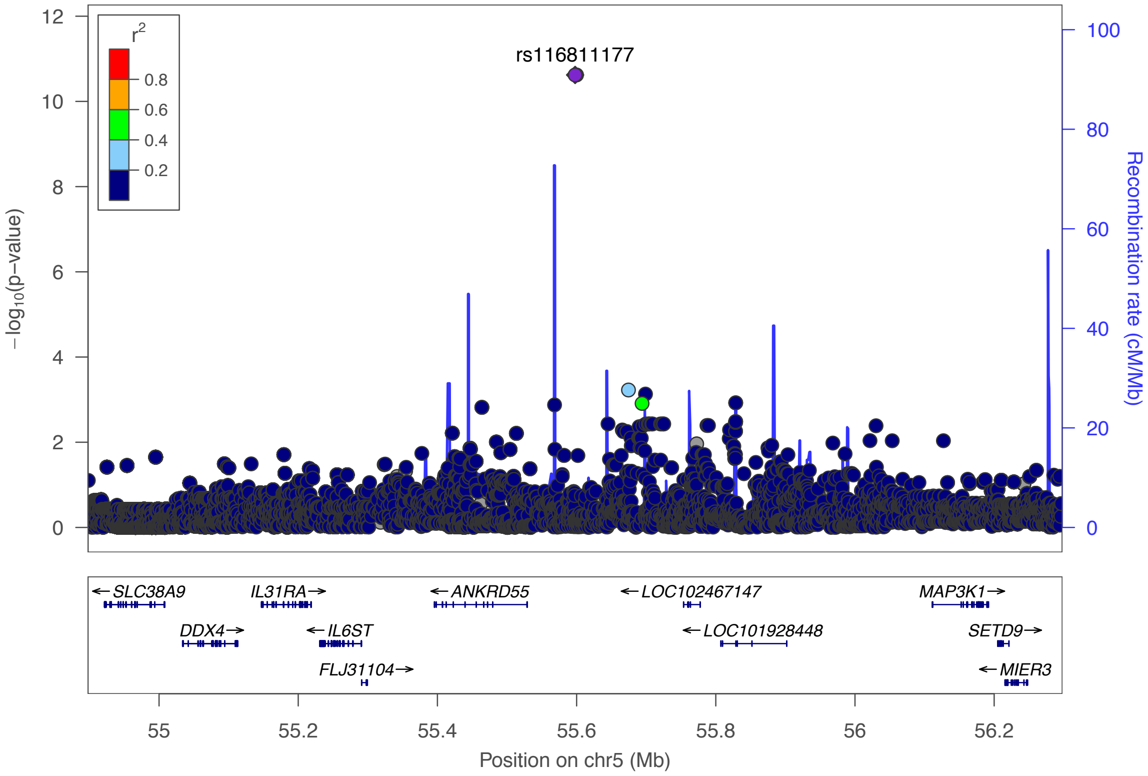


Figure 3.14: Genic location of the most significant signal associated with neutrophil respiratory burst

Regional plot of association for variants associated with respiratory burst RFU 6 in the region (+/- 700 kb) of the lead variants, rs116811177/rs115109232.

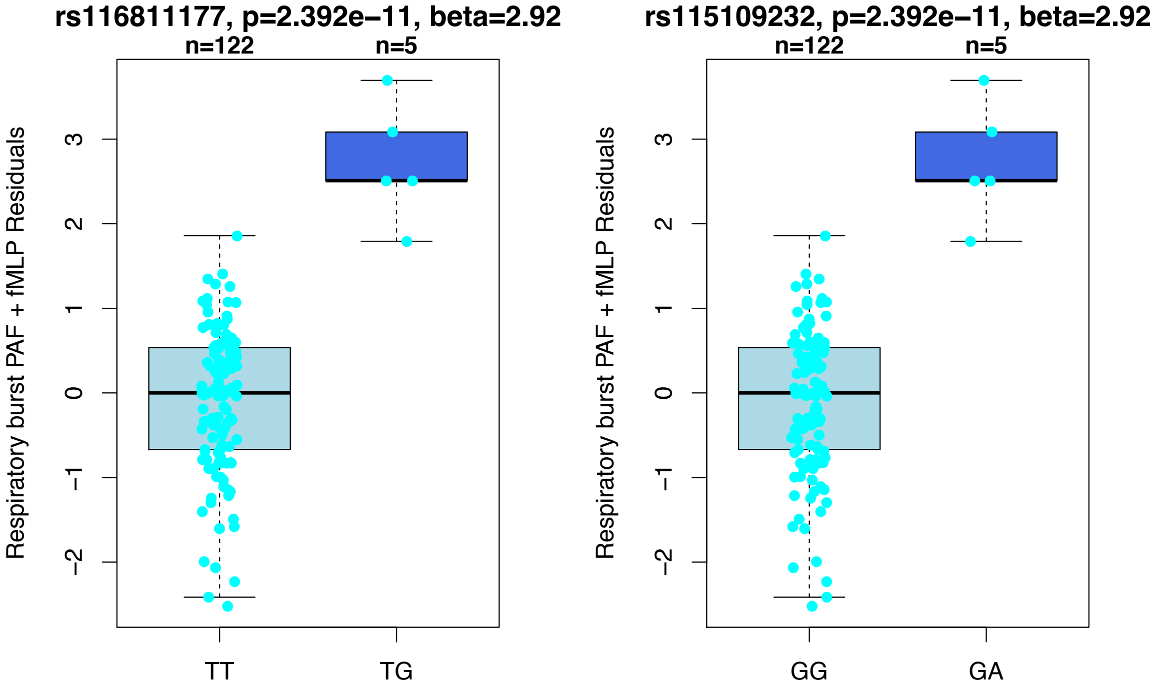


Figure 3.15: rs116811177/rs115109232 increases respiratory burst response after neutrophils are stimulated with PAF and fMLP

Residualised values of the respiratory burst response stimulated by PAF and fMLP at RFU 6 cycles are stratified by genotype of the two SNPs of the top genetic signal associated with this trait. Trait values of each individual are demonstrated as light blue dots.

3.3.3 Functional annotation of the PAF & fMLP respiratory burst rs116811177/rs115109232 locus

Figure 3.14 shows that the rs116811177/rs115109232 SNPs are intergenic, complicating identification of functional mechanism. In order to identify potential target genes, I used a range of publicly available datasets including gene expression QTLs (eQTLs) and promoter capture HiC data to query long-range interactions (Ward and Kellis, 2012, Chen et al., 2016a, Javierre et al., 2016).

Recently, it was demonstrated that complex-trait associated variants are enriched in regions of the genome connected through long-range chromatin interactions (Javierre et al., 2016). Potential distal gene targets of non-coding variants were identified using specific chromatin interaction data (Javierre et al., 2016). Therefore, to identify possible gene targets, I intersected variant positions (Table 3.3) with the fragment locations of promoter-capture HiC data from primary human neutrophils (Javierre et al., 2016, Schofield et al., 2016). I identified that both of the top most significant variants, rs116811177 and rs115109232, overlap a fragment that connected these variants to the promoter (and gene body) of two protein-coding genes; *MAP3K1* and *AC022431.2* as well as an interaction that connected these variants to the promoter of the *MIER3* gene (Javierre et al., 2016) (Figure 3.16). The latter gene, *AC022431.2*, which has no known genic annotation, was lowly expressed in neutrophils (median FPKM < 3 (Chen et al., 2016a)). In contrast, *MIER3* (the mesoderm induction early response protein 3) and *MAP3K1* (mitogen-activated kinase kinase kinase 1) were both expressed in neutrophils with *MAP3K1* being the most highly expressed (> 9 median FPKM).

Having identified possible gene targets, I next investigated whether the SNPs were associated with differential expression of these genes. I queried both publicly available gene expression data (HaploReg v4.1 (Ward and Kellis, 2012) and the primary human neutrophil RNA-seq data from the BLUEPRINT cohort (Chen et al., 2016a). I used data from unstimulated cells, rather than stimulated, as the PAF & fMLP respiratory burst response was known to be a fast reaction (minutes) rather than over several hours. Therefore, any gene expression effect associated with these SNPs would have to be present in unstimulated neutrophils. The low-frequency rs116811177/rs115109232 SNPs were not tested as part of the BLUEPRINT, therefore I retrieved the normalised FPKM gene expression values and the genotypes of individuals discordant at these SNPs in order to identify any evidence of gene expression effects. Within this cohort, nine individuals were heterozygous for rs116811177 or rs115109232. I also considered genes in the proximal region with a median neutrophil gene expression value greater than 3 FPKM and those with a known function. The included: *IL31RA, IL6ST, ANKRD55, FLJ31104, MAP3K1, SETD9, MIER3, SLC38A9* (Figure 3.14). I did not detect any significant differences in gene expression for any gene including MAP3K1, despite the significant long-range chromatin interactions (Figure 3.16).

I did not observe any further epigenomic intersections with the rs116811177/rs115109232 SNPs and other datasets such as histone modification peaks or transcription factor binding (H3K4me1, H3K27ac, H3K4me3, PU.1, C/EBPβ, C/EBPε, cohesin) using the bedtools intersect function (Materials and Methods).

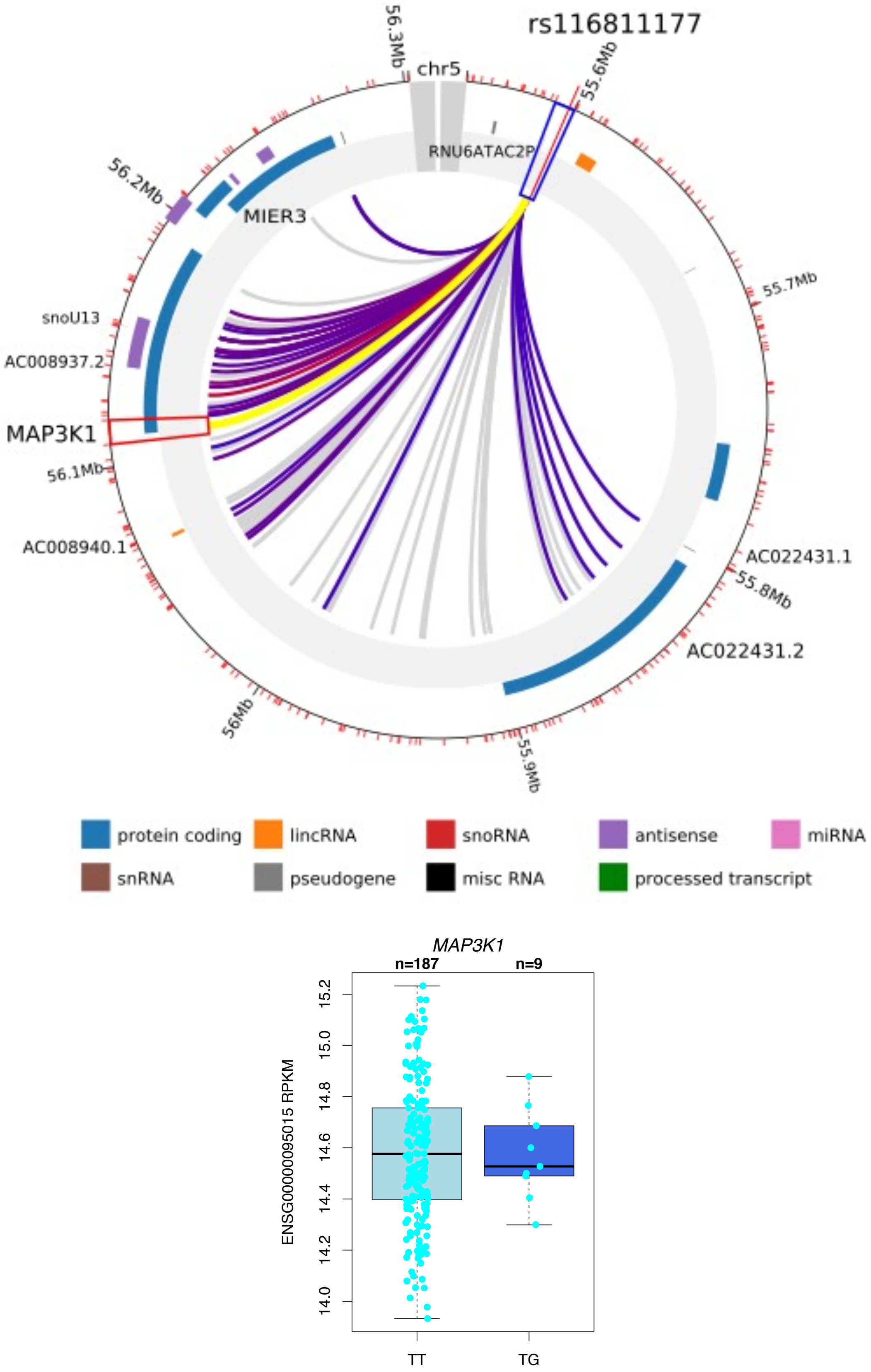


Figure 3.16: Functional annotation of the rs116811177/rs115109232 (*MAP3K1*) locus
Top panel depicts long-range interactions of the rs116811177/rs115109232 locus with various genes (labelled). Significant neutrophil interactions are shown in purple, with the *MAP3K1* promoter interaction in yellow with interaction score of 12 (> 5 is a significant interaction over background) (Cairns et al., 2016, Dryden et al., 2014). This figure was produced using the Capture HiC Plotter (CHiCP) accessed in July 2017 (Schofield et al., 2016). The bottom panel shows there is no evidence of differential *MAP3K1* gene expression with respect to genotype of the rs116811177 SNP.

3.3.4 Replication of the rs116811177/rs115109232 locus in an independent cohort

In order to investigate whether we could replicate the association at the rs116811177/rs115109232 locus, we established an independent cohort of healthy individuals from Sanquin Research, University of Amsterdam, the Netherlands (Materials and Methods). The DNA from healthy individuals was genotyped at the Wellcome Trust Sanger Institute and the final number of individuals in this cohort was 140 (Materials and Methods).

We aimed to use this cohort to replicate the increase in respiratory burst response within donors that were either heterozygous or homozygous for the effect allele of the rs116811177/rs115109232 locus. We decided that individuals of all three genotypes would be recalled and tested on the same day, in part to mitigate against the observed source of co-variation that can be introduced in measuring function on different days. Within the Sanquin cohort, there was one individual homozygous for the increasing PAF/fMLP respiratory burst effect allele (GG) and five heterozygous individuals. The homozygous increasing (GG) individual was recalled on the same day as individuals of either the heterozygous (GT) and homozygous decreasing (TT) genotype. These experiments are technically complex to perform and therefore a maximum of six individuals could be recalled on one day. The experiments were performed at Sanquin Research by Anton Tool.

Figure 3.17 shows the initial replication effort where the respiratory burst response was measured in nmol H_2O_2/min per million neutrophils, where the raw measured RFU was converted to nmol using a calibration curve of known concentrations of H_2O_2. A higher concentration of H_2O_2 corresponds to a higher RFU and therefore higher response that could be compared to the Cambridge discovery cohort. Figure 3.17 shows a higher PAF fMLP respiratory burst response in a homozygous increasing individual (rs116811177, GG) compared to an individual of TT genotype (non-carrier, Figure 3.17). No similar difference in response was observed for the other conditions tested (Figure 3.17). The high PAF and fMLP response associated with increasing copies of the G allele was consistent with the observations of the effect of this locus in the Cambridge discovery cohort (Figure 3.15).

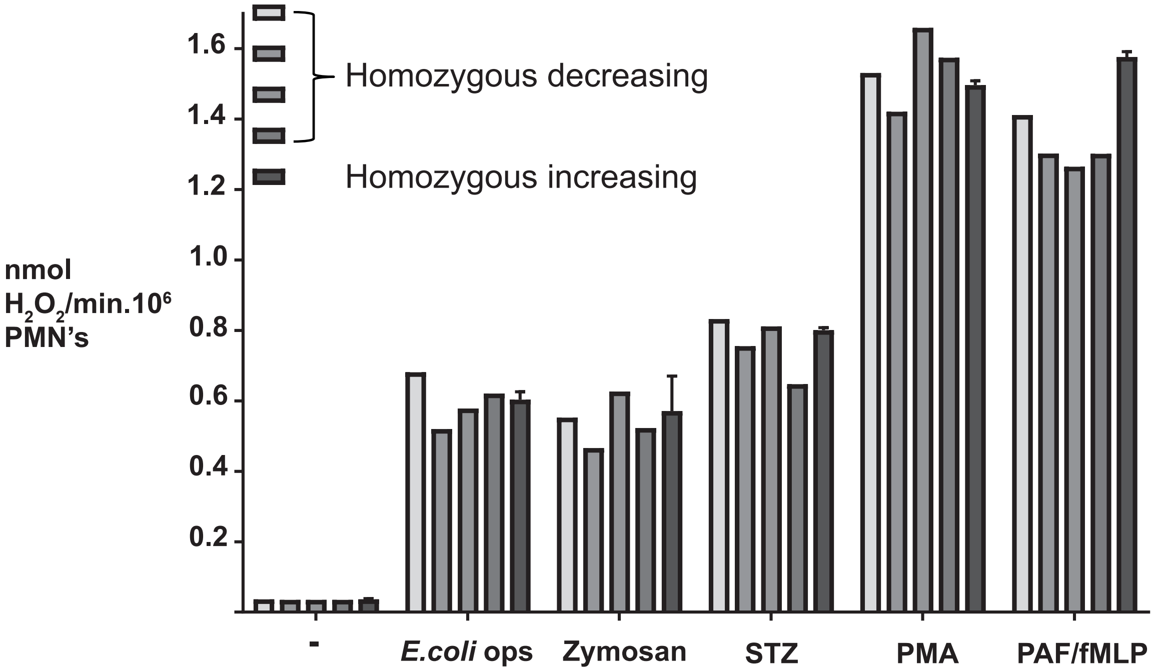


Figure 3.17: The rs116811177/rs115109232 (*MAP3K1*) locus in an independent cohort
Neutrophil function responses for different stimuli and donors (bars) as measured at Sanquin Research by Anton Tool. The different bars are coloured based on the genotype of the rs116811177/rs115109232 locus. Genotype groups refer to the predicted direction of PAF/fMLP respiratory burst effect from the original Cambridge cohort (Figure 3.15). A similar increase in PAF and fMLP response is seen in the homozygous-increasing individual compared to the non-carriers. The response is measured in nmol H_2O_2 per minute per million neutrophils by conversion using a calibration curve. Error bars on the plots demonstrate the difference between technical replicates. PMNs = polymorphonuclear neutrophils. Figure adapted from Anton Tool.

3.4 Discussion

We demonstrated, through thorough exploration and visualisation, how functional phenotypes are highly complex in both acquisition and sensitivity to environmental and technical factors, although good agreement was observed between technical replicates. I observed that after applying linear regression, the residuals, which contain the remaining variation after removal of specified sources of co-variation, did not show clear batch stratification (Figure 3.10). Some neutrophil responses were affected by external factors such as season (Figure 3.9, Supplementary Figure 3.1).

Previously the effect of annual seasonality on cytokine production from stimulated PBMCs and macrophages has been shown in the Human functional genomics project (Ter Horst et al., 2016). In this study, the authors also tested the effect of season on cytokine immune responses using a linear regression approach and demonstrated that the production of TNFα, IL-1β and IL-6 are highest in the summer. Conversely, the anti-inflammatory alpha-1-antitrypsin (AAT) production was highest in the winter. AAT seasonality was inversely correlated with the incidence of joint-disorder gout in a cohort of 800 patients. AAT inhibits IL-1β corresponding to lower cell infiltration into mouse joints, potentially explaining how the drop of AAT increases the inflammatory environment leading to exacerbation of gout (Ter Horst et al., 2016). There is, therefore, precedence for the effect of season on immune responses. I observed that in most cases, neutrophil functional responses were lowest in winter. The biological reasons and physiological implications for this are unclear and could reflect seasonality or differences in environmental temperatures in experimental processing. The decreased inflammatory response in winter was in agreement with that observed by Ter Horst *et al.* (2016). In future, it would be interesting to investigate any available clinical data in order to assess any physiological correlations with this observation. In this study, the effect of season was removed using linear regression prior to genetic association tests.

Individual's age and gender did not have a large significant effect on most neutrophil responses as shown by the small R^2 values and non-significant p values from ANOVA testing (Figures 3.8-3.9). Previously, differences in neutrophil differential gene expression have been observed between males and females (Ecker et al., 2017). Genes with higher expression levels in females were enriched in immune response pathways but those increased in males were enriched in basic cellular processes, which the authors suggested could explain the higher incidence of autoimmune diseases in women. Therefore, gender (and age) effects could have an effect on some responses, but in comparison to experimental factors and season the R^2 estimates are lower. Indeed, gender was found to be a significant covariate for adhesion stimulated with TNF (p value = 3.9 x 10^{-03}), Pam3Cys (p value = 0.029), LPS (p value = 6.6×10^{-04}) and fMLP (p value = 0.020) but was not significant for any degranulation or respiratory burst traits.

Encouragingly, after correction for sources of co-variation, the correlation of trait measurements seemed to recapitulate aspects of previously observed biological relationships between these functional responses. Degranulation and respiratory burst traits formed a separate cluster compared to adhesion responses, which perhaps suggests the former biological processes share biological pathways or could be co-regulated. Degranulation responses and adhesion responses showed the lowest correlation of all comparisons. In the degranulation assay, we stimulated neutrophils with the soluble stimulus, fMLP. When using soluble stimuli, we know that DQBSA is cleaved by the elastase and Cathepsin G proteases, which are released into the neutrophil supernatant (Anton Tool, personal communication). These two proteases are contained in the azurophilic granules, which are known to be released at later stages in the neutrophil activation process (Table 3.1) (Amulic et al., 2012). Therefore, this may explain the low correlation with adhesion processes, which occur earlier in neutrophil activation.

There is evidence for adhesion-stimulated degranulation (and ROS production) in an “outside-in” signalling process mediated by integrins. Integrin interaction with surface-bound anti-β2 and anti-β3 monoclonal antibodies alone, without inflammatory stimulus activation, induces neutrophil respiratory burst (Berton et al., 1992, Lowell et al., 1996, Berton and Lowell, 1999). Indeed, regulation of these processes is required to prevent aberrant inflammation. An initial inhibition of ROS production following neutrophil adherence to the extra-cellular matrix proteins such as fibronectin, fibrinogen and laminin, acts as a mechanism to prevent inappropriate tissue damage as these cells migrate towards inflamed tissues (Zhao et al., 2003, Al Laham et al., 2010). However, currently, the full pathways involved in these responses are not yet known so further experimental evidence is required to thoroughly interpret these correlations.

We were able to sufficiently correct for *known* technical and biological sources of covariation. However, the experimental design of this study did not enable correction for unknown sources of covariation or day-to-day effects. Studies with higher sample sizes have demonstrated a vast number of factors that lead to variation in trait values, for example in the large GWAS (N ~ 174000) of mature blood cell counts co-variation sources included time of measurement, menopausal age or smoking status (Astle et al., 2016). However, the detection of these effects in our study is limited by the low sample numbers, low number of traits and low number of individuals measured in one day (owing to the technical complexity of the assays). In most cases (56%), only one donor was measured per day, prohibiting approaches such as within-day standardisation to correct for day-to-day changes.

These issues highlight the importance of extensive exploration of association signals using other sources of biological data. In the original Cambridge discovery cohort, we identified a single genetic locus of two highly correlated, low-frequency SNPs rs116811177 (EA = G, EAF = 0.02, beta = 2.92, SE = 0.398, p value = 2.39 x 10^{-11}) and rs115109232 (EA = A, EAF = 0.02, beta = 2.92, SE = 0.398, p value = 2.39 x 10^{-11}) reaching genome-wide significance associated with the fMLP/PAF respiratory burst response. There were five heterozygous donors for the rs116811177/rs115109232 (*MAP3K1*) locus predicted have an increased respiratory burst response. Four of these donors had genetic data available from whole-genome sequencing (mean PAF + fMLP RFU 6 residualised trait value = 2.12) while the remaining donor had data available from a genotyping chip (PAF + fMLP RFU 6 residualised trait value = 0.93). We identified that two of these donors had been assayed on the same day (26th January 2015). I identified that there were no reagent batch changes between the previous days and when these two donors were measured and both donors were female. This suggests that the increased responses were not due to any known sources of covariation. Given this limitation in our study design, we designed the Sanquin replication study (described below) so that donors of different genotypes will be measured on the same day, removing the possibility of day-to-day variation in affecting these responses.

Despite these limitations, there is evidence that the proposed target gene of this locus, *MAP3K1*, is important in the respiratory burst response. MAP3K1, also known as MEKK, is a signalling kinase that activates the ERK and JNK pathways to coordinate downstream cellular processes such as proliferation, differentiation and stress (Chadee et al., 2002). It has also been observed that MAP3K1 is activated in response to fMLP in primary neutrophils, which is the same chemoattractant stimulating the respiratory burst response to which the rs116811177/rs115109232 is associated (Avdi et al., 1996). This could indicate that, through an effect not mediated by gene expression, the variants influence *MAP3K1* functioning in the respiratory burst response. For example, the phosphorylation status of *MAP3K1* could be altered as a function of genotype, which could be experimentally measured and is under discussion as future work in the replication cohort. Further annotations provided by the HaploReg resource (Ward and Kellis, 2012) also provided evidence that this region may be important in cellular signalling. The signalling compound, c-FOS has been shown in HUVEC cells to be bound at the r116811177 variant site using ENCODE ChIP-seq approaches. The second variant at this locus, rs115109232, is predicted to disrupt a Myc transcription factor motif, which is another important signalling molecule.

Subsequent to this analysis, it was identified that the Buffer 3 used in neutrophil preparation contained citrate, likely to facilitate measurement of full blood count in the Sysmex analyser. Citrate acts to chelates extracellular calcium, which could reduce calcium influx and in turn diminish neutrophil functional responses. Subsequent experimental repeats of these assays with and without calcium in the buffer demonstrated a decrease in functional responses. This could result in a decrease in power to detect signals and places greater importance on fully replicating the genome-wide significant locus identified in the Cambridge discovery cohort.

In conclusion, despite the small number of samples, this study showed that GWAS of neutrophil functional traits could provide a means for discovering novel regulatory regions controlling immune responses, in this particular case of the neutrophil fMLP-respiratory burst response. However, the use of such datasets to annotate disease-risk loci that could be explained by dysregulated neutrophil function would require greater sample sizes as we were limited in power in this present study to broadly annotate disease-risk loci.

Chapter 4

Dissecting the functional relationship between neutrophil count and surface expression of cellular receptors

Collaboration Note

The recall organisation of individuals and collection of blood samples was performed by staff at the NHS Blood and Transplant Department, Addenbrooke's Hospital, University of Cambridge.

Experimental work and assay development was supervised by Kate Downes as part of Willem Ouwehand's team at the Department of Haematology, University of Cambridge. Genotype QC and processing were performed by Heather Elding and the merging of the genotypes relevant for this study was performed by Klaudia Walter. All other analyses were performed by myself.

The Sanquin cohort was coordinated in conjunction with Anton Tool, Taco Kuijpers and Judy Geissler. Replication and further experiments were performed by Evelien Sprenkeler and Anton Tool at Sanquin Research, The Netherlands.

4 Dissecting the functional relationship between neutrophil count and surface expression of cellular receptors

4.1 Introduction

In the previous chapter, I discussed how studying genetic regulation of cellular traits could be complementary to studying molecular phenotypes in discovering novel pathways or genes involved in immune responses. Heritable genetic variation of a broad range of blood cell frequencies has also been previously identified using quantitative FACS-based immunophenotyping (Orru et al., 2013). In this study, immune cell frequencies of 95 cell types were profiled and the genetic contribution of 272 immune traits in a cohort of 1,629 Sardinian individuals was evaluated (Orru et al., 2013). The authors identified 23 independent variants at 13 loci and found three loci were known autoimmune risk genes (*HLA, IL2RA* and *SH2B3/ATXN2*). Hierarchical gating using antibody-labelled cellular receptors enabled the assessment of cell types such as regulatory T (T_{REG}) cells, which are mostly characterised by the surface expression of CD39. These cells play important roles in regulating immune responses and preventing autoimmunity and from this study, were observed to be the most heritable traits (mean 55%).

In addition to investigating cell frequencies, studying the cellular surface expression levels of proteins adds a further layer of functional insight into immune responses (Roederer et al., 2015). The heritability of 78,000 immunophenotype traits in 669 female twins was evaluated, and 11 genetic loci explaining up to 36% of the variation in 19 traits (from the top 151 heritable traits genetically assessed) were identified (Roederer et al., 2015). In this study, associations with two different mechanisms are described; the homeostatic regulation of cell levels through proliferation or elimination of a certain cell type and the regulation of the expression of the protein, presumably through variants affecting promoter or enhancer activity (Roederer et al., 2015). The authors examined the same highly heritable T_{REG} *ENTPD1/CD39* locus, through a variant in LD with the first identified by Orru *et al.* (2013), and found the association was explained by the phenotype (i.e. expression level of CD39), rather the frequency of T_{REG} cells (Roederer et al., 2015). Therefore, cell population frequencies and the surface expression of phenotypic receptors are both highly heritable and thus integrating multiple sources of functional data aids our understanding of biological systems. In both studies, significant loci were also known to be associated with disease risk.

I discussed in Chapter 1 how using an automated haematology analyser enables the measurement of mature blood cell counts, albeit a lower range of cell types compared to a FACs-based approach but the method is amenable for large cohorts, in this case of 173,480

European-ancestry individuals (Table 1.2) (Astle et al., 2016). This GWAS exemplified how studying the variation of haematological traits provided insight into blood cell biology and also the general architecture of complex traits. An unprecedented 2,706 independent variants associated with variation in 36 indices were identified (Astle et al., 2016). The functional importance of both coding and regulatory variants was demonstrated. Enhancer variants explained 19%-46% of heritable variation, similar to that of transcribed regions (4-30%). Coding variants were enriched in Mendelian genes, and medically important observations were made in leukocyte subsets, which previously had been investigated in studies of limited power (Astle et al., 2016). The associations from this GWAS provide a rich resource where in-depth functional interrogation of this dataset offers the opportunity to further dissect haematopoietic cellular biology.

From an in-depth investigation of genetic loci associated with neutrophil count from the Astle *et al.* (2016) blood GWAS, I observed many cases of significant variants located within or nearby to genes encoding cellular receptor proteins. Roederer *et al.* (2015) posit that the surface expression of proteins represents an independent mechanism compared to cell levels. The variants from the Astle *et al.* (2016) GWAS were associated with neutrophil count but located in genes encoding receptors not traditionally used to quantify cell phenotypes. In addition, these receptors were known to be involved in neutrophil biology. Therefore, I postulated that variation in receptor levels associated with these loci, could be functionally linked to cell count. Differences in the surface expression of the receptor protein could result in altered receptor signalling, which in turn influence the numbers of circulating cells.

After applying prioritisation methods described in Materials and Methods, I investigated variants associated with neutrophil count located in genes encoding two receptors, the granulocyte colony-stimulating factor receptor (GCSFR) and the urokinase receptor (PLAUR).

GCSFR is expressed on the surface of progenitor and mature neutrophil granulocytes, with higher surface receptor expression levels detected at later stages of development, on more mature neutrophils (Nicola and Metcalf, 1985, Panopoulos and Watowich, 2008). Differentiation from haematopoietic stem cells to granulocytes is highly dependent on G-CSF and to a lesser extent, GM-CSF (Mehta et al., 2015). The main motivation for investigating a possible functional relationship between count and surface expression is that GCSFR is essential for granulopoiesis and the cognate ligand to the receptor, G-CSF, is the principal cytokine-regulator of neutrophils (Panopoulos and Watowich, 2008).
Both G-CSF and GCSFR- deficient mice have chronic neutropenia, with significant reductions in the levels of peripheral neutrophils and granulocytic precursors in the bone

marrow (Lieschke et al., 1994, Liu et al., 1996). In normal conditions, G-CSF stimulates proliferation of all stages of granulopoiesis and increases neutrophil survival (Lord et al., 1989, Liu et al., 1996). This increased proliferation and survival is important to meet the demand of infection but is also exploited in clinical administration of G-CSF to neutropenic patients with very low neutrophil numbers (Panopoulos and Watowich, 2008, Sung and Dror, 2007). Both effects are termed ‘emergency’ granulopoiesis. (Semerad et al., 2002). Resolving the mechanism whereby variants located in the *CSF3R* gene are associated with neutrophil count is therefore clinically relevant.

In addition to stimulating neutrophil production, G-CSF also affects function (Betsuyaku et al., 1999). The ROS response to fMLP is increased in neutrophils primed with G-CSF (Betsuyaku et al., 1999). The residual neutrophils that remain in GCSFR-deficient mice show impaired functionality such as adhesion and migration in response to IL8 (Betsuyaku et al., 1999). GCSFR signalling is therefore required for particular functions of normal neutrophils.

Acquired GCSFR mutations increase the risk of acute myeloid leukaemia (AML) in patients with severe chronic neutropenia (SCN), for example due to C-terminal truncations that impair internalisation, increase surface receptor levels and stimulate proliferation over differentiation (Touw, 2015, Liongue and Ward, 2014, Ward et al., 1999, Aarts et al., 2004). Hereditary autosomal GCSFR mutations have also been observed, for example, T617N results in chronic neutrophilia due to constitutive activation of GCSFR as a result of dimerisation of the transmembrane domain (Plo et al., 2009). Clearly, GCSFR plays a key role in controlling neutrophil numbers and differentiation.

I investigated a second receptor to evaluate whether a functional relationship between surface expression and neutrophil count may be widespread. The plasminogen urokinase receptor (PLAUR/uPAR) receptor missense variant, rs4760, was highly associated with neutrophil count. PLAUR, associates with the plasma membrane via a GPI-anchor, and binds and activates the extracellular urokinase-type plasminogen activator (uPA/urokinase) (Smith and Marshall, 2010). Active uPA then generates the protease plasmin, which in turn cleaves extracellular matrix components. Through this process, uPAR regulates proteolysis, cell survival, growth and migration (Smith and Marshall, 2010). Expression of PLAUR is increased during inflammation and tissue remodelling and is correlated with poor prognosis in cancer making this receptor a potential therapeutic target (Smith and Marshall, 2010, Del Rosso et al., 2008).

Neutrophils express uPA and its receptor, PLAUR/uPAR (Gyetko et al., 1995). PLAUR functionality plays an important role in leukocyte adhesion and migration. PLAUR expression is increased during differentiation and activation of leukocytes, suggesting this receptor plays an important role in immune function (May et al., 1998, Nusrat and Chapman, 1991, Smith and Marshall, 2010). Treatment with an anti-CD87 (PLAUR) antibody inhibited chemotaxis in PMNs but was unaffected by anti-uPA antibodies, implicating the receptor in neutrophil function and migration (Gyetko et al., 1995). During specific infection with the bacteria *S.pneumoniae,* PLAUR-deficient mice had diminished granulocyte accumulation in the lungs and reduced survival, clearly providing evidence of the importance of PLAUR in neutrophil inflammatory responses (Rijneveld et al., 2002).

4.1.1 Aims of this chapter

I performed flow cytometry experiments in a recall-by-genotype (RbG) study (Section 1.6) to test whether pre-selected significant neutrophil count variants could also affect the surface expression of these receptors, potentially reflecting a functional link between this and neutrophil count. I also integrated my findings with available data sources including neutrophil molecular traits from the BLUEPRINT consortium (Chen et al., 2016a). In comparison to the functional GWAS I carried out in Chapter 3, which was performed over a year, this study took place over only two weeks. Blood was collected by the same nurses, and the same machine and reagent batches were used. Increase control over covariates in this shorter study allowed greater control over technical variability such as that observed in Chapter 3.

Using such a RbG study design, I collected a panel of 2 cell surface markers of specific gene candidates (see below) measured in up to 70 individuals divergent for alleles of the independent neutrophil-count associated variants (Figure 4.6 and Section 4.2.1). I measured the mean fluorescence intensity of selected receptors, PLAUR and GCSFR on the surface of neutrophils in whole blood from healthy donors. I then tested the association of SNPs located within the receptor gene (and +/- 500 kB) with the level of these receptors as measured using flow cytometry. I then further integrated molecular sources of information and other external datasets to explain my observations and provide functional hypotheses.

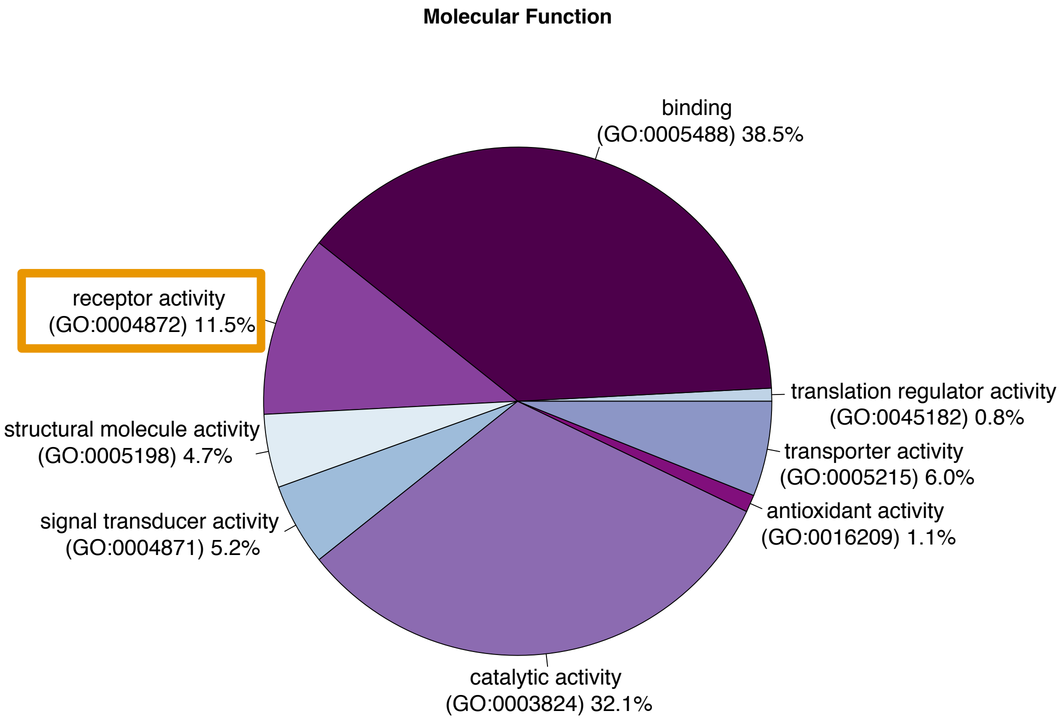
4.2 Materials and Methods

4.2.1 Selection of receptor and genetic variant candidates for experimental follow up

This hypothesis-driven investigation was the result of my efforts to thoroughly integrate genetic information from a neutrophil count trait GWAS (Astle et al., 2016) with epigenetic and transcriptional information from the BLUEPRINT consortium (Chen et al., 2016a).

As of December 2015, there were 17,673 variants associated with neutrophil count, which were clustered into 134 high LD groups (Astle et al., 2016). For the selection of candidates, I focused on common variants to ensure sufficient donor stratification across genotype groups in a cohort of less than 100. Applying this filter reduced the number of variants to 17175 and LD groups to 124. Genes were assigned to each variant if a variant overlapped a gene(s) and also the nearest upstream and downstream gene was also assigned. This resulted in a total of 567 unique assigned genes. Although these genes were assigned solely on proximity, a number of the most significant gene ontology terms identified using g:profiler suggested relevance to neutrophil biology, such as phagosome (p value = 9.92×10^{-13}), immune response (p value = 3.42×10^{-11}) and cytokine-mediated signalling pathways (p value = 6.40×10^{-10}) (Reimand et al., 2016). Of these 364 had a known function using the PANTER classification gene ontology system (Version 12.0) (Mi et al., 2017). 11.5% of genes with a known function were annotated with “receptor activity” (Figure 4.1). I found, for example, that of the variants located directly within a gene, 21% had known receptor function including *CSF3R*, *SLC25A24, FCGR2B, SLC12A7, SLC22A4, HLA-A, HLA-C, SLC44A4, HLA-DRB1, PLAUR, LY75, ACKR2, MYO1G, ACVRL1* and *VMP1.*

I speculated that there may be a relationship between receptor cellular surface expression and neutrophil count, likely underpinning the role of signalling and cell communication during differentiation of mature neutrophils. Therefore, I set to test the hypothesis that genetic variants affecting neutrophil count could exert their effect through a change in receptor expression on the cell surface. I established several additional criteria for the selection of downstream variants from variants annotated with a known receptor gene. There must be a known function of the receptor in neutrophil biology. The region of association must be resolved to one or a few SNPs so there is a higher chance of identifying a causal signal linked to the experimental measurement.



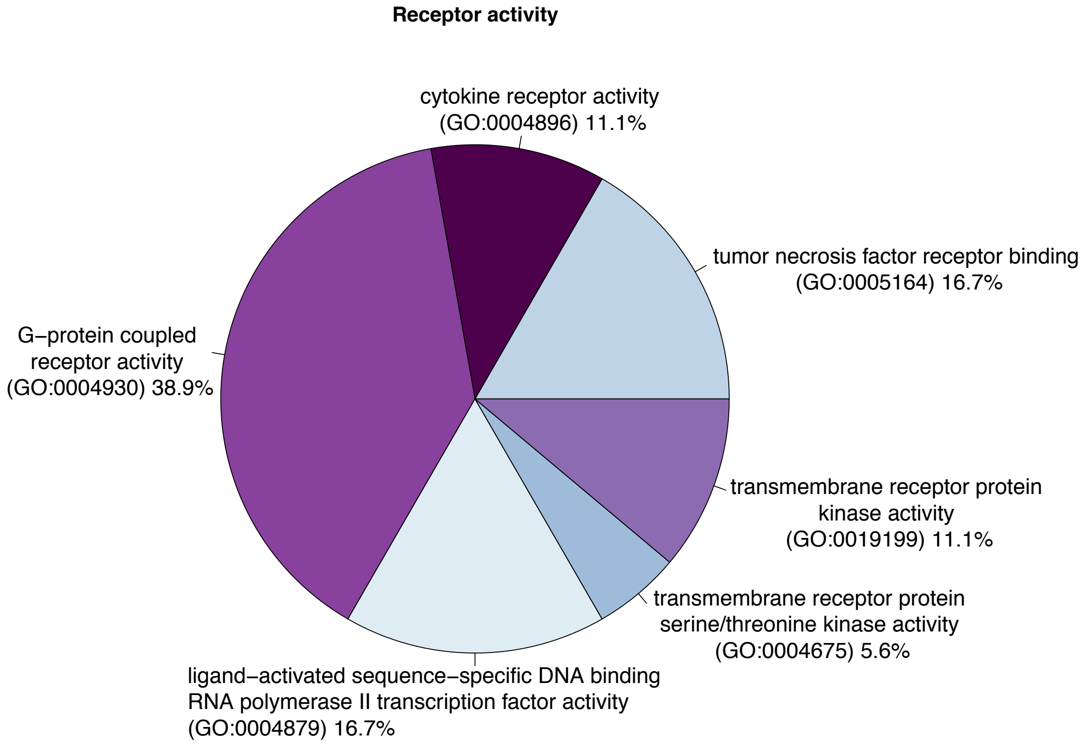


Figure 4.1: Molecular function of neutrophil count variant assigned genes
Pie charts show the proportions of molecular function annotations of the 364 genes with functional hits in the PANTER version 12 classification system. Receptor activity (11.5%) is highlighted in orange. The second pie chart (below) shows the proportions of molecular function terms within the highlighted receptor activity annotation.

I used neutrophil promoter-capture HiC data to verify that the assigned gene was the likely gene targeted by the variant (Javierre et al., 2016). If the presence of long-range chromosome interactions suggested the variants could be interacting with other likely candidate genes, then these loci were excluded. Criteria for gene functionality included expression in neutrophils or a known role of the gene in neutrophil biology. There must be available an antibody against the candidate receptor that has been previously validated in neutrophils. The antibody must be available conjugated to the fluorophore, phycoerythrin (PE), which emits a bright and stable fluorescence. The experimental design of the recall study, with up to 13 donors per day, limited the number of candidates I could feasibly investigate to a maximum of three.

I applied these criteria to annotated genes as described above and prioritised two receptor candidates for downstream analysis; the granulocyte colony stimulating factor 3 receptor (GCSFR) and the urokinase receptor (PLAUR/uPAR). The existing genetic evidence for these loci is discussed below.

4.2.1.1 G-CSF receptor

GCSFR is encoded by the *CSF3R* gene (Entrez Gene:1441, Ensembl:ENSG00000119535). I have described how GCSFR plays a key role in differentiation leading to the production of mature neutrophils. I postulated that variation in surface expression levels could impact on neutrophil count or vice versa.

An overview of the neutrophil count association in the Astle *et al.* (2016) study in this region is shown in Figure 4.2. The variant rs3917932 had the strongest association with neutrophil count (EAF =0.42, beta = 0.048, SE = 3.63 x 10^{-03}, p value = 2.06 x 10^{-39}, N = 173,480, Table 4.1) and is also associated with granulocyte count, myeloid count, white blood cell count, monocyte percentage, neutrophil percentage, lymphocyte percentage, granulocyte percentage of myeloid cells, neutrophil + eosinophil count and basophil + neutrophil count. Conditional analysis was performed at the locus as part of the Astle *et al.* (2016) study. The authors also found a second low-frequency variant, rs3917914 (EAF = 0.01, beta = 0.16, SE =1.69e-02, p value = 1.69 x 10^{-22}) in the same locus by regressing out the common signal (Astle et al., 2016). rs3917914 is also associated with granulocyte count, myeloid count, white blood cell count, neutrophil + eosinophil count and basophil + neutrophil count.

To the best of my knowledge, there is no previous study investigating whether genetic variation in the neutrophil receptor surface expression also influences neutrophil count in healthy individuals. The experimental process I designed in this chapter tests the surface expression of GCSFR on mature circulating neutrophils. The count measured in the Astle *et* *al.* (2016) paper describes the numbers of mature circulating neutrophils in a unit of whole blood. Therefore, this allows me to test my hypothesis that the expression of surface receptors on mature neutrophils is related to the numbers of mature neutrophils in the circulation using genetics as an anchor for causality.

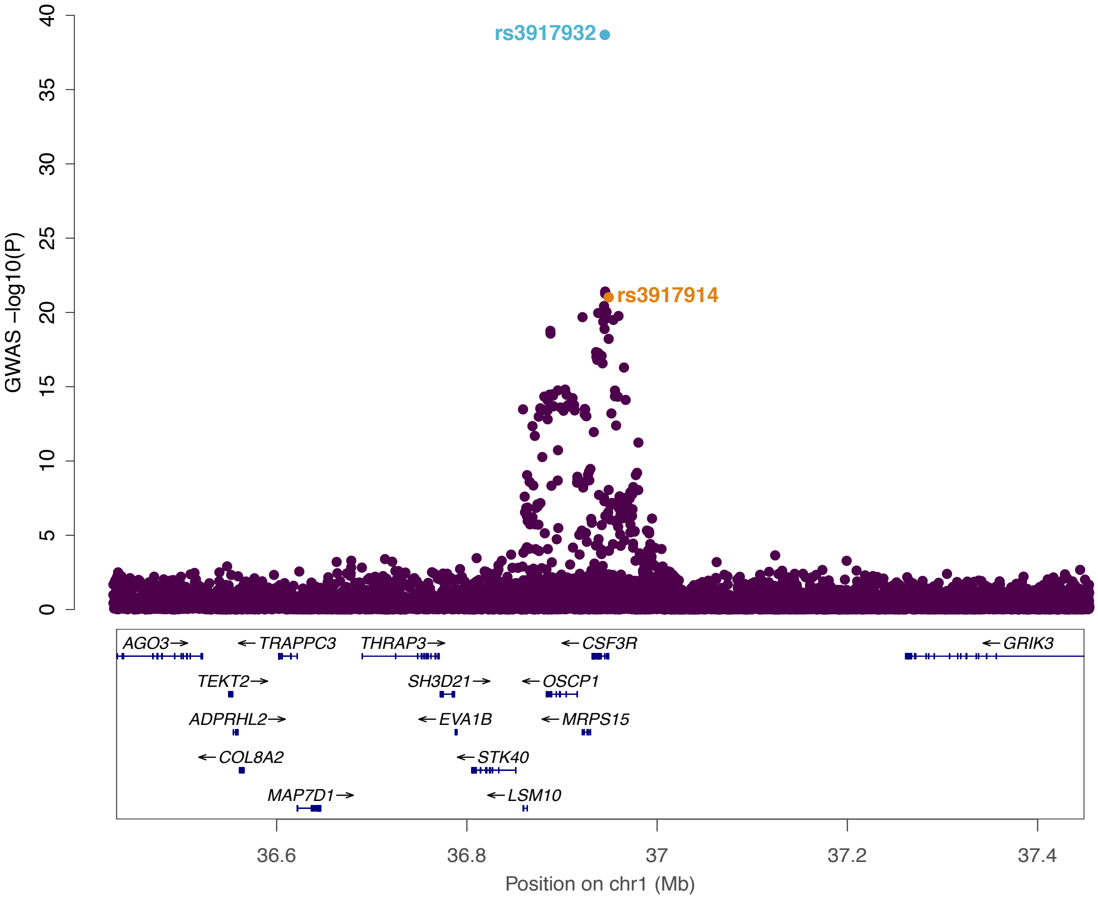


Figure 4.2: Genetic variants associated with neutrophil count in and around the *CSF3R* gene

Regional association plot of variants associated with neutrophil count around the *CSF3R* locus. Conditional analysis performed as part of the analysis in Astle *et al.* (2016) revealed two independent Intronic signals, rs3917932 (common) and rs3917914 (rare/low frequency). The chromosomal location (hg19) is shown on the x-axis, the left-hand y-axis is the -log$_{10}$(p-value) of association with the trait.

4.2.1.2 PLAUR

The second candidate under investigate was the PLAUR receptor, encoded by the *PLAUR* gene (Entrez Gene:5329, Ensembl:ENSG00000011422). One missense variant, rs4760, is located within an exon of the *PLAUR* gene and is significantly associated with neutrophil count (EAF=0.84, beta = 8.61 x 10^{-02}, SE= 4.92 x 10^{-03}, p value = 1.43 x 10^{-68}) (Table 4.1). rs4760 is also associated with granulocyte count, myeloid count, white blood cell count, monocyte percentage, neutrophil percentage, lymphocyte percentage, granulocyte percentage of myeloid cells, neutrophil percentage of granulocytes, neutrophil + eosinophil count and basophil + neutrophil count (Astle et al., 2016). rs4760 is associated with neutrophil count approximately 40 orders of magnitude more strongly than remaining SNPs in the proximal region. These proximal SNPs are also located within the neighbouring gene, *CADM4* (Figure 4.3). Therefore, it is highly likely that rs4760 is the causal SNP in this locus.

The rs4760 missense variant causes a leucine to proline amino acid change at residue 317 (L317P). rs4760 is predicted to be possibly damaging in the angiogenesis pathway by PolyPhen prediction from the Cancer Genome Anatomy Project Genetic Annotation (Savas et al., 2006). Residue 317 is located within a PLAUR protein isoform 1 domain predicted to be non-cytoplasmic using Phobius (Kall et al., 2007). The probability of that this domain is transmembrane decreased when I manually substituted the leucine 317 for a proline (0.4 to 0.1 where 1 is maximum likelihood) (Kall et al., 2007). This may suggest that this residue could impact the transmembrane domain or non-cytoplasmic domain functionality.

Therefore, rs4760 and the PLAUR receptor were included in this study given the role of the PLAUR receptor in neutrophil biology and particular disorders as well as the prediction that this variant could affect receptor stability. The inclusion of PLAUR and GCSFR also enables dissection of the effects of both intronic (*CSF3R*) and missense (*PLAUR*) SNPs.

In comparing the two receptor candidates studied here, GCSFR plays a pivotal role in neutrophil differentiation. PLAUR has been implicated in macrophage differentiation and phorbol-ester mediated differentiation of the neutrophil model cell line, HL60 (Rao et al., 1995, Nusrat and Chapman, 1991). Therefore, demonstrating a shared relationship between neutrophil count and PLAUR surface expression might indicate a possible role in neutrophil development. In the case of GCSFR, indication of a shared relationship could highlight the importance of receptor number on the surface in differentiation by directly linking receptor signalling to neutrophil count.

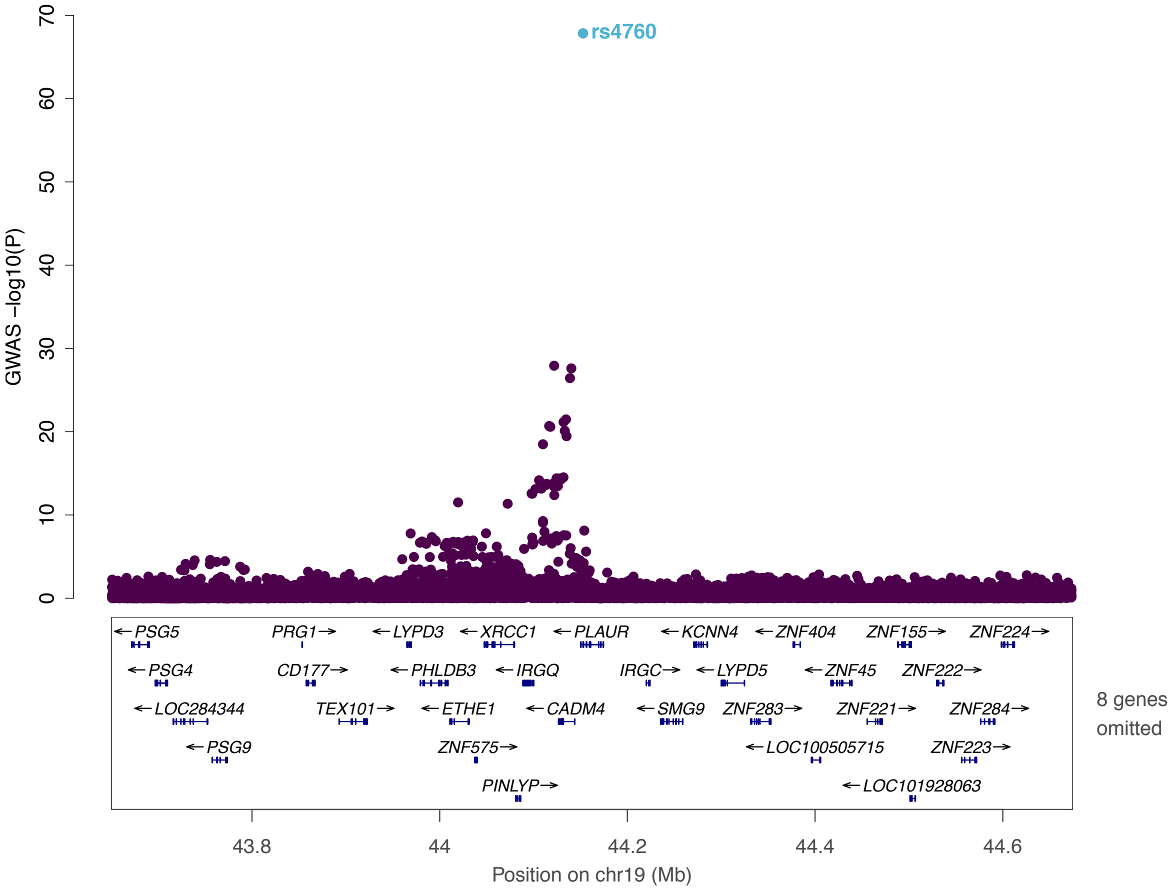


Figure 4.3: Genetic variants associated with neutrophil count in and around the *PLAUR* receptor gene

Regional association plot of variants associated with neutrophil count from the Astle *et al.* (2016) study. The most significant SNP, rs4760 is a missense SNP located in an exon of the *PLAUR* gene. rs4760 is more than 40 orders of magnitude more significant than remaining SNPs in the locus and is therefore predicted to be causal for the neutrophil count association in this locus.

rsID	Chr:pos (hg19)	EA/OA	EAF	Effect (SE)	Astle Neu count P	Gene	Type
rs4760	19:44153100	G/A	0.16	-8.61 x 10^{-2} (4.92 x 10^{-03})	1.43 x 10^{-68}	*PLAUR*	Missense L317P
rs3917932	1:36943916	C/G	0.42	4.80 x 10^{-02} (3.63 x 10^{-03})	2.06 x 10^{-39}	*CSF3R*	Intron 3 full transcript (First intron of CSF3R-204)
rs3917914	1:36947888	A/G	0.01	1.60 x 10^{-01} (1.69 x 10^{-02})	9.54 x 10^{-22}	*CSF3R*	Intron 1 of full transcript (and truncated CSF3R-008

Table 4.1: Candidate variants selected for functional follow-up
Neutrophil-count (number of neutrophils per nL per unit volume of blood) associated variants from the blood count GWAS (Astle et al., 2016). The effect size is expressed in SD of the trait. The conditionally independent variant(s) for the *CSF3R* (one common, one rare) and *PLAUR* (one common) loci are listed. *CSF3R* gene transcripts, as referred to in the table, are shown in Figure 4.10. EA = the effect allele, here defined as corresponding to the decreasing receptor expression allele for the *CSF3R* locus (see Table 4.5). SE = standard error. EAF = effect allele frequency. P = p value of neutrophil count.

4.2.4 Study samples

Individuals were recalled from the NIHR Cambridge BioResource (http://www.cambridgebioresource.group.cam.ac.uk) as part of the UNICORN study organised at the NHSBT, Addenbrooke's Hospital. All individuals were of blood group O. Peripheral blood samples were collected from healthy donors with informed consent (A Blueprint of blood cells, REC 12/EE/0040, East of England-Hertfordshire Research Ethics committee). The cohort used for functional interrogation of the relationship between receptor surface expression and neutrophil count comprised of 70 individuals of European descent. The mean age of the individuals was 61.27 years, and the range was from 25 to 79 years. The cohort consisted of 27 males and 43 females. After integrating with available genotype data, and exclusion of outliers, the final sample numbers used for association were 66 (GCSFR) and 65 (PLAUR).

4.2.5 Flow cytometry assessment of receptor surface expression

Blood was collected from 70 individuals across seven days. The number of donors processed per day ranged from six to thirteen. Blood samples were experimentally processed in batches of three-four donors. In the experimental processing, antibody volumes (Table 4.2) were first pipetted into the bottom of each tube. 100 μl of blood from each donor was then added to each tube. For each individual, three separate tubes were prepared, two receptor tests and one unstained sample (no antibody) (Table 4.2). In tubes 1-2 (receptor tests), the sample was labelled with specific antibodies for CD16 (APC), CD66b (FITC), which were used to identify the double positive neutrophil population in the gating strategy (Figure 4.4). In tube 1 the antibody against CD86 (PLAUR-PE) was added and in tube 2 the antibody against CD193 (GCSFR-PE) was added. After adding blood, tubes were mixed by inverting three times. After 20 minutes of incubation in the dark at room temperature, the red blood cells were lysed by addition of 2 ml of lysing solution and vortexed for three seconds. Following a six-minute incubation in the dark at room temperature, samples were then centrifuged at 600 x g for six minutes at 4°C (accelerator 9 and break 9). The supernatant was discarded and the pellet was re-suspended gently in 500 μl of PBS. Samples were vortexed for three seconds and stored at 4°C until ready for flow cytometry analysis. The time of labelling, lysing, fixing was recorded and investigated as potential covariates (see below). The samples were measured using a Beckman Coulter Gallios^{TM} Flow Cytometer system.

Tube	Blood Vol (μl)	Antibody	Ab Vol (μl)	Lysis Vol (ml)	Other	Tube Type
Tube 1	100	CD16 CD66b PLAUR/CD87	1.25 5 18	2	-	PLAUR test
Tube 2	100	CD16 CD66b CSF3R/CD114	1.25 5 9	2	-	GCSFR test
Tube 3	100	-	-	2	-	Unstained
Tube 4	0	CD16 APC	1.25	-	One drop of each negative/positive compensation beads	Compensation Control
Tube 5	0	CD66b FITC	5	-		
Tube 6	0	PE test/CD193	4	-		

Table 4.2: Contents of each sample tube per donor
Table summarises the experimental design where six tubes were made per donor, two tubes contained different combinations of antibodies, one was the unstained sample and three were compensation controls.

Specific controls were prepared for each experiment. First, for every donor, an unstained sample was processed in parallel except without antibody addition (Figure 4.4, Table 4.2). Second, compensation beads for PE, FITC and APC were prepared once per week and measured on the Gallios^{TM} every day. One drop of each compensation beads was added to each of three tubes (Table 4.2 and 4.4). Compensation beads provide two distinct polystyrene micro-particle populations; the positive population which binds to mouse κ light chain immunoglobulins and the negative population of beads with no binding capacity. Compensation was required to adjust for any spectral overlap in the three colours, PE, FITC and APC used in these experiments. 1.25 μl of anti-APC, 5 μl of anti-FITC and 4 μl PE were added to tubes 4, 5 and 6 respectively. Tubes were incubated in the dark at RT for five minutes. 1 mL of PBS was then added to each tube and these were stored at 4°C until use. The FlowJo software auto-compensation procedure was used to adjust for the multicolour flow cytometry data (PE, FITC and APC were used all in one tube to detect GCSFR/CD87, CD66b and CD16 respectively).

Third, control beads were run daily to verify the optical alignment of the lasers and functionality of the fluidics system. Flow-Check^{TM} Pro Fluorospheres, a suspension of fluorescent microspheres, were analysed on the Gallios^{TM} machine. Prior to sample ascertainment, mean fluorescence values for each laser were measured to detect any deviations from the pre-selected ranges that may require re-alignment of the lasers. Laser voltage adjustment was required only on the second day of experiments and correct functionality was confirmed after adjustment using the Flow-Check^{TM} Pro Fluorospheres. Lastly, isotype controls against the functional receptors (GCSFR and CD87) (Table 4.3) were run for one donor and confirmed the lack of a high level of non-specific background signal (data not shown).

Protein	Antibody	Colour	Source	Volume (µl)
CD16	Mouse anti-human, VEP13, mouse IgM	APC	Miltenyi, 130-091-246	1.25
CD66b	Mouse anti-human, G10F5, mouse IgM, k	FITC	BD Pharmigen, 555724	5
CD87/PLAUR	Mouse anti-human, VIM5, mouse IgG1, k	PE	BD Pharmigen, 555768	18
CD114/CSF3R	Mouse anti-human, LMM741 (RUO), mouse IgG1, k	PE	BD Pharmigen, 554538	9
Isotype control for CD87 and CD114	Mouse Isotype IgG1, k	PE	BD Pharmigen, 555749. Lot: 3046675	20
Isotype control for CD66b	Mouse IgM K, Clone G155-228	PE	BD Pharmigen, 555584	20

Table 4.3: Antibodies used for each marker in flow cytometry assays
Evidence of previous use in neutrophils for each clone type was assessed before final selection of the antibody (CD87 VIM5 (Elghetany et al., 2003), CD114 LMM741 (Piper et al., 2010)).

Bead/Reagent	Supplier	Code
Flow-Check Pro Fluorospheres	Beckman Coulter	A63493
Anti-mouse Ig, κ/negative control compensation particles set	BD Bioscience (BD^{TM}) CompBead	552843
Set-up beads for green/yellow laser	Life Technologies	C16508 LOT: 1438512
Set-up beads for blue laser	Life Technologies	C16509 LOT: 1438509
Set-up beads for red laser	Life Technologies	C16507
Lysing (and fixing) solution 10X Concentrate	BD Biosciences, BD FACS^{TM}	349202

Table 4.4: Reagents and beads
This table lists the beads used for the control experiments, the supplier and catalogue number along with the reference for the lysing solution used in the protocol to remove red blood cells.

Gating Strategy: Figure 4.4 shows the gating process used for all individuals to identify neutrophil populations. First, the fluorescence signal from the FL2 laser (PE) across time of measurement was assessed. This gate was used to remove potential machine technical factors such as debris in the fluidics system. The time gate was selected manually for each sample and judged to remove regions of lower density fluorescence measurements at the start of the reaction. A similar gating procedure was applied in a recent paper, also measuring median fluorescence intensity (MFI) of defined receptors on the surface of immune cells (Roederer et al., 2015).

The granulocyte population of neutrophils was selected using a plot of forward scatter height (FS-H) and side-scatter height (SS-H) (Figure 4.4). This population contains both granular cells, neutrophils and eosinophils. Doublet cells were then removed and following this, the double positive CD16+CD66b+ neutrophil population was selected. CD66b was used as a marker of granulocytes and is expressed on the surface of both eosinophils and neutrophils (Yoon et al., 2007). CD16 is not expressed on the surface of naïve eosinophils, therefore allowing specific selection of neutrophil populations (CD66+CD16+) (Davoine et al., 2002). 10,000 events/cells of the double-positive neutrophil population were collected for each sample. The median fluorescence intensity (MFI) of the receptors (PE, FL2 signal) was calculated from this population and used as an estimate of the surface protein expression.

4.2.6 Phenotype processing and genetic association

Phenotype processing: Receptors were treated separately as tube 1 (PLAUR) and tube 2 (GCSFR) respectively. The receptor MFI level of the unstained population (per donor) was subtracted from the receptor MFI level of the double positive CD16+CD66b+ neutrophil population and referred to as the normalised MFI value. The data were analysed and MFI statistics calculated using the FlowJo analysis suite version 10.1r5.

Outlier removal: the mean and standard deviation of the normalised MFI of a total of 67 donors was calculated, and any MFI value outside of the mean +/- 3 standard deviations thresholds were excluded. Overall, there were 66 individuals with genotype data for GCSFR and 65 for PLAUR, shown in Figure 4.5.

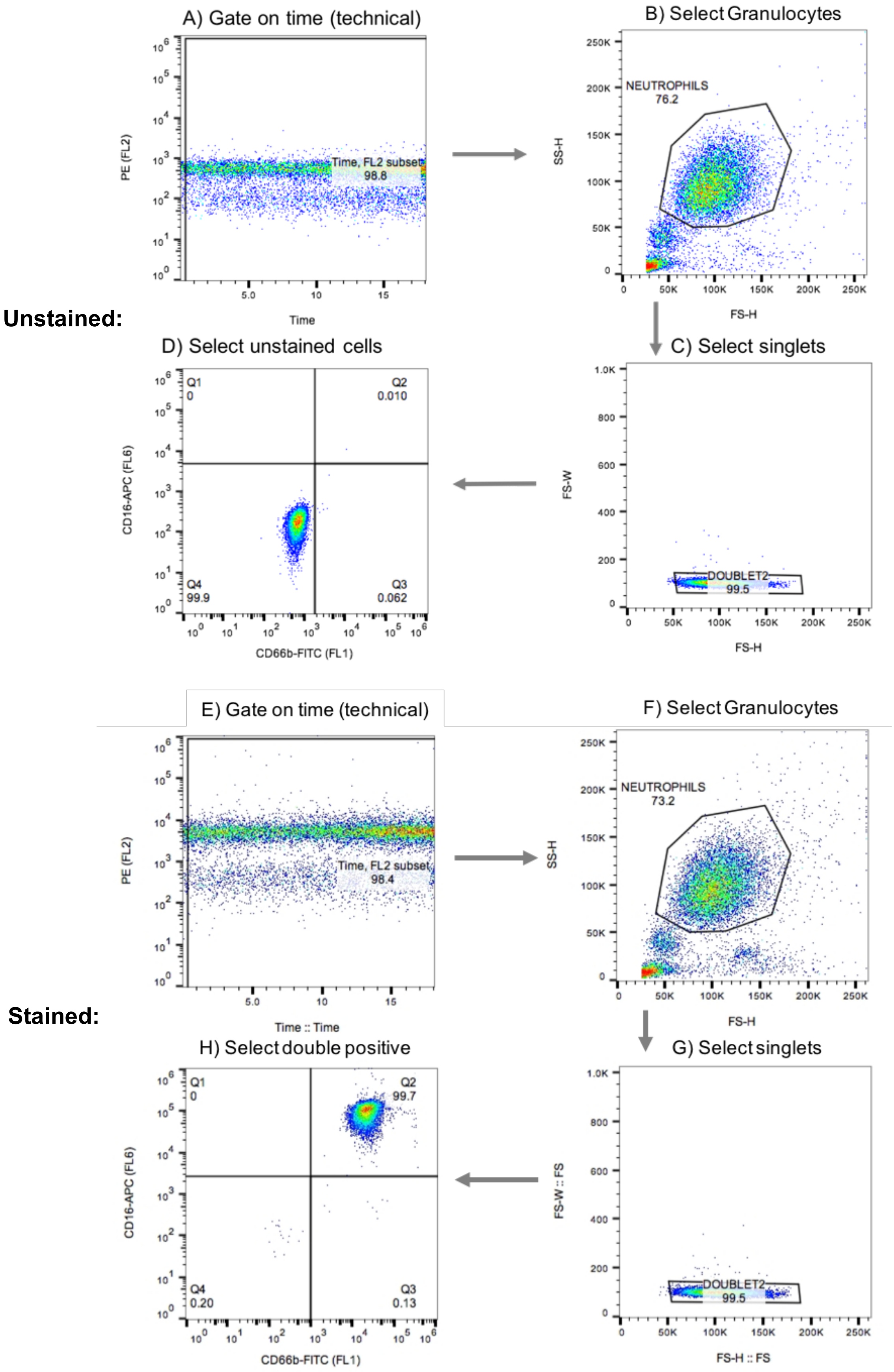


Figure 4.4: Gating strategy
Fluorescence plots used to measure neutrophil surface receptor MFI from whole blood. A)-D) Unstained E)-H) Stained**.** First the time gate removed technical artefacts. The granule population was selected using FS-H and SS-H (forward scatter height and side scatter height). Doublets are removed and finally the double-positive neutrophils are selected and MFI of either PLAUR or GCSFR was measured for this population using the PE signal.

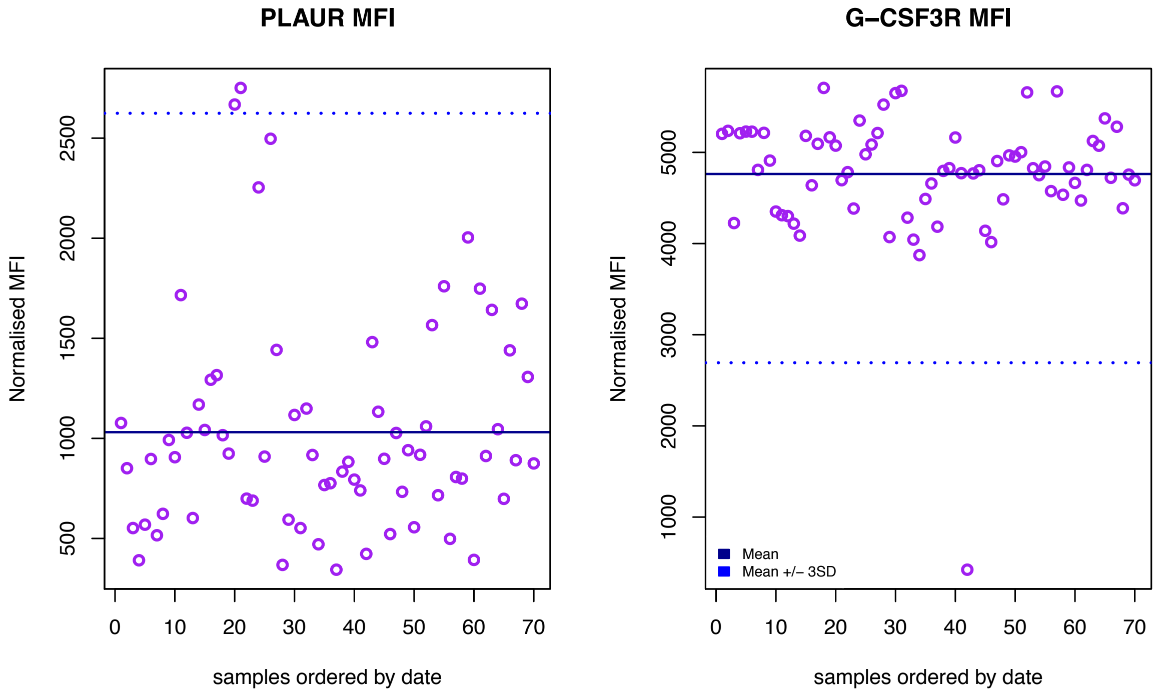


Figure 4.5: MFI by time including outlier exclusion thresholds
Plot of normalised MFI levels (unstained subtracted) in order of acquisition (samples by date). The mean of the MFI of all individuals is shown by the solid line. The mean +/- 3 standard deviations (SD) is shown by the dashed line. Donors with MFI values falling outside of these thresholds were excluded from the downstream analysis.

Covariate investigation and correction: Before association testing, I investigated potential sources of technical co-variation and stratification. Known covariates included age, sex, full blood count measures, date and experimental information that I recorded. These included time of bleed, labelling, and fixing and experimental processing batch. Given the compressed time-frame of this study, no reagent or antibody batch changes occurred. Up to 13 donors were processed per day prohibiting experimental processing of all samples together. Individuals were bled at 30-minute intervals; therefore, batches of three-to-five donors were processed (labelled and fixed) together. This minimised the time between bleed and processing. Groups of samples were designated into an experimental processing batch if there was an unforeseen change in the experimental protocol that could have introduced variation. For the time covariates, if there were too many levels (in some cases, there was only one donor with a specific time), bins of ten minutes were created.

Full blood count was measured using both a Sysmex XE-5000 and XN-1000 haematology analyser for each donor (collected at the time of blood collection by NHSBT collaborators). Results from both XE-5000 and XN-1000 were considered, and in the case of covariates, the most significant association was used to select the measurement for covariate selection.
I used ANOVA to assess whether covariates had a significant effect on the mean values of inverse normalised (and unstained normalised) MFI across different levels. No significant covariates (p value < 0.05) were found for GCSFR MFI levels. However, given that the experimental processing covariate reflects known experimental variation, I took the approach of conservatively adjusting for this batch effect using a linear model with processing batch as a predictor variable and GCSFR as the response variable, which marginally improved the strength of association.

I identified several significant covariates for PLAUR MFI values including date, experimental processing batch, sex, eosinophil count, time from bleed to labelling and time from bleed to fixing. Some of these covariates potentially measured the same effects, for example, the two time-period covariates and date and experimental processing batch. I corrected for each covariate in a linear model and used the residuals to test if the second covariate was still significant. In all cases the covariates were no longer significant in the second iteration of linear regression. Therefore, I corrected for the covariate with the lowest ANOVA p value.

In summary, date, sex, time from bleed to labelling and eosinophil count were used as covariates in a linear regression model to correct for these effects on the PLAUR surface expression levels. After outlier removal, I inverse normalised the unstained subtracted MFI values and then used the lm() function in R to correct for the specified covariates. Covariates were input as factors (date, sex, experimental processing batch) or as numeric for covariates such as time from bleed to labelling and eosinophil count. The effect of these covariates on PLAUR and GCSFR MFI levels are shown in Supplementary Figures 4.1 and 4.2. The corrected residuals were then standardised and used as an input into the genetic association tests described below.

I calculated the adjusted r^2 parameter of the linear model for each covariate independently against PLAUR MFI. The following adjusted r^2 values were obtained: Date (0.15), Sex (0.05), absolute eosinophil count (EO) (0.06) and time from bleed to labelling (0.20), indicating that date and time from bleed to labelling explained the most amount of variation in PLAUR MFI levels. The adjusted r^2 value for the overall model, correcting for all four covariates was 0.39.

Investigation of neutrophil size: I next investigated whether MFI was affected by neutrophil size where a larger neutrophil could express a greater number of receptors on the surface and vice versa for a smaller neutrophil. Currently, it is not known whether there is significant variation in the size of neutrophils in a general population. Measurements using a DxH 800 Coulter haematology analyser showed mean neutrophil volume was larger in sepsis patients and has been suggested as an additional clinical diagnostic measure for acute bacterial infection (Lee and Kim, 2013, Chaves et al., 2005). Neutrophil size may, however, be highly regulated in healthy populations to ensure the circulation of neutrophils through blood vessels of defined diameters.

I used an independent cell size parameter, NE-FSC, measured by the Sysmex haematology analyser (see above). NE-FSC is the forward light intensity of the NEUT area and gives an indication of cell size. I also investigated NE-SSC, which is the side scattered light intensity and represents the internal complexity or granularity of neutrophils (Buoro et al., 2016, Sysmex Corporation, 2010-2012). I assessed the correlation of these two parameters with the surface expression of both receptors and found no strong or significant correlations of GCSFR or PLAUR MFI and neutrophil parameters (p value <0.05) (data not shown). The low non-significant correlations suggested, that within the power limitations of this study, there was no evidence of a relationship between cell size and receptor expression (represented by MFI). Therefore, I did not implement a further size-normalisation for the receptor MFI values in addition to the steps described above.

Genotype processing: Individuals were genotyped as part of a large cohort by the NIHR Cambridge BioResource. Genotype processing was performed by Heather Elding, and standard genotype quality control procedures were followed as described in (Anderson et al., 2010). Genotypes were imputed against the combined reference set of UK10K and 1000 Genomes Phase 1. Imputed genotypes were used for the single variant genetic association tests described below. As described in Chapter 3, I am currently validating genotypes of all lead SNPs in this cohort with Sanger sequencing, particularly to confirm genotypes of the significant rare SNPs.

Single-variant genetic association and conditional analysis: The variants were extracted from gen files containing all included genome-wide variants from the genotyped cohort of 80 individuals. The region for the *PLAUR* locus was set from chr19:43650247 to chr19:44674699 (500 kb upstream and downstream of the gene start and end positions). The region for the *CSF3R* locus was set from chr1:36431644 to chr1:37448879 (500 kb upstream and downstream of the gene start and end positions). Phenotype values were inverse normalised after outlier removal and corrected for covariates as detailed above. Standardised residuals were formatted in gen .sample format and input into SNPTEST v2.5, a slightly newer version that enables analysis of fewer than 100 samples (66 for GCSFR and

65 for PLAUR). The same linear model was fitted for each variant as described in Chapter 2 Materials and Methods using the same options except with the additional (-use_lower_sample_limit). For conditional analysis, this was run separately by conditioning on (condition_ on additive model) rs3917924 and rs3917912 as the top common SNP and rare SNP respectively. The analysis was also run by conditioning on both SNPs.

Variant pruning: PLINK v2 was used to generate a list of independent variants using the --indep option. The variants tested (gene +/- 500kb) were pruned using a pairwise r^2 threshold of 0.1 with a variant count window of 500. Pairs of variants in the current window with a squared correlation greater than the threshold are pruned until no such pairs in the window remain.

Minor allele frequency calculation: QCTOOL v1.4 was used to calculate minor allele frequency using the –snp-stats function using the final sample sizes after outlier removal.

Linkage disequilibrium calculation: The correlation between variants, r^2, was used to indicate the linkage disequilibrium between variants. For this functional dataset, genotypes were available for 80 individuals. This data was used to calculate r^2 for the 14 most significant variants using PLINK v2 --flag with input data in gen format (--data and --oxford-single-chr 1). To avoid duplicate IDs the flag --set-missing-var-ids @:# was used. In addition, the Astle *et al.* (2016) cohort was used to calculate the LD between variants reaching the significance threshold as detailed in Chapter 2 Materials and Methods.

Locus zoom plot: the LocusZoom website was used to generate initial locus zoom plots and cis-gene locations (Pruim et al., 2010) using summary statistics. 500 kb flanking of the gene was plotted. Custom scripts were used to generate the regional association plots with highlighted SNPs for the same genomic region.

Gene annotation: In order to annotate the genic location of variants, the same annotation as used in BLUEPRINT (Chen et al., 2016a), GENCODE version 15 (ENSEMBL release 70), genome version hg19, was used. Custom scripts were developed to calculate intronic regions from the GENCODE annotation to specifically locate which intron each transcript the variants were located in. Introns were annotated as all regions between each exon from the start and end of the *GCSFR* gene which falls on the negative strand. HAVANA and ENSEMBL annotated transcripts were treated separately.

Transcript abundance and visualisation: Transcript abundance was previously assessed within the BLUEPRINT consortium (Chen et al., 2016a). Briefly, transcripts were quantified with Cufflinks v2.2.1 using RNA-seq data, Gencode v15 annotation and without *de novo* transcript assembly. Transcripts were visualised using the R package, ggplot2 (Wickham, 2009). sQTLseekeR was used for isoform splicing QTL mapping as part of the published BLUEPRINT study (Monlong et al., 2014). For evaluation of absolute transcript expression, rather than transcript ratios, the Cufflinks quantified expression in FPKM was used (Chen et al., 2016a).

Epigenome intersection: bedtools intersect version 2-2.23.0 was used to intersect molecular data with variant positions. Bed files for variant locations were generated by subtracting 1 from the position of the variant to act as the start position of the bed file.

Replication cohort: An independent cohort of 140 healthy individuals was established at Sanquin Research as described in Chapter 2. Donors heterozygous for rs3917912, rs3917914 and heterozygous and homozygous individuals for rs3917924 and rs3917932 were identified and used for planning replication experiments (Section 4.5, Future Work). These individuals, along with donors with non-effect haplotypes for all SNPs were/are being recalled at Sanquin Research and MFI of GCSFR on neutrophils is being measured by Anton Tool and Evelien Sprenkeler.

Phasing haplotypes: haplotypes were estimated in the original cohort using genotypes of three SNPs, rs3917932 (neutrophil count lead), rs3917924 (GCSFR lead) and rs3917914 (lead rare SNP), using SHAPEIT v2 and genotype data from 80 individuals (Delaneau et al., 2011). Three-SNP haplotypes were also estimated in the Sanquin replication cohort (above) using genetic data of the same three SNPs. In both cases, the HapMap phase II b37 genetic map was used to provide recombination estimates, as recommended. rs3917932 is missing from HapMap phase II cohort, in this case, SHAPEIT internally determines the genetic position. Haplotypes were phased using rs3917914 instead of rs3917912 as all genotype probabilities for rs3917914 heterozygous donors were above the threshold of 0.9. One donor heterozygous for rs3917912 with genotype probabilities less than 0.9 (approximately 0.8), SHAPEIT would therefore incorrectly call this genotype homozygous.

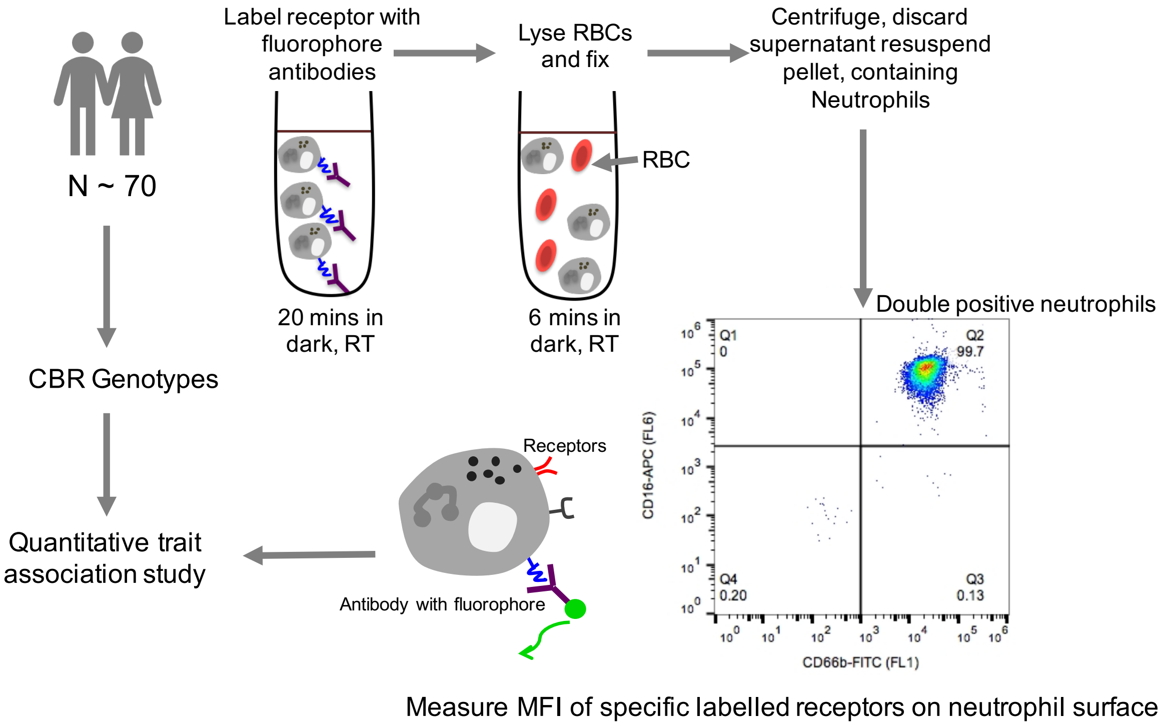
Genotype validation: I used Sanger sequencing with probes for the SNPs rs3917914 and rs3917912 to confirm that the heterozygous individuals carried at least one GCSFR-surface level-decreasing allele at this locus (as shown in Figure 4.8 for the discovery cohort association). The genotyping assay was designed by Agena Bioscience using the MassARRAY® System with the iPLEX® chemistry.

4.3 Results

4.3.1 Measurement of surface expression level in a cohort of 70 individuals

I investigated predicted target genes for neutrophil-count association variants and identified that 11.5% of genes with a known function were annotated with receptor activity. I hypothesised that given the role of receptors in cell signalling and the importance of these processes in differentiation that there might exist a relationship between significant neutrophil-count variants located within receptor genes and the neutrophil surface expression of these corresponding receptors. I predicted a possible functional role between these two traits for the G-CSF cytokine receptor based on the importance of the signalling pathways of this receptor in controlling neutrophil differentiation and mature neutrophil counts (Panopoulos and Watowich, 2008). I also investigated the PLAUR receptor missense variant given the role of this receptor in neutrophil function and as a comparison to GCSFR, a receptor with a well-established role in neutrophil development.

Figure 4.6 summarises the study design to test these hypotheses. Briefly, peripheral blood mononuclear cells (PBMCs) were stained with antibodies against CD16, CD66b, CD114 (GCSFR) and CD87 (PLAUR) and the surface expression of the latter two receptors was measured in a CD16+CD66b+ neutrophil population. The mean fluorescence intensity (MFI) of the population is used in this study to represent a quantitative measurement of receptor surface expression.



Label receptor with fluorophore antibodies
Lyse RBCs and fix
Centrifuge, discard supernatant resuspend pellet, containing Neutrophils
N ~ 70
RBC
20 mins in dark, RT
6 mins in dark, RT
CBR Genotypes
Double positive neutrophils
Q1
0
Q2
99.7
Q4
0.20
Q3
0.13
CD16-APC (FL6)
CD66b-FITC (FL1)
Receptors
Quantitative trait association study
Antibody with fluorophore
Measure MFI of specific labelled receptors on neutrophil surface

Figure 4.6: Project and experimental design
Overview of study and experimental design. Peripheral blood samples from 70 individuals were collected and labelled using specific antibodies. Mean fluorescence intensities (MFI) were collected for two receptors for surface expression on the neutrophil population using flow cytometry. These values for each individual were then used as a quantitative trait in genetic association tests to assess whether variation in surface expression was genetically controlled.

4.3.2 Two independent genetic signals are associated with GCSFR surface expression levels

Single-variant association tests were performed using GCSFR residualised MFI receptor values across 66 individuals. 14698 variants within the gene and 500 kb upstream and downstream of the start and end gene positions were tested for association with the receptor levels. These variants were pruned using a pairwise r^2 threshold of 0.1 to generate a list of independent variants in the region (159). To correct for multiple testing, a stringent Bonferroni correction was used to correct for the number of independent variants tested (0.05/159) resulting in a significance p-value threshold of 3.14 x 10^{-04} for GCSFR. In the *CSF3R* locus, 14 genetic variants reached significant levels of association with surface expression levels of G-CSF receptor (Figure 4.7 and Table 4.5). Four of these variants were low-frequency, the rest were common.

	rsID	chr 1 Pos (hg19)	EA/OA	EAF	P	Beta, (SE)	EAF NEU	NEU effect (SE)	UKBB P	BP beta (SE)	BP P	R^2 Astle	Genic location
Rare signal	**rs3917912**	36947936	T/C	0.02	1.53 x 10^{-05}	-2.48 (0.53)	0.01	0.16 (0.02)	6.04 x 10^{-19}	-	-	-	Intron 1 (CSF3R-008)
	rs3917914	36947888	A/G	0.02	1.62 x 10^{-05}	-2.37 (0.51)	0.01	0.16 (0.02)	9.54 x 10^{-22}	-	-	0.99	Intron 1 (CSF3R-008)
	rs116668546	36952851	A/G	0.02	1.69 x 10^{-05}	-2.32 (0.50)	-	-	-	-	-	0.71	Upstream (-)
	rs3917922	36945713	T/C	0.02	1.70 x 10^{-05}	-2.32 (0.50)	0.01	0.16 (0.02)	9.32 x 10^{-21}	-	-	0.85	Intron 2 (CSF3R-008)
Common signal	**rs3917924**	36945653	G/A	0.56	1.92 x 10^{-05}	-0.68 (0.15)	0.61	0.03 (0.004)	2.49 x 10^{-20}	1.10	5.51 x 10^{-35}	-	Intron 2 (CSF3R-008)
	rs3917931	36944054	C/T	0.56	1.93 x 10^{-05}	-0.68 (0.15)	0.61	0.03 (0.004)	1.96 x 10^{-20}	1.12	9.98 x 10^{-36}	1.00	Intron 3 (Intron 1 of CSF3R-204)
	rs3832027	36945307	AG/A	0.56	1.96 x 10^{-05}	-0.68 (0.15)	0.61	0.03 (0.004)	2.34 x 10^{-20}	-	-	1.00	Intron 2 (CSF3R-008)
	rs3917925	36945559	G/A	0.56	1.97 x 10^{-05}	-0.68 (0.15)	0.61	0.03 (0.004)	2.42 x 10^{-20}	1.12	9.98 x 10^{-36}	0.99	Intron 2 (CSF3-008)
	rs3917932	36943916	C/G	0.39	2.32 x 10^{-05}	-0.71 (0.16)	0.42	0.05 (0.004)	2.06 x 10^{-39}	-	-	0.47	Intron 3 (Intron 1 of CSF3R-204)
	rs199833813	36954487	AT/A	0.59	3.01 x 10^{-05}	-0.67 (0.15)	0.65	0.03 (0.004)	4.41 x 10^{-15}	-	-	0.78	Upstream (-)
	rs6667127	36957501	C/T	0.59	3.43 x 10^{-05}	-0.67 (0.15)	0.65	0.03 (0.004)	4.69 x 10^{-15}	0.97	1.24 x 10^{-19}	0.78	Upstream (-)
	rs955115	36858145	C/A	0.77	1.53e x 10^{-04}	-0.93 (0.23)	-	-	-	0.59	4.95 x 10^{-06}	0.11	Downstream (-)
	rs3917933	36943655	G/A	0.55	1.62e x 10^{-04}	-0.61 (0.15)	0.61	0.03 (0.004)	1.3 x 10^{-19}	1.12	9.98 x 10^{-36}	1.00	Intron 3 (Intron 1 of CSF3R-204)
	rs11295216	36947906	C/CG	0.53	2.86 x 10^{-04}	-0.65 (0.17)	-	-	-	-	-	0.49	Intron 3 (CSF3R-008)

Table 4.5: 14 Significant variants associated with the G-CSF receptor surface levels
Summary statistics from the GCSFR surface level and Astle *et al.* (2016) neutrophil count (NEU). Bold SNPs are the lead GCSFR MFI or NEU SNPs. EAF, effect allele frequency derived from MAF calculated using 66 donors. Standardised beta and standard error (SE) is given. EAF NEU was calculated for 173,480 individuals in Astle *et al.* (2016) and effect in SD of the trait. Location is relative to the most abundant transcript, CSF3R-204 or the second most abundant truncated transcript, CSF3R-008. EA = effect allele. OA= other allele. BP EA and BP direction relate to the significant blueprint exon effect (corrected for local SNPs as qvalue). Variants not tested in the study are shown by missing values (-). LD in r^2 between the variant and lead variant per common and rare signal respectively was calculated using the Astle *et al.* (2016) cohort.

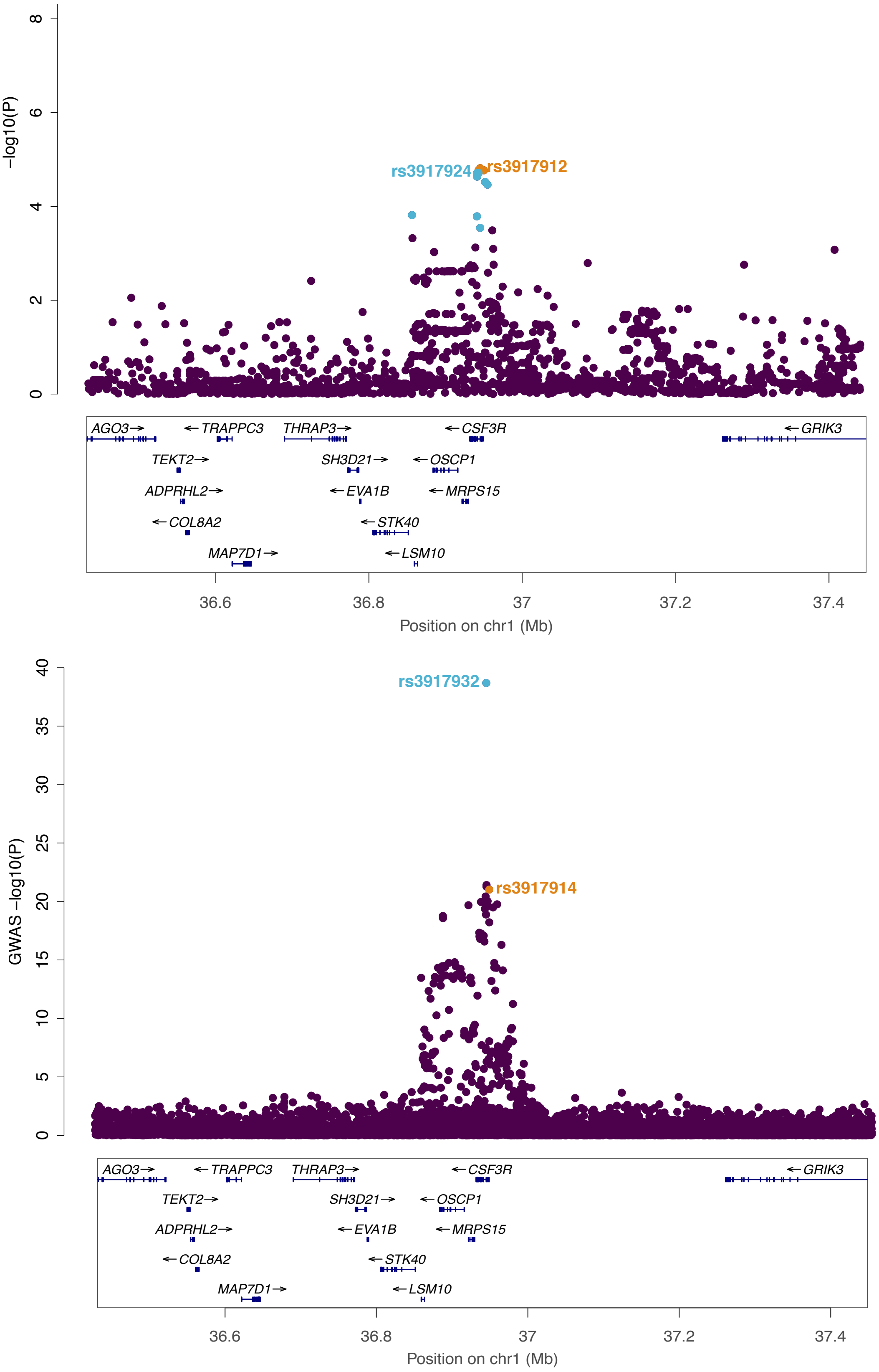


Figure 4.7: G-CSF surface expression association results

Regional association plot of the *CSF3R* locus. The top panel shows associations with the GCSFR surface expression levels. Variants reaching significance are highlighted in blue (common) or orange (rare) (Table 4.5). Lead common and rare SNPs are designated with the rsID. The regional association plot for the neutrophil count association is repeated from Figure 4.10 for comparison.

I carried out conditional analysis including the lead rare and common SNPs as covariates in an association model to test if these were independent signals. The beta and p value of rs3917912 remained similar to the univariate model and when regressing out the common, rs3917924 signal in the association model (univariate beta = -2.48 (SE = 0.53), conditional beta = -2.10 (SE = 0.48), Table 4.6). Similarly, when testing for association of rs3917924 with GCSFR MFI while conditioning on rs3917912, the common SNP remained significant (Table 4.6). After conditioning on both rs3917912 and rs3917924, no SNPs remained significant. Theses analyses provided evidence that the GCSFR surface expression association signal consists of two independent signals of different frequencies. Within the cohort of 66, three individuals carried the heterozygous genotype (T/C) for the low-frequency SNP, rs3917912. For clarity, I will refer to the significant signals that are described by lead SNPs, rs3917924 and rs3917912 as the common and rare signals respectively.

	Model	EA/OA	EAF	Beta (SE)	P value
rs3917912 (rare)	Univariate	T/C	0.03	-2.48 (0.53)	1.53×10^{-05}
	Conditional (rs3917924)	T/C	0.03	-2.10 (0.48)	4.42×10^{-05}
rs3917924 (common)	Univariate	G/A	0.57	-0.68 (0.15)	1.92×10^{-05}
	Conditional (rs3917912)	G/A	0.57	-0.57 (0.13)	5.53×10^{-05}

Table 4.6: Conditional analysis demonstrated the *CSF3R* locus contains two independent signals
Results of the two lead SNPs (common and rare) for association with GCSFR surface levels are given. The univariate beta and p values are from initial genetic association tests. Association tests were repeated but with conditioning on the respective top SNPs i.e. condition on rare and test for common association and vice versa. The conditional beta and p values are given and show that in both cases the remaining signal is still significant, confirming the common and rare signals are independent. The association analysis was also performed conditioning on both common and rare SNPs and no variants remained significantly associated using the before mentioned threshold (data not shown).

4.3.3 The relationship between GCSFR MFI and neutrophil count

The neutrophil count association signal in the *CSF3R* locus is described by two independent signals; a common signal with lead SNP rs3917932 (EA = C, EAF = 0.42, beta = 0.048, SE = 3.63e-03, p value = 2.06 x 10^{-39}) and a rare signal with lead SNP rs3917914 (EA = A, beta = 0.16, SE = 1.62 x 10^{-02}, p value = 9.54 x 10^{-22}) (Table 4.1) (Astle et al., 2016).

The lead common neutrophil count variant, rs3917932 is significant in the GCSFR data, with a slightly less significant p value but larger beta than the GCSFR lead SNP, rs3917924 (Table 4.5). These two SNPs are practically indistinguishable in the GCSFR data. In the Astle *et al.* (2016) study of neutrophil count, rs3917932 is 20 orders of magnitude more significant than rs3917924 (2.49 x 10^{-39}, 2.49 x 10^{-20} respectively). There was also a reasonable correlation between the two SNPs; using HaploReg v4.1 1000 Genomes the r^2 between rs3917932 and rs3917924 is 0.46 (Ward and Kellis, 2012). I confirmed that rs3917932 and rs3917924 were not independently associated with GCSFR MFI using conditional analysis (data not shown). Combined, this was evidence that the neutrophil count and GCSFR surface expression common signal can be explained by the same causal variant(s), where the Astle *et al.* (2016) study has greater power to distinguish the putative causal variant (rs3917932).

Both rare SNPs, rs3917914 (neutrophil count lead) and rs3917912 (GCSFR surface expression lead) are significant in the GCSFR data (p value 1.62 x 10^{-05} and 1.53 x 10^{-05} respectively). The p values are similar in the neutrophil count association also (9.54 x 10^{-22} and 6.04 x 10^{-19} respectively) (Table 4.5). These variants were perfectly correlated with r^2 of 1 (1000G). Therefore, it is highly likely that the neutrophil count and GCSFR surface expression rare signal can be explained by the same causal variant(s).

I observed an unexpected inverse relationship between the associations of both variants with GCSFR surface expression and neutrophil count. For both common and rare signals, the GCSFR MFI-increasing allele was associated with decreased neutrophil count (Figure 4.8). I had predicted, based on the role of GCSFR in stimulating neutrophil differentiation, that there would exist a positive relationship between surface expression and neutrophil count, where more receptor would result in increased stimulation and higher neutrophil numbers (Panopoulos and Watowich, 2008, Mehta et al., 2015). My results suggest a more complex relationship where rather than the total surface expression of the receptor, possible structural or functional changes could link receptor activity to neutrophil count.

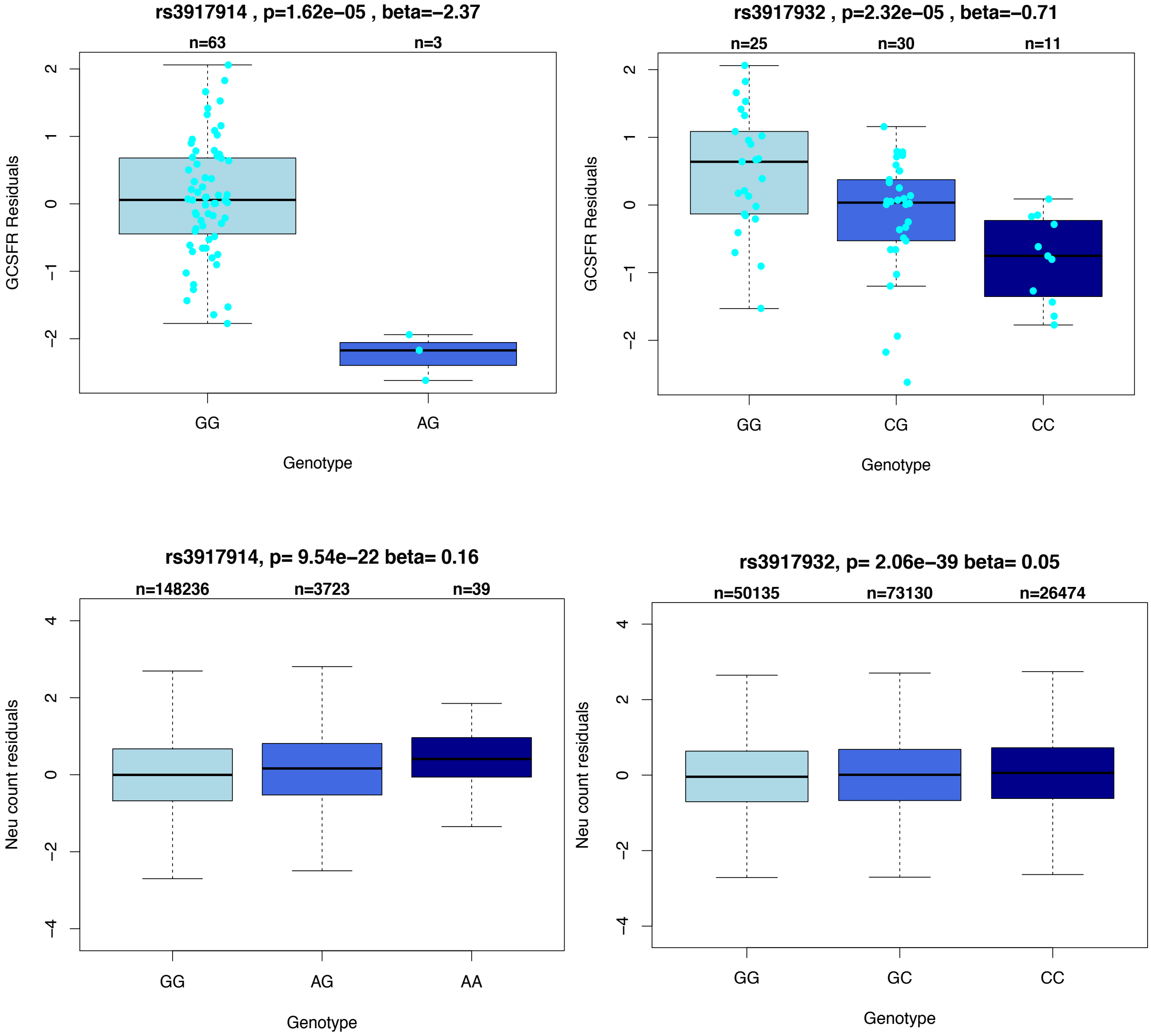


Figure 4.8: Directions of effect of the two independent genetic signals associated with GCSFR surface expression and neutrophil count

Trait residuals are stratified by genotype of the rare (left) and common (right) SNPs. The top panel shows the GCSFR MFI residuals and the bottom panel shows the Astle *et al.* (2016) neutrophil count residuals. GCSFR trait values for each individual in the study (N = 66) are shown but not for neutrophil count (bottom panel) as the number of individuals tested was too high.

4.3.4 Evidence of high molecular functionality in the *CSF3R* genic locus

I next investigated whether using molecular and epigenomic data could aid the interpretation of the functional consequences at this locus and explain the inverse effect. In this study, the most significant SNPs were located in the introns of the *CSF3R* gene. It is known that enhancers, which are regulatory elements that control transcription can be located within introns in addition to upstream and downstream of genes (Pennacchio et al., 2013). Therefore, I investigated and reanalysed several public genomic resources and unpublished data to further characterise the potential molecular and epigenomic characteristics of the *CSF3R* locus.

I found that most GCSFR-associated variants intersected with ROADMAP/ENCODE/BLUEPRINT primary neutrophil chromatin regulatory state data (Supplementary Table 4.1) (Ward and Kellis, 2012, Chen et al., 2016a, Carrillo-de-Santa-Pau et al., 2017). The common and rare SNPs intersected with active TSS (H3K4me3) chromatin states (rs3917924, rs3917912/4) and rs3917932 intersected with an active enhancer state (H3K27ac, H3K4me1). Using a combination of TF ChIP-seq data from primary neutrophils as well as the HL60 differentiated and undifferentiated cell line models described in Chapter 2, I identified that the region also showed evidence of PU.1, C/EBPβ, C/EBPε and the enhancer-associated co-factor, P300 (Supplementary Table 4.1). Examples of binding of some factors in the *CSF3R* locus is shown in Figure 4.9 with representative signal peaks from one BLUEPRINT individual.

The intersection with epigenomic data demonstrated that this was a complex genic region with high molecular functionality, also confirmed by the varied transcript architecture, with 18 different annotated transcripts in GENCODE v15 (Figures 4.9-4.10). While informative for general regional functionality, intersections are prone to chance overlaps and it is difficult to identify a potentially causal SNP or mechanism based solely on the overlap of epigenomic functionality. This is particularly the case, if multiple variants overlap the same or different peaks. In comparison, molecular quantitative trait loci (QTLs) afford the opportunity to dissect individual variant associations with specific molecular marks with statistical confidence. I next investigated such evidence, described below.

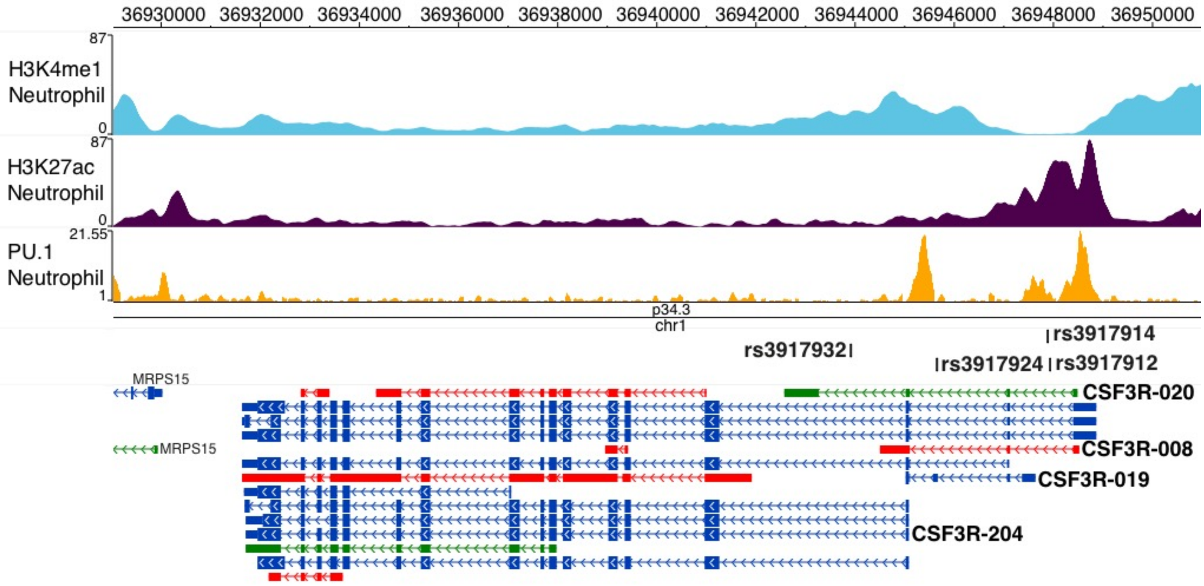


Figure 4.9 Epigenetic and transcript architecture of the *CSF3R* locus
Regional genome plot of the *CSF3R* locus. H3K4me1, H3K27ac and PU.1 peaks are shown for primary mature neutrophils. The lead neutrophil count and GCSFR SNPs are shown. All *CSF3R* transcripts are given with the key transcripts discussed in this thesis labelled. Red transcripts are those with an issue such as retained intron, green transcripts are processed transcripts and blue protein-coding as predicted by GENCODE v17, the earliest version available in the Washu browser (Zhou et al., 2011).

4.3.5 Molecular QTL effects of the common GCSFR MFI association

I accessed the Blueprint consortium human variation panel dataset (Chen et al., 2016a), which provides molecular QTLs and allele-specific events in CD16^{+} neutrophils (Chapter 2 Materials and Methods). The different molecular traits and analytical approaches are summarised in Figure 2.2. In Chapter 2, I focused on investigating gene expression, histone modification and percent-splice in QTLs that were measured in up to 197 individuals. Here, I also assessed the transcript isoform ratio QTLs as well as allele-specific gene and histone effects. Allele-specific approaches evaluate differences in expression or modification signals that occur within an individual heterozygous at the locus of interest and as a result can increase the power to detect genetic effects (Chen et al., 2016a). Transcript isoform effects also indicate splicing events that result in a differential ratio between two transcripts (summarised in Figure 1.1). As part of the main study, expression levels were quantified across all known transcripts normalised by the length of the transcript. The ratio of the two isoforms that exhibited the highest expression change and showed symmetrical changes were then evaluated for significant QTLs (Chen et al., 2016a). The size of the isoform effect is quantified as the maximum difference (MD) in relative expression between SNP genotype groups where a 20% shift in the relative expression of one transcript across genotypes is reflected by an MD of 0.2 (Chen et al., 2016a). I was also able to access additional expression quantifications that were not part of the main analysis including exon and splicing junction QTLs, which were variants associated with different levels of RNA-seq reads specifically at exons and splicing junctions.

SNPs with MAF of less than 1% (allele count = 4), were not included in the blueprint study. In addition, the index neutrophil count SNP, rs3917932, was not included in the final variant set. Therefore, from this point, I investigated effects of the common GCSFR signal and evaluated evidence that the GCSFR index SNP (rs3917924) was associated with specific molecular QTLs or phenotypes.

rs3917924 was significantly associated with the allele-specific expression of the whole *CSF3R* gene (EA = G, p value = 5.59 x 10^{-33}, beta = 0.10) (Chen et al., 2016a). Aligning the effects showed that reduced GCSFR surface expression corresponds with a small increase in GCSFR allele-specific gene expression and increased neutrophil count. This suggests that the reduced surface expression effect is not due to a reduction in the expression of full-length gene.

There were no other associations of rs3917924 in the QTLs assessed as part of the main BLUEPRINT Chen *et al.* (2016a) study. However, I identified highly significant exon and splicing junction associations with rs3917924 (EA = G, p value = 1.09 x 10^{-37}, beta = 1.10, Table 4.5). Interestingly, the exon corresponding to this effect was the third exon located in a truncated transcript, CSF3R-019, not the transcript encoding the full-length receptor (Figure 4.10). rs3917924 is also located within exon 3 of CSF3R-019 (36,945,681-36,945,588). I next investigated whether these effects could reflect regulation at the level of individual transcripts.

4.3.6 Differential *CSF3R* transcript expression associated with rs3917924

Figure 4.10 shows there are 18 *CSF3R* transcripts of varying lengths included in the GENCODE v15 annotation (used in the BLUEPRINT study (Chen et al., 2016a)). Figure 4.10 also shows how each exon contributes to the GCSFR protein domain structure, where the canonical third exon contributes to the N-terminal start of the protein. CSF3R-019 is a truncated transcript with a different third exon compared to the longer protein-coding transcripts such as CSF3R-024.

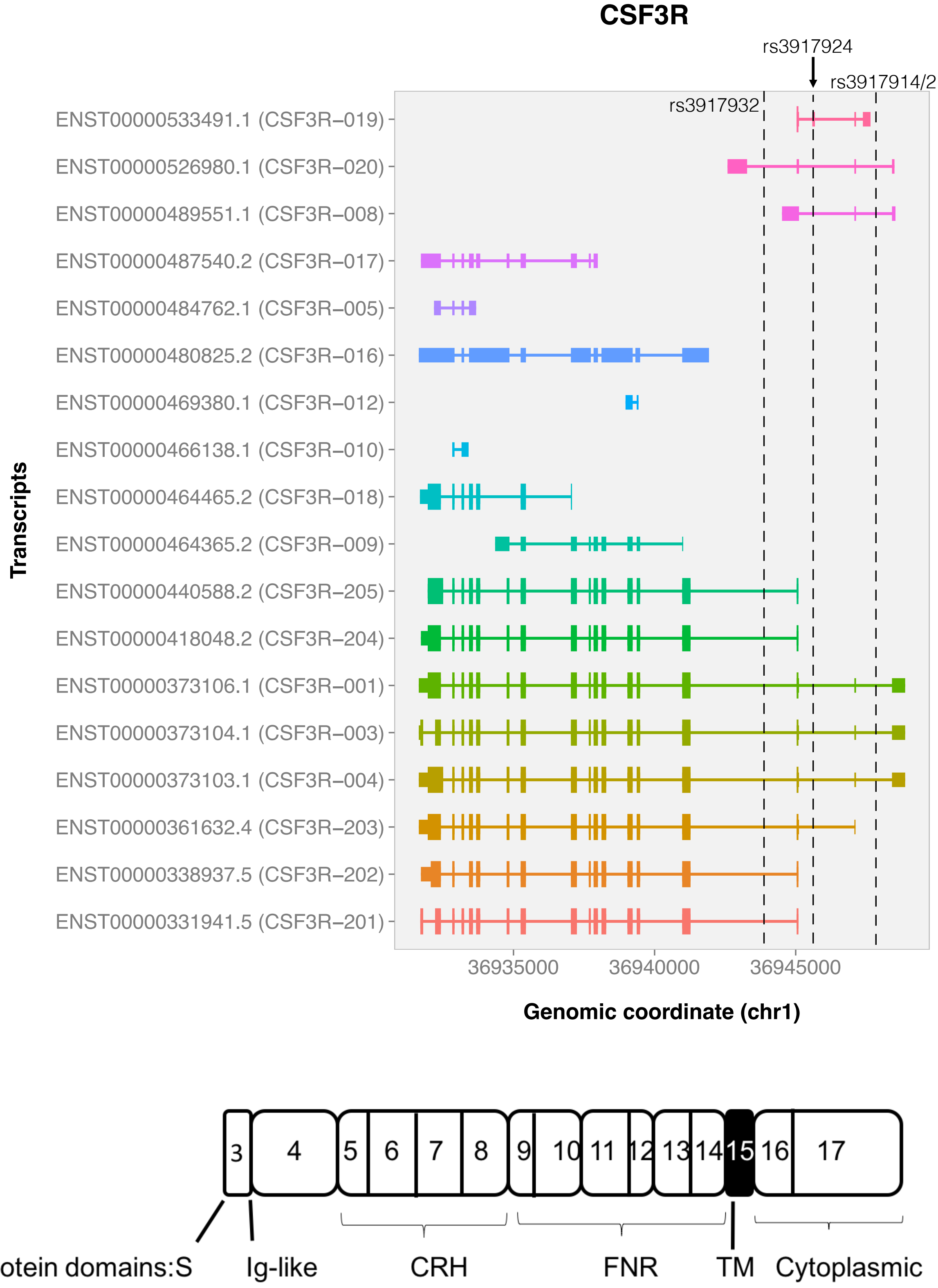


Figure 4.10: GCSFR transcript and protein structure
Upper panel: All possible *CSF3R* transcripts (GENCODE v15) shown with lead SNP positions. The *CSF3R* gene is located on the reverse strand (right to left). Lower panel: Protein domains of the GCSFR protein with respect to contributing exons (exons 3-17 of CSF3R-204, which is most abundantly expressed and generates a functional protein). Different protein domains include an Ig-like domain, a cytokine receptor homologous domain (CRH), three fibronectin type III domains (FNR), a transmembrane domain (TM) and a cytoplasmic domain (Seto et al., 1992). S, the signal peptide, is required for direction of membrane proteins to the cell surface. Figure adapted from (Seto et al., 1992).

I evaluated the relative abundance expression of these 18 transcripts using the transcript expression quantifications from the BLUEPRINT project (Figure 4.11). The CSF3R-204 transcript is the most abundant in neutrophils, followed by the 3' truncated transcripts, CSF3R-008 and CSF3R-020.

ENSEMBL (GRCh37 release 75 2014) predicts that the CSF3R-204 transcript encodes a protein product of 836 amino acids. CSF3R-008 retains an intron and is not predicted to be protein-coding (Figure 4.10). CSF3R-020 is also not predicted to be protein-coding, instead a processed non-coding transcript. In neutrophils and monocytes, the protein-coding CSF3R-204 transcript was the most abundant, and GCSFR is known to be expressed on the surface (higher in neutrophils). In contrast, no protein-coding transcripts are expressed in T cells and the highest expressed transcripts are the truncated CSF3R-020 and CSF3R-008 (Supplementary Figure 4.3) Interestingly, the receptor is not expressed on the surface of lymphocytes, suggesting the switch to higher truncated abundance may reflect a regulatory mechanism generating cell-type specific protein abundance (Christopher et al., 2011).

I next investigated whether there was evidence that transcript level was affected by genotype of the common signal, possibly explaining the significant differential exon expression. I observed a visible change in expression level of the second most abundant transcripts, CSF3R-008 and CSF3R-020 (Figure 4.11). The ratio of these two transcripts was tested as part of the Chen *et al.* (2016) study based on the criteria I explained above. A cautioned, conservative approach for evaluating the significance of associations was used in this study; despite not meeting the stringent significant threshold corrected for multiple testing, there was some evidence of an effect. rs3917924 just missed the significance level for a single test ($p < 0.05$) for association with the CSF3R-008/CSF3R-020 ratio (p value = 0.060, MD = 0.033). Other highly correlated SNPs that I also identified as significantly associated with GCSFR MFI, perhaps showed evidence of a small effect (rs3917933 p = 0.045, rs3917931, p = 0.035).

I aimed to further resolve the transcript based effects, rather than evaluating effects on transcript ratios, I investigated genetic effects on the absolute expression of each *CSF3R* transcript expressed in fragments per kilobase of transcript per million fragments sequenced (FPKM). Using a standard linear regression approach on inverse normalised FPKM, I identified significant associations with rs3917924 and the three truncated transcripts, CSF3R-020, CSF3R-008 and CSF3R-019 (Figure 4.12). The association with CSF3R-019 (containing the exon identified as a significant QTL) was the most significant (p value = 6.002 x 10^{-44}, beta = 1.155, SE = 0.063). This association was significant in monocytes, but reduced compared to neutrophils (p value = 9.148 x 10^{-05}, beta = 0.399, SE = 0.100, Supplementary Figure 4.4). No significant association was found with the expression level of the protein-coding transcript, CSF3R-204 in neutrophils (p value = 0.319) or in monocytes (Supplementary Figure 4.4).

In ENSEMBL, the CSF3R-019 transcript is predicted to generate a very short protein of 21 amino acids in length. Further, in the human cell atlas database, CSF3R-019 is not predicted to contain a transmembrane region and may represent a secreted protein (predicted by SPOCTOPUS (www.proteinatlas.org) (Uhlen et al., 2015, Viklund et al., 2008). The relative abundance shows low expression of CSF3R-019 (Figure 4.11). However, the absolute FPKM ranged from 1.70 to 20.08 across individuals with a median expression of 9.37 FPKM. Other CSF3R transcripts had even lower expression, some with 9 FPKM in neutrophils (CSF3R-001, CSF3R-003, CSF3R-004, CSF3R-009, CSF3R-010). The FPKM cut-off used for transcripts in the BLUEPRINT study was 0.1, suggesting that CSF3R-019 may be moderately expressed, but at much lower levels than the dominant CSF3R-204 transcript (Chen et al., 2016a).

In summary, reduced (allele-specific) gene-expression corresponds to increased truncated CSF3R-019 expression levels, reduced surface GCSFR expression and increased neutrophil count. Whether there is a regulatory role at the transcript level for the three differential expressed transcripts or at the protein-level with respect to the predicted truncated protein from CSF3R-019 requires further functional investigation.

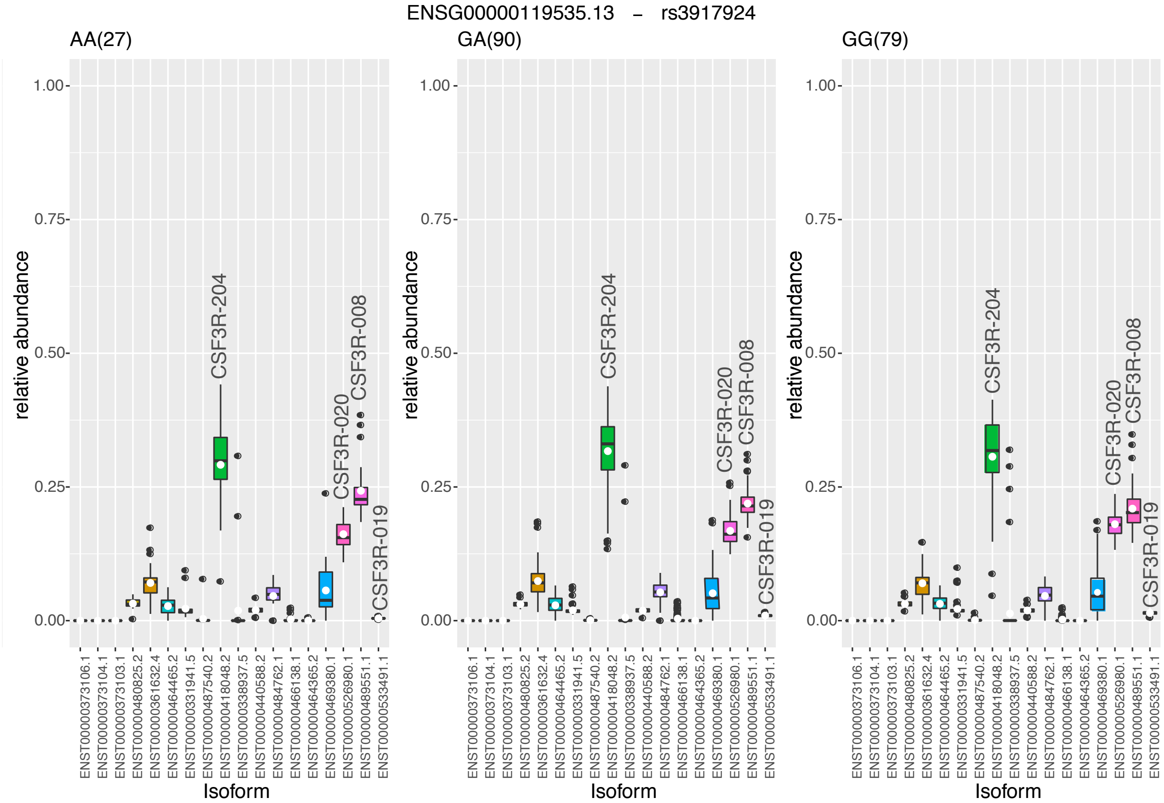


Figure 4.11: Relative abundance of all *CSF3R* transcripts, stratified by genotype of rs3917924, may suggest a marginal genetic splicing effect

Relative *CSF3R* transcript abundances stratified by the genotype of the common index GCSFR SNP identified in the BLUEPRINT study (Chen et al., 2016a). This figure demonstrates that there may be a difference in abundance of some transcripts across genotypes. This figure is adapted from the original produced by Diego Garrido Martin using data analysed as part of the BLUEPRINT consortium (Chen et al., 2016a).

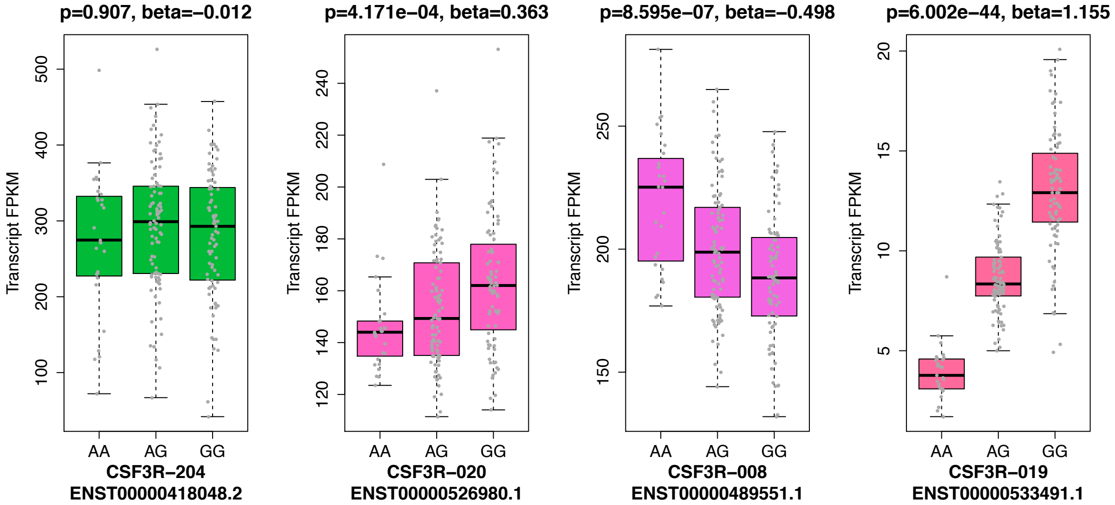


Figure 4.12: Association of the common SNP, rs3917924, with *CSF3R* transcript expression levels

Absolute expression of the transcripts is tested for association with the lead GCSFR surface expression level SNP, rs3917924. Transcripts where a significant association with rs3917924 was identified are shown stratified by genotype. Transcript expression is measured in FPKM (expected fragments per kilobase of transcript per million fragments sequenced) and was estimated as part of the BLUEPRINT consortium using transcripts defined by the Cufflinks tool, without *de novo* assembly (Chen et al., 2016a). Regression was performed using inverse normalised FPKM values. The boxplot colour matches that of the corresponding transcript in Figure 4.10 and Figure 4.11.

4.3.7 Investigation of the rs4760, the PLAUR missense SNP

I also investigated the surface expression of a second receptor, PLAUR, which has a role in neutrophil function (Section 4.1). I tested all variants for association with PLAUR MFI that were within the length of the gene and 500 kb upstream and downstream (N = 16155) including the predicted neutrophil count causal missense SNP, rs4760. The threshold I used for evaluating significantly associated variants was 2.2 x 10^{-04}, which was based on correcting for 218 pruned, independent variants in the region (pairwise r^2 threshold of 0.1).

No significant associations with the PLAUR receptor level were identified (Figure 4.13). The missense SNP, rs4760, did not meet the stringent significance threshold (p value = 0.01). This could be suggestive evidence of association using the p value threshold for a single test (p < 0.05), but confirmation may require further testing.

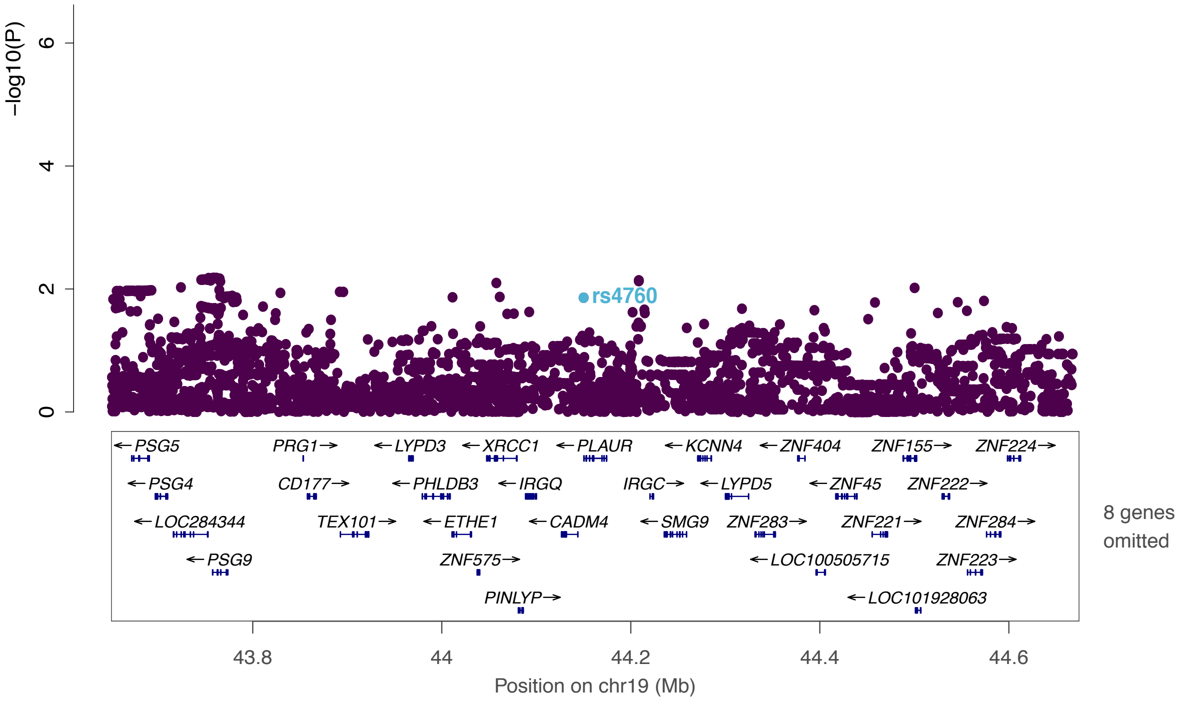


Figure 4.13: PLAUR surface expression association results
Regional association plot with PLAUR receptor surface expression. Variants within the gene and 500kb upstream and downstream of the index SNP are shown with the –$\log_{10}$(p-value) of the association with receptor level (y-axis). There are many genes in this region, with the plot centred around the PLAUR gene. No associations were found to be significant in this region after correction for multiple testing.

Given the complexity of the relationship between GCSFR surface expression levels and neutrophil count, I explored other possible functionality of rs4760 in additional unpublished neutrophil-relevant datasets. I observed a significant association between rs4760 with two additional Sysmex haematological analyser traits, NE-FSC (EA = G, p value = 8.6 x 10^{-15}, beta = 0.075, SE = 0.010) and NE-SFL (EA = G, p value 2.9 x 10^{-17}, beta = 0.077, SE = 0.010) that had been analysed by Parsa Akbari using the INTERVAL cohort of approximately 50,000 individuals (unpublished observations). NE-FSC is a forward scatter parameter that is used as an estimated of neutrophil size. NE-SFL is neutrophil side fluorescence, which increases with a higher amount of cellular DNA and RNA (Buoro et al., 2016, Sysmex Corporation, 2010-2012). These two associations could be related as it is possible that a larger cell could contain a higher amount of nucleic acid. None of the significant SNPs associated with GCSFR surface levels were found to be significantly associated with neutrophil cell size or other additional granularity traits. Interestingly, when I tested for an association of rs4760 and NE-FSC that was also measured using a Sysmex haematology analyser within this recall study (N = 65), the association was not significant (p value = 0.508). The significant association in the larger cohort perhaps suggests that this recall study was limited in power to detect associations of rs4760 with NE-FSC and possibly also PLAUR MFI. Using the pwr R package, I estimated that with a cohort size of 100 and p value threshold of 2.2 x 10^{-04}, the study would be powered to detect variants of similar frequency (rs4760 MAF = 0.16) with a beta > 1, confirming my functional cohort was not powered to detect small effect sizes of associations (Champely, 2012).

The effects of these associations demonstrate rs4760 causes a decrease in neutrophil count (EA = G, p value = 1.428 x 10^{-68}, beta = -8.615 x 10^{-02}, SE = 4.923 x 10^{-03}) and an increase in neutrophil cell size (EA = G, p value = 8.6 x 10^{-15}, beta = 0.075, SE = 0.010). The association with receptor expression from this study demonstrates a decrease in PLAUR surface expression, although this misses the stringent significance threshold applied (p value = 0.01, beta = -0.59, SE = 0.233). Given that little is known about the role of neutrophil size in development or function of neutrophils, this inverse relationship would need further investigation.

4.4 Discussion

In this chapter, I used a recall-by-genotype (RbG) design to test a hypothesis that significant neutrophil count variants located in protein receptor genes were also associated with the surface expression of those receptors. I demonstrated in Chapter 3 how performing a QTL or GWAS study using neutrophil functional phenotypes is technically challenging. RbG studies provide an alternative but highly efficient design to test specific hypotheses based on previous biological and genetic observations, thus requiring a smaller sample size.

Using a cohort of 66 healthy individuals, I identified common and rare signals located in the *CSF3R* locus that are significantly associated with the level of GCSFR neutrophil surface expression. Both signals were also independently associated with neutrophil count from a large GWAS study in 173,480 individuals (Astle et al., 2016). I identified an inverse relationship between these two traits; a decrease in GCSFR at the surface corresponded to an increase in neutrophil count. I also observed other molecular effects, including an increase in expression of the truncated transcript, CSF3R-019.

The opposing direction of effects is not initially intuitive as signalling through the G-CSF receptor is known to increase neutrophil numbers and promote neutrophil differentiation (Lord et al., 1989, Lord et al., 1991, Panopoulos and Watowich, 2008). Naively, a positive relationship between level of GCSFR surface expression and neutrophil numbers could be expected. In this scenario, more signalling through the receptor could lead to a higher number of neutrophils during both differentiation and in the increased release of mature neutrophils from the bone marrow. However, this prediction does not take into account other more complex possibilities. Upon activation, GCSFR is internalised into the cell, a process which is dependent on the C-terminal internalisation motif and functions to regulate signalling preventing over-activation (Kindwall-Keller et al., 2008). Increased surface expression of GCSFR could, therefore, be a result of reduced internalisation due to less activation, perhaps reflecting a difference in protein functionality due to genotype (not tested here). A similar recall-by-genotype study where neutrophils are stimulated by G-CSF overnight, followed by measurement of the phosphorylation of downstream signalling targets such as STAT3 could help in assessing whether GCSFR activation or functionality is altered due to genotype. STAT3 is necessary for the increased neutrophil numbers and maturation during emergency granulopoiesis in response to G-CSF (Zhang et al., 2010). Experiments testing neutrophil responses to stimuli (as described in Chapter 3), could also indicate if despite lower surface receptor levels, the receptor is more sensitive to stimulation in individuals carrying the receptor lowering allele. For example, G-CSF has been shown to prime fMLP-dependent ROS production (Yuo et al., 1990, Khwaja et al., 1992).

Alternatively, GCSFR surface expression could have important ramifications for neutrophil count at the progenitor stage rather than the mature circulating neutrophils as tested here. To investigate this further, CD34$^+$ progenitor cells could be collected from donors and differentiated *in vitro* allowing assessment of GCSFR surface expression at earlier and later stages of neutrophil development.

From this study, it is unclear whether the truncated transcript, CSF3R-019, also plays a regulatory role given the strong association of the common signal with increased expression of this transcript. CSF3R-019 is predicted to generate a short protein of 21 amino acids that is missing a transmembrane domain and eventually secreted (Uhlen et al., 2015, Viklund et al., 2008). Predicted proteins generated from other *CSF3R* transcripts are associated with GO terms such as receptor activity and integral plasma membrane component, but no such terms are associated with the CSF3R-019 transcript. There is precedence for soluble receptor protein forms regulating membrane-bound receptor activity, either in an antagonistic or agonistic manner (Xing et al., 2003). Soluble IL6R (sIL6R) is generated from proteolytic cleavage of the membrane-bound form or alternative splicing (Farahi et al., 2017). sIL6R forms a complex with the ligand, IL6 and activates gp130, in turn leading to increased expression and nuclear translocation of STAT3 (Hawkins et al., 2012, Farahi et al., 2017). The GCSFR Ig-like domain (Figure 4.10) is a close homologue of gp130, located in the N-terminal region and is required for G-CSF binding (Layton et al., 2001, Yorke-Smith et al., 2011). However, all previous experimental evidence for truncated *CSF3R* mRNA and soluble GCSFR (sGCSFR), have been for receptors that are larger than that predicted to be encoded by CSF3R-019, for example, 80 and 85 kDa (Iwasaki et al., 1999, Fukunaga et al., 1990). In addition, CSF3R-019 is expressed to a much lower level than the dominant *CSF3R* transcripts. To test if CSF3R-019 produces a soluble protein, plasma from recalled individuals of different genotypes could be tested for the existence of different soluble forms or qPCR of neutrophil RNA would help assessment of possible transcripts.

Understanding the functional mechanism of the relationship between receptor surface expression and neutrophil count also has the potential for clinical benefit. rs3917924 was previously associated with mobilisation potential and recovery of granulocytes in patients receiving a transplantation of autologous peripheral blood progenitor cells (PBPCT) (Bogunia-Kubik et al., 2012). Peripheral blood progenitor cells (PBPCs) are a source of hematopoietic stem cells that can be used in place of bone marrow for transplantation (Jansen et al., 2002). Mobilisation is the increase in steady-state concentrations of PBPCs by inducing migration of hematopoietic cells into the periphery, which can be supplemented with injection of G-CSF prior to cell collection. Mobilisation also indicates recovery after PBPCT, where recovery is evaluated by the number of granulocytes per μL (Jansen et al., 2002, Bogunia-Kubik et al., 2012). rs3917924 was associated with higher mobilisation potential and a faster recovery of granulocytes in patients after transplant, but in this study corresponded to a lower GCSFR surface expression (Bogunia-Kubik et al., 2012). The authors state that this effect may be due to the alternative allele (A) resulting in an impaired interaction between G-CSF and its receptor leading to a lower response to G-CSF. This could, therefore, be explained by the evidence listed above, that a higher surface level of the receptor is reflective of a lower activity leading and a lower level of internalisation.

I did not identify any significant associations with the PLAUR receptor, but the neutrophil count lead variant, rs4760 was associated with measures of neutrophil size and nucleic acid content in a larger study (N = 50,000, Parsa Akbari, unpublished). This association was also not significant in my cohort of 65 individuals, suggesting that to further evaluate the relationship between neutrophil count, size and PLAUR MFI, larger sample sizes would be required. The lead SNPs associated with GCSFR MFI were not associated with any additional neutrophil measures studied in this larger GWAS, suggesting that the relationship of receptor surface expression and neutrophil count may be specific to each type or functionality of receptor studied and certainly seems more complex than I initially predicted.

Given the well-known examples of other soluble receptors regulating receptor function, I queried my variants in a GWAS of the human plasma proteome including nearly 3000 protein levels in a cohort of 3,301 individuals (Sun et al., 2017). In this, a soluble form of the PLAUR receptor was studied, but there was no equivalent soluble version of GCSFR included in the study. Plasma PLAUR levels were significantly associated, not with rs4760, but with an independent SNP, rs36229204 (EA = T, OA = C, EAF = 0.038, beta = -0.48, SE = 0.07, p value = 5.2×10^{-13}) (Sun et al., 2017). The rare SNP, rs36229204 (CEU MAF = 0.038) was not significantly associated with neutrophil count (Astle et al., 2016) and not correlated with rs4760. This suggests these SNPs are two independent associations within the *PLAUR* locus with different functional consequences, indeed soluble PLAUR has been suggested to competitively inhibit PLAU protein binding to the membrane-anchored PLAUR receptor form (Sloand et al., 2008). Two independent genetic regulatory effects on different stages within the same receptor function pathway further demonstrates how complex these receptor functions and their relationship to neutrophils count are.

In conclusion, genetic analyses can aid the development of functional hypotheses, but further experimental investigation is required. Examples have been discussed here, such as investigating the activity of the receptor as a function of genotype, or investigating the relationship between GCSFR receptor expression and the numbers of neutrophil progenitors. Below, I describe ongoing efforts to replicate the GCSFR receptor signal.

4.5 Future work: Replication of the GCSFR MFI effect

Replication of the common and rare associations with GCSFR MFI is currently being explored using the Sanquin cohort described in Chapter 3. It was possible to recall only a maximum of 20 individuals from this cohort. I therefore designed the study to maximise the power to replicate the MFI effect. In order to select the donors predicted to have the maximal difference between GCSFR expression levels and to investigate the relationship between common and rare signals, I estimated the haplotypes using the genotypes of the three lead SNPs (rs3917914, rs3917924, rs3917932) using SHAPEIT (Materials and Methods). Both common index SNPs were phased to ascertain whether the effect alleles were located on the same haplotype, which would negate the need to select one of the common SNPs and potentially enable an assessment of the causality between the two common SNPs. One rare SNP was used in the haplotype analysis as there was no difference between heterozygous donors given the perfect correlation between rs3917914 and rs3917912. rs3917914 was selected as all of the genotype probabilities of the heterozygote donors were above 0.9, the threshold implemented in SHAPEIT.

Figure 4.14 shows the GCSFR MFI stratified by haplotype in the original RbG study I performed. A similar decrease in surface GCSFR was associated with haplotypes that carry two copies of the common decreasing alleles (**CG**G/**CG**G) than with haplotypes that carry an additional rare decrease allele (CG**A**/GGG), where the decreasing alleles were in order of rs3917932 (C), rs3917924 (G) and rs3917914 (A). Interestingly, a bigger decrease in GCSFR MFI occurred in individuals with haplotypes that were homozygous for the rs3917932 decreasing alleles (CGG/CGG) than those that were homozygous for the rs3917924 decreasing alleles (CGG/GGG). This confirms the larger observed beta estimate for rs3917932 from the association with GCSFR MFI (rs3917932 beta = -0.71, SE = 0.16, rs3917924 beta = -0.68, SE = 0.15) and is evidence that rs3917932 is likely causal common SNP.

I also estimated the haplotypes in the Sanquin cohort using the genotype data from the same SNPs (Table 4.7). Although missing from the original cohort, within the Sanquin cohort four individuals were also homozygous for both the common lead SNPs, CGG/CGA and heterozygous for the decreasing rare allele (A). Based on my association evidence, this haplotype combination would be associated with the lowest GCSFR MFI. I suggested recalling five individuals homozygous for the GAG/GAG haplotype (highest GCSFR MFI) and all CGG/CGA individuals (lowest GCSFR MFI) (Table 4.7). Comparison of individuals with these haplotypes should give the greatest power to replicate the receptor effect. This is currently ongoing. Experiments assessing STAT3 phosphorylation (described above) and measuring neutrophil function responses (Chapter 3) are also being considered.

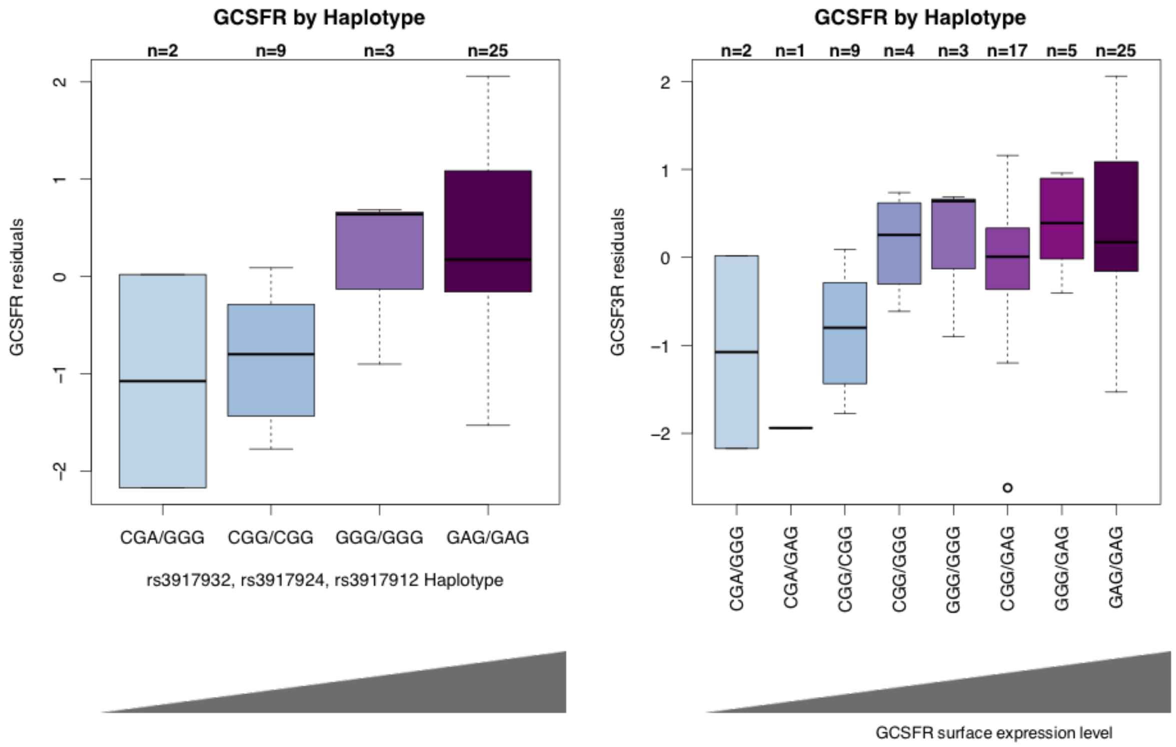


Figure 4.14 GCSFR MFI stratified by haplotype
GCSFR MFI residuals (original Cambridge cohort) stratified by haplotype of rs3917932 (decreasing allele = C), rs3917924 (decreasing allele = G), rs3917914 (decreasing allele = A). The lowest receptor expression was associated with individuals who are heterozygous for the rare variant and the neutrophil count lead variant, rs3917932 (CGA/GGG). CGA contains the lowering effect alleles for all three SNPs. There were no individuals homozygous for the rare lowering haplotype (CGA) in the cohort. The left panel shows the effects in individuals with homozygous haplotypes. The right panel shows MFI for all of the haplotypes estimated using 66 individuals.

Haplotype (estimated frequency)	**Haplotype Genotypes**	**No of Individuals recalled**
CG**A** (3%)	CGG/CG**A**	4
	CG**A**/GGG	2
	CG**A**/GAG	
CGG (43%)	CGG/CGG	4
	CGG/CG**A**	-
	CGG/GAG	
	CGG/GGG	
GGG (22%)	CGG/GGG	
	CG**A**/GGG	-
	GGG/GAG	
	GGG/GGG	
GAG (33%)	CGG/GAG	
	CG**A**/GAG	
	GGG/GAG	
	GAG/GAG	5

Table 4.7: Haplotype frequencies of the Sanquin replication cohort
Haplotype frequencies of the four main haplotypes for the Sanquin replication cohort are shown, along with all the haplotype combinations in 140 individuals. There are 9 unique haplotype genotype combinations and the frequencies matched those from UK10K (data not shown). The suggested number of individuals to be recalled is listed and based on haplotypes that will demonstrate the biggest difference in receptor levels as predicted from the discovery cohort.

Chapter 5

Conclusion and outlook

5 Conclusion and outlook

5.1 The compromise between high-throughput and in-depth functional insights

With recent technological advances and the falling cost of whole-genome sequencing, the main challenge we now face is not the generation of genetic data but in the interpretation of biological mechanisms linking GWAS loci to function. We must consider what are the best approaches in understanding the different steps involved in the pathway from sequence variation to organismal phenotypes such as disease susceptibility. Addressing these challenges will aid the translation of GWAS findings to the clinic and capitalise on the power of genetics to predict and identify drug targets.

In this thesis, I have used a combination of approaches and phenotypic traits in an attempt to understand the cellular and functional consequences of genetic variation. These methods fall into two broad categories. First, large-scale annotation efforts such as those from the ENCODE and BLUEPRINT consortia provide broad functional insight into multiple loci across the genome. These data can be used to either annotate variants with epigenomic functionality or directly linking molecular phenotypes to variants in formal QTL association tests. Second, detailed bespoke investigations such as recall-by-genotype studies or targeted genome-editing provide in-depth insight but are generally lower-throughput and focus on a small number of loci. An important question as we endeavour to move from GWAS to function, is should our focus be on the application of GWAS in larger sample sizes (perhaps in millions of individuals) using existing traits or on increasing the phenotype complexity by continuing recent efforts to apply genetic approaches to functional and cellular traits?

In Chapter 1, I provided further demonstration of the power of molecular QTLs in providing workable mechanistic hypotheses for disease- and complex trait-associated variants as well as suggesting a relevant and specific experimental cellular model. I also provided further support for how the identification of genes dysregulated by variants can highlight potential therapeutic targets for disease treatment and management. The use of such phenotypes to provide functional insight through colocalisation and enrichment methods, as employed here, is a vast improvement on early methods of gene target identification, which relied on proximity to the sentinel variant with no indication of causal cell types involved. Indeed, QTL discovery is now being extended to a wider range of cell types and cellular contexts and will vastly improve our ability to search for the molecular consequences of significant loci. However, it is important to note that large-scale annotation efforts in multiple cell types and large cohorts require substantial financial, logistical and analytical investment, as I experienced through working as part of the BLUEPRINT and UK Biobank consortia (Chen et al., 2016a, Astle et al., 2016). In addition, for certain cell types, it is challenging to access human samples. In some cases, important cell populations are present in low numbers and therefore provide technical challenges in both isolation and the application of genomic approaches. Some advances have already been made in applying ChIP-seq to small populations, such as haematopoietic progenitors (Lara-Astiaso et al., 2014). Many protocols for the differentiation of induced pluripotent stem cells (iPSCs) to different cell populations are already available and advances in this area will also facilitate the study of a wider range of cell types, particularly when coupled to genome-engineering approaches such as CRISPR. Extensively characterised iPSC cell lines are available to research groups through The Human Induced Pluripotent Stem Cells Initiative (HipSci, http://www.hipsci.org).

Even with the already vast amount of epigenome and QTL data available to the wider community, I have shown that to reach full mechanistic understanding still requires detailed, painstaking and often manual integration and annotation of important loci. Creating unified databases to summarise and visualise all available genetic, functional data and the multiple associations for each locus would greatly facilitate these efforts. Certainly, platforms such as HaploReg (Ward and Kellis, 2012), the Open Targets Platform (Koscielny et al., 2017) and DrugBank (Law et al., 2014) already improve the efficiency of this process, as I showed throughout my thesis. However, a unified browser of all available data from multiple cell types would greatly improve the formulation of functional hypotheses of individual genetic loci.

In Chapters 2 and 3 of this thesis, I moved beyond molecular processes, first performing a GWAS on novel neutrophil function traits and second performing an in-depth functional investigation into neutrophil count and surface receptor expression, which in both cases provided functional insight into the biology of the important immune cell type, neutrophils. Recall-by-genotype studies leverage the power of previous GWAS, in this case of nearly 174,000 individuals, to design follow up experiments that delve deeper into the functional processes.

My implementation of functional phenotypes here was not the first example of the utilisation of these traits in genetics (Orru et al., 2013, Roederer et al., 2015, Steri et al., 2017, Li et al., 2016b, Astle et al., 2016, Ahola-Olli et al., 2017). Indeed, for some cell types (particularly PBMCs and monocytes/macrophages) and assays (cytokine production), these measurements are as tractable as the genomic approaches used in molecular QTL studies. However, in Chapter 3, my work demonstrated that for some phenotypes and cell types, measurement of function can be technically complex and subject to many sources of co-variation. Specifically, I outlined the difficulties in working with neutrophil function, which may represent a particularly challenging cell to work with compared to other haematopoietic cell lineages and I discussed possible reasons for this in Chapter 3. However, I gleaned mechanistic insight from small cohorts (10s-100s) using other functional measurements on these cells, namely flow cytometry of the surface expression of receptor proteins. In studying a specific neutrophil count associated locus, the important neutrophil receptor, GCSFR, I identified that while there was no evidence of association of the same common variant with the standard molecular phenotypes (gene expression, histone modifications), the locus was significantly associated with surface receptor expression, allele-specific expression and the expression level of certain *CSF3R* transcripts. Clearly, additional insight can be obtained by combining multiple layers of molecular, cellular and functional information.

Combining the lessons learnt from this thesis and from other similar studies, it is clear that many trait-associated variants affect not just one functional phenotype but many processes from the epigenome, to gene expression, post-transcriptional processes, protein levels, cell function, cell abundances and beyond. In reality, we cannot restrict our efforts to just one type of approach. Collating multiple phenotypes from unified large populations will enable phenome-wide association studies that identify multiple phenotypes affected by a single locus. Thus, molecular and functional networks could be constructed and provide insight into the complexity of interconnections in biological processes. Expanding sample sizes of these cohorts will increase power and enable detection of *trans* effects, which from our currently limited knowledge of these affects, seem to be even more subject to changes in cellular contexts (Delaneau et al., 2017). Providing dense, if not complete, genetic maps using high quality imputation or whole-genome sequencing combined with rich phenotype data will potentially lead to a full description of the genetic architecture of complex traits and perhaps eventually, prediction of an individual's risk based on their genetics. Such initiatives are seen with the recent biobank studies from, for example, the UK Biobank (Collins, 2012), INTERVAL (Moore et al., 2014) and the Precision Medicine Initiative in the United States (Sankar and Parker, 2017).

Below, I discuss two key challenges that we face in understanding functional genetic variation; that is the ever-increasing complexity of the regulatory epigenome and dissecting causal variants and causal relationships between different traits. To end, I highlight particular ongoing efforts to address these challenges and highlight a particular area in human disease that could also reveal interesting interactions with human disease; the microbiome.

5.2 The ever-increasing complexity of the epigenome and the regulatory code

Despite the vast amount we have learnt from genome annotation methods and large-scale consortia, we are still unable to read the regulatory code. We cannot yet predict epigenomic function directly from sequence alone. What we do know is that the regulatory genome is more complex than originally predicted. For example, the genomic regulatory function is highly cell type and context- specific as well as controlled in the three-dimensional space in addition to the early two-dimensional models of regulation of protein binding to a linear DNA sequence. Transcription factors and other regulators bind DNA with multiple cofactors forming large functional clusters, as I set out in Chapter 1. Complex interactions between cofactors located at varying genomic distances could perhaps underlie the observations of distal SNPs affecting TF binding even when they are not located within or nearby to the binding site (Wang et al., 2017b). Context specificity and the complexity of multiple layers of regulation suggests that in order to answer these questions we would require genome-wide binding profiles of *all* possible TFs and cofactors in *all* possible cell types and contexts and connect local effects to *all* interactions in the 3D space. While this seems like a daunting task, observed high functional correlation in genomic domains such as TADs, sub-TADs or variable chromatin modules may reduce the dimensionality of the regulatory genome thereby allowing us to study only the key “seed” factors that explain the majority of variability in the genomic region (Grubert et al., 2015, Waszak et al., 2015). Indeed, in this thesis, I demonstrated that different variant target genes between myeloid and hepatic cells at the CAD *SORT1* locus seemed to be due to the binding of lineage-specific pioneer factors, PU.1 and FOXA1 respectively. It seems remarkable that a small number of factors may be able to control cell-type specific transcription, but the challenge is establishing which layer of functionality underpins the causality at a particular locus and whether causal TF binding, as has been suggested, is a general genome-wide phenomenon or applies to specific cases (Wei et al., 2017).

Indeed, as we build our understanding of the regulatory genome, we find that the genome is even more complex and often challenges previously established functional paradigms. This, in turn, further complicates the interpretation of non-coding sequence variation. For example, I have described in this thesis how the histone modification, H3K4me1, is generally associated with poised or active enhancers and correlates with cell-type specific gene expression (Heintzman et al., 2009). Recently, however, a role for H3K4me1 bound at promoters was observed in inducible gene repression, rather than activation from a distal enhancer (Cheng et al., 2014). This repression seems to be mediated by the methyltransferase, MLL3/4 and appeared to restrict access to readers of the active promoter mark, H3K4me3 (Cheng et al., 2014). It was also recently shown that intragenic enhancers can attenuate gene expression rather than activate it. By using CRISPR-cas9 knock-down, the authors showed that deleting an intragenic enhancer from the mouse ESC gene, *Meis1,* led to de-repression in the region and phenotypes consistent with ESC differentiation (Cinghu et al., 2017). Therefore, intragenic-enhancer repression appeared to have a physiological role. Interestingly, this effect was evident for genes that were not highly expressed but were expressed at medium-to-low levels, which may suggest that weaker intragenic enhancers could be repressive and stronger intragenic enhancers remain active (Cinghu et al., 2017).

Both of these examples show the importance of considering the full genomic context when interpreting variant function. They also demonstrate a need to increase our efforts to study all aspects of genomic regulation to help us ascribe function to important genetic variants.

5.3 Causal variants and causal relationships

Identifying causal variant(s) is an important step in fully understand the mechanism of action of genetic loci. In Chapter 2, I briefly discussed how the colocalisation method provides a posterior probability estimate of each variant being causal that can be used to fine-map potential causal variant(s) based on association evidence (Section 2.3.5.2) (Pickrell et al., 2016). Although not applied in this thesis, there are also other statistical methods available for fine-mapping causal variants (Spain and Barrett, 2015). In addition, I demonstrated how epigenomic information can facilitate fine-mapping, for example, if a particular variant disrupts a TF binding motif (CAD *SORT1,* Figure 2.10) or lies directly under a histone peak (AMD *TNFRSF10A*, Figure 2.17). However, for loci that contain many variants in LD that overlap the same epigenomic marks, we are still limited in detecting the causal variant. In this case, further experimental approaches could be employed to help dissect complex loci. For example, the combination of genome engineering and high-throughput production of induced human pluripotent stem cells (iPSCs) (Kilpinen et al., 2017) allows single nucleotide knock-out in a wide range of differentiated cell types. Coupling CRISPR-Cas9 approaches with tractable experimental read-outs, could help distinguish between closely correlated variants by experimentally comparing their effect sizes on an intermediate phenotype. In addition, using denser genetic information in GWAS also increases the chances of identifying the true causal variant. In Chapter 2, when I used the denser BLUEPRINT phase 2 cohort that included the predicted causal SNP, the associations at the AMD *TNFRSF10A* locus were more significant for all phenotypes of interest.

Establishing causality between different traits is also a complex challenge in human genetics. I described this concept in Chapter 1, where establishing whether a particular trait is a causal risk factor for a disease requires more than just identifying a high correlation between the two factors. Establishing causality between intermediate (including molecular) phenotypes and disease risk provide useful readouts for monitoring disease progression, for potential therapeutic targets and also for experimental use in querying functional relationships at genetic loci. A clear therapeutic success started with the observation that variants within the HMG-CoA reductase gene, *HMGCR*, are associated with lipid levels. Now, *HMGCR* is targeted by cholesterol-lowering statins (Kathiresan et al., 2009). Cholesterol and LDL levels are known risk factors for coronary artery disease (Khera and Kathiresan, 2017).

In this thesis, I used a colocalisation method to provide a statistical assessment of regions of the genome that were associated with two different traits but this method does not provide a definitive demonstration of causality between traits. Mendelian randomization (MR) approaches have been successfully used to assess the causality between traits such as blood counts and complex diseases or LDL and CAD risk (Astle et al., 2016, Khera and Kathiresan, 2017, Holmes et al., 2017). MR application to molecular traits is complex. Often there is extensive QTL pleiotropy, where a variant affects multiple genes, which could all exert different effects on the outcome/disease. I also investigated examples where a genetic variant was associated with multiple molecular phenotypes such as TF binding, enhancer activity, and gene expression. In these cases, there may be multiple molecular routes through which a change in gene expression could affect an outcome. There also could be inter-relationships between the molecular function, which would essentially be described as reverse causation. In Figure 5.1, I highlight possible interactions and directionality between multiple epigenomic functions, partly based on my observations in Chapter 2. These complex molecular relationships can violate MR assumptions (Evans and Davey Smith, 2015).

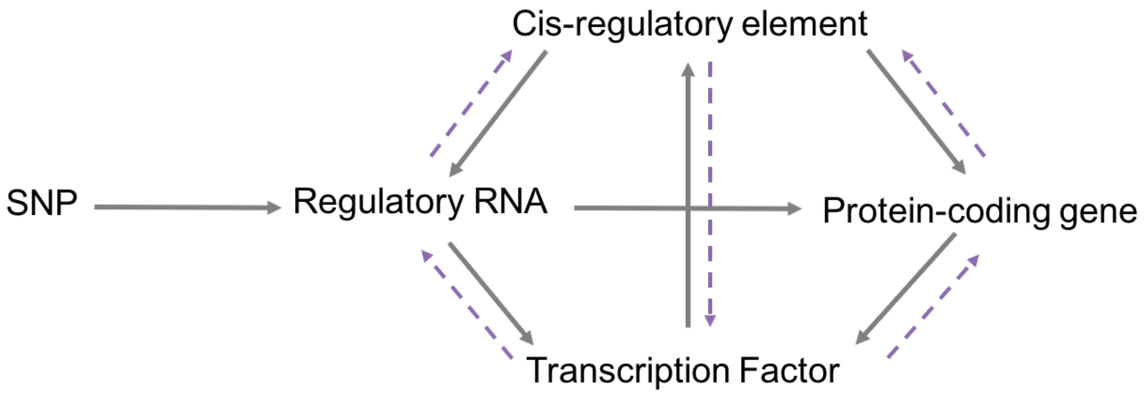


Figure 5.1: Complex molecular regulatory mechanisms
This schematic summarises possible complex regulatory mechanisms that could occur at one genetic locus. Block arrows represent the direction of regulatory effect and dashed arrows represent possible feedback mechanisms or relationships between functions. The generally accepted model is that pioneer factors first bind and then stimulate further chromatin modifications to form cis-regulatory elements, but for some TFs, open chromatin or certain functionality is required (hence the dashed arrow). There could also be feedback mechanisms between the level of gene expression and molecular regulators.

In light of this, I would argue that currently, integrating QTLs with disease GWAS SNPs is state of the art in forming molecular hypotheses, but currently demonstrating a causal relationship must come from downstream experimental testing. Many common variants, however, have small effect sizes on diseases or complex traits, making it difficult to experimentally demonstrate an effect on disease as a result of manipulation of gene function. Mouse models can highlight the consequences of extreme cases of gene knock-out or overexpression, which can identify relevant phenotypes. However, if a target has effects on many pathways, such as *TNFRSF10A*, this could be difficult to ascertain. Intermediate phenotypes that have been shown to be causal risk factors for a particular disease can be used as tractable experimental readouts to establish causal mechanisms. For coronary artery disease, our knowledge of causal intermediate risk factors is quite advanced due to extensive MR analysis and experimental evidence. In particular, LDL levels can be measured in mouse models and the effect of variant knock-down or overexpression on these levels can and have been evaluated (Musunuru et al., 2010). Indeed, cellular responses are also tractable experimental measures. I have already discussed the example of measuring the propensity of macrophages to form foam cells after exposure to oxidised LDL (Reschen et al., 2015).

However, amenable intermediate risk factors have not yet been established for all diseases, particularly for diseases that involve tissues that are difficult to access such as age-related macular degeneration, Alzheimer's disease or schizophrenia. If there is a causal role for peripheral blood cells in these diseases, it would be interesting the perform MR analysis using blood counts and these diseases to assess whether these are causal risk factors. Future work must involve efforts to identify tractable and causal disease intermediates. This, combined with experimental investigation would enable definitive assessment of the causality between peripheral immune function and for example AMD.

While a definitive causal relationship between immune factors and some of the diseases studied here has yet to be determined, therapeutics targeting dysregulated immunological processes that reduce disease severity or progression or manage severe symptoms. Therefore, there is a potential benefit to the management of these disorders through identifying pathways that increase disease severity for example.

5.4 Ongoing efforts and future goals in functional genomics

The next few years promise to be an exciting era for human genetics both in studies of vastly increased sample sizes and in new efforts to understand cellular function. In one of the largest single genetic cohorts and most comprehensive resource, genetic data and detailed phenotypes of 500,000 individuals has been released by UK Biobank project. The increased power of this dataset will allow identification of many more rare variants and will transform our knowledge of the allelic architecture of complex traits (Vasquez et al., 2016).

There are also exciting ongoing efforts in improving our understanding of human cellular biology. The Human Cell Atlas is an international collaboration aiming to use cutting-edge single-cell approaches to classify all human cells (Rozenblatt-Rosen et al., 2017). Although we have learnt much from genomic approaches applied to bulk tissue samples, the results provide an average picture across all possible cellular sub-types. Accurate molecular blueprints on a single cell basis for every type of cell in the human body will provide unprecedented insight into cellular interactions, different cellular states, transitions involved in differentiation, and potentially uncover previously unknown cell subtypes. Integration of these data with GWAS variants will allow us to gain a more detailed understanding of how genetic variation affects cellular phenotypes ultimately influencing disease risk (Rozenblatt-Rosen et al., 2017).

Another important factor in modulating immune responses that was not explored in this thesis is the gut microbiome. Changes in the composition of these bacteria and their taxonomy have been implicated in multiple human diseases such as inflammatory bowel disease, multiple sclerosis, rheumatoid arthritis and asthma (Hall et al., 2017, Berer et al., 2017). A role for host genetic variation in bacterial composition alternations was revealed through using microbiome analysis such as 16S rRNA or metagenomics sequencing as the phenotype in GWAS (Hall et al., 2017). Identified loci had a clear role in disease, for example, 48 IBD risk genes were also associated with altered gut microbiome composition (Hall et al., 2017). The interaction between the host and microbiome also affects the production of host cytokines. Up to 9.7% of the variation in cytokine production was explained by the gut microbiome (Netea et al., 2016). Clearly, the influence of the microbiome in immune responses and genetic disease is an important area to consider. Large cohorts including the Framingham 4000 cohort (Mahmood et al., 2014) and TEDDY (N = 10,000) (Group, 2007) will perform microbiome GWAS and enable further exploration of the host-microbiome inter-relationship. As ever, there is the challenge of assessing causality between these factors and disease outcome.

I discussed how experimental approaches using causal risk factors as tractable readouts and MR approaches are currently employed to assess causality between different phenotypes. In future, longitudinal studies where susceptible individuals are tracked prior to disease onset would be invaluable in assessing the causality between immune function, microbiome composition or molecular changes with disease. This requires the establishment of detailed population-based biobanks that collate a wide-range of rich phenotype data including molecular, cellular and functional measurements as well as lifestyle factors (Leading Edge Voices, 2017). The INTERVAL study is such an example where the design is analogous to a randomised clinical trial. Here the aim is to link genetic determinants to the propensity of individuals to develop anaemia after repeat blood donations (Moore et al., 2014). Rich phenotypic data will be collected at several time points.

For complex diseases with later-in-life onset, such as age-related macular degeneration or Alzheimer’s disease, realising the potential for causality assessments may take several generations of data collection and analysis. However, currently these resources will allow us to build our knowledge of pleiotropy, heritability and genetic architecture of many traits. Biobank resources may also help us understand why some individuals who carry the risk variants do not develop disease (Leading Edge Voices, 2017). Importantly, these biobanks also provide the opportunity of engaging the public in scientific endeavours as we move to collating data from larger and larger populations. Such large and potentially dynamic data resources bring with them consent and ethical challenges, which will need to be addressed (Caulfield and Murdoch, 2017).

In summary, although there remain challenges in interpreting the function of GWAS associations, a decade after the initial GWA studies, we have seen many examples of the power of genetics to uncover novel biological paradigms and potentially improve the success of therapeutic candidates in clinical trials. Indeed, GSK and Regeneron recently committed to sequence the first 50,000 UK Biobank samples, providing a denser variant set than the original array genotyping (GlaxoSmithKline plc., 2017). AstraZeneca announced efforts to build an integrated genomic database consisting of two million genomes as well as clinical trial and electronic health records data (AstraZeneca, 2016). It has even been estimated that by 2025, between 100 million and two billion human genomes would have been sequenced (Stephens et al., 2015). This wealth of genetic data may well enable us to completely resolve the genetic heritability and allelic architecture of complex traits and in doing so transforming our knowledge of basic science and the genetic risk of complex disease.

References

1000 GENOMES PROJECT CONSORTIUM, AUTON, A., BROOKS, L. D., DURBIN, R. M., GARRISON, E. P., KANG, H. M., KORBEL, J. O., MARCHINI, J. L., MCCARTHY, S., MCVEAN, G. A., et al. 2015. A global reference for human genetic variation. *Nature,* 526**,** 68-74.

AARTS, L. H., ROOVERS, O., WARD, A. C. & TOUW, I. P. 2004. Receptor activation and 2 distinct COOH-terminal motifs control G-CSF receptor distribution and internalization kinetics. *Blood,* 103**,** 571-9.

ADAMS, C. C. & WORKMAN, J. L. 1995. Binding of disparate transcriptional activators to nucleosomal DNA is inherently cooperative. *Mol Cell Biol,* 15**,** 1405-21.

ADAMS, D., ALTUCCI, L., ANTONARAKIS, S. E., BALLESTEROS, J., BECK, S., BIRD, A., BOCK, C., BOEHM, B., CAMPO, E., CARICASOLE, A., et al. 2012. BLUEPRINT to decode the epigenetic signature written in blood. *Nat Biotechnol,* 30**,** 224-6.

AHOLA-OLLI, A. V., WURTZ, P., HAVULINNA, A. S., AALTO, K., PITKANEN, N., LEHTIMAKI, T., KAHONEN, M., LYYTIKAINEN, L. P., RAITOHARJU, E., SEPPALA, I., et al. 2017. Genome-wide Association Study Identifies 27 Loci Influencing Concentrations of Circulating Cytokines and Growth Factors. *Am J Hum Genet,* 100**,** 40-50.

AL LAHAM, F., KALSCH, A. I., HEINRICH, L., BIRCK, R., KALLENBERG, C. G., HEERINGA, P. & YARD, B. 2010. Inhibition of neutrophil-mediated production of reactive oxygen species (ROS) by endothelial cells is not impaired in anti-neutrophil cytoplasmic autoantibodies (ANCA)-associated vasculitis patients. *Clin Exp Immunol,* 161**,** 268-75.

ALASOO, K., RODRIGUES, J., MUKHOPADHYAY, S., KNIGHTS, A. J., MANN, A. L., KUNDU, K., CONSORTIUM., H., HALE, C., DOUGAN, G. & GAFFNEY, D. J. 2017. [Pre-print] Shared genetic effects on chromatin and gene expression reveal widespread enhancer priming in immune response [Accessed May 2017]. bioRxiv.

ALI, T., RENKAWITZ, R. & BARTKUHN, M. 2016. Insulators and domains of gene expression. *Curr Opin Genet Dev,* 37**,** 17-26.

ALLEN, E. K., RANDOLPH, A. G., BHANGALE, T., DOGRA, P., OHLSON, M., OSHANSKY, C. M., ZAMORA, A. E., SHANNON, J. P., FINKELSTEIN, D., DRESSEN, A., et al. 2017. SNP-mediated disruption of CTCF binding at the IFITM3 promoter is associated with risk of severe influenza in humans. *Nat Med,* 23**,** 975-983.

AMULIC, B., CAZALET, C., HAYES, G. L., METZLER, K. D. & ZYCHLINSKY, A. 2012. Neutrophil function: from mechanisms to disease. *Annu Rev Immunol,* 30**,** 459-89.

ANDERSON, C. A., PETTERSSON, F. H., CLARKE, G. M., CARDON, L. R., MORRIS, A. P. & ZONDERVAN, K. T. 2010. Data quality control in genetic case-control association studies. *Nat Protoc,* 5**,** 1564-73.

ASSMANN, G., CULLEN, P. & SCHULTE, H. 2002. Simple scoring scheme for calculating the risk of acute coronary events based on the 10-year follow-up of the prospective cardiovascular Munster (PROCAM) study. *Circulation,* 105**,** 310-5.

ASTLE, W. J., ELDING, H., JIANG, T., ALLEN, D., RUKLISA, D., MANN, A. L., MEAD, D., BOUMAN, H., RIVEROS-MCKAY, F., KOSTADIMA, M. A., et al. 2016. The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. *Cell,* 167**,** 1415-1429 e19.

ASTRAZENECA. 2016. *AstraZeneca lauches integrated genomics approach to transform drug discovery and development* [Online]. www.astrazeneca.com. Available: https://www.astrazeneca.com/media-centre/press-releases/2016/AstraZeneca-launches-integrated-genomics-approach-to-transform-drug-discovery-and-development-22042016.html [Accessed 20th November 2017].

AVDI, N. J., WINSTON, B. W., RUSSEL, M., YOUNG, S. K., JOHNSON, G. L. & WORTHEN, G. S. 1996. Activation of MEKK by formyl-methionyl-leucyl-phenylalanine in human neutrophils. Mapping pathways for mitogen-activated protein kinase activation. *J Biol Chem,* 271**,** 33598-606.

BANERJI, J., OLSON, L. & SCHAFFNER, W. 1983. A lymphocyte-specific cellular enhancer is located downstream of the joining region in immunoglobulin heavy chain genes. *Cell,* 33**,** 729-40.

BANERJI, J., RUSCONI, S. & SCHAFFNER, W. 1981. Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. *Cell,* 27**,** 299-308.

BANNISTER, A. J. & KOUZARIDES, T. 2011. Regulation of chromatin by histone modifications. *Cell Res,* 21**,** 381-95.

BANNISTER, A. J., ZEGERMAN, P., PARTRIDGE, J. F., MISKA, E. A., THOMAS, J. O., ALLSHIRE, R. C. & KOUZARIDES, T. 2001. Selective recognition of methylated lysine 9 on histone H3 by the HP1 chromo domain. *Nature,* 410**,** 120-4.

BARDOEL, B. W., KENNY, E. F., SOLLBERGER, G. & ZYCHLINSKY, A. 2014. The balancing act of neutrophils. *Cell Host Microbe,* 15**,** 526-36.

BARNES, D. E. & LINDAHL, T. 2004. Repair and genetic consequences of endogenous DNA base damage in mammalian cells. *Annu Rev Genet,* 38**,** 445-76.

BARREIRO, L. B., TAILLEUX, L., PAI, A. A., GICQUEL, B., MARIONI, J. C. & GILAD, Y. 2012. Deciphering the genetic architecture of variation in the immune response to Mycobacterium tuberculosis infection. *Proc Natl Acad Sci U S A,* 109**,** 1204-9.

BARRETT, J. C. & CARDON, L. R. 2006. Evaluating coverage of genome-wide association studies. *Nat Genet,* 38**,** 659-62.

BARRETT, J. C., DUNHAM, I. & BIRNEY, E. 2015. Using human genetics to make new medicines. *Nat Rev Genet,* 16**,** 561-2.

BARSKI, A., CUDDAPAH, S., CUI, K., ROH, T. Y., SCHONES, D. E., WANG, Z., WEI, G., CHEPELEV, I. & ZHAO, K. 2007. High-resolution profiling of histone methylations in the human genome. *Cell,* 129**,** 823-37.

BASS JDSWCFAJ, D. A. A. R. D. 2015. qvalue: Q-value estimation for false discovery rate control. R package version 2.8.0 ed.

BATTLE, A., KHAN, Z., WANG, S. H., MITRANO, A., FORD, M. J., PRITCHARD, J. K. & GILAD, Y. 2015. Genomic variation. Impact of regulatory variation from RNA to protein. *Science,* 347**,** 664-7.

BATTLE, A. & MONTGOMERY, S. B. 2014. Determining causality and consequence of expression quantitative trait loci. *Hum Genet,* 133**,** 727-35.

BATTLE, A., MOSTAFAVI, S., ZHU, X., POTASH, J. B., WEISSMAN, M. M., MCCORMICK, C., HAUDENSCHILD, C. D., BECKMAN, K. B., SHI, J., MEI, R., et al. 2014. Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome Res,* 24**,** 14-24.

BENTHAM, J., MORRIS, D. L., GRAHAM, D. S. C., PINDER, C. L., TOMBLESON, P., BEHRENS, T. W., MARTIN, J., FAIRFAX, B. P., KNIGHT, J. C., CHEN, L., et al. 2015. Genetic association analyses implicate aberrant regulation of innate and adaptive immunity genes in the pathogenesis of systemic lupus erythematosus. *Nat Genet,* 47**,** 1457-1464.

BENTO, A. P., GAULTON, A., HERSEY, A., BELLIS, L. J., CHAMBERS, J., DAVIES, M., KRUGER, F. A., LIGHT, Y., MAK, L., MCGLINCHEY, S., et al. 2014. The ChEMBL bioactivity database: an update. *Nucleic Acids Res,* 42**,** D1083-90.

BERER, K., GERDES, L. A., CEKANAVICIUTE, E., JIA, X., XIAO, L., XIA, Z., LIU, C., KLOTZ, L., STAUFFER, U., BARANZINI, S. E., et al. 2017. Gut microbiota from multiple sclerosis patients enables spontaneous autoimmune encephalomyelitis in mice. *Proc Natl Acad Sci U S A,* 114**,** 10719-10724.

BERTON, G., LAUDANNA, C., SORIO, C. & ROSSI, F. 1992. Generation of signals activating neutrophil functions by leukocyte integrins: LFA-1 and gp150/95, but not CR3, are able to stimulate the respiratory burst of human neutrophils. *J Cell Biol,* 116**,** 1007-17.

BERTON, G. & LOWELL, C. A. 1999. Integrin signalling in neutrophils and macrophages. *Cell Signal,* 11**,** 621-35.

BETSUYAKU, T., LIU, F., SENIOR, R. M., HAUG, J. S., BROWN, E. J., JONES, S. L., MATSUSHIMA, K. & LINK, D. C. 1999. A functional granulocyte colony-stimulating factor receptor is required for normal chemoattractant-induced neutrophil activation. *J Clin Invest,* 103**,** 825-32.

BIELCZYK-MACZYNSKA, E., SERBANOVIC-CANIC, J., FERREIRA, L., SORANZO, N., STEMPLE, D. L., OUWEHAND, W. H. & CVEJIC, A. 2014. A loss of function screen of identified genome-wide association study Loci reveals new genes controlling hematopoiesis. *PLoS Genet,* 10**,** e1004450.

BINTU, L., ISHIBASHI, T., DANGKULWANICH, M., WU, Y. Y., LUBKOWSKA, L., KASHLEV, M. & BUSTAMANTE, C. 2012. Nucleosomal elements that control the topography of the barrier to transcription. *Cell,* 151**,** 738-49.

BIRNIE, G. D. 1988. The HL60 cell line: a model system for studying human myeloid cell differentiation. *Br J Cancer Suppl,* 9**,** 41-5.

BLAKE, J. A., BULT, C. J., EPPIG, J. T., KADIN, J. A., RICHARDSON, J. E. & MOUSE GENOME DATABASE, G. 2014. The Mouse Genome Database: integration of and access to knowledge about the laboratory mouse. *Nucleic Acids Res,* 42**,** D810-7.

BOGUNIA-KUBIK, K., GIERYNG, A., GEBURA, K. & LANGE, A. 2012. Genetic variant of the G-CSF receptor gene is associated with lower mobilization potential and slower recovery of granulocytes after transplantation of autologous peripheral blood progenitor cells. *Cytokine,* 60**,** 463-7.

BOMBA, L., WALTER, K. & SORANZO, N. 2017. The impact of rare and low-frequency genetic variants in common disease. *Genome Biol,* 18**,** 77.

BOOS, C. J. & LIP, G. Y. 2007. Assessment of mean platelet volume in coronary artery disease - what does it mean? *Thromb Res,* 120**,** 11-3.

BORREGAARD, N. 2010. Neutrophils, from marrow to microbes. *Immunity,* 33**,** 657-70.

BOUMA, G., ANCLIFF, P. J., THRASHER, A. J. & BURNS, S. O. 2010. Recent advances in the understanding of genetic defects of neutrophil number and function. *Br J Haematol,* 151**,** 312-26.

BREITMAN, T. R., SELONICK, S. E. & COLLINS, S. J. 1980. Induction of differentiation of the human promyelocytic leukemia cell line (HL-60) by retinoic acid. *Proc Natl Acad Sci U S A,* 77**,** 2936-40.

BRINKMANN, V., REICHARD, U., GOOSMANN, C., FAULER, B., UHLEMANN, Y., WEISS, D. S., WEINRAUCH, Y. & ZYCHLINSKY, A. 2004. Neutrophil extracellular traps kill bacteria. *Science,* 303**,** 1532-5.

BRODIN, P., JOJIC, V., GAO, T., BHATTACHARYA, S., ANGEL, C. J., FURMAN, D., SHEN-ORR, S., DEKKER, C. L., SWAN, G. E., BUTTE, A. J., et al. 2015. Variation in the human immune system is largely driven by non-heritable influences. *Cell,* 160**,** 37-47.

BRUSSELLE, G. G., JOOS, G. F. & BRACKE, K. R. 2011. New insights into the immunology of chronic obstructive pulmonary disease. *Lancet,* 378**,** 1015-26.

BUORO, S., SEGHEZZI, M., VAVASSORI, M., DOMINONI, P., APASSITI ESPOSITO, S., MANENTI, B., MECCA, T., MARCHESI, G., CASTELLUCCI, E., AZZARA, G., et al. 2016. Clinical significance of cell population data (CPD) on Sysmex XN-9000 in septic patients with our without liver impairment. *Ann Transl Med,* 4**,** 418.

BUSH, W. S. & MOORE, J. H. 2012. Chapter 11: Genome-wide association studies. *PLoS Comput Biol,* 8**,** e1002822.

CAIRNS, J., FREIRE-PRITCHETT, P., WINGETT, S. W., VARNAI, C., DIMOND, A., PLAGNOL, V., ZERBINO, D., SCHOENFELDER, S., JAVIERRE, B. M., OSBORNE, C., et al. 2016. CHiCAGO: robust detection of DNA looping interactions in Capture Hi-C data. *Genome Biol,* 17**,** 127.

CALISKAN, M., BAKER, S. W., GILAD, Y. & OBER, C. 2015. Host genetic variation influences gene expression response to rhinovirus infection. *PLoS Genet,* 11**,** e1005111.

CARRILLO-DE-SANTA-PAU, E., JUAN, D., PANCALDI, V., WERE, F., MARTIN-SUBERO, I., RICO, D., VALENCIA, A. & CONSORTIUM, B. 2017. Automatic identification of informative regions with epigenomic changes associated to hematopoiesis. *Nucleic Acids Res,* 45**,** 9244-9259.

CASCI, T. 2010. Population genetics: SNPs that come in threes. *Nat Rev Genet,* 11**,** 8.

CAULFIELD, T. & MURDOCH, B. 2017. Genes, cells, and biobanks: Yes, there's still a consent problem. *PLoS Biol,* 15**,** e2002654.

CHADEE, D. N., YUASA, T. & KYRIAKIS, J. M. 2002. Direct activation of mitogen-activated protein kinase kinase kinase MEKK1 by the Ste20p homologue GCK and the adapter protein TRAF2. *Mol Cell Biol,* 22**,** 737-49.

CHAMI, N. & LETTRE, G. 2014. Lessons and Implications from Genome-Wide Association Studies (GWAS) Findings of Blood Cell Phenotypes. *Genes (Basel),* 5**,** 51-64.

CHAMPELY, S. 2012. *pwr: Basic functions for power analysis. R package version 1.1.1.* [Online]. Available: http://cran.r-project.org/package=pwr [Accessed].

CHANG, H. H., OH, P. Y., INGBER, D. E. & HUANG, S. 2006. Multistable and multistep dynamics in neutrophil differentiation. *BMC Cell Biol,* 7**,** 11.

CHAVES, F., TIERNO, B. & XU, D. 2005. Quantitative determination of neutrophil VCS parameters by the Coulter automated hematology analyzer: new and reliable indicators for acute bacterial infection. *Am J Clin Pathol,* 124**,** 440-4.

CHEN, L., GE, B., CASALE, F. P., VASQUEZ, L., KWAN, T., GARRIDO-MARTIN, D., WATT, S., YAN, Y., KUNDU, K., ECKER, S., et al. 2016a. Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells. *Cell,* 167**,** 1398-1414 e24.

CHEN, L., KOSTADIMA, M., MARTENS, J. H. A., CANU, G., GARCIA, S. P., TURRO, E., DOWNES, K., MACAULAY, I. C., BIELCZYK-MACZYNSKA, E., COE, S., et al. 2014. Transcriptional diversity during lineage commitment of human blood progenitors. *Science,* 345**,** 1251033.

CHEN, M., LECHNER, J., ZHAO, J., TOTH, L., HOGG, R., SILVESTRI, G., KISSENPFENNIG, A., CHAKRAVARTHY, U. & XU, H. 2016b. STAT3 Activation in Circulating Monocytes Contributes to Neovascular Age-Related Macular Degeneration. *Curr Mol Med,* 16**,** 412-23.

CHENG, J., BLUM, R., BOWMAN, C., HU, D., SHILATIFARD, A., SHEN, S. & DYNLACHT, B. D. 2014. A role for H3K4 monomethylation in gene repression and partitioning of chromatin readers. *Mol Cell,* 53**,** 979-92.

CHRISTOPHER, M. J., RAO, M., LIU, F., WOLOSZYNEK, J. R. & LINK, D. C. 2011. Expression of the G-CSF receptor in monocytic cells is sufficient to mediate hematopoietic progenitor mobilization by G-CSF in mice. *J Exp Med,* 208**,** 251-60.

CHUN, S., CASPARINO, A., PATSOPOULOS, N. A., CROTEAU-CHONKA, D. C., RABY, B. A., DE JAGER, P. L., SUNYAEV, S. R. & COTSAPAS, C. 2017. Limited statistical evidence for shared genetic effects of eQTLs and autoimmune-disease-associated loci in three major immune-cell types. *Nat Genet,* 49**,** 600-605.

CINGHU, S., YANG, P., KOSAK, J. P., CONWAY, A. E., KUMAR, D., OLDFIELD, A. J., ADELMAN, K. & JOTHI, R. 2017. Intragenic Enhancers Attenuate Host Gene Expression. *Mol Cell,* 68**,** 104-117 e6.

CLARKE, R., PEDEN, J. F., HOPEWELL, J. C., KYRIAKOU, T., GOEL, A., HEATH, S. C., PARISH, S., BARLERA, S., FRANZOSI, M. G., RUST, S., et al. 2009. Genetic variants associated with Lp(a) lipoprotein level and coronary disease. *N Engl J Med,* 361**,** 2518-28.

COLLINS, R. 2012. What makes UK Biobank special? *Lancet,* 379**,** 1173-4.

COOPER, D. L., MARTIN, S. G., ROBINSON, J. I., MACKIE, S. L., CHARLES, C. J., NAM, J., CONSORTIUM, Y., ISAACS, J. D., EMERY, P. & MORGAN, A. W. 2012. FcgammaRIIIa expression on monocytes in rheumatoid arthritis: role in immune-complex stimulated TNF production and non-response to methotrexate therapy. *PLoS One,* 7**,** e28918.

CORBIN, L. J., TAN, V. Y., HUGHES, D. A., WADE, K. H., PAUL, D. S., TANSEY, K. E., BUTCHER, F., DUDBRIDGE, F., HOWSON, J. M., JALLOW, M. W., et al. 2017. Causal Analyses, Statistical Efficiency And Phenotypic Precision Through Recall-By-Genotype Study Design. *bioRxiv.*

COUZIN-FRANKEL, J. 2017. Anti-inflammatory prevents heart attacks. *Science,* 357**,** 855.

CREYGHTON, M. P., CHENG, A. W., WELSTEAD, G. G., KOOISTRA, T., CAREY, B. W., STEINE, E. J., HANNA, J., LODATO, M. A., FRAMPTON, G. M., SHARP, P. A., et al. 2010. Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc Natl Acad Sci U S A,* 107**,** 21931-6.

CUI, K., ZANG, C., ROH, T. Y., SCHONES, D. E., CHILDS, R. W., PENG, W. & ZHAO, K. 2009. Chromatin signatures in multipotent human hematopoietic stem cells indicate the fate of bivalent genes during differentiation. *Cell Stem Cell,* 4**,** 80-93.

CZIRR, E. & WYSS-CORAY, T. 2012. The immunology of neurodegeneration. *J Clin Invest,* 122**,** 1156-63.

DAI, C., DENG, Y., QUINLAN, A., GASKIN, F., TSAO, B. P. & FU, S. M. 2014. Genetics of systemic lupus erythematosus: immune responses and end organ resistance to damage. *Curr Opin Immunol,* 31**,** 87-96.

DALY, M. J., RIOUX, J. D., SCHAFFNER, S. F., HUDSON, T. J. & LANDER, E. S. 2001. High-resolution haplotype structure in the human genome. *Nat Genet,* 29**,** 229-32.

DAVOINE, F., LAVIGNE, S., CHAKIR, J., FERLAND, C., BOULAY, M. E. & LAVIOLETTE, M. 2002. Expression of FcgammaRIII (CD16) on human peripheral blood eosinophils increases in allergic conditions. *J Allergy Clin Immunol,* 109**,** 463-9.

DE LANGE, K. M. & BARRETT, J. C. 2015. Understanding inflammatory bowel disease via immunogenetics. *J Autoimmun,* 64**,** 91-100.

DE LANGE, K. M., MOUTSIANAS, L., LEE, J. C., LAMB, C. A., LUO, Y., KENNEDY, N. A., JOSTINS, L., RICE, D. L., GUTIERREZ-ACHURY, J., JI, S. G., et al. 2017. Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *Nat Genet,* 49**,** 256-261.

DE WIT, E. & DE LAAT, W. 2012. A decade of 3C technologies: insights into nuclear organization. *Genes Dev,* 26**,** 11-24.

DEGNER, J. F., PAI, A. A., PIQUE-REGI, R., VEYRIERAS, J. B., GAFFNEY, D. J., PICKRELL, J. K., DE LEON, S., MICHELINI, K., LEWELLEN, N., CRAWFORD, G. E., et al. 2012. DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature,* 482**,** 390-4.

DEKKER, J., RIPPE, K., DEKKER, M. & KLECKNER, N. 2002. Capturing chromosome conformation. *Science,* 295**,** 1306-11.

DEL ROSSO, M., FIBBI, G., PUCCI, M., MARGHERI, F. & SERRATI, S. 2008. The plasminogen activation system in inflammation. *Front Biosci,* 13**,** 4667-86.

DEL ZOPPO, G. J. 1998. The role of platelets in ischemic stroke. *Neurology,* 51**,** S9-14.

DELANEAU, O., MARCHINI, J. & ZAGURY, J. F. 2011. A linear complexity phasing method for thousands of genomes. *Nat Methods,* 9**,** 179-81.

DELANEAU, O., ZAZHYTSKA, M., BOREL, C., HOWALD, C., KUMAR, S., ONGEN, H., POPADIN, K., MARBACH, D., AMBROSINI, G., BIELSER, D., et al. 2017. [Pre-print] Intra- and inter-chromosomal chromatin interactions mediate genetic effects on regulatory networks. *biorxiv.*

DELGADO-VEGA, A. M., ALARCON-RIQUELME, M. E. & KOZYREV, S. V. 2010. Genetic associations in type I interferon related pathways with autoimmunity. *Arthritis Res Ther,* 12 Suppl 1**,** S2.

DELIGEZER, U. & DALAY, N. 2007. Expression of the TRAIL receptors in blood mononuclear cells in leukemia. *Pathol Oncol Res,* 13**,** 290-4.

DENG, W., RUPON, J. W., KRIVEGA, I., BREDA, L., MOTTA, I., JAHN, K. S., REIK, A., GREGORY, P. D., RIVELLA, S., DEAN, A., et al. 2014. Reactivation of developmentally silenced globin genes by forced chromatin looping. *Cell,* 158**,** 849-860.

DENKER, A. & DE LAAT, W. 2015. A Long-Distance Chromatin Affair. *Cell,* 162**,** 942-3.

DERRIEN, T., JOHNSON, R., BUSSOTTI, G., TANZER, A., DJEBALI, S., TILGNER, H., GUERNEC, G., MARTIN, D., MERKEL, A., KNOWLES, D. G., et al. 2012. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res,* 22**,** 1775-89.

DEVARAJ, S., CHEN, X., ADAMS-HUET, B. & JIALAL, I. 2013. Increased expression of Fc-gamma receptors on monocytes in patients with nascent metabolic syndrome. *J Clin Endocrinol Metab,* 98**,** E1510-5.

DIEHL, G. E., YUE, H. H., HSIEH, K., KUANG, A. A., HO, M., MORICI, L. A., LENZ, L. L., CADO, D., RILEY, L. W. & WINOTO, A. 2004. TRAIL-R as a negative regulator of innate immune cell responses. *Immunity,* 21**,** 877-89.

DIMAS, A. S., DEUTSCH, S., STRANGER, B. E., MONTGOMERY, S. B., BOREL, C., ATTAR-COHEN, H., INGLE, C., BEAZLEY, C., GUTIERREZ ARCELUS, M., SEKOWSKA, M., et al. 2009. Common regulatory variation impacts gene expression in a cell type-dependent manner. *Science,* 325**,** 1246-50.

DING, Z., NI, Y., TIMMER, S. W., LEE, B. K., BATTENHOUSE, A., LOUZADA, S., YANG, F., DUNHAM, I., CRAWFORD, G. E., LIEB, J. D., et al. 2014. Quantitative genetics of CTCF binding reveal local sequence effects and different modes of X-chromosome association. *PLoS Genet,* 10**,** e1004798.

DIXON, J. R., SELVARAJ, S., YUE, F., KIM, A., LI, Y., SHEN, Y., HU, M., LIU, J. S. & REN, B. 2012. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature,* 485**,** 376-80.

DO, R., WILLER, C. J., SCHMIDT, E. M., SENGUPTA, S., GAO, C., PELOSO, G. M., GUSTAFSSON, S., KANONI, S., GANNA, A., CHEN, J., et al. 2013. Common variants associated with plasma triglycerides and risk for coronary artery disease. *Nat Genet,* 45**,** 1345-52.

DOBBYN, A., HUCKINS;, L. M., BOOCOCK;, J., SLOOFMAN;, L. G., GLICKSBERG;, B. S., GIAMBARTOLOMEI;, C., HOFFMAN;, G., PERUMAL;, T., GIRDHAR;, K., JIANG;, Y., et al. 2017. [Pre-print] Co-localization of Conditional eQTL and GWAS Signatures in Schizophrenia [Accessed August 2017]. bioRxiv.

DOMINICAL, V. M., BERTOLO, M. B., ALMEIDA, C. B., GARRIDO, V. T., MIGUEL, L. I., COSTA, F. F. & CONRAN, N. 2011. Neutrophils of rheumatoid arthritis patients on anti-TNF-alpha therapy and in disease remission present reduced adhesive functions in association with decreased circulating neutrophil-attractant chemokine levels. *Scand J Immunol,* 73**,** 309-18.

DRUGBANK. 2017. *DrugBank v5.0.0* [Online]. Available: https://www.drugbank.ca/ [Accessed October 2017].

DRYDEN, N. H., BROOME, L. R., DUDBRIDGE, F., JOHNSON, N., ORR, N., SCHOENFELDER, S., NAGANO, T., ANDREWS, S., WINGETT, S., KOZAREWA, I., et al. 2014. Unbiased analysis of potential targets of breast cancer susceptibility loci by Capture Hi-C. *Genome Res,* 24**,** 1854-68.

DUBOIS, P. C., TRYNKA, G., FRANKE, L., HUNT, K. A., ROMANOS, J., CURTOTTI, A., ZHERNAKOVA, A., HEAP, G. A., ADANY, R., AROMAA, A., et al. 2010. Multiple common variants for celiac disease influencing immune gene expression. *Nat Genet,* 42**,** 295-302.

ECKER, S., CHEN, L., PANCALDI, V., BAGGER, F. O., FERNANDEZ, J. M., CARRILLO DE SANTA PAU, E., JUAN, D., MANN, A. L., WATT, S., CASALE, F. P., et al. 2017. Genome-wide analysis of differential transcriptional and epigenetic variability across human immune cell types. *Genome Biol,* 18**,** 18.

EGTEX PROJECT 2017. Enhancing GTEx by bridging the gaps between genotype, gene expression, and disease. *Nat Genet*.

EL-BENNA, J., DANG, P. M. & GOUGEROT-POCIDALO, M. A. 2008. Priming of the neutrophil NADPH oxidase activation: role of p47phox phosphorylation and NOX2 mobilization to the plasma membrane. *Semin Immunopathol,* 30**,** 279-89.

ELGHETANY, M. T., PATEL, J., MARTINEZ, J. & SCHWAB, H. 2003. CD87 as a marker for terminal granulocytic maturation: assessment of its expression during granulopoiesis. *Cytometry B Clin Cytom,* 51**,** 9-13.

EMERY, P., LOPEZ, A. F., BURNS, G. F. & VADAS, M. A. 1988. Synovial fluid neutrophils of patients with rheumatoid arthritis have membrane antigen changes that reflect activation. *Ann Rheum Dis,* 47**,** 34-9.

ENCODE PROJECT CONSORTIUM 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature,* 489**,** 57-74.

ENSRUD, K. & GRIMM, R. H., JR. 1992. The white blood cell count and risk for coronary heart disease. *Am Heart J,* 124**,** 207-13.

EVANS, D. M. & DAVEY SMITH, G. 2015. Mendelian Randomization: New Applications in the Coming Age of Hypothesis-Free Causality. *Annu Rev Genomics Hum Genet,* 16**,** 327-50.

EVANS, D. M., FRAZER, I. H. & MARTIN, N. G. 1999. Genetic and environmental causes of variation in basal levels of blood cells. *Twin Res,* 2, 250-7.

FAIRFAX, B. P., HUMBURG, P., MAKINO, S., NARANBHAI, V., WONG, D., LAU, E., JOSTINS, L., PLANT, K., ANDREWS, R., MCGEE, C., et al. 2014. Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. *Science,* 343, 1246949.

FALSCHLEHNER, C., SCHAEFER, U. & WALCZAK, H. 2009. Following TRAIL's path in the immune system. *Immunology,* 127, 145-54.

FARAHI, N., PAIGE, E., BALLA, J., PRUDENCE, E., FERREIRA, R. C., SOUTHWOOD, M., APPLEBY, S. L., BAKKE, P., GULSVIK, A., LITONJUA, A. A., et al. 2017. Neutrophil-mediated IL-6 receptor trans-signaling and the risk of chronic obstructive pulmonary disease and asthma. *Hum Mol Genet,* 26, 1584-1596.

FARH, K. K., MARSON, A., ZHU, J., KLEINEWIETFELD, M., HOUSLEY, W. J., BEIK, S., SHORESH, N., WHITTON, H., RYAN, R. J., SHISHKIN, A. A., et al. 2015. Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature,* 518, 337-43.

FERNANDEZ-RUIZ, I. 2016. Immune system and cardiovascular disease. *Nat Rev Cardiol,* 13, 503.

FERREIRA, R. C., FREITAG, D. F., CUTLER, A. J., HOWSON, J. M., RAINBOW, D. B., SMYTH, D. J., KAPTOGE, S., CLARKE, P., BOREHAM, C., COULSON, R. M., et al. 2013. Functional IL6R 358Ala allele impairs classical IL-6 receptor signaling and influences risk of diverse inflammatory diseases. *PLoS Genet,* 9, e1003444.

FILIPPAKOPOULOS, P. & KNAPP, S. 2014. Targeting bromodomains: epigenetic readers of lysine acetylation. *Nat Rev Drug Discov,* 13, 337-56.

FINAN, C., GAULTON, A., KRUGER, F. A., LUMBERS, R. T., SHAH, T., ENGMANN, J., GALVER, L., KELLEY, R., KARLSSON, A., SANTOS, R., et al. 2017. The druggable genome and support for target identification and validation in drug development. *Sci Transl Med,* 9.

FINUCANE, H. K., BULIK-SULLIVAN, B., GUSEV, A., TRYNKA, G., RESHEF, Y., LOH, P. R., ANTTILA, V., XU, H., ZANG, C., FARH, K., et al. 2015. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat Genet,* 47, 1228-35.

FLANAGAN, J. F., MI, L. Z., CHRUSZCZ, M., CYMBOROWSKI, M., CLINES, K. L., KIM, Y., MINOR, W., RASTINEJAD, F. & KHORASANIZADEH, S. 2005. Double chromodomains cooperate to recognize the methylated histone H3 tail. *Nature,* 438, 1181-5.

FLETCHER, J. M., LALOR, S. J., SWEENEY, C. M., TUBRIDY, N. & MILLS, K. H. 2010. T cells in multiple sclerosis and experimental autoimmune encephalomyelitis. *Clin Exp Immunol,* 162, 1-11.

FRANSEN, K., VISSCHEDIJK, M. C., VAN SOMMEREN, S., FU, J. Y., FRANKE, L., FESTEN, E. A., STOKKERS, P. C., VAN BODEGRAVEN, A. A., CRUSIUS, J. B., HOMMES, D. W., et al. 2010. Analysis of SNPs with an effect on gene expression identifies UBE2L3 and BCL3 as potential new risk genes for Crohn's disease. *Hum Mol Genet,* 19, 3482-8.

FRAZER, K. A., MURRAY, S. S., SCHORK, N. J. & TOPOL, E. J. 2009. Human genetic variation and its contribution to complex traits. *Nat Rev Genet,* 10, 241-51.

FRITSCHE, L. G., IGL, W., BAILEY, J. N., GRASSMANN, F., SENGUPTA, S., BRAGG-GRESHAM, J. L., BURDON, K. P., HEBBRING, S. J., WEN, C., GORSKI, M., et al. 2016. A large genome-wide association study of age-related macular degeneration highlights contributions of rare and common variants. *Nat Genet,* 48, 134-43.

FUKUNAGA, R., SETO, Y., MIZUSHIMA, S. & NAGATA, S. 1990. Three different mRNAs encoding human granulocyte colony-stimulating factor receptor. *Proc Natl Acad Sci U S A,* 87, 8702-6.

FULLGRABE, J., KAVANAGH, E. & JOSEPH, B. 2011. Histone onco-modifications. *Oncogene,* 30, 3391-403.

G. TEX CONSORTIUM 2015. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science,* 348, 648-60.

G. TEX CONSORTIUM 2017. Genetic effects on gene expression across human tissues. *Nature,* 550**,** 204-213.

GAFFNEY, P. M., KEARNS, G. M., SHARK, K. B., ORTMANN, W. A., SELBY, S. A., MALMGREN, M. L., ROHLF, K. E., OCKENDEN, T. C., MESSNER, R. P., KING, R. A., et al. 1998. A genome-wide search for susceptibility genes in human systemic lupus erythematosus sib-pair families. *Proc Natl Acad Sci U S A,* 95**,** 14875-9.

GARNER, C., TATU, T., REITTIE, J. E., LITTLEWOOD, T., DARLEY, J., CERVINO, S., FARRALL, M., KELLY, P., SPECTOR, T. D. & THEIN, S. L. 2000. Genetic influences on F cells and other hematologic variables: a twin heritability study. *Blood,* 95**,** 342-6.

GASZNER, M. & FELSENFELD, G. 2006. Insulators: exploiting transcriptional and epigenetic mechanisms. *Nat Rev Genet,* 7**,** 703-13.

GATZ, M., REYNOLDS, C. A., FRATIGLIONI, L., JOHANSSON, B., MORTIMER, J. A., BERG, S., FISKE, A. & PEDERSEN, N. L. 2006. Role of genes and environments for explaining Alzheimer disease. *Arch Gen Psychiatry,* 63**,** 168-74.

GENNERY, A. 2017. Recent advances in understanding and treating chronic granulomatous disease. *F1000Res,* 6**,** 1427.

GEUVADIS. 2010. *Percentage Splicing Index* [Online]. GeuvadisWiki. Available: http://geuvadiswiki.crg.es/index.php/Percentage_Splicing_Index [Accessed October 2017].

GHATTAS, A., GRIFFITHS, H. R., DEVITT, A., LIP, G. Y. & SHANTSILA, E. 2013. Monocytes in coronary artery disease and atherosclerosis: where are we now? *J Am Coll Cardiol,* 62**,** 1541-51.

GHISLETTI, S., BAROZZI, I., MIETTON, F., POLLETTI, S., DE SANTA, F., VENTURINI, E., GREGORY, L., LONIE, L., CHEW, A., WEI, C. L., et al. 2010. Identification and characterization of enhancers controlling the inflammatory gene expression program in macrophages. *Immunity,* 32**,** 317-28.

GIEGER, C., RADHAKRISHNAN, A., CVEJIC, A., TANG, W., PORCU, E., PISTIS, G., SERBANOVIC-CANIC, J., ELLING, U., GOODALL, A. H., LABRUNE, Y., et al. 2011. New gene functions in megakaryopoiesis and platelet formation. *Nature,* 480**,** 201-8.

GLAXOSMITHKLINE PLC. 2017. *UK Biobank, GSK and Regeneron announce largest gene sequencing initiative on world's most detailed health database to improve drug discovery and disease diagnosis* [Online]. www.gsk.com. Available: http://www.gsk.com/en-gb/media/press-releases/uk-biobank-gsk-and-regeneron-announce-largest-gene-sequencing-initiative-on-world-s-most-detailed-health-database-to-improve-drug-discovery-and-disease-diagnosis/ [Accessed 20th November 2017].

GLINOS, D. A., SOSKIC, B. & TRYNKA, G. 2017. Immunogenomic approaches to understand the function of immune disease variants. *Immunology*.

GRECO, C. M. & CONDORELLI, G. 2015. Epigenetic modifications and noncoding RNAs in cardiac hypertrophy and failure. *Nat Rev Cardiol,* 12**,** 488-97.

GRONTVED, L., JOHN, S., BAEK, S., LIU, Y., BUCKLEY, J. R., VINSON, C., AGUILERA, G. & HAGER, G. L. 2013. C/EBP maintains chromatin accessibility in liver and facilitates glucocorticoid receptor recruitment to steroid response elements. *EMBO J,* 32**,** 1568-83.

GROUP, T. S. 2007. The Environmental Determinants of Diabetes in the Young (TEDDY) study: study design. *Pediatr Diabetes,* 8**,** 286-98.

GRUBERT, F., ZAUGG, J. B., KASOWSKI, M., URSU, O., SPACEK, D. V., MARTIN, A. R., GREENSIDE, P., SRIVAS, R., PHANSTIEL, D. H., PEKOWSKA, A., et al. 2015. Genetic Control of Chromatin States in Humans Involves Local and Distal Chromosomal Interactions. *Cell,* 162**,** 1051-65.

GRUNDBERG, E., SMALL, K. S., HEDMAN, A. K., NICA, A. C., BUIL, A., KEILDSON, S., BELL, J. T., YANG, T. P., MEDURI, E., BARRETT, A., et al. 2012. Mapping cis- and trans-regulatory effects across multiple tissues in twins. *Nat Genet,* 44**,** 1084-9.

GRUNIN, M., BURSTYN-COHEN, T., HAGBI-LEVI, S., PELED, A. & CHOWERS, I. 2012. Chemokine receptor expression in peripheral blood monocytes from patients with neovascular age-related macular degeneration. *Invest Ophthalmol Vis Sci,* 53**,** 5292-300.

GRUNIN, M., HAGBI-LEVI, S., RINSKY, B., SMITH, Y. & CHOWERS, I. 2016. Transcriptome Analysis on Monocytes from Patients with Neovascular Age-Related Macular Degeneration. *Sci Rep,* 6**,** 29046.

GUENTHER, M. G., LEVINE, S. S., BOYER, L. A., JAENISCH, R. & YOUNG, R. A. 2007. A chromatin landmark and transcription initiation at most promoters in human cells. *Cell,* 130**,** 77-88.

GUICCIARDI, M. E. & GORES, G. J. 2009. Life and death by death receptors. *FASEB J,* 23**,** 1625-37.

GUIRO, J. & MURPHY, S. 2017. Regulation of expression of human RNA polymerase II-transcribed snRNA genes. *Open Biol,* 7.

GUO, H., FORTUNE, M. D., BURREN, O. S., SCHOFIELD, E., TODD, J. A. & WALLACE, C. 2015. Integration of disease association and eQTL data using a Bayesian colocalisation approach highlights six candidate causal genes in immune-mediated diseases. *Hum Mol Genet,* 24**,** 3305-13.

GUPTA, S. & KAPLAN, M. J. 2016. The role of neutrophils and NETosis in autoimmune and renal diseases. *Nat Rev Nephrol,* 12**,** 402-13.

GUTIERREZ-ARCELUS, M., RICH, S. S. & RAYCHAUDHURI, S. 2016. Autoimmune diseases - connecting risk alleles with molecular traits of the immune system. *Nat Rev Genet,* 17**,** 160-74.

GYETKO, M. R., SITRIN, R. G., FULLER, J. A., TODD, R. F., 3RD, PETTY, H. & STANDIFORD, T. J. 1995. Function of the urokinase receptor (CD87) in neutrophil chemotaxis. *J Leukoc Biol,* 58**,** 533-8.

HAGEMAN, G. S., LUTHERT, P. J., VICTOR CHONG, N. H., JOHNSON, L. V., ANDERSON, D. H. & MULLINS, R. F. 2001. An integrated hypothesis that considers drusen as biomarkers of immune-mediated processes at the RPE-Bruch's membrane interface in aging and age-related macular degeneration. *Prog Retin Eye Res,* 20**,** 705-32.

HALE, C., YEUNG, A., GOULDING, D., PICKARD, D., ALASOO, K., POWRIE, F., DOUGAN, G. & MUKHOPADHYAY, S. 2015. Induced pluripotent stem cell derived macrophages as a cellular system to study salmonella and other pathogens. *PLoS One,* 10**,** e0124307.

HALL, A. B., TOLONEN, A. C. & XAVIER, R. J. 2017. Human genetic variation and the gut microbiome in disease. *Nat Rev Genet,* 18**,** 690-699.

HALLETT, M. B. & LLOYDS, D. 1995. Neutrophil priming: the cellular signals that say 'amber' but not 'green'. *Immunol Today,* 16**,** 264-8.

HAMEED, I., MASOODI, S. R., MIR, S. A., NABI, M., GHAZANFAR, K. & GANAI, B. A. 2015. Type 2 diabetes mellitus: From a metabolic disorder to an inflammatory condition. *World J Diabetes,* 6**,** 598-612.

HARDISON, R. C. & TAYLOR, J. 2012. Genomic approaches towards finding cis-regulatory modules in animals. *Nat Rev Genet,* 13**,** 469-83.

HARRINGTON, R. A. 2017. Targeting Inflammation in Coronary Artery Disease. *N Engl J Med,* 377**,** 1197-1198.

HARROW, J., FRANKISH, A., GONZALEZ, J. M., TAPANARI, E., DIEKHANS, M., KOKOCINSKI, F., AKEN, B. L., BARRELL, D., ZADISSA, A., SEARLE, S., et al. 2012. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res,* 22**,** 1760-74.

HAWKINS, G. A., ROBINSON, M. B., HASTIE, A. T., LI, X., LI, H., MOORE, W. C., HOWARD, T. D., BUSSE, W. W., ERZURUM, S. C., WENZEL, S. E., et al. 2012. The IL6R variation Asp(358)Ala is a potential modifier of lung function in subjects with asthma. *J Allergy Clin Immunol,* 130**,** 510-5 e1.

HAYTER, S. M. & COOK, M. C. 2012. Updated assessment of the prevalence, spectrum and case definition of autoimmune disease. *Autoimmun Rev,* 11**,** 754-65.

HEINTZMAN, N. D., HON, G. C., HAWKINS, R. D., KHERADPOUR, P., STARK, A., HARP, L. F., YE, Z., LEE, L. K., STUART, R. K., CHING, C. W., et al. 2009. Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature,* 459**,** 108-12.

HEINZ, S., BENNER, C., SPANN, N., BERTOLINO, E., LIN, Y. C., LASLO, P., CHENG, J. X., MURRE, C., SINGH, H. & GLASS, C. K. 2010. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell,* 38**,** 576-89.

HEINZ, S., ROMANOSKI, C. E., BENNER, C., ALLISON, K. A., KAIKKONEN, M. U., OROZCO, L. D. & GLASS, C. K. 2013. Effect of natural genetic variation on enhancer selection and function. *Nature,* 503**,** 487-92.

HEINZ, S., ROMANOSKI, C. E., BENNER, C. & GLASS, C. K. 2015. The selection and function of cell type-specific enhancers. *Nat Rev Mol Cell Biol,* 16**,** 144-54.

HENEKA, M. T., GOLENBOCK, D. T. & LATZ, E. 2015. Innate immunity in Alzheimer's disease. *Nat Immunol,* 16**,** 229-36.

HILLER, M., HUSE, K., SZAFRANSKI, K., JAHN, N., HAMPE, J., SCHREIBER, S., BACKOFEN, R. & PLATZER, M. 2006. Single-nucleotide polymorphisms in NAGNAG acceptors are highly predictive for variations of alternative splicing. *Am J Hum Genet,* 78**,** 291-302.

HOFFMAN, M., BLUM, A., BARUCH, R., KAPLAN, E. & BENJAMIN, M. 2004. Leukocytes and coronary heart disease. *Atherosclerosis,* 172**,** 1-6.

HOLLENBACH, J. A. & OKSENBERG, J. R. 2015. The immunogenetics of multiple sclerosis: A comprehensive review. *J Autoimmun,* 64**,** 13-25.

HOLMES, M. V., ALA-KORPELA, M. & SMITH, G. D. 2017. Mendelian randomization in cardiometabolic disease: challenges in evaluating causality. *Nat Rev Cardiol,* 14**,** 577-590.

HON, G. C., HAWKINS, R. D. & REN, B. 2009. Predictive chromatin signatures in the mammalian genome. *Hum Mol Genet,* 18**,** R195-201.

HONEYCUTT, P. J. & NIEDEL, J. E. 1986. Cytochalasin B enhancement of the diacylglycerol response in formyl peptide-stimulated neutrophils. *J Biol Chem,* 261**,** 15900-5.

HRDLICKOVA, B., DE ALMEIDA, R. C., BOREK, Z. & WITHOFF, S. 2014. Genetic variation in the non-coding genome: Involvement of micro-RNAs and long non-coding RNAs in disease. *Biochim Biophys Acta,* 1842**,** 1910-1922.

HUANG, G., XU, X. C., ZHOU, J. S., LI, Z. Y., CHEN, H. P., WANG, Y., LI, W., SHEN, H. H. & CHEN, Z. H. 2017a. Neutrophilic Inflammation in the Immune Responses of Chronic Obstructive Pulmonary Disease: Lessons from Animal Models. *J Immunol Res,* 2017**,** 7915975.

HUANG, H., FANG, M., JOSTINS, L., UMICEVIC MIRKOV, M., BOUCHER, G., ANDERSON, C. A., ANDERSEN, V., CLEYNEN, I., CORTES, A., CRINS, F., et al. 2017b. Fine-mapping inflammatory bowel disease loci to single-variant resolution. *Nature,* 547**,** 173-178.

HUANG, J., HOWIE, B., MCCARTHY, S., MEMARI, Y., WALTER, K., MIN, J. L., DANECEK, P., MALERBA, G., TRABETTI, E., ZHENG, H. F., et al. 2015. Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. *Nat Commun,* 6**,** 8111.

HUANG, K. L., MARCORA, E., PIMENOVA, A. A., DI NARZO, A. F., KAPOOR, M., JIN, S. C., HARARI, O., BERTELSEN, S., FAIRFAX, B. P., CZAJKOWSKI, J., et al. 2017c. A common haplotype lowers PU.1 expression in myeloid cells and delays onset of Alzheimer's disease. *Nat Neurosci,* 20**,** 1052-1061.

HUGOT, J. P., CHAMAILLARD, M., ZOUALI, H., LESAGE, S., CEZARD, J. P., BELAICHE, J., ALMER, S., TYSK, C., O'MORAIN, C. A., GASSULL, M., et al. 2001. Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature,* 411**,** 599-603.

HUKKINEN, M., KAPRIO, J., BROMS, U., VILJANEN, A., KOTZ, D., RANTANEN, T. & KORHONEN, T. 2011. Heritability of lung function: a twin study among never-smoking elderly women. *Twin Res Hum Genet,* 14**,** 401-7.

HUMPHRIES, C., KOHLI, M. A., WHITEHEAD, P., MASH, D. C., PERICAK-VANCE, M. A. & GILBERT, J. 2015. Alzheimer disease (AD) specific transcription, DNA methylation and splicing in twenty AD associated loci. *Mol Cell Neurosci,* 67**,** 37-45.

IL R. GENETICS CONSORTIUM EMERGING RISK FACTORS COLLABORATION, SARWAR, N., BUTTERWORTH, A. S., FREITAG, D. F., GREGSON, J., WILLEIT, P., GORMAN, D. N., GAO, P., SALEHEEN, D., RENDON, A., et al. 2012. Interleukin-6 receptor pathways in coronary heart disease: a collaborative meta-analysis of 82 studies. *Lancet,* 379**,** 1205-13.

IMMUNOBASE. 2017. *ImmunoBase* [Online]. Available: https://www.immunobase.org/ [Accessed October 2016].

INTERLEUKIN-6 RECEPTOR MENDELIAN RANDOMISATION ANALYSIS CONSORTIUM, SWERDLOW, D. I., HOLMES, M. V., KUCHENBAECKER, K. B., ENGMANN, J. E., SHAH, T., SOFAT, R., GUO, Y., CHUNG, C., PEASEY, A., et al. 2012. The interleukin-6 receptor as a target for prevention of coronary heart disease: a mendelian randomisation analysis. *Lancet,* 379**,** 1214-24.

INTERNATIONAL HAPMAP CONSORTIUM 2005. A haplotype map of the human genome. *Nature,* 437**,** 1299-320.

IOTCHKOVA, V., RITCHIE, G. R. S., GEIHS, M., MORGANELLA, S., MIN, J. L., WALTER, K., TIMPSON, N. J., CONSORTIUM., U. K., DUNHAM, I., BIRNEY, E., et al. 2016. [Pre-print] GARFIELD - GWAS Analysis of Regulatory or Functional Information Enrichment with LD correction [Accessed October 2017]. bioRxivs.

IWAFUCHI-DOI, M., DONAHUE, G., KAKUMANU, A., WATTS, J. A., MAHONY, S., PUGH, B. F., LEE, D., KAESTNER, K. H. & ZARET, K. S. 2016. The Pioneer Transcription Factor FoxA Maintains an Accessible Nucleosome Configuration at Enhancers for Tissue-Specific Gene Activation. *Mol Cell,* 62**,** 79-91.

IWASAKI, H., SHIMODA, K., OKAMURA, S., OTSUKA, T., NAGAFUJI, K., HARADA, N., OHNO, Y., MIYAMOTO, T., AKASHI, K., HARADA, M., et al. 1999. Production of soluble granulocyte colony-stimulating factor receptors from myelomonocytic cells. *J Immunol,* 163**,** 6907-11.

JACOB, C., LEPORT, M., SZILAGYI, C., ALLEN, J. M., BERTRAND, C. & LAGENTE, V. 2002. DMSO-treated HL60 cells: a model of neutrophil-like cells mainly expressing PDE4B subtype. *Int Immunopharmacol,* 2**,** 1647-56.

JAKOBSEN, J. S., WAAGE, J., RAPIN, N., BISGAARD, H. C., LARSEN, F. S. & PORSE, B. T. 2013. Temporal mapping of CEBPA and CEBPB binding during liver regeneration reveals dynamic occupancy and specific regulatory codes for homeostatic and cell cycle gene batteries. *Genome Res,* 23**,** 592-603.

JANSEN, J., THOMPSON, J. M., DUGAN, M. J., NOLAN, P., WIEMANN, M. C., BIRHIRAY, R., HENSLEE-DOWNEY, P. J. & AKARD, L. P. 2002. Peripheral blood progenitor cell transplantation. *Ther Apher,* 6**,** 5-14.

JAVIERRE, B. M., BURREN, O. S., WILDER, S. P., KREUZHUBER, R., HILL, S. M., SEWITZ, S., CAIRNS, J., WINGETT, S. W., VARNAI, C., THIECKE, M. J., et al. 2016. Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. *Cell,* 167**,** 1369-1384 e19.

JENSEN, H. K., DONSKOV, F., MARCUSSEN, N., NORDSMARK, M., LUNDBECK, F. & VON DER MAASE, H. 2009. Presence of intratumoral neutrophils is an independent prognostic factor in localized renal cell carcinoma. *J Clin Oncol,* 27**,** 4709-17.

JOEHANES, R., ZHANG, X., HUAN, T., YAO, C., YING, S. X., NGUYEN, Q. T., DEMIRKALE, C. Y., FEOLO, M. L., SHAROPOVA, N. R., STURCKE, A., et al. 2017. Integrated genome-wide analysis of expression quantitative trait loci aids interpretation of genomic association studies. *Genome Biol,* 18**,** 16.

JONES, P. A. 2012. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat Rev Genet,* 13**,** 484-92.

JONSSON, S., SVEINBJORNSSON, G., DE LAPUENTE PORTILLA, A. L., SWAMINATHAN, B., PLOMP, R., DEKKERS, G., AJORE, R., ALI, M., BENTLAGE, A. E. H., ELMER, E., et al. 2017. Identification of sequence variants influencing immunoglobulin levels. *Nat Genet,* 49**,** 1182-1191.

JOSTINS, L., RIPKE, S., WEERSMA, R. K., DUERR, R. H., MCGOVERN, D. P., HUI, K. Y., LEE, J. C., SCHUMM, L. P., SHARMA, Y., ANDERSON, C. A., et al. 2012. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature,* 491**,** 119-24.

KADENER, S., CRAMER, P., NOGUES, G., CAZALLA, D., DE LA MATA, M., FEDEDA, J. P., WERBAJH, S. E., SREBROW, A. & KORNBLIHTT, A. R. 2001. Antagonistic effects of T-Ag and VP16 reveal a role for RNA pol II elongation on alternative splicing. *EMBO J,* 20**,** 5759-68.

KAGEY, M. H., NEWMAN, J. J., BILODEAU, S., ZHAN, Y., ORLANDO, D. A., VAN BERKUM, N. L., EBMEIER, C. C., GOOSSENS, J., RAHL, P. B., LEVINE, S. S., et al. 2010. Mediator and cohesin connect gene expression and chromatin architecture. *Nature,* 467**,** 430-5.

KALL, L., KROGH, A. & SONNHAMMER, E. L. 2007. Advantages of combined transmembrane topology and signal peptide prediction--the Phobius web server. *Nucleic Acids Res,* 35**,** W429-32.

KAMOHARA, H., MATSUYAMA, W., SHIMOZATO, O., ABE, K., GALLIGAN, C., HASHIMOTO, S., MATSUSHIMA, K. & YOSHIMURA, T. 2004. Regulation of tumour necrosis factor-related apoptosis-inducing ligand (TRAIL) and TRAIL receptor expression in human neutrophils. *Immunology,* 111**,** 186-94.

KAPLAN, M. J. 2013. Role of neutrophils in systemic autoimmune diseases. *Arthritis Res Ther,* 15**,** 219.

KASOWSKI, M., GRUBERT, F., HEFFELFINGER, C., HARIHARAN, M., ASABERE, A., WASZAK, S. M., HABEGGER, L., ROZOWSKY, J., SHI, M., URBAN, A. E., et al. 2010. Variation in transcription factor binding among humans. *Science,* 328**,** 232-5.

KASOWSKI, M., KYRIAZOPOULOU-PANAGIOTOPOULOU, S., GRUBERT, F., ZAUGG, J. B., KUNDAJE, A., LIU, Y., BOYLE, A. P., ZHANG, Q. C., ZAKHARIA, F., SPACEK, D. V., et al. 2013. Extensive variation in chromatin states across humans. *Science,* 342**,** 750-2.

KATHIRESAN, S., WILLER, C. J., PELOSO, G. M., DEMISSIE, S., MUSUNURU, K., SCHADT, E. E., KAPLAN, L., BENNETT, D., LI, Y., TANAKA, T., et al. 2009. Common variants at 30 loci contribute to polygenic dyslipidemia. *Nat Genet,* 41**,** 56-65.

KAUPPINEN, A., PATERNO, J. J., BLASIAK, J., SALMINEN, A. & KAARNIRANTA, K. 2016. Inflammation and its role in age-related macular degeneration. *Cell Mol Life Sci,* 73**,** 1765-86.

KE, X. 2012. Presence of multiple independent effects in risk loci of common complex human diseases. *Am J Hum Genet,* 91**,** 185-92.

KHERA, A. V. & KATHIRESAN, S. 2017. Genetics of coronary artery disease: discovery, biology and clinical translation. *Nat Rev Genet,* 18**,** 331-344.

KHERADPOUR, P. & KELLIS, M. 2014. Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. *Nucleic Acids Res,* 42**,** 2976-87.

KHWAJA, A., CARVER, J. E. & LINCH, D. C. 1992. Interactions of granulocyte-macrophage colony-stimulating factor (CSF), granulocyte CSF, and tumor necrosis factor alpha in the priming of the neutrophil respiratory burst. *Blood,* 79**,** 745-53.

KILPINEN, H., GONCALVES, A., LEHA, A., AFZAL, V., ALASOO, K., ASHFORD, S., BALA, S., BENSADDEK, D., CASALE, F. P., CULLEY, O. J., et al. 2017. Common genetic variation drives molecular heterogeneity in human iPSCs. *Nature,* 546**,** 370-375.

KILPINEN, H., WASZAK, S. M., GSCHWIND, A. R., RAGHAV, S. K., WITWICKI, R. M., ORIOLI, A., MIGLIAVACCA, E., WIEDERKEHR, M., GUTIERREZ-ARCELUS, M., PANOUSIS, N. I., et al. 2013. Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription. *Science,* 342**,** 744-7.

KIM-HELLMUTH, S., BECHHEIM, M., PUTZ, B., MOHAMMADI, P., NEDELEC, Y., GIANGRECO, N., BECKER, J., KAISER, V., FRICKER, N., BEIER, E., et al. 2017. Genetic regulatory effects modified by immune activation contribute to autoimmune disease associations. *Nat Commun,* 8**,** 266.

KIM-HELLMUTH, S. & LAPPALAINEN, T. 2016. Concerted Genetic Function in Blood Traits. *Cell,* 167**,** 1167-1169.

KINDWALL-KELLER, T. L., DRUHAN, L. J., AI, J., HUNTER, M. G., MASSULLO, P., LOVELAND, M. & AVALOS, B. R. 2008. Role of the proteasome in modulating native G-CSFR expression. *Cytokine,* 43**,** 114-23.

KJOLBY, M., NIELSEN, M. S. & PETERSEN, C. M. 2015. Sortilin, encoded by the cardiovascular risk gene SORT1, and its suggested functions in cardiovascular disease. *Curr Atheroscler Rep,* 17**,** 496.

KLARIN, D., ZHU, Q. M., EMDIN, C. A., CHAFFIN, M., HORNER, S., MCMILLAN, B. J., LEED, A., WEALE, M. E., SPENCER, C. C. A., AGUET, F., et al. 2017. Genetic analysis in UK Biobank links insulin resistance and transendothelial migration pathways to coronary artery disease. *Nat Genet,* 49**,** 1392-1397.

KOBAYASHI, S. D., VOYICH, J. M., BURLAK, C. & DELEO, F. R. 2005. Neutrophils in the innate immune response. *Arch Immunol Ther Exp (Warsz),* 53**,** 505-17.

KOCH, L. 2015. Genomics: Adding another dimension to gene regulation. *Nat Rev Genet,* 16**,** 563.

KOSCIELNY, G., AN, P., CARVALHO-SILVA, D., CHAM, J. A., FUMIS, L., GASPARYAN, R., HASAN, S., KARAMANIS, N., MAGUIRE, M., PAPA, E., et al. 2017. Open Targets: a platform for therapeutic target identification and validation. *Nucleic Acids Res,* 45**,** D985-D994.

KRIJGER, P. H. & DE LAAT, W. 2016. Regulation of disease-associated gene expression in the 3D genome. *Nat Rev Mol Cell Biol,* 17**,** 771-782.

KUMASAKA, N., KNIGHTS, A. J. & GAFFNEY, D. J. 2016. Fine-mapping cellular QTLs with RASQUAL and ATAC-seq. *Nat Genet,* 48**,** 206-13.

LAMBERT, J. C., IBRAHIM-VERBAAS, C. A., HAROLD, D., NAJ, A. C., SIMS, R., BELLENGUEZ, C., DESTAFANO, A. L., BIS, J. C., BEECHAM, G. W., GRENIER-BOLEY, B., et al. 2013. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat Genet,* 45**,** 1452-8.

LAPPALAINEN, T., SAMMETH, M., FRIEDLANDER, M. R., T HOEN, P. A., MONLONG, J., RIVAS, M. A., GONZALEZ-PORTA, M., KURBATOVA, N., GRIEBEL, T., FERREIRA, P. G., et al. 2013. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature,* 501**,** 506-11.

LARA-ASTIASO, D., WEINER, A., LORENZO-VIVAS, E., ZARETSKY, I., JAITIN, D. A., DAVID, E., KEREN-SHAUL, H., MILDNER, A., WINTER, D., JUNG, S., et al. 2014. Immunogenetics. Chromatin state dynamics during blood formation. *Science,* 345**,** 943-9.

LAVELLE, C. 2014. Pack, unpack, bend, twist, pull, push: the physical side of gene expression. *Curr Opin Genet Dev,* 25**,** 74-84.

LAW, V., KNOX, C., DJOUMBOU, Y., JEWISON, T., GUO, A. C., LIU, Y., MACIEJEWSKI, A., ARNDT, D., WILSON, M., NEVEU, V., et al. 2014. DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Res,* 42**,** D1091-7.

LAYTON, J. E., HALL, N. E., CONNELL, F., VENHORST, J. & TREUTLEIN, H. R. 2001. Identification of ligand-binding site III on the immunoglobulin-like domain of the granulocyte colony-stimulating factor receptor. *J Biol Chem,* 276**,** 36779-87.

LEADING EDGE VOICES 2017. Human Genetics: Questions, Challenges, and the Future. *Cell,* 171**,** 259-260.

LECHNER, J., CHEN, M., HOGG, R. E., TOTH, L., SILVESTRI, G., CHAKRAVARTHY, U. & XU, H. 2015. Alterations in Circulating Immune Cells in Neovascular Age-Related Macular Degeneration. *Sci Rep,* 5**,** 16754.

LECHNER, J., CHEN, M., HOGG, R. E., TOTH, L., SILVESTRI, G., CHAKRAVARTHY, U. & XU, H. 2017. Peripheral blood mononuclear cells from neovascular age-related macular degeneration patients produce higher levels of chemokines CCL2 (MCP-1) and CXCL8 (IL-8). *J Neuroinflammation,* 14**,** 42.

LEE, A. J. & KIM, S. G. 2013. Mean cell volumes of neutrophils and monocytes are promising markers of sepsis in elderly patients. *Blood Res,* 48**,** 193-7.

LEE, J. T. 2012. Epigenetic regulation by long noncoding RNAs. *Science,* 338**,** 1435-9.

LEE, J. W., LEE, Y. K., YUK, D. Y., CHOI, D. Y., BAN, S. B., OH, K. W. & HONG, J. T. 2008. Neuro-inflammation induced by lipopolysaccharide causes cognitive impairment through enhancement of beta-amyloid generation. *J Neuroinflammation,* 5**,** 37.

LEE, K. H., CHANG, M. Y., AHN, J. I., YU, D. H., JUNG, S. S., CHOI, J. H., NOH, Y. H., LEE, Y. S. & AHN, M. J. 2002. Differential gene expression in retinoic acid-induced differentiation of acute promyelocytic leukemia cells, NB4 and HL-60 cells. *Biochem Biophys Res Commun,* 296**,** 1125-33.

LEE, M. N., YE, C., VILLANI, A. C., RAJ, T., LI, W., EISENHAURE, T. M., IMBOYWA, S. H., CHIPENDO, P. I., RAN, F. A., SLOWIKOWSKI, K., et al. 2014. Common genetic variants modulate pathogen-sensing responses in human dendritic cells. *Science,* 343**,** 1246980.

LEEK, J. T., SCHARPF, R. B., BRAVO, H. C., SIMCHA, D., LANGMEAD, B., JOHNSON, W. E., GEMAN, D., BAGGERLY, K. & IRIZARRY, R. A. 2010. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat Rev Genet,* 11**,** 733-9.

LI, M., JIA, C., KAZMIERKIEWICZ, K. L., BOWMAN, A. S., TIAN, L., LIU, Y., GUPTA, N. A., GUDISEVA, H. V., YEE, S. S., KIM, M., et al. 2014. Comprehensive analysis of gene expression in human retina and supporting tissues. *Hum Mol Genet,* 23**,** 4001-14.

LI, Y., OOSTING, M., DEELEN, P., RICANO-PONCE, I., SMEEKENS, S., JAEGER, M., MATZARAKI, V., SWERTZ, M. A., XAVIER, R. J., FRANKE, L., et al. 2016a. Inter-individual variability and genetic influences on cytokine responses to bacteria and fungi. *Nat Med,* 22**,** 952-60.

LI, Y., OOSTING, M., SMEEKENS, S. P., JAEGER, M., AGUIRRE-GAMBOA, R., LE, K. T., DEELEN, P., RICANO-PONCE, I., SCHOFFELEN, T., JANSEN, A. F., et al. 2016b. A Functional Genomics Approach to Understand Variation in Cytokine Production in Humans. *Cell,* 167**,** 1099-1110 e14.

LI, Y. I., VAN DE GEIJN, B., RAJ, A., KNOWLES, D. A., PETTI, A. A., GOLAN, D., GILAD, Y. & PRITCHARD, J. K. 2016c. RNA splicing is a primary link between genetic variation and disease. *Science,* 352**,** 600-4.

LIDEN, M., ROMERT, A., TRYGGVASON, K., PERSSON, B. & ERIKSSON, U. 2001. Biochemical defects in 11-cis-retinol dehydrogenase mutants associated with fundus albipunctatus. *J Biol Chem,* 276**,** 49251-7.

LIEBERMAN-AIDEN, E., VAN BERKUM, N. L., WILLIAMS, L., IMAKAEV, M., RAGOCZY, T., TELLING, A., AMIT, I., LAJOIE, B. R., SABO, P. J., DORSCHNER, M. O., et al. 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science,* 326**,** 289-93.

LIESCHKE, G. J., GRAIL, D., HODGSON, G., METCALF, D., STANLEY, E., CHEERS, C., FOWLER, K. J., BASU, S., ZHAN, Y. F. & DUNN, A. R. 1994. Mice lacking granulocyte colony-stimulating factor have chronic neutropenia, granulocyte and macrophage progenitor cell deficiency, and impaired neutrophil mobilization. *Blood,* 84**,** 1737-46.

LIGUORI, M., BURACCHI, C., PASQUALINI, F., BERGOMAS, F., PESCE, S., SIRONI, M., GRIZZI, F., MANTOVANI, A., BELGIOVINE, C. & ALLAVENA, P. 2016. Functional TRAIL receptors in monocytes and tumor-associated macrophages: A possible targeting pathway in the tumor microenvironment. *Oncotarget,* 7**,** 41662-41676.

LIONGUE, C. & WARD, A. C. 2014. Granulocyte colony-stimulating factor receptor mutations in myeloid malignancy. *Front Oncol,* 4**,** 93.

LIU, F., WU, H. Y., WESSELSCHMIDT, R., KORNAGA, T. & LINK, D. C. 1996. Impaired production and increased apoptosis of neutrophils in granulocyte colony-stimulating factor receptor-deficient mice. *Immunity,* 5**,** 491-501.

LOPEZ, P. F., GROSSNIKLAUS, H. E., LAMBERT, H. M., AABERG, T. M., CAPONE, A., JR., STERNBERG, P., JR. & L'HERNAULT, N. 1991. Pathologic features of surgically excised subretinal neovascular membranes in age-related macular degeneration. *Am J Ophthalmol,* 112**,** 647-56.

LORD, B. I., BRONCHUD, M. H., OWENS, S., CHANG, J., HOWELL, A., SOUZA, L. & DEXTER, T. M. 1989. The kinetics of human granulopoiesis following treatment with granulocyte colony-stimulating factor in vivo. *Proc Natl Acad Sci U S A,* 86**,** 9499-503.

LORD, B. I., MOLINEUX, G., POJDA, Z., SOUZA, L. M., MERMOD, J. J. & DEXTER, T. M. 1991. Myeloid cell kinetics in mice treated with recombinant interleukin-3, granulocyte colony-stimulating factor (CSF), or granulocyte-macrophage CSF in vivo. *Blood,* 77**,** 2154-9.

LOWELL, C. A., FUMAGALLI, L. & BERTON, G. 1996. Deficiency of Src family kinases p59/61hck and p58c-fgr results in defective adhesion-dependent neutrophil functions. *J Cell Biol,* 133**,** 895-910.

LUGER, K., MADER, A. W., RICHMOND, R. K., SARGENT, D. F. & RICHMOND, T. J. 1997. Crystal structure of the nucleosome core particle at 2.8 A resolution. *Nature,* 389**,** 251-60.

LUO, C., ZHAO, J., MADDEN, A., CHEN, M. & XU, H. 2013. Complement expression in retinal pigment epithelial cells is modulated by activated macrophages. *Exp Eye Res,* 112**,** 93-101.

LUO, X., YANG, W., YE, D. Q., CUI, H., ZHANG, Y., HIRANKARN, N., QIAN, X., TANG, Y., LAU, Y. L., DE VRIES, N., et al. 2011. A functional variant in microRNA-146a promoter modulates its expression and confers disease risk for systemic lupus erythematosus. *PLoS Genet,* 7**,** e1002128.

LYNCH, K. W. & WEISS, A. 2001. A CD45 polymorphism associated with multiple sclerosis disrupts an exonic splicing silencer. *J Biol Chem,* 276**,** 24341-7.

MACARTHUR, J., BOWLER, E., CEREZO, M., GIL, L., HALL, P., HASTINGS, E., JUNKINS, H., MCMAHON, A., MILANO, A., MORALES, J., et al. 2017. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res,* 45**,** D896-D901.

MAHMOOD, S. S., LEVY, D., VASAN, R. S. & WANG, T. J. 2014. The Framingham Heart Study and the epidemiology of cardiovascular disease: a historical perspective. *Lancet,* 383**,** 999-1008.

MANOLIO, T. A., COLLINS, F. S., COX, N. J., GOLDSTEIN, D. B., HINDORFF, L. A., HUNTER, D. J., MCCARTHY, M. I., RAMOS, E. M., CARDON, L. R., CHAKRAVARTI, A., et al. 2009. Finding the missing heritability of complex diseases. *Nature,* 461**,** 747-53.

MARANVILLE, J. C., LUCA, F., RICHARDS, A. L., WEN, X., WITONSKY, D. B., BAXTER, S., STEPHENS, M. & DI RIENZO, A. 2011. Interactions between glucocorticoid treatment and cis-regulatory polymorphisms contribute to cellular response phenotypes. *PLoS Genet,* 7**,** e1002162.

MARIONI, J. C., MASON, C. E., MANE, S. M., STEPHENS, M. & GILAD, Y. 2008. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res,* 18**,** 1509-17.

MATHIS, T., HOUSSET, M., EANDI, C., BEGUIER, F., TOUHAMI, S., REICHMAN, S., AUGUSTIN, S., GONDOUIN, P., SAHEL, J. A., KODJIKIAN, L., et al. 2017. Activated monocytes resist elimination by retinal pigment epithelium and downregulate their OTX2 expression via TNF-alpha. *Aging Cell,* 16**,** 173-182.

MAURANO, M. T., HUMBERT, R., RYNES, E., THURMAN, R. E., HAUGEN, E., WANG, H., REYNOLDS, A. P., SANDSTROM, R., QU, H., BRODY, J., et al. 2012. Systematic localization of common disease-associated variation in regulatory DNA. *Science,* 337**,** 1190-5.

MAY, A. E., KANSE, S. M., LUND, L. R., GISLER, R. H., IMHOF, B. A. & PREISSNER, K. T. 1998. Urokinase receptor (CD87) regulates leukocyte recruitment via beta 2 integrins in vivo. *J Exp Med,* 188**,** 1029-37.

MAYADAS, T. N., CULLERE, X. & LOWELL, C. A. 2014. The multifaceted functions of neutrophils. *Annu Rev Pathol,* 9**,** 181-218.

MCCARTHY, M. I., ABECASIS, G. R., CARDON, L. R., GOLDSTEIN, D. B., LITTLE, J., IOANNIDIS, J. P. & HIRSCHHORN, J. N. 2008. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet,* 9**,** 356-69.

MCDANIELL, R., LEE, B. K., SONG, L., LIU, Z., BOYLE, A. P., ERDOS, M. R., SCOTT, L. J., MORKEN, M. A., KUCERA, K. S., BATTENHOUSE, A., et al. 2010. Heritable individual-specific and allele-specific chromatin signatures in humans. *Science,* 328**,** 235-9.

MCEVER, R. P. & CUMMINGS, R. D. 1997. Role of PSGL-1 binding to selectins in leukocyte recruitment. *J Clin Invest,* 100**,** S97-103.

MCGOVERN, A., SCHOENFELDER, S., MARTIN, P., MASSEY, J., DUFFUS, K., PLANT, D., YARWOOD, A., PRATT, A. G., ANDERSON, A. E., ISAACS, J. D., et al. 2016.

Capture Hi-C identifies a novel causal gene, IL20RA, in the pan-autoimmune genetic susceptibility region 6q23. *Genome Biol,* 17**,** 212.

MCLAREN, W., GIL, L., HUNT, S. E., RIAT, H. S., RITCHIE, G. R., THORMANN, A., FLICEK, P. & CUNNINGHAM, F. 2016. The Ensembl Variant Effect Predictor. *Genome Biol,* 17**,** 122.

MCVICKER, G., VAN DE GEIJN, B., DEGNER, J. F., CAIN, C. E., BANOVICH, N. E., RAJ, A., LEWELLEN, N., MYRTHIL, M., GILAD, Y. & PRITCHARD, J. K. 2013. Identification of genetic variants that affect histone modifications in human cells. *Science,* 342**,** 747-9.

MEEUWSEN, J. A. L., WESSELING, M., HOEFER, I. E. & DE JAGER, S. C. A. 2017. Prognostic Value of Circulating Inflammatory Cells in Patients with Stable and Acute Coronary Artery Disease. *Front Cardiovasc Med,* 4**,** 44.

MEHTA, H. M., MALANDRA, M. & COREY, S. J. 2015. G-CSF and GM-CSF in Neutropenia. *J Immunol,* 195**,** 1341-9.

MI, H., HUANG, X., MURUGANUJAN, A., TANG, H., MILLS, C., KANG, D. & THOMAS, P. D. 2017. PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. *Nucleic Acids Res,* 45**,** D183-D189.

MIFSUD, B., TAVARES-CADETE, F., YOUNG, A. N., SUGAR, R., SCHOENFELDER, S., FERREIRA, L., WINGETT, S. W., ANDREWS, S., GREY, W., EWELS, P. A., et al. 2015. Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nat Genet,* 47**,** 598-606.

MOHR, W. & WESSINGHAGE, D. 1978. The relationship between polymorphonuclear granulocytes and cartilage destruction in rheumatoid arthritis. *Z Rheumatol,* 37**,** 81-6.

MONLONG, J., CALVO, M., FERREIRA, P. G. & GUIGO, R. 2014. Identification of genetic variants associated with alternative splicing using sQTLseekeR. *Nat Commun,* 5**,** 4698.

MONTGOMERY, S. B. & DERMITZAKIS, E. T. 2011. From expression QTLs to personalized transcriptomics. *Nat Rev Genet,* 12**,** 277-82.

MOORE, C., SAMBROOK, J., WALKER, M., TOLKIEN, Z., KAPTOGE, S., ALLEN, D., MEHENNY, S., MANT, J., DI ANGELANTONIO, E., THOMPSON, S. G., et al. 2014. The INTERVAL trial to determine whether intervals between blood donations can be safely and acceptably decreased to optimise blood supply: study protocol for a randomised controlled trial. *Trials,* 15**,** 363.

MORA, J. R., IWATA, M. & VON ANDRIAN, U. H. 2008. Vitamin effects on the immune system: vitamins A and D take centre stage. *Nat Rev Immunol,* 8**,** 685-98.

MORRIS, A. P., VOIGHT, B. F., TESLOVICH, T. M., FERREIRA, T., SEGRE, A. V., STEINTHORSDOTTIR, V., STRAWBRIDGE, R. J., KHAN, H., GRALLERT, H., MAHAJAN, A., et al. 2012. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nat Genet,* 44**,** 981-90.

MORTENSEN, M. B., KJOLBY, M., GUNNERSEN, S., LARSEN, J. V., PALMFELDT, J., FALK, E., NYKJAER, A. & BENTZON, J. F. 2014. Targeting sortilin in immune cells reduces proinflammatory cytokines and atherosclerosis. *J Clin Invest,* 124**,** 5317-22.

MUELLER, H., STADTMANN, A., VAN AKEN, H., HIRSCH, E., WANG, D., LEY, K. & ZARBOCK, A. 2010. Tyrosine kinase Btk regulates E-selectin-mediated integrin activation and neutrophil recruitment by controlling phospholipase C (PLC) gamma2 and PI3Kgamma pathways. *Blood,* 115**,** 3118-27.

MULLEN, A. C., ORLANDO, D. A., NEWMAN, J. J., LOVEN, J., KUMAR, R. M., BILODEAU, S., REDDY, J., GUENTHER, M. G., DEKOTER, R. P. & YOUNG, R. A. 2011. Master transcription factors determine cell-type-specific responses to TGF-beta signaling. *Cell,* 147**,** 565-76.

MULLER, N., WEIDINGER, E., LEITNER, B. & SCHWARZ, M. J. 2015. The role of inflammation in schizophrenia. *Front Neurosci,* 9**,** 372.

MUMBACH, M. R., SATPATHY, A. T., BOYLE, E. A., DAI, C., GOWEN, B. G., CHO, S. W., NGUYEN, M. L., RUBIN, A. J., GRANJA, J. M., KAZANE, K. R., et al. 2017. Enhancer connectome in primary human cells identifies target genes of disease-associated DNA elements. *Nat Genet,* 49**,** 1602-1612.

MUSUNURU, K., STRONG, A., FRANK-KAMENETSKY, M., LEE, N. E., AHFELDT, T., SACHS, K. V., LI, X., LI, H., KUPERWASSER, N., RUDA, V. M., et al. 2010. From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature,* 466**,** 714-9.

NAKANO, A., HARADA, T., MORIKAWA, S. & KATO, Y. 1990. Expression of leukocyte common antigen (CD45) on various human leukemia/lymphoma cell lines. *Acta Pathol Jpn,* 40**,** 107-15.

NARANBHAI, V., FAIRFAX, B. P., MAKINO, S., HUMBURG, P., WONG, D., NG, E., HILL, A. V. & KNIGHT, J. C. 2015. Genomic modulators of gene expression in human neutrophils. *Nat Commun,* 6**,** 7545.

NELSON, J., MCFERRAN, N. V., PIVATO, G., CHAMBERS, E., DOHERTY, C., STEELE, D. & TIMSON, D. J. 2008. The 67 kDa laminin receptor: structure, function and role in disease. *Biosci Rep,* 28**,** 33-48.

NELSON, M. R., TIPNEY, H., PAINTER, J. L., SHEN, J., NICOLETTI, P., SHEN, Y., FLORATOS, A., SHAM, P. C., LI, M. J., WANG, J., et al. 2015. The support of human genetic evidence for approved drug indications. *Nat Genet,* 47**,** 856-60.

NEPOM, G. T. 1998. Major histocompatibility complex-directed susceptibility to rheumatoid arthritis. *Adv Immunol,* 68**,** 315-32.

NETEA, M. G., JOOSTEN, L. A., LI, Y., KUMAR, V., OOSTING, M., SMEEKENS, S., JAEGER, M., TER HORST, R., SCHIRMER, M., VLAMAKIS, H., et al. 2016. Understanding human immune function using the resources from the Human Functional Genomics Project. *Nat Med,* 22**,** 831-3.

NICA, A. C. & DERMITZAKIS, E. T. 2013. Expression quantitative trait loci: present and future. *Philos Trans R Soc Lond B Biol Sci,* 368**,** 20120362.

NICOLA, N. A. & METCALF, D. 1985. Binding of 125I-labeled granulocyte colony-stimulating factor to normal murine hemopoietic cells. *J Cell Physiol,* 124**,** 313-21.

NIETO MORENO, N., GIONO, L. E., CAMBINDO BOTTO, A. E., MUNOZ, M. J. & KORNBLIHTT, A. R. 2015. Chromatin, DNA structure and alternative splicing. *FEBS Lett,* 589**,** 3370-8.

NIKPAY, M., GOEL, A., WON, H. H., HALL, L. M., WILLENBORG, C., KANONI, S., SALEHEEN, D., KYRIAKOU, T., NELSON, C. P., HOPEWELL, J. C., et al. 2015. A comprehensive 1,000 Genomes-based genome-wide association meta-analysis of coronary artery disease. *Nat Genet,* 47**,** 1121-1130.

NILSEN, T. W. & GRAVELEY, B. R. 2010. Expansion of the eukaryotic proteome by alternative splicing. *Nature,* 463**,** 457-63.

NISTICO, L., BUZZETTI, R., PRITCHARD, L. E., VAN DER AUWERA, B., GIOVANNINI, C., BOSI, E., LARRAD, M. T., RIOS, M. S., CHOW, C. C., COCKRAM, C. S., et al. 1996. The CTLA-4 gene region of chromosome 2q33 is linked to, and associated with, type 1 diabetes. Belgian Diabetes Registry. *Hum Mol Genet,* 5**,** 1075-80.

NUSRAT, A. R. & CHAPMAN, H. A., JR. 1991. An autocrine role for urokinase in phorbol ester-mediated differentiation of myeloid cell lines. *J Clin Invest,* 87**,** 1091-7.

NUSSENBLATT, R. B., BYRNES, G., SEN, H. N., YEH, S., FAIA, L., MEYERLE, C., WROBLEWSKI, K., LI, Z., LIU, B., CHEW, E., et al. 2010. A randomized pilot study of systemic immunosuppression in the treatment of age-related macular degeneration with choroidal neovascularization. *Retina,* 30**,** 1579-87.

NUSSENBLATT, R. B. & FERRIS, F., 3RD 2007. Age-related macular degeneration and the immune response: implications for therapy. *Am J Ophthalmol,* 144**,** 618-26.

ODHAMS, C. A., CORTINI, A., CHEN, L., ROBERTS, A. L., VINUELA, A., BUIL, A., SMALL, K. S., DERMITZAKIS, E. T., MORRIS, D. L., VYSE, T. J., et al. 2017. Mapping eQTLs with RNA-seq reveals novel susceptibility genes, non-coding RNAs and alternative-splicing events in systemic lupus erythematosus. *Hum Mol Genet,* 26**,** 1003-1017.

ODOM, D. T., DOWELL, R. D., JACOBSEN, E. S., NEKLUDOVA, L., ROLFE, P. A., DANFORD, T. W., GIFFORD, D. K., FRAENKEL, E., BELL, G. I. & YOUNG, R. A. 2006. Core transcriptional regulatory circuitry in human hepatocytes. *Mol Syst Biol,* 2**,** 2006 0017.

OHRADANOVA-REPIC, A., MACHACEK, C., FISCHER, M. B. & STOCKINGER, H. 2016. Differentiation of human monocytes and derived subsets of macrophages and dendritic cells by the HLDA10 monoclonal antibody panel. *Clin Transl Immunology,* 5**,** e55.

OKADA, Y., WU, D., TRYNKA, G., RAJ, T., TERAO, C., IKARI, K., KOCHI, Y., OHMURA, K., SUZUKI, A., YOSHIDA, S., et al. 2014. Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature,* 506**,** 376-81.

ONG, C. T. & CORCES, V. G. 2011. Enhancer function: new insights into the regulation of tissue-specific gene expression. *Nat Rev Genet,* 12**,** 283-93.

OPEN TARGETS. 2017. *Open Targets Platform* [Online]. Available: https://www.targetvalidation.org/ [Accessed October 2017].

ORKIN, S. H. & ZON, L. I. 2008. Hematopoiesis: an evolving paradigm for stem cell biology. *Cell,* 132**,** 631-44.

ORRU, V., STERI, M., SOLE, G., SIDORE, C., VIRDIS, F., DEI, M., LAI, S., ZOLEDZIEWSKA, M., BUSONERO, F., MULAS, A., et al. 2013. Genetic variants regulating immune cell levels in health and disease. *Cell,* 155**,** 242-56.

PAI, A. A., PRITCHARD, J. K. & GILAD, Y. 2015. The genetic and mechanistic basis for variation in gene regulation. *PLoS Genet,* 11**,** e1004857.

PANOPOULOS, A. D. & WATOWICH, S. S. 2008. Granulocyte colony-stimulating factor: molecular mechanisms of action during steady state and 'emergency' hematopoiesis. *Cytokine,* 42**,** 277-88.

PAPAYANNOPOULOS, V. 2017. Neutrophil extracellular traps in immunity and disease. *Nat Rev Immunol*.

PARMEGGIANI, F., ROMANO, M. R., COSTAGLIOLA, C., SEMERARO, F., INCORVAIA, C., D'ANGELO, S., PERRI, P., DE PALMA, P., DE NADAI, K. & SEBASTIANI, A. 2012. Mechanism of inflammation in age-related macular degeneration. *Mediators Inflamm,* 2012**,** 546786.

PENNACCHIO, L. A., BICKMORE, W., DEAN, A., NOBREGA, M. A. & BEJERANO, G. 2013. Enhancers: five essential questions. *Nat Rev Genet,* 14**,** 288-95.

PENNINGTON, K. L. & DEANGELIS, M. M. 2016. Epidemiology of age-related macular degeneration (AMD): associations with cardiovascular disease phenotypes and lipid factors. *Eye Vis (Lond),* 3**,** 34.

PERSAD, P. J., HEID, I. M., WEEKS, D. E., BAIRD, P. N., DE JONG, E. K., HAINES, J. L., PERICAK-VANCE, M. A., SCOTT, W. K. & INTERNATIONAL AGE-RELATED MACULAR DEGENERATION GENOMICS, C. 2017. Joint Analysis of Nuclear and Mitochondrial Variants in Age-Related Macular Degeneration Identifies Novel Loci TRPM1 and ABHD2/RLBP1. *Invest Ophthalmol Vis Sci,* 58**,** 4027-4038.

PHAM, T. H., MINDERJAHN, J., SCHMIDL, C., HOFFMEISTER, H., SCHMIDHOFER, S., CHEN, W., LANGST, G., BENNER, C. & REHLI, M. 2013. Mechanisms of in vivo binding site selection of the hematopoietic master transcription factor PU.1. *Nucleic Acids Res,* 41**,** 6391-402.

PHILLIPS-CREMINS, J. E., SAURIA, M. E., SANYAL, A., GERASIMOVA, T. I., LAJOIE, B. R., BELL, J. S., ONG, C. T., HOOKWAY, T. A., GUO, C., SUN, Y., et al. 2013. Architectural protein subclasses shape 3D organization of genomes during lineage commitment. *Cell,* 153**,** 1281-95.

PICKRELL, J. K. 2014. Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *Am J Hum Genet,* 94**,** 559-73.

PICKRELL, J. K., BERISA, T., LIU, J. Z., SEGUREL, L., TUNG, J. Y. & HINDS, D. A. 2016. Detection and interpretation of shared genetic influences on 42 human traits. *Nat Genet,* 48**,** 709-17.

PICKRELL, J. K., MARIONI, J. C., PAI, A. A., DEGNER, J. F., ENGELHARDT, B. E., NKADORI, E., VEYRIERAS, J. B., STEPHENS, M., GILAD, Y. & PRITCHARD, J. K. 2010. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature,* 464**,** 768-72.

PILIA, G., CHEN, W. M., SCUTERI, A., ORRU, M., ALBAI, G., DEI, M., LAI, S., USALA, G., LAI, M., LOI, P., et al. 2006. Heritability of cardiovascular and personality traits in 6,148 Sardinians. *PLoS Genet,* 2**,** e132.

PIMENOVA, A. A., RAJ, T. & GOATE, A. M. 2017. Untangling Genetic Risk for Alzheimer's Disease. *Biol Psychiatry*.

PIPER, M. G., MASSULLO, P. R., LOVELAND, M., DRUHAN, L. J., KINDWALL-KELLER, T. L., AI, J., COPELAN, A. & AVALOS, B. R. 2010. Neutrophil elastase downmodulates native G-CSFR expression and granulocyte-macrophage colony formation. *J Inflamm (Lond),* 7**,** 5.

PLENGE, R. M., PADYUKOV, L., REMMERS, E. F., PURCELL, S., LEE, A. T., KARLSON, E. W., WOLFE, F., KASTNER, D. L., ALFREDSSON, L., ALTSHULER, D., et al. 2005. Replication of putative candidate-gene associations with rheumatoid arthritis in >4,000 samples from North America and Sweden: association of susceptibility with PTPN22, CTLA4, and PADI4. *Am J Hum Genet,* 77**,** 1044-60.

PLO, I., ZHANG, Y., LE COUEDIC, J. P., NAKATAKE, M., BOULET, J. M., ITAYA, M., SMITH, S. O., DEBILI, N., CONSTANTINESCU, S. N., VAINCHENKER, W., et al. 2009. An activating mutation in the CSF3R gene induces a hereditary chronic neutrophilia. *J Exp Med,* 206**,** 1701-7.

POITOU, C., DALMAS, E., RENOVATO, M., BENHAMO, V., HAJDUCH, F., ABDENNOUR, M., KAHN, J. F., VEYRIE, N., RIZKALLA, S., FRIDMAN, W. H., et al. 2011. CD14dimCD16+ and CD14+CD16+ monocytes in obesity and during weight loss: relationships with fat mass and subclinical atherosclerosis. *Arterioscler Thromb Vasc Biol,* 31**,** 2322-30.

PRICE, A. L., HELGASON, A., THORLEIFSSON, G., MCCARROLL, S. A., KONG, A. & STEFANSSON, K. 2011. Single-tissue and cross-tissue heritability of gene expression via identity-by-descent in related or unrelated individuals. *PLoS Genet,* 7**,** e1001317.

PROITSI, P., LEE, S. H., LUNNON, K., KEOHANE, A., POWELL, J., TROAKES, C., AL-SARRAJ, S., FURNEY, S., SOININEN, H., KLOSZEWSKA, I., et al. 2014. Alzheimer's disease susceptibility variants in the MS4A6A gene are associated with altered levels of MS4A6A expression in blood. *Neurobiol Aging,* 35**,** 279-90.

PRUIM, R. J., WELCH, R. P., SANNA, S., TESLOVICH, T. M., CHINES, P. S., GLIEDT, T. P., BOEHNKE, M., ABECASIS, G. R. & WILLER, C. J. 2010. LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics,* 26**,** 2336-7.

RAHL, P. B., LIN, C. Y., SEILA, A. C., FLYNN, R. A., MCCUINE, S., BURGE, C. B., SHARP, P. A. & YOUNG, R. A. 2010. c-Myc regulates transcriptional pause release. *Cell,* 141**,** 432-45.

RAJ, T., ROTHAMEL, K., MOSTAFAVI, S., YE, C., LEE, M. N., REPLOGLE, J. M., FENG, T., LEE, M., ASINOVSKI, N., FROHLICH, I., et al. 2014. Polarization of the effects of autoimmune and neurodegenerative risk alleles in leukocytes. *Science,* 344**,** 519-23.

RAO, N. K., SHI, G. P. & CHAPMAN, H. A. 1995. Urokinase receptor is a multifunctional protein: influence of receptor occupancy on macrophage gene expression. *J Clin Invest,* 96**,** 465-74.

REBHAN, M., CHALIFA-CASPI, V., PRILUSKY, J. & LANCET, D. 1998. GeneCards: a novel functional genomics compendium with automated data mining and query reformulation support. *Bioinformatics,* 14**,** 656-64.

RECORD, J., MALINOVA, D., ZENNER, H. L., PLAGNOL, V., NOWAK, K., SYED, F., BOUMA, G., CURTIS, J., GILMOUR, K., CALE, C., et al. 2015. Immunodeficiency and severe susceptibility to bacterial infection associated with a loss-of-function homozygous mutation of MKL1. *Blood,* 126**,** 1527-35.

REIMAND, J., ARAK, T., ADLER, P., KOLBERG, L., REISBERG, S., PETERSON, H. & VILO, J. 2016. g:Profiler-a web server for functional interpretation of gene lists (2016 update). *Nucleic Acids Res,* 44**,** W83-9.

RESCHEN, M. E., GAULTON, K. J., LIN, D., SOILLEUX, E. J., MORRIS, A. J., SMYTH, S. S. & O'CALLAGHAN, C. A. 2015. 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**,** e1005061.

RICH, S. S., WEITKAMP, L. R. & BARBOSA, J. 1984. Genetic heterogeneity of insulin-dependent (type I) diabetes mellitus: evidence from a study of extended haplotypes. *Am J Hum Genet,* 36**,** 1015-23.

RIDKER, P. M., EVERETT, B. M., THUREN, T., MACFADYEN, J. G., CHANG, W. H., BALLANTYNE, C., FONSECA, F., NICOLAU, J., KOENIG, W., ANKER, S. D., et al. 2017. Antiinflammatory Therapy with Canakinumab for Atherosclerotic Disease. *N Engl J Med,* 377**,** 1119-1131.

RIJNEVELD, A. W., LEVI, M., FLORQUIN, S., SPEELMAN, P., CARMELIET, P. & VAN DER POLL, T. 2002. Urokinase receptor is necessary for adequate host defense against pneumococcal pneumonia. *J Immunol,* 168**,** 3507-11.

RISCH, N. & MERIKANGAS, K. 1996. The future of genetic studies of complex human diseases. *Science,* 273**,** 1516-7.

ROADMAP EPIGENOMICS CONSORTIUM, KUNDAJE, A., MEULEMAN, W., ERNST, J., BILENKY, M., YEN, A., HERAVI-MOUSSAVI, A., KHERADPOUR, P., ZHANG, Z., WANG, J., et al. 2015. Integrative analysis of 111 reference human epigenomes. *Nature,* 518**,** 317-30.

ROEDER, R. G. 1996. The role of general initiation factors in transcription by RNA polymerase II. *Trends Biochem Sci,* 21**,** 327-35.

ROEDERER, M., QUAYE, L., MANGINO, M., BEDDALL, M. H., MAHNKE, Y., CHATTOPADHYAY, P., TOSI, I., NAPOLITANO, L., TERRANOVA BARBERIO, M., MENNI, C., et al. 2015. The genetic architecture of the human immune system: a bioresource for autoimmunity and disease pathogenesis. *Cell,* 161**,** 387-403.

ROY, A. L. & SINGER, D. S. 2015. Core promoters in transcription: old problem, new insights. *Trends Biochem Sci,* 40**,** 165-71.

ROZENBLATT-ROSEN, O., STUBBINGTON, M. J. T., REGEV, A. & TEICHMANN, S. A. 2017. The Human Cell Atlas: from vision to reality. *Nature,* 550**,** 451-453.

SACHIDANANDAM, R., WEISSMAN, D., SCHMIDT, S. C., KAKOL, J. M., STEIN, L. D., MARTH, G., SHERRY, S., MULLIKIN, J. C., MORTIMORE, B. J., WILLEY, D. L., et al. 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature,* 409**,** 928-33.

SANKAR, P. L. & PARKER, L. S. 2017. The Precision Medicine Initiative's All of Us Research Program: an agenda for research on its ethical, legal, and social issues. *Genet Med,* 19**,** 743-750.

SATO, T., HONGU, T., SAKAMOTO, M., FUNAKOSHI, Y. & KANAHO, Y. 2013. Molecular mechanisms of N-formyl-methionyl-leucyl-phenylalanine-induced superoxide generation and degranulation in mouse neutrophils: phospholipase D is dispensable. *Mol Cell Biol,* 33**,** 136-45.

SAVAS, S., SCHMIDT, S., JARJANAZI, H. & OZCELIK, H. 2006. Functional nsSNPs from carcinogenesis-related genes expressed in breast tissue: potential breast cancer risk alleles and their distribution across human populations. *Hum Genomics,* 2**,** 287-96.

SCHMIDT, D., SCHWALIE, P. C., ROSS-INNES, C. S., HURTADO, A., BROWN, G. D., CARROLL, J. S., FLICEK, P. & ODOM, D. T. 2010. A CTCF-independent role for cohesin in tissue-specific transcription. *Genome Res,* 20**,** 578-88.

SCHMIDT, D., WILSON, M. D., SPYROU, C., BROWN, G. D., HADFIELD, J. & ODOM, D. T. 2009. ChIP-seq: using high-throughput sequencing to discover protein-DNA interactions. *Methods,* 48**,** 240-8.

SCHOFIELD, E. C., CARVER, T., ACHUTHAN, P., FREIRE-PRITCHETT, P., SPIVAKOV, M., TODD, J. A. & BURREN, O. S. 2016. CHiCP: a web-based tool for the integrative and interactive visualization of promoter capture Hi-C datasets. *Bioinformatics,* 32**,** 2511-3.

SCHOR, I. E., RASCOVAN, N., PELISCH, F., ALLO, M. & KORNBLIHTT, A. R. 2009. Neuronal cell depolarization induces intragenic chromatin modifications affecting NCAM alternative splicing. *Proc Natl Acad Sci U S A,* 106**,** 4325-30.

SEDDON, J. M., COTE, J., PAGE, W. F., AGGEN, S. H. & NEALE, M. C. 2005. The US twin study of age-related macular degeneration: relative roles of genetic and environmental influences. *Arch Ophthalmol,* 123**,** 321-7.

SEGAL, B. H., LETO, T. L., GALLIN, J. I., MALECH, H. L. & HOLLAND, S. M. 2000. Genetic, biochemical, and clinical features of chronic granulomatous disease. *Medicine (Baltimore),* 79**,** 170-200.

SELMI, C., LU, Q. & HUMBLE, M. C. 2012. Heritability versus the role of the environment in autoimmunity. *J Autoimmun,* 39**,** 249-52.

SELVATICI, R., FALZARANO, S., MOLLICA, A. & SPISANI, S. 2006. Signal transduction pathways triggered by selective formylpeptide analogues in human neutrophils. *Eur J Pharmacol,* 534**,** 1-11.

SEMERAD, C. L., LIU, F., GREGORY, A. D., STUMPF, K. & LINK, D. C. 2002. G-CSF is an essential regulator of neutrophil trafficking from the bone marrow to the blood. *Immunity,* 17**,** 413-23.

SENGUPTA, K., ARANDA-ESPINOZA, H., SMITH, L., JANMEY, P. & HAMMER, D. 2006. Spreading of neutrophils: from activation to migration. *Biophys J,* 91**,** 4638-48.

SETO, Y., FUKUNAGA, R. & NAGATA, S. 1992. Chromosomal gene organization of the human granulocyte colony-stimulating factor receptor. *J Immunol,* 148**,** 259-66.

SHIINA, T., HOSOMICHI, K., INOKO, H. & KULSKI, J. K. 2009. The HLA genomic loci map: expression, interaction, diversity and disease. *J Hum Genet,* 54**,** 15-39.

SLATKIN, M. 2008. Linkage disequilibrium--understanding the evolutionary past and mapping the medical future. *Nat Rev Genet,* 9**,** 477-85.

SLOAND, E. M., PFANNES, L., SCHEINBERG, P., MORE, K., WU, C. O., HORNE, M. & YOUNG, N. S. 2008. Increased soluble urokinase plasminogen activator receptor (suPAR) is associated with thrombosis and inhibition of plasmin generation in paroxysmal nocturnal hemoglobinuria (PNH) patients. *Exp Hematol,* 36**,** 1616-24.

SMITH, H. W. & MARSHALL, C. J. 2010. Regulation of cell signalling by uPAR. *Nat Rev Mol Cell Biol,* 11**,** 23-36.

SOLLID, L. M., MARKUSSEN, G., EK, J., GJERDE, H., VARTDAL, F. & THORSBY, E. 1989. Evidence for a primary association of celiac disease to a particular HLA-DQ alpha/beta heterodimer. *J Exp Med,* 169**,** 345-50.

SORANZO, N., SPECTOR, T. D., MANGINO, M., KUHNEL, B., RENDON, A., TEUMER, A., WILLENBORG, C., WRIGHT, B., CHEN, L., LI, M., et al. 2009. A genome-wide meta-analysis identifies 22 loci associated with eight hematological parameters in the HaemGen consortium. *Nat Genet,* 41**,** 1182-90.

SPAIN, S. L. & BARRETT, J. C. 2015. Strategies for fine-mapping complex traits. *Hum Mol Genet,* 24**,** R111-9.

STENBERG, A., SEHLIN, J. & OLDENBORG, P. A. 2013. Neutrophil apoptosis is associated with loss of signal regulatory protein alpha (SIRPalpha) from the cell surface. *J Leukoc Biol,* 93**,** 403-12.

STEPHENS, Z. D., LEE, S. Y., FAGHRI, F., CAMPBELL, R. H., ZHAI, C., EFRON, M. J., IYER, R., SCHATZ, M. C., SINHA, S. & ROBINSON, G. E. 2015. Big Data: Astronomical or Genomical? *PLoS Biol,* 13**,** e1002195.

STERI, M., ORRU, V., IDDA, M. L., PITZALIS, M., PALA, M., ZARA, I., SIDORE, C., FAA, V., FLORIS, M., DEIANA, M., et al. 2017. Overexpression of the Cytokine BAFF and Autoimmunity Risk. *N Engl J Med,* 376**,** 1615-1626.

STRANGER, B. E., NICA, A. C., FORREST, M. S., DIMAS, A., BIRD, C. P., BEAZLEY, C., INGLE, C. E., DUNNING, M., FLICEK, P., KOLLER, D., et al. 2007. Population genomics of human gene expression. *Nat Genet,* 39**,** 1217-24.

STUNNENBERG, H. G., INTERNATIONAL HUMAN EPIGENOME, C. & HIRST, M. 2016. The International Human Epigenome Consortium: A Blueprint for Scientific Collaboration and Discovery. *Cell,* 167**,** 1897.

SUN, B. B., MARANVILLE, J. C., PETERS, J. E., STACEY, D., STALEY, J. R., BLACKSHAW, J., BURGESS, S., JIANG, T., PAIGE, E., SURENDRAN, P., et al. 2017. Consequences Of Natural Perturbations In The Human Plasma Proteome. *bioRxiv.*

SUNG, L. & DROR, Y. 2007. Clinical applications of granulocyte-colony stimulating factor. *Front Biosci,* 12**,** 1988-2002.

SYMMONS, O., USLU, V. V., TSUJIMURA, T., RUF, S., NASSARI, S., SCHWARZER, W., ETTWILLER, L. & SPITZ, F. 2014. Functional and topological characteristics of mammalian regulatory domains. *Genome Res,* 24**,** 390-400.

SYSMEX CORPORATION. 2010-2012. *Automated Hematology Analyzer XN series Administrator's Guide* [Online]. portal.sysmex.co.uk. Available: http://portal.sysmex.co.uk/resources/content/XN_ADM_1202_en XN3000.pdf [Accessed November 2017 2017].

TAN, P. L., BOWES RICKMAN, C. & KATSANIS, N. 2016. AMD and the alternative complement pathway: genetics and functional implications. *Hum Genomics,* 10**,** 23.

TANG, Y., LUO, X., CUI, H., NI, X., YUAN, M., GUO, Y., HUANG, X., ZHOU, H., DE VRIES, N., TAK, P. P., et al. 2009. MicroRNA-146A contributes to abnormal activation of the type I interferon pathway in human lupus by targeting the key signaling proteins. *Arthritis Rheum,* 60**,** 1065-75.

TANG, Z., LUO, O. J., LI, X., ZHENG, M., ZHU, J. J., SZALAJ, P., TRZASKOMA, P., MAGALSKA, A., WLODARCZYK, J., RUSZCZYCKI, B., et al. 2015. CTCF-Mediated Human 3D Genome Architecture Reveals Chromatin Topology for Transcription. *Cell,* 163**,** 1611-27.

TEHRANCHI, A. K., MYRTHIL, M., MARTIN, T., HIE, B. L., GOLAN, D. & FRASER, H. B. 2016. Pooled ChIP-Seq Links Variation in Transcription Factor Binding to Complex Disease Risk. *Cell,* 165**,** 730-41.

TER HORST, R., JAEGER, M., SMEEKENS, S. P., OOSTING, M., SWERTZ, M. A., LI, Y., KUMAR, V., DIAVATOPOULOS, D. A., JANSEN, A. F., LEMMERS, H., et al. 2016. Host and Environmental Factors Influencing Individual Human Cytokine Responses. *Cell,* 167**,** 1111-1124 e13.

THERMOFISHER SCIENTIFIC. 2017. *Substrates for Oxidases, Including Amplex Red Kits-Section 10.5* [Online]. www.thermofisher.com: ThermoFisher Scientific. Available: https://www.thermofisher.com/uk/en/home/references/molecular-probes-the-handbook/enzyme-substrates/substrates-for-oxidases-including-amplex-red-kits.html [Accessed 20th November 2017].

TOUW, I. P. 2015. Game of clones: the genomic evolution of severe congenital neutropenia. *Hematology Am Soc Hematol Educ Program,* 2015**,** 1-7.

TROWSDALE, J. & KNIGHT, J. C. 2013. Major histocompatibility complex genomics and human disease. *Annu Rev Genomics Hum Genet,* 14**,** 301-23.

TRYNKA, G. 2017. Enhancers looping to target genes. *Nat Genet,* 49**,** 1564-1565.

TRYNKA, G., SANDOR, C., HAN, B., XU, H., STRANGER, B. E., LIU, X. S. & RAYCHAUDHURI, S. 2013. Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nat Genet,* 45**,** 124-30.

TUFEKCI, K. U., MEUWISSEN, R., GENC, S. & GENC, K. 2012. Inflammation in Parkinson's disease. *Adv Protein Chem Struct Biol,* 88**,** 69-132.

TURNER, S. D. 2014. [Pre-print] qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. *bioRxiv.*

UHLEN, M., FAGERBERG, L., HALLSTROM, B. M., LINDSKOG, C., OKSVOLD, P., MARDINOGLU, A., SIVERTSSON, A., KAMPF, C., SJOSTEDT, E., ASPLUND, A., et al. 2015. Proteomics. Tissue-based map of the human proteome. *Science,* 347**,** 1260419.

UK10K. CONSORTIUM, WALTER, K., MIN, J. L., HUANG, J., CROOKS, L., MEMARI, Y., MCCARTHY, S., PERRY, J. R., XU, C., FUTEMA, M., et al. 2015. The UK10K project identifies rare variants in health and disease. *Nature,* 526**,** 82-90.

URIARTE, S. M., POWELL, D. W., LUERMAN, G. C., MERCHANT, M. L., CUMMINS, T. D., JOG, N. R., WARD, R. A. & MCLEISH, K. R. 2008. Comparison of proteins expressed on secretory vesicle membranes and plasma membranes of human neutrophils. *J Immunol,* 180**,** 5575-81.

URIARTE, S. M., RANE, M. J., LUERMAN, G. C., BARATI, M. T., WARD, R. A., NAUSEEF, W. M. & MCLEISH, K. R. 2011. Granule exocytosis contributes to priming and activation of the human neutrophil respiratory burst. *J Immunol,* 187**,** 391-400.

VAN DE VEN, J. P., NILSSON, S. C., TAN, P. L., BUITENDIJK, G. H., RISTAU, T., MOHLIN, F. C., NABUURS, S. B., SCHOENMAKER-KOLLER, F. E., SMAILHODZIC, D., CAMPOCHIARO, P. A., et al. 2013. A functional variant in the CFI gene confers a high risk of age-related macular degeneration. *Nat Genet,* 45**,** 813-7.

VAN DER HARST, P., ZHANG, W., MATEO LEACH, I., RENDON, A., VERWEIJ, N., SEHMI, J., PAUL, D. S., ELLING, U., ALLAYEE, H., LI, X., et al. 2012. Seventy-five genetic loci influencing the human red blood cell. *Nature,* 492**,** 369-75.

VASQUEZ, L. J., MANN, A. L., CHEN, L. & SORANZO, N. 2016. From GWAS to function: lessons from blood cells. *ISBT Sci Ser,* 11**,** 211-219.

VENTERS, B. J. & PUGH, B. F. 2009. How eukaryotic genes are transcribed. *Crit Rev Biochem Mol Biol,* 44**,** 117-41.

VIKLUND, H., BERNSEL, A., SKWARK, M. & ELOFSSON, A. 2008. SPOCTOPUS: a combined predictor of signal peptides and membrane protein topology. *Bioinformatics,* 24**,** 2928-9.

VISSCHER, P. M., WRAY, N. R., ZHANG, Q., SKLAR, P., MCCARTHY, M. I., BROWN, M. A. & YANG, J. 2017. 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am J Hum Genet,* 101**,** 5-22.

VOIGHT, B. F., PELOSO, G. M., ORHO-MELANDER, M., FRIKKE-SCHMIDT, R., BARBALIC, M., JENSEN, M. K., HINDY, G., HOLM, H., DING, E. L., JOHNSON, T., et al. 2012. Plasma HDL cholesterol and risk of myocardial infarction: a mendelian randomisation study. *Lancet,* 380**,** 572-80.

WAIN, L. V., SHRINE, N., MILLER, S., JACKSON, V. E., NTALLA, I., SOLER ARTIGAS, M., BILLINGTON, C. K., KHEIRALLAH, A. K., ALLEN, R., COOK, J. P., et al. 2015. Novel insights into the genetics of smoking behaviour, lung function, and chronic obstructive pulmonary disease (UK BiLEVE): a genetic association study in UK Biobank. *Lancet Respir Med,* 3**,** 769-81.

WALL, J. D. & PRITCHARD, J. K. 2003. Haplotype blocks and linkage disequilibrium in the human genome. *Nat Rev Genet,* 4**,** 587-97.

WANG, J., CHAI, X., ERIKSSON, U. & NAPOLI, J. L. 1999. Activity of human 11-cis-retinol dehydrogenase (Rdh5) with steroids and retinoids and expression of its mRNA in extra-ocular human tissue. *Biochem J,* 338 (Pt 1)**,** 23-7.

WANG, L., PITTMAN, K. J., BARKER, J. R., SALINAS, R. E., STANAWAY, I. B., CARROLL, R. J., BALMAT, T., INGHAM, A., GOPALAKRISHNAN, A. M., GIBBS, K. D., et al. 2017a. [Pre-print] An atlas of genetic variation for linking pathogen-induced cellular traits to human disease. *bioRxiv.*

WANG, X., HE, L., GOGGIN, S., SAADAT, A., WANG, L., CLAUSSNITZER, M. & KELLIS, M. 2017b. [Pre-print] High-resolution genome-wide functional dissection of transcriptional regulatory regions in human. *bioRxiv.*

WARD, A. C., VAN AESCH, Y. M., SCHELEN, A. M. & TOUW, I. P. 1999. Defective internalization and sustained activation of truncated granulocyte colony-stimulating factor receptor found in severe congenital neutropenia/acute myeloid leukemia. *Blood,* 93**,** 447-58.

WARD, L. D. & KELLIS, M. 2012. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res,* 40**,** D930-4.

WARREN, C. R., O'SULLIVAN, J. F., FRIESEN, M., BECKER, C. E., ZHANG, X., LIU, P., WAKABAYASHI, Y., MORNINGSTAR, J. E., SHI, X., CHOI, J., et al. 2017. Induced Pluripotent Stem Cell Differentiation Enables Functional Validation of GWAS Variants in Metabolic Disease. *Cell Stem Cell,* 20**,** 547-557 e7.

WASZAK, S. M., DELANEAU, O., GSCHWIND, A. R., KILPINEN, H., RAGHAV, S. K., WITWICKI, R. M., ORIOLI, A., WIEDERKEHR, M., PANOUSIS, N. I., YUROVSKY, A., et al. 2015. Population Variation and Genetic Control of Modular Chromatin Architecture in Humans. *Cell,* 162**,** 1039-50.

WATKINS, H. & FARRALL, M. 2006. Genetic susceptibility to coronary artery disease: from promise to progress. *Nat Rev Genet,* 7**,** 163-73.

WEI, B., JOLMA, A., SAHU, B., ORRE, L. M., ZHONG, F., ZHU, F., KIVIOJA, T., KAUR SUR, I., LEHTIO, J., TAIPALE, M., et al. 2017. Strong binding activity of few transcription factors is a major determinant of open chromatin. *bioRxiv.* bioRxiv.

WEIDENBUSCH, M., KULKARNI, O. P. & ANDERS, H. J. 2017. The innate immune system in human systemic lupus erythematosus. *Clin Sci (Lond),* 131**,** 625-634.

WEISS, S. T. 2010. Lung function and airway diseases. *Nat Genet,* 42**,** 14-6.

WESTERTERP, M. & TALL, A. R. 2015. SORTILIN: many headed hydra. *Circ Res,* 116**,** 764-6.

WESTRA, H. J., PETERS, M. J., ESKO, T., YAGHOOTKAR, H., SCHURMANN, C., KETTUNEN, J., CHRISTIANSEN, M. W., FAIRFAX, B. P., SCHRAMM, K., POWELL, J. E., et al. 2013. Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat Genet,* 45**,** 1238-1243.

WHEELER, J. G., MUSSOLINO, M. E., GILLUM, R. F. & DANESH, J. 2004. Associations between differential leucocyte count and incident coronary heart disease: 1764 incident cases from seven prospective studies of 30,374 individuals. *Eur Heart J,* 25**,** 1287-92.

WICKHAM, H. 2009. ggplot2: Elegant Graphics for Data Analysis. *Springer-Verlag New York*.

WILD, P. S., ZELLER, T., SCHILLERT, A., SZYMCZAK, S., SINNING, C. R., DEISEROTH, A., SCHNABEL, R. B., LUBOS, E., KELLER, T., ELEFTHERIADIS, M. S., et al. 2011. A genome-wide association study identifies LIPA as a susceptibility gene for coronary artery disease. *Circ Cardiovasc Genet,* 4**,** 403-12.

WON, H. H., NATARAJAN, P., DOBBYN, A., JORDAN, D. M., ROUSSOS, P., LAGE, K., RAYCHAUDHURI, S., STAHL, E. & DO, R. 2015. Disproportionate Contributions of Select Genomic Compartments and Cell Types to Genetic Risk for Coronary Artery Disease. *PLoS Genet,* 11**,** e1005622.

WOODS, B. A. & LEVINE, R. L. 2015. The role of mutations in epigenetic regulators in myeloid malignancies. *Immunol Rev,* 263**,** 22-35.

WRIGHT, H. L., MOOTS, R. J., BUCKNALL, R. C. & EDWARDS, S. W. 2010. Neutrophil function in inflammation and inflammatory diseases. *Rheumatology (Oxford),* 49**,** 1618-31.

WRIGHT, H. L., MOOTS, R. J. & EDWARDS, S. W. 2014. The multifactorial role of neutrophils in rheumatoid arthritis. *Nat Rev Rheumatol,* 10**,** 593-601.

WYSS-CORAY, T. & ROGERS, J. 2012. Inflammation in Alzheimer disease-a brief review of the basic science and clinical literature. *Cold Spring Harb Perspect Med,* 2**,** a006346.

XING, Y., XU, Q. & LEE, C. 2003. Widespread production of novel soluble protein isoforms by alternative splicing removal of transmembrane anchoring domains. *FEBS Lett,* 555**,** 572-8.

YAN, J., ENGE, M., WHITINGTON, T., DAVE, K., LIU, J., SUR, I., SCHMIERER, B., JOLMA, A., KIVIOJA, T., TAIPALE, M., et al. 2013. Transcription factor binding in human cells occurs in dense clusters formed around cohesin anchor sites. *Cell,* 154**,** 801-13.

YANG, J., FERREIRA, T., MORRIS, A. P., MEDLAND, S. E., GENETIC INVESTIGATION OF, A. T. C., REPLICATION, D. I. G., META-ANALYSIS, C., MADDEN, P. A., HEATH, A. C., MARTIN, N. G., et al. 2012. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat Genet,* 44**,** 369-75, S1-3.

YANG, J., LEE, S. H., GODDARD, M. E. & VISSCHER, P. M. 2011. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet,* 88**,** 76-82.

YEHOSHUA, Z., DE AMORIM GARCIA FILHO, C. A., NUNES, R. P., GREGORI, G., PENHA, F. M., MOSHFEGHI, A. A., ZHANG, K., SADDA, S., FEUER, W. & ROSENFELD, P. J. 2014. Systemic complement inhibition with eculizumab for geographic atrophy in age-related macular degeneration: the COMPLETE study. *Ophthalmology,* 121**,** 693-701.

YOON, J., TERADA, A. & KITA, H. 2007. CD66b regulates adhesion and activation of human eosinophils. *J Immunol,* 179**,** 8454-62.

YORKE-SMITH, M., PIGNI, A. & SA, M. S. 2011. *CSF3R polypeptides and uses thereof*. US patent application US 12/304,427.

YUO, A., KITAGAWA, S., OHSAKA, A., SAITO, M. & TAKAKU, F. 1990. Stimulation and priming of human neutrophils by granulocyte colony-stimulating factor and granulocyte-macrophage colony-stimulating factor: qualitative and quantitative differences. *Biochem Biophys Res Commun,* 171**,** 491-7.

YUSUF, S., HAWKEN, S., OUNPUU, S., DANS, T., AVEZUM, A., LANAS, F., MCQUEEN, M., BUDAJ, A., PAIS, P., VARIGOS, J., et al. 2004. Effect of potentially modifiable risk factors associated with myocardial infarction in 52 countries (the INTERHEART study): case-control study. *Lancet,* 364**,** 937-52.

ZABIDI, M. A., ARNOLD, C. D., SCHERNHUBER, K., PAGANI, M., RATH, M., FRANK, O. & STARK, A. 2015. Enhancer-core-promoter specificity separates developmental and housekeeping gene regulation. *Nature,* 518**,** 556-9.

ZARBOCK, A. & LEY, K. 2008. Mechanisms and consequences of neutrophil interaction with the endothelium. *Am J Pathol,* 172**,** 1-7.

ZARET, K. S., WATTS, J., XU, J., WANDZIOCH, E., SMALE, S. T. & SEKIYA, T. 2008. Pioneer factors, genetic competence, and inductive signaling: programming liver and pancreas progenitors from the endoderm. *Cold Spring Harb Symp Quant Biol,* 73**,** 119-26.

ZELLER, T., WILD, P., SZYMCZAK, S., ROTIVAL, M., SCHILLERT, A., CASTAGNE, R., MAOUCHE, S., GERMAIN, M., LACKNER, K., ROSSMANN, H., et al. 2010. Genetics and beyond--the transcriptome of human monocytes and disease susceptibility. *PLoS One,* 5**,** e10693.

ZENARO, E., PIACENTINO, G. & CONSTANTIN, G. 2017. The blood-brain barrier in Alzheimer's disease. *Neurobiol Dis,* 107**,** 41-56.

ZENARO, E., PIETRONIGRO, E., DELLA BIANCA, V., PIACENTINO, G., MARONGIU, L., BUDUI, S., TURANO, E., ROSSI, B., ANGIARI, S., DUSI, S., et al. 2015. Neutrophils promote Alzheimer's disease-like pathology and cognitive decline via LFA-1 integrin. *Nat Med,* 21**,** 880-6.

ZHANG, H., NGUYEN-JACKSON, H., PANOPOULOS, A. D., LI, H. S., MURRAY, P. J. & WATOWICH, S. S. 2010. STAT3 controls myeloid progenitor growth during emergency granulopoiesis. *Blood,* 116**,** 2462-71.

ZHANG, Y., LIU, T., MEYER, C. A., EECKHOUTE, J., JOHNSON, D. S., BERNSTEIN, B. E., NUSBAUM, C., MYERS, R. M., BROWN, M., LI, W., et al. 2008. Model-based analysis of ChIP-Seq (MACS). *Genome Biol,* 9**,** R137.

ZHAO, T., BENARD, V., BOHL, B. P. & BOKOCH, G. M. 2003. The molecular basis for adhesion-mediated suppression of reactive oxygen species generation by human neutrophils. *J Clin Invest,* 112**,** 1732-40.

ZHENG, S., CHEN, H., WANG, Y., GAO, W., FU, Z., ZHOU, Q., JIANG, Y., LIN, Q., TAN, L., YE, H., et al. 2016. Long non-coding RNA LOC389641 promotes progression of pancreatic ductal adenocarcinoma and increases cell invasion by regulating E-cadherin in a TNFRSF10A-related manner. *Cancer Lett,* 371**,** 354-65.

ZHERNAKOVA, A., STAHL, E. A., TRYNKA, G., RAYCHAUDHURI, S., FESTEN, E. A., FRANKE, L., WESTRA, H. J., FEHRMANN, R. S., KURREEMAN, F. A., THOMSON, B., et al. 2011. Meta-analysis of genome-wide association studies in celiac disease and rheumatoid arthritis identifies fourteen non-HLA shared loci. *PLoS Genet,* 7**,** e1002004.

ZHOU, Q., LI, T. & PRICE, D. H. 2012. RNA polymerase II elongation control. *Annu Rev Biochem,* 81**,** 119-43.

ZHOU, X., MARICQUE, B., XIE, M., LI, D., SUNDARAM, V., MARTIN, E. A., KOEBBE, B. C., NIELSEN, C., HIRST, M., FARNHAM, P., et al. 2011. The Human Epigenome Browser at Washington University. *Nat Methods,* 8**,** 989-90.

ZHU, Z., ZHANG, F., HU, H., BAKSHI, A., ROBINSON, M. R., POWELL, J. E., MONTGOMERY, G. W., GODDARD, M. E., WRAY, N. R., VISSCHER, P. M., et al. 2016. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat Genet,* 48**,** 481-7.

ZIEGLER-HEITBROCK, L. 2007. The CD14+ CD16+ blood monocytes: their role in infection and inflammation. *J Leukoc Biol,* 81**,** 584-92.

Supplementary Information

Chapter 2 Supplementary Information

GWAS summary statistics contributions

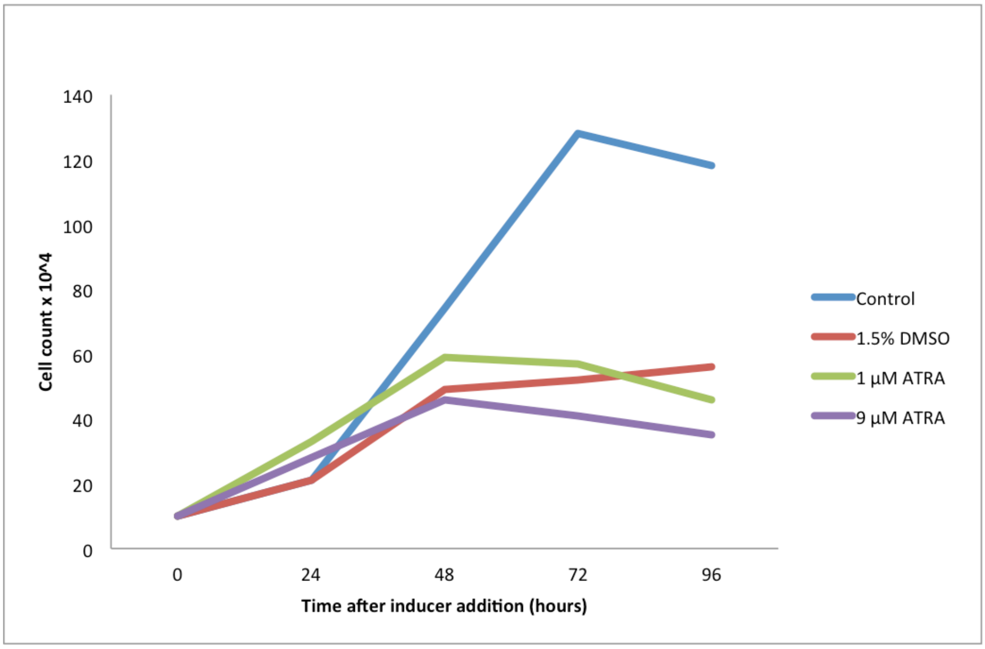
We thank the International Genomics of Alzheimer's Project (IGAP) for providing summary results data for these analyses. The investigators within IGAP contributed to the design and implementation of IGAP and/or provided data but did not participate in analysis or writing of this report. IGAP was made possible by the generous participation of the control subjects, the patients, and their families. The iSelect chips was funded by the French National Foundation on Alzheimer's disease and related disorders. EADI was supported by the LABEX (laboratory of excellence program investment for the future) DISTALZ grant, Inserm, Institut Pasteur de Lille, Universite de Lille 2 and the Lille University Hospital. GERAD was supported by the Medical Research Council (Grant no 503480), Alzheimer's Research UK (Grant no 503176), the Wellcome Trust (Grant no 082604/2/07/Z) and German Federal Ministry of Education and Research (BMBF): Competence Network Dementia (CND) grant no 01GI0102, 01GI0711, 01GI0420. CHARGE was partly supported by the NIH/NIA grant R01 AG033193 and the NIA AG081220 and AGES contract N01AG12100, the NHLBI grant R01 HL105756, the Icelandic Heart Association, and the Erasmus Medical Center and Erasmus University. ADGC was supported by the NIH/NIA grants: U01 AG032984, U24 AG021886, U01 AG016976, and the Alzheimer's Association grant ADGC10196728. International Genomics of Alzheimer's Project (IGAP) is a large two-stage study based upon genome-wide association studies (GWAS) on individuals of European ancestry. In stage 1, IGAP used genotyped and imputed data on 7,055,881 single nucleotide polymorphisms (SNPs) to meta-analyse four previously-published GWAS datasets consisting of 17,008 Alzheimer's disease cases and 37,154 controls (The European Alzheimer's disease Initiative EADI the Alzheimer Disease Genetics Consortium ADGC The Cohorts for Heart and Aging Research in Genomic Epidemiology consortium CHARGE The Genetic and Environmental Risk in AD consortium GERAD). In stage 2, 11,632 SNPs were genotyped and tested for association in an independent set of 8,572 Alzheimer's disease cases and 11,312 controls. Finally, a meta-analysis was performed combining results from stages 1 & 2. The text above was reproduced at request from the International Genomics of Alzheimer's Project (IGAP).

We thank the IAMDGC consortium for sharing the summary statistics and effect sizes of advanced age-related macular degeneration GWAS from Fritsche *et al.* (2016). The list of consortia members can be found in (Persad et al., 2017).

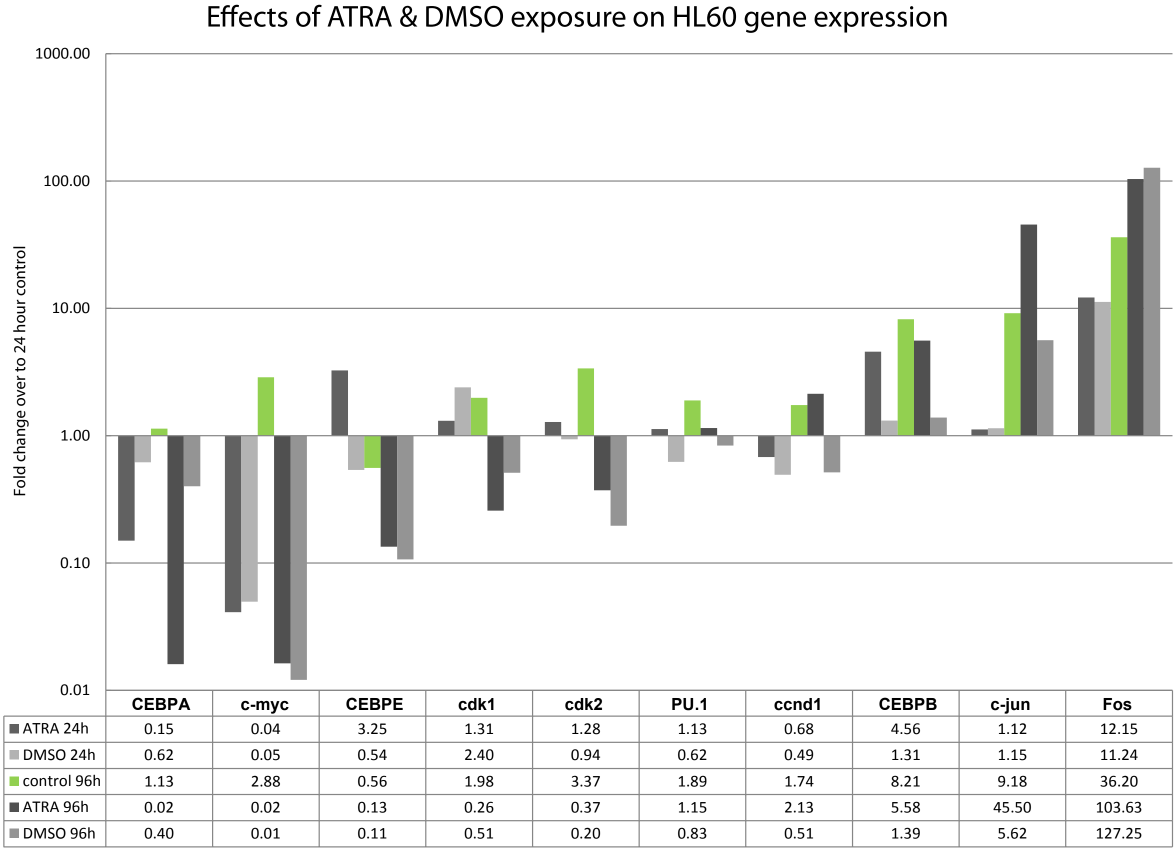
Applications and queries can be made through the following contacts: goncalo@umich.edu, larsf@umich.edu

Further information on AMD statistics: http://amdgenetics.org

Chapter 2 Supplementary Figures



Supplementary Figure 2.1: Proliferation of dividing and differentiated HL60 cells Cell count of HL60 cells shown for each 24-hour period after addition of either DMSO or ATRA at different concentrations shown. The control HL60 cells had no addition and continued to proliferate until reaching a maximum, likely due to limited nutrients as medium was not changed in this period. The reduced proliferation of ATRA and DMSO conditions is evidence of differentiation.



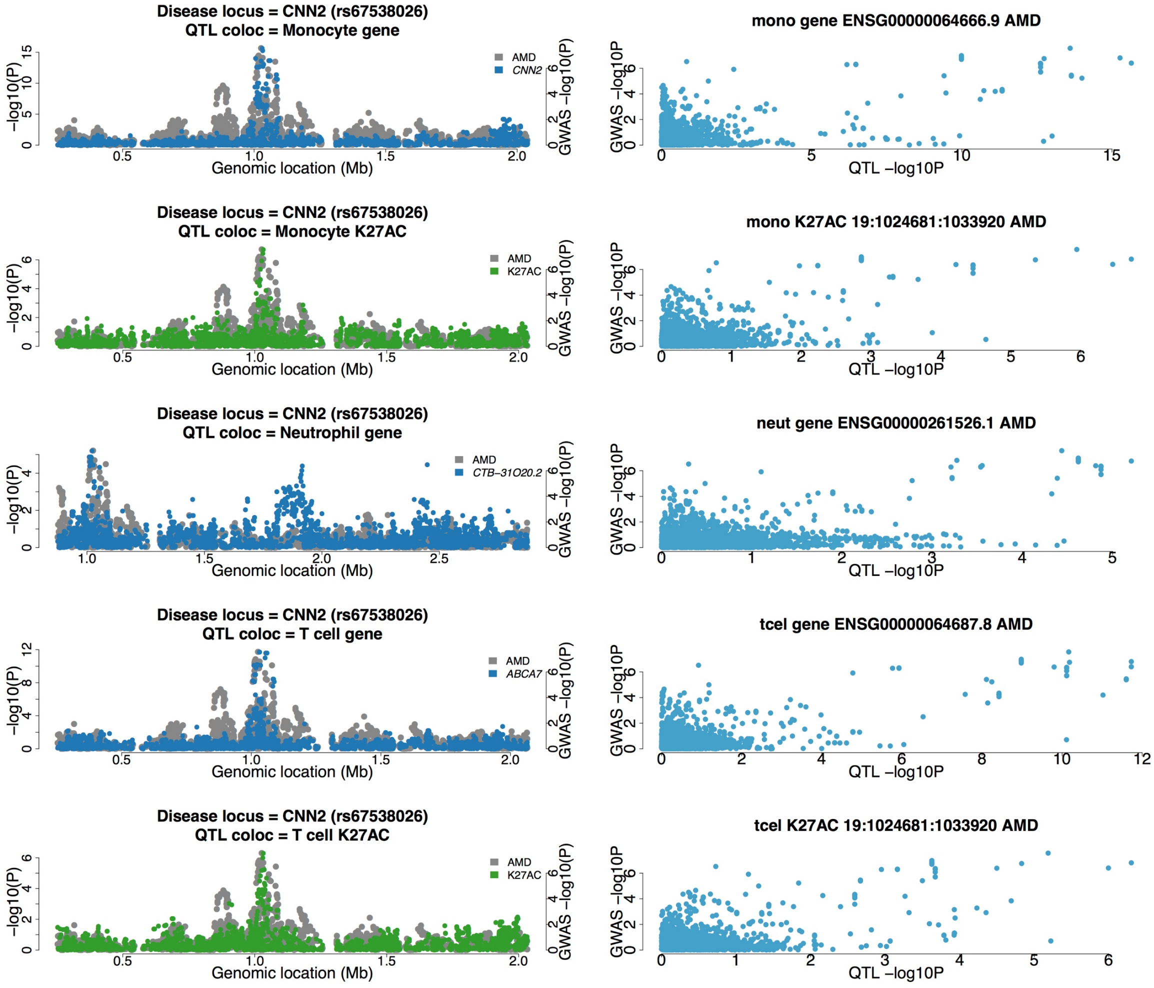
	CEBPA	c-myc	CEBPE	cdk1	cdk2	PU.1	ccnd1	CEBPB	c-jun	Fos
■ ATRA 24h	0.15	0.04	3.25	1.31	1.28	1.13	0.68	4.56	1.12	12.15
■ DMSO 24h	0.62	0.05	0.54	2.40	0.94	0.62	0.49	1.31	1.15	11.24
■ control 96h	1.13	2.88	0.56	1.98	3.37	1.89	1.74	8.21	9.18	36.20
■ ATRA 96h	0.02	0.02	0.13	0.26	0.37	1.15	2.13	5.58	45.50	103.63
■ DMSO 96h	0.40	0.01	0.11	0.51	0.20	0.83	0.51	1.39	5.62	127.25

Supplementary Figure 2.2: Changes in gene expression as a result of HL60 gene expression Barplot shows changes in expression of 10 genes known to be affected by HL60 differentiation. Expression was measured by real-time qPCR and the change in expression is evaluated with respect to HL60 cells grown for 24 hours. Figure produced by Stephen Watt.

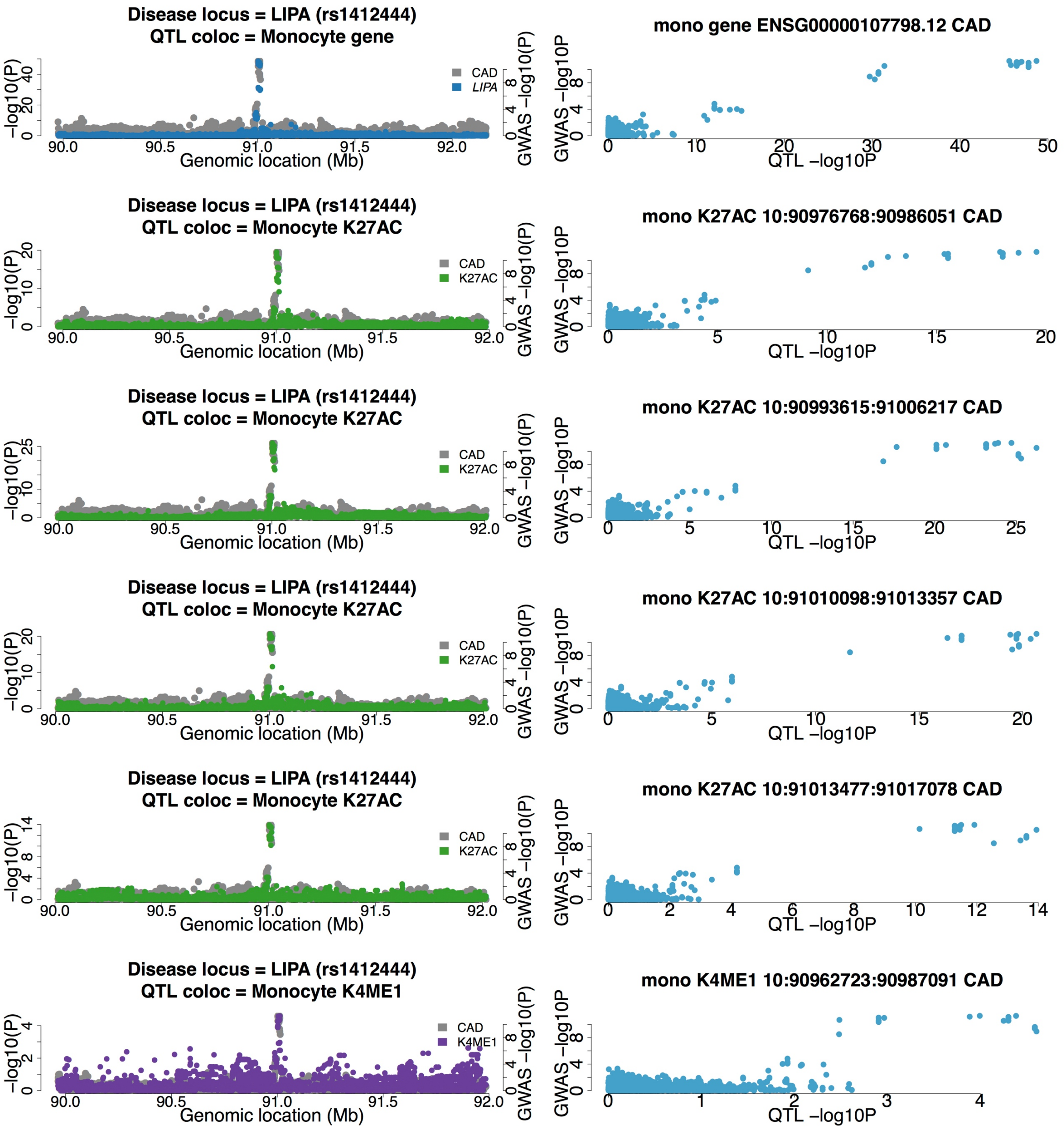
Supplementary Figures 2.3: Examples of disease loci and colocalised molecular features

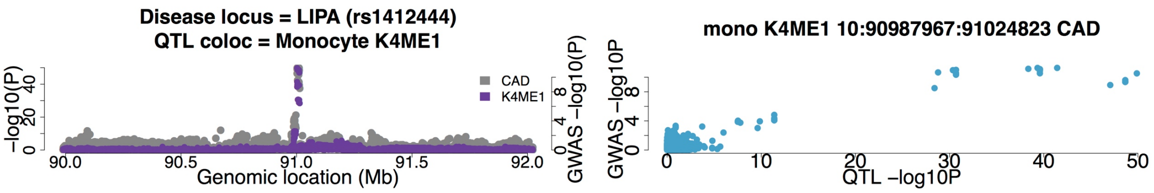
Regional association plots for particular colocalised loci mentioned in the main text of this chapter (left plot). The disease locus as defined by the GWAS study and the lead SNP in brackets is included, as well as the cell type and feature that is colocalised. The eQTL gene is also included in the legend. The correlation between p values of the two traits is also given (right plot). If a locus is colocalised, then the p values in the locus should be correlated. These also provide an initial indication of independent signals that may not be colocalised if the correlation is lower.

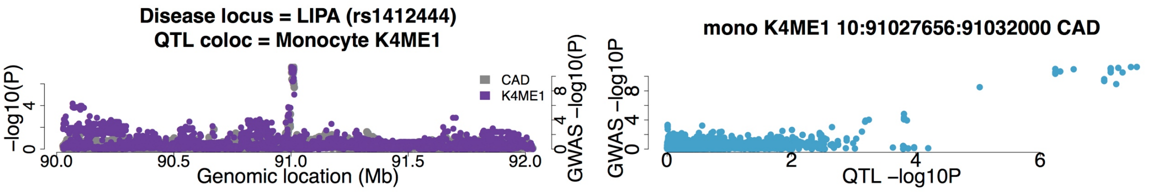
A) Advanced age-related macular degeneration *CNN2* locus

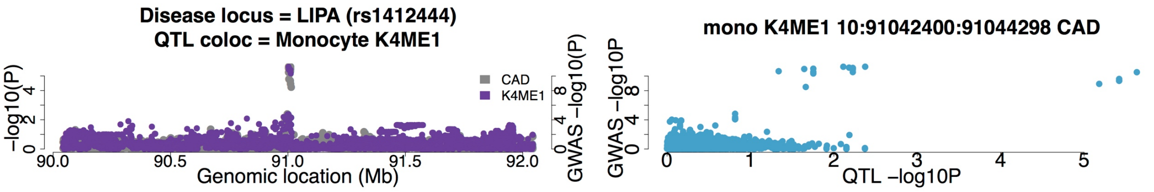


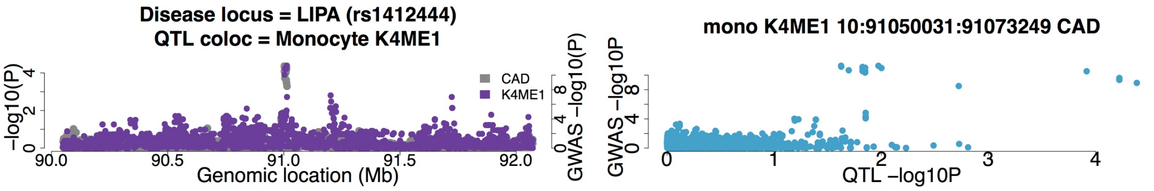
B) Coronary artery disease *LIPA* locus

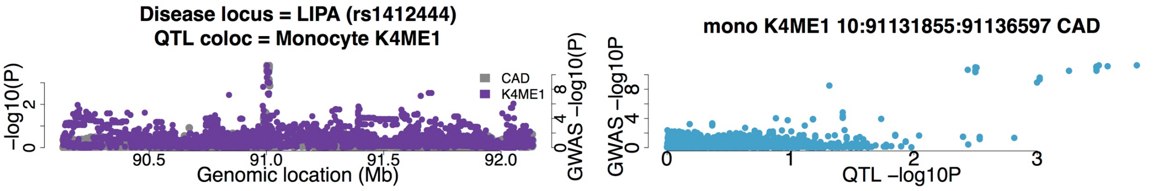


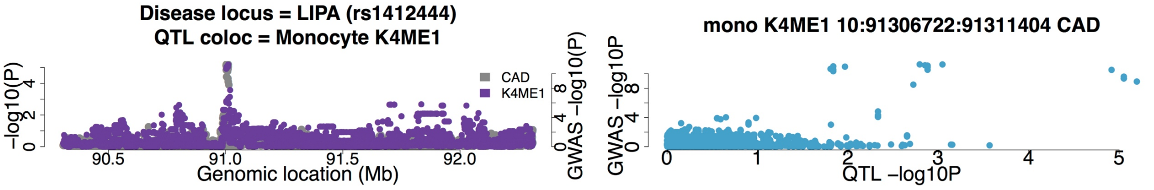

Disease locus = LIPA (rs1412444)
QTL coloc = Monocyte K4ME1
−log10(P)
0 20 40
90.0 90.5 91.0 91.5 92.0
Genomic location (Mb)
CAD
K4ME1
GWAS −log10(P)
0 4 8
mono K4ME1 10:90987967:91024823 CAD
GWAS −log10P
0 4 8
0 10 20 30 40 50
QTL −log10P

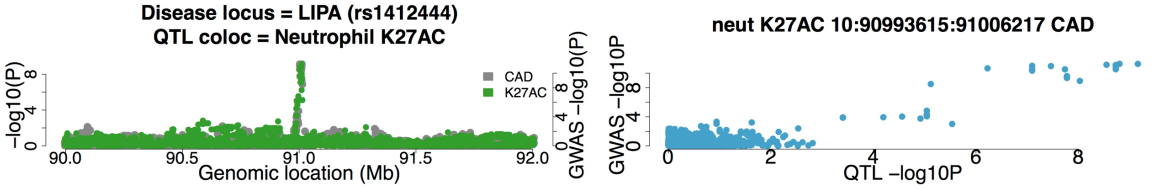

Disease locus = LIPA (rs1412444)
QTL coloc = Monocyte K4ME1
−log10(P)
0 4
90.0 90.5 91.0 91.5 92.0
Genomic location (Mb)
CAD
K4ME1
GWAS −log10(P)
0 4 8
mono K4ME1 10:91027656:91032000 CAD
GWAS −log10P
0 4 8
0 2 4 6
QTL −log10P

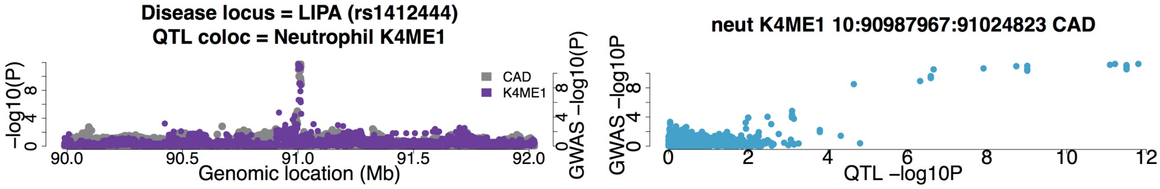

Disease locus = LIPA (rs1412444)
QTL coloc = Monocyte K4ME1
−log10(P)
0 2 4
90.0 90.5 91.0 91.5 92.0
Genomic location (Mb)
CAD
K4ME1
GWAS −log10(P)
0 4 8
mono K4ME1 10:91042400:91044298 CAD
GWAS −log10P
0 4 8
0 1 2 3 4 5
QTL −log10P

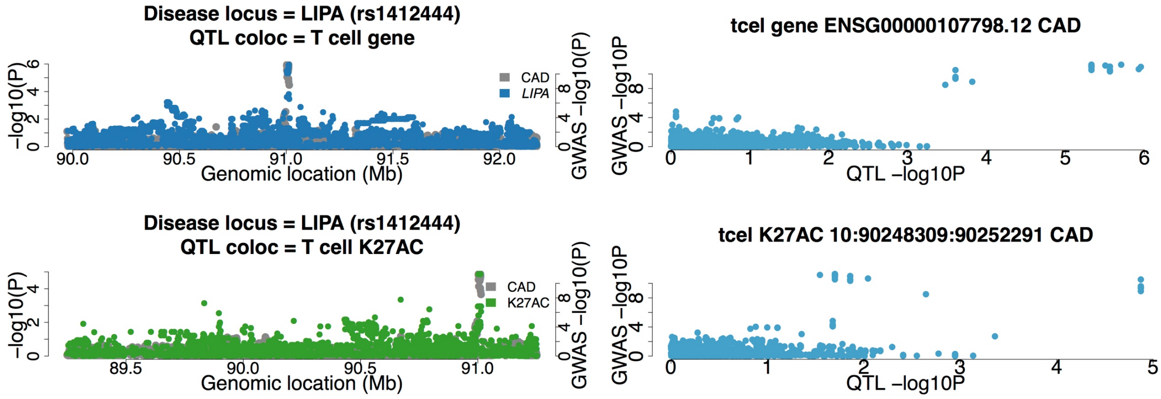

Disease locus = LIPA (rs1412444)
QTL coloc = Monocyte K4ME1
−log10(P)
0 2 4
90.0 90.5 91.0 91.5 92.0
Genomic location (Mb)
CAD
K4ME1
GWAS −log10(P)
0 4 8
mono K4ME1 10:91050031:91073249 CAD
GWAS −log10P
0 4 8
0 1 2 3 4
QTL −log10P


Disease locus = LIPA (rs1412444)
QTL coloc = Monocyte K4ME1
−log10(P)
0 2
90.5 91.0 91.5 92.0
Genomic location (Mb)
CAD
K4ME1
GWAS −log10(P)
0 4 8
mono K4ME1 10:91131855:91136597 CAD
GWAS −log10P
0 4 8
0 1 2 3
QTL −log10P


Disease locus = LIPA (rs1412444)
QTL coloc = Monocyte K4ME1
−log10(P)
0 2 4
90.5 91.0 91.5 92.0
Genomic location (Mb)
CAD
K4ME1
GWAS −log10(P)
0 4 8
mono K4ME1 10:91306722:91311404 CAD
GWAS −log10P
0 4 8
0 1 2 3 4 5
QTL −log10P


Disease locus = LIPA (rs1412444)
QTL coloc = Neutrophil K27AC
−log10(P)
0 4 8
90.0 90.5 91.0 91.5 92.0
Genomic location (Mb)
CAD
K27AC
GWAS −log10(P)
0 4 8
neut K27AC 10:90993615:91006217 CAD
GWAS −log10P
0 4 8
0 2 4 6 8
QTL −log10P


Disease locus = LIPA (rs1412444)
QTL coloc = Neutrophil K4ME1
−log10(P)
0 4 8
90.0 90.5 91.0 91.5 92.0
Genomic location (Mb)
CAD
K4ME1
GWAS −log10(P)
0 4 8
neut K4ME1 10:90987967:91024823 CAD
GWAS −log10P
0 4 8
0 2 4 6 8 10 12
QTL −log10P



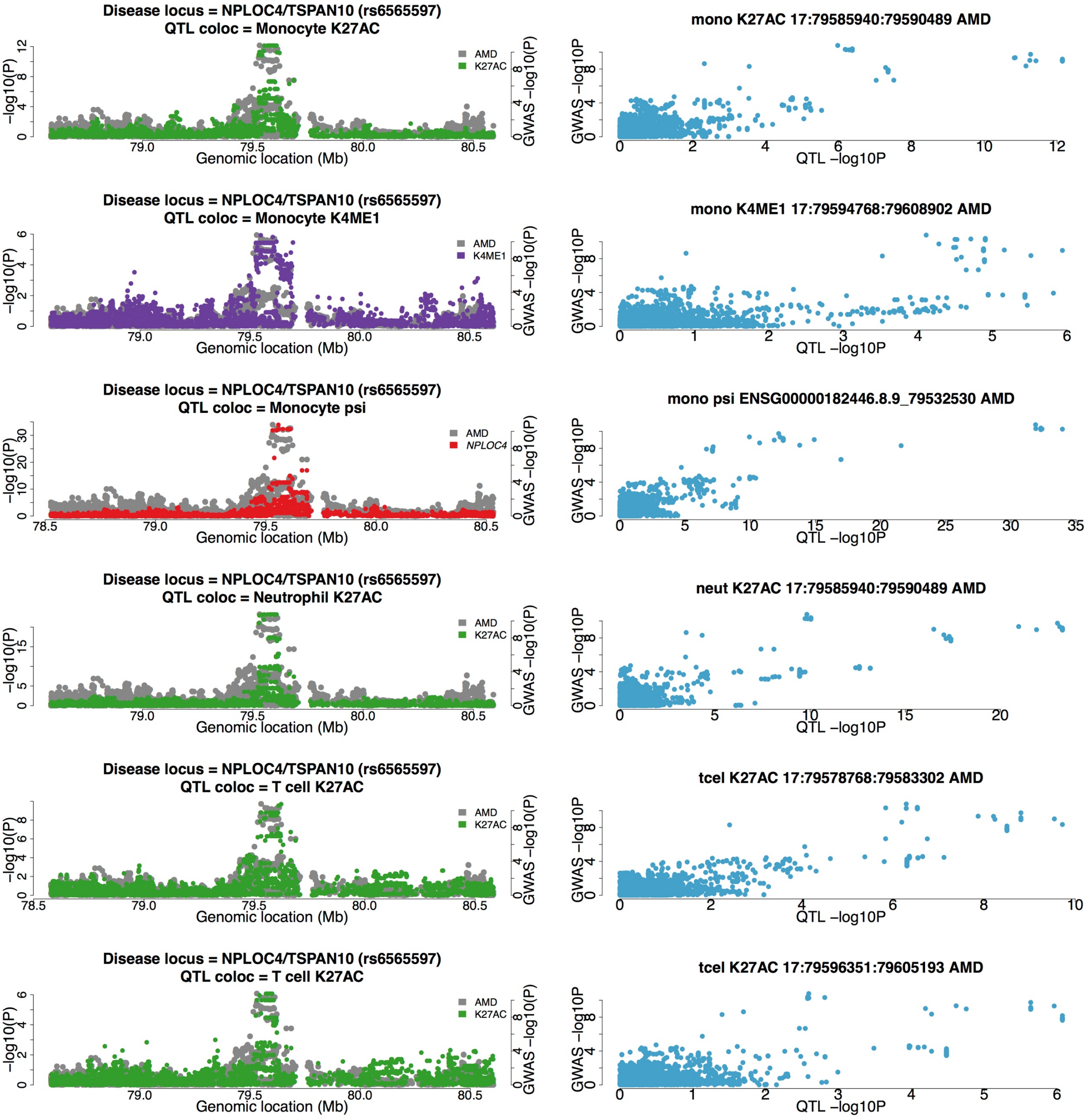
Disease locus = LIPA (rs1412444)
QTL coloc = T cell gene

tcel gene ENSG00000107798.12 CAD

Disease locus = LIPA (rs1412444)
QTL coloc = T cell K27AC

tcel K27AC 10:90248309:90252291 CAD

C) Advanced age-related macular degeneration *NPLOC4* locus



Disease locus = NPLOC4/TSPAN10 (rs6565597)
QTL coloc = Monocyte K27AC

mono K27AC 17:79585940:79590489 AMD

Disease locus = NPLOC4/TSPAN10 (rs6565597)
QTL coloc = Monocyte K4ME1

mono K4ME1 17:79594768:79608902 AMD

Disease locus = NPLOC4/TSPAN10 (rs6565597)
QTL coloc = Monocyte psi

mono psi ENSG00000182446.8.9_79532530 AMD

Disease locus = NPLOC4/TSPAN10 (rs6565597)
QTL coloc = Neutrophil K27AC

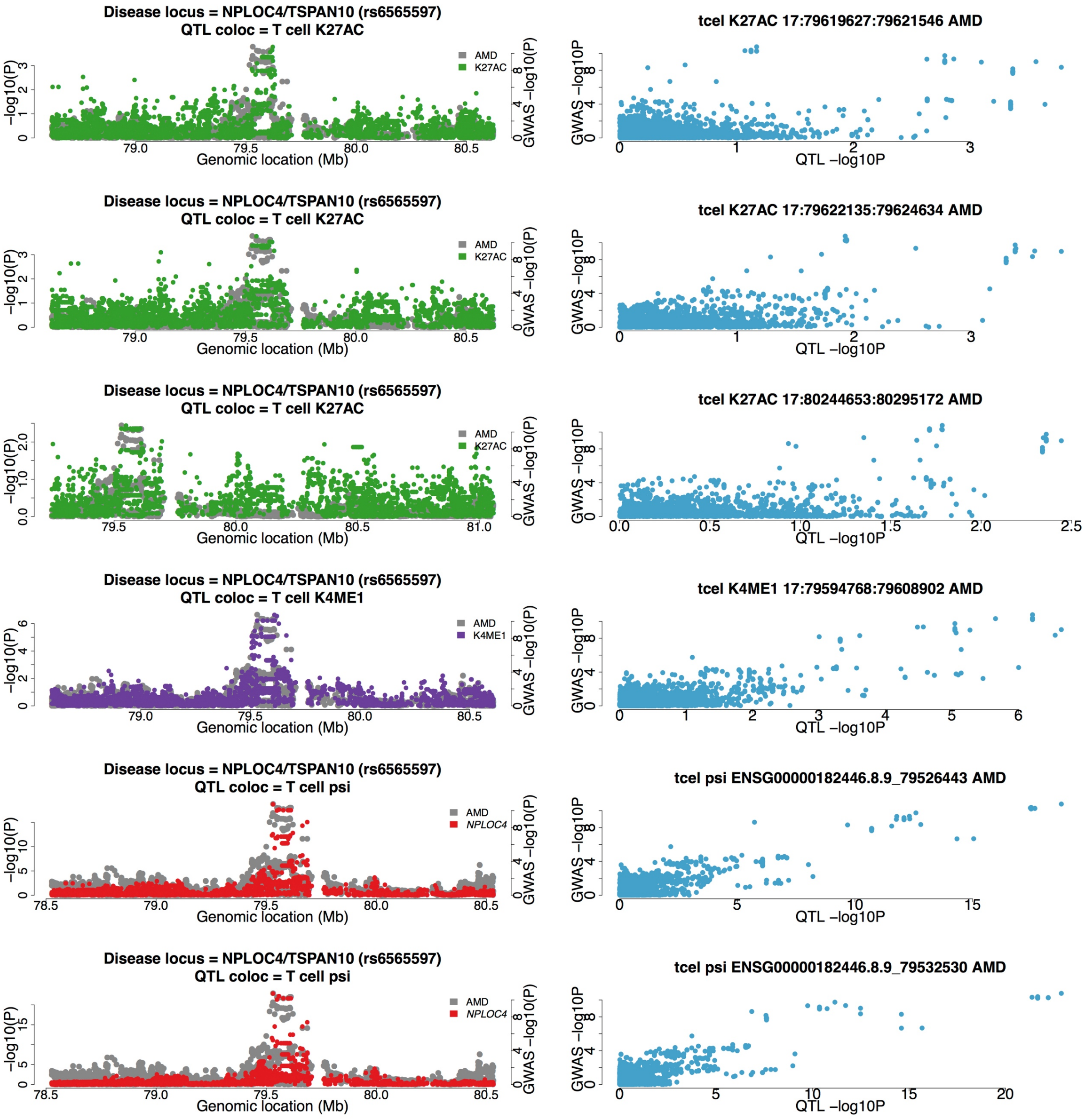
neut K27AC 17:79585940:79590489 AMD

Disease locus = NPLOC4/TSPAN10 (rs6565597)
QTL coloc = T cell K27AC

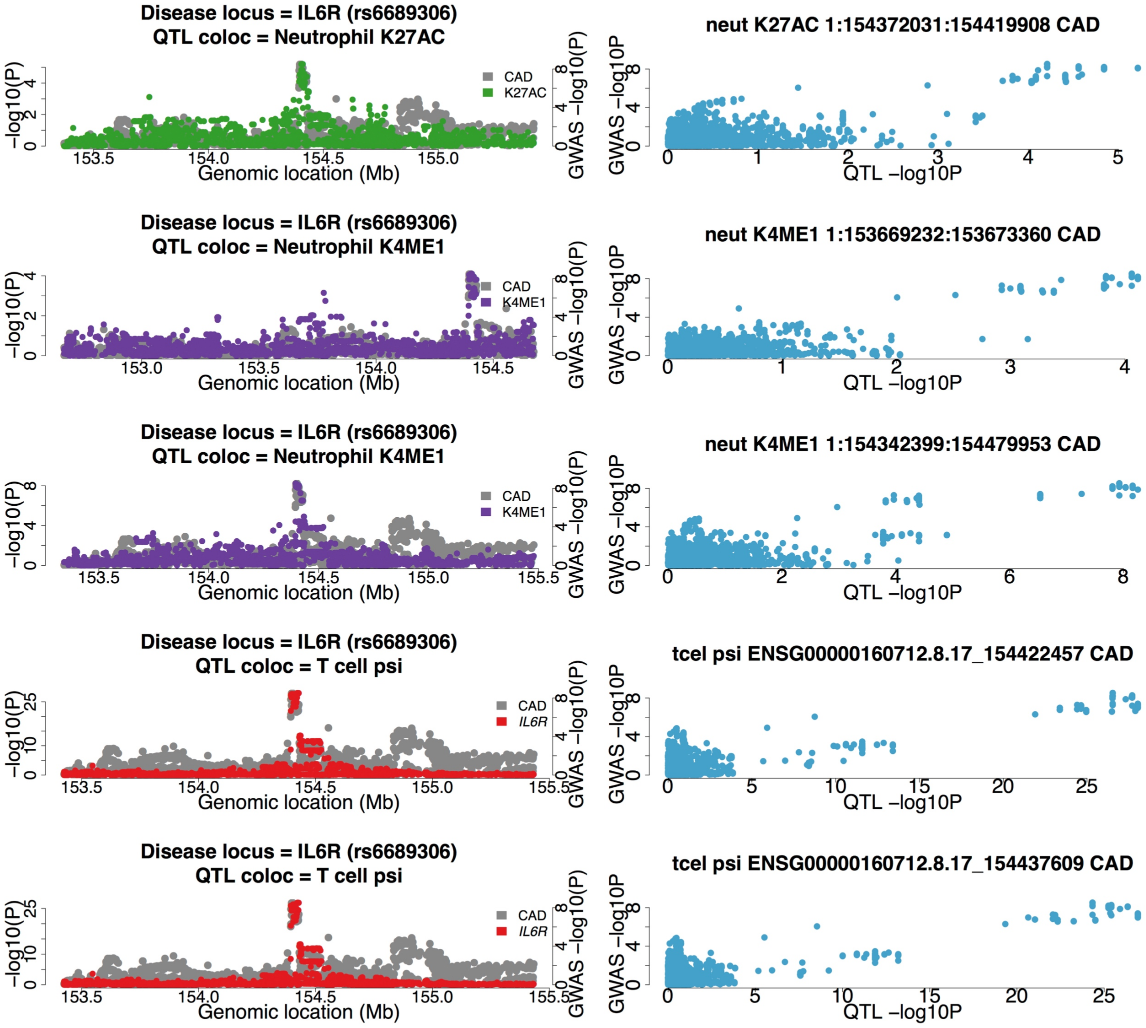
tcel K27AC 17:79578768:79583302 AMD

Disease locus = NPLOC4/TSPAN10 (rs6565597)
QTL coloc = T cell K27AC

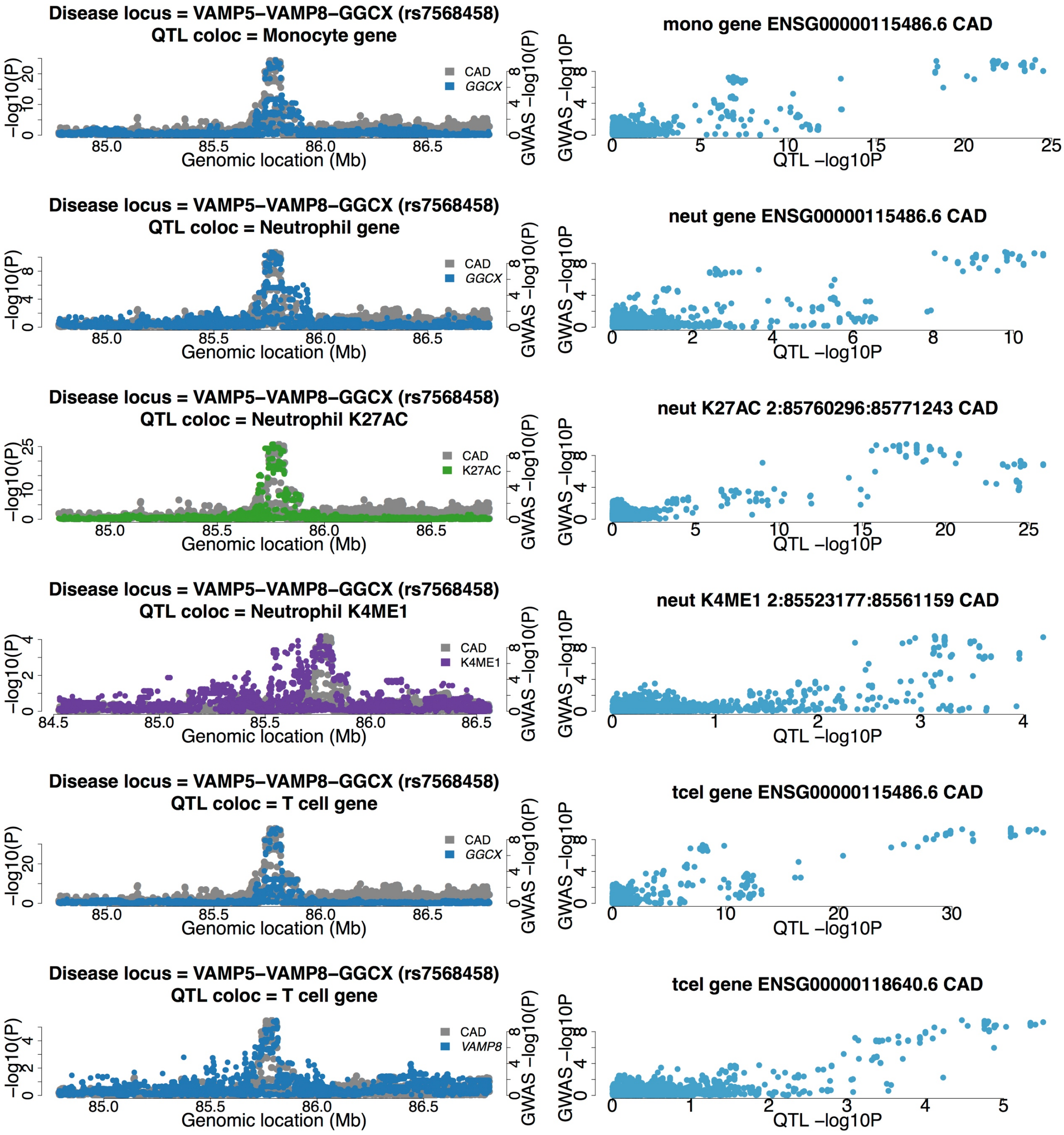
tcel K27AC 17:79596351:79605193 AMD


Disease locus = NPLOC4/TSPAN10 (rs6565597)
QTL coloc = T cell K27AC
AMD
K27AC
−log10(P)
GWAS −log10(P)
Genomic location (Mb)
79.0
79.5
80.0
80.5
tcel K27AC 17:79619627:79621546 AMD
GWAS −log10P
QTL −log10P
Disease locus = NPLOC4/TSPAN10 (rs6565597)
QTL coloc = T cell K27AC
tcel K27AC 17:79622135:79624634 AMD
Disease locus = NPLOC4/TSPAN10 (rs6565597)
QTL coloc = T cell K27AC
81.0
tcel K27AC 17:80244653:80295172 AMD
Disease locus = NPLOC4/TSPAN10 (rs6565597)
QTL coloc = T cell K4ME1
K4ME1
tcel K4ME1 17:79594768:79608902 AMD
Disease locus = NPLOC4/TSPAN10 (rs6565597)
QTL coloc = T cell psi
NPLOC4
78.5
tcel psi ENSG00000182446.8.9_79526443 AMD
Disease locus = NPLOC4/TSPAN10 (rs6565597)
QTL coloc = T cell psi
NPLOC4
tcel psi ENSG00000182446.8.9_79532530 AMD

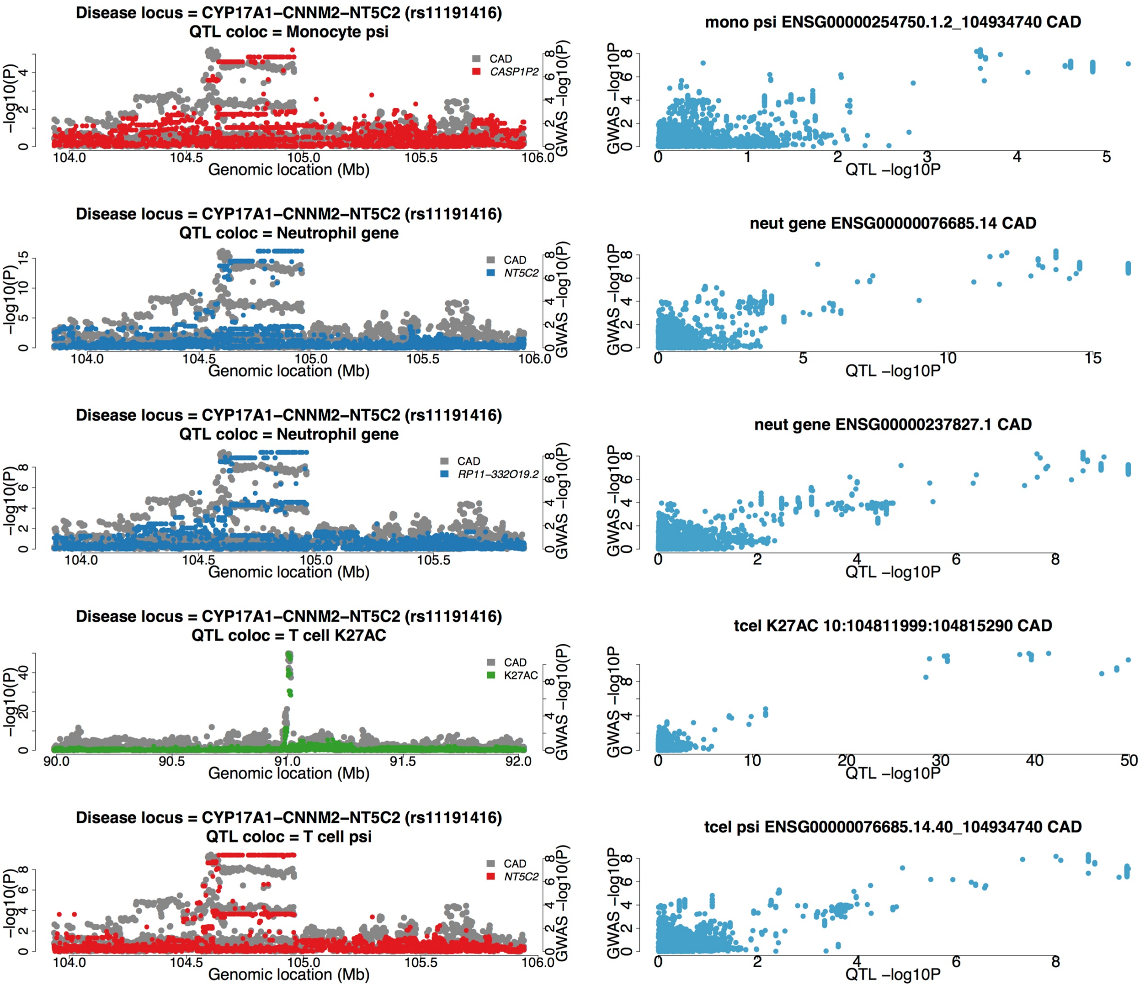
D) Coronary artery disease *IL6R* locus



E) Coronary artery disease *VAMP5-VAMP8-GGCX* locus



F) Coronary artery disease *CYP17A1-CNNM2-NT5C2* locus



Disease locus = CYP17A1–CNNM2–NT5C2 (rs11191416)
QTL coloc = Monocyte psi

mono psi ENSG00000254750.1.2_104934740 CAD

Disease locus = CYP17A1–CNNM2–NT5C2 (rs11191416)
QTL coloc = Neutrophil gene

neut gene ENSG00000076685.14 CAD

Disease locus = CYP17A1–CNNM2–NT5C2 (rs11191416)
QTL coloc = Neutrophil gene

neut gene ENSG00000237827.1 CAD

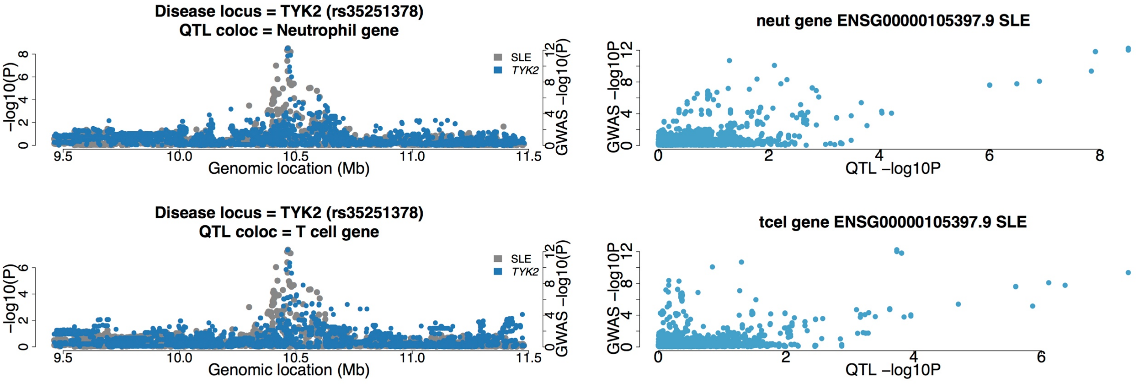
Disease locus = CYP17A1–CNNM2–NT5C2 (rs11191416)
QTL coloc = T cell K27AC

tcel K27AC 10:104811999:104815290 CAD

Disease locus = CYP17A1–CNNM2–NT5C2 (rs11191416)
QTL coloc = T cell psi

tcel psi ENSG00000076685.14.40_104934740 CAD

G) Systemic lupus erythematosus *TYK* locus



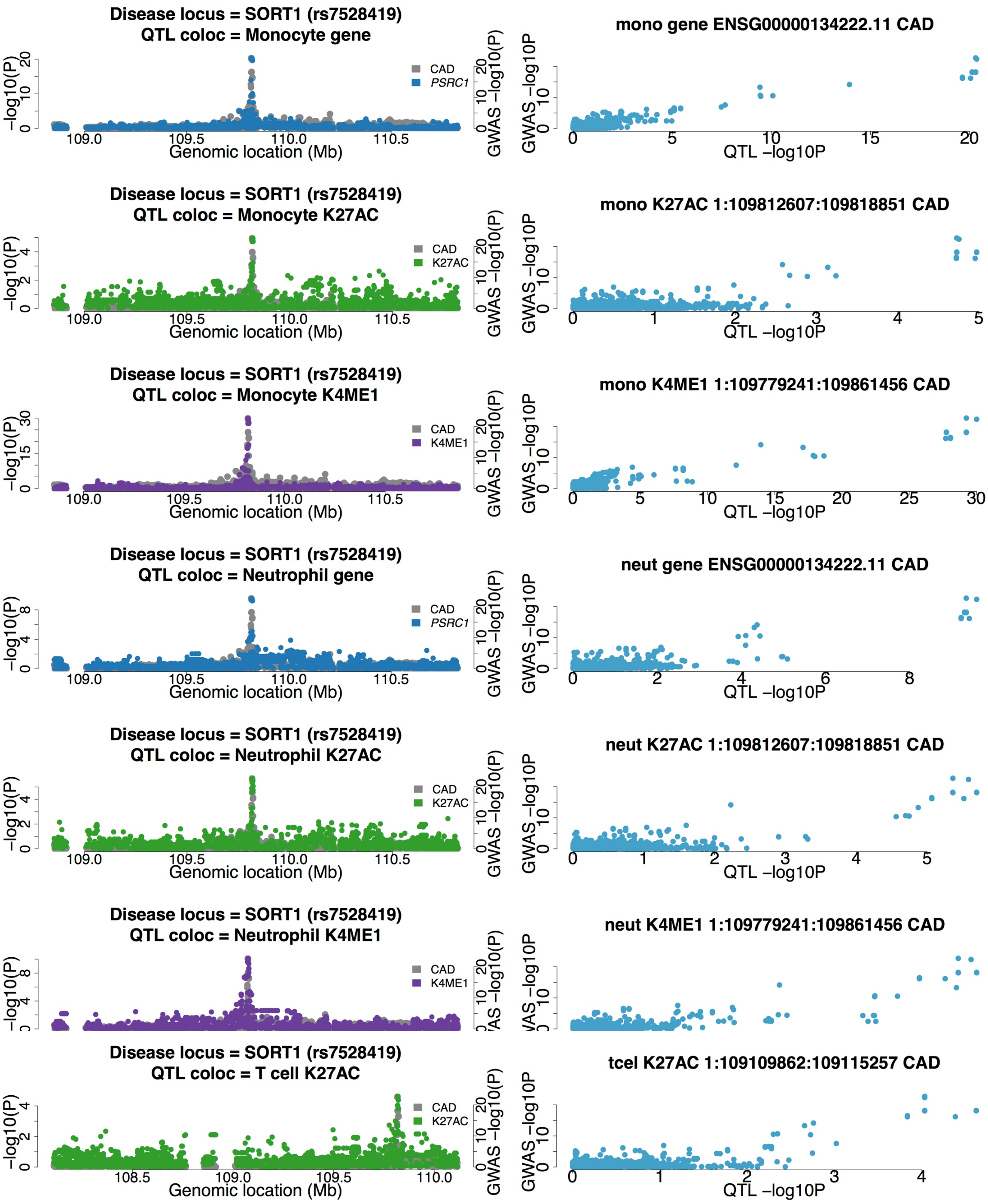
Disease locus = TYK2 (rs35251378)
QTL coloc = Neutrophil gene

neut gene ENSG00000105397.9 SLE

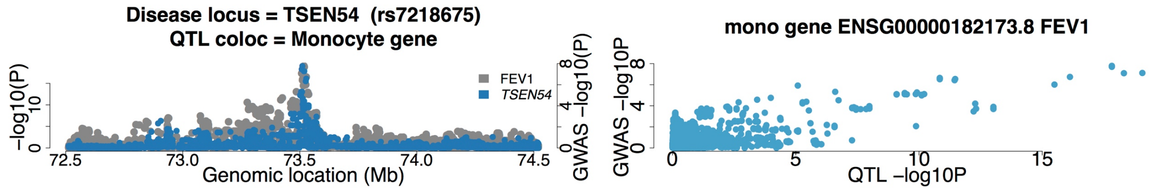
Disease locus = TYK2 (rs35251378)
QTL coloc = T cell gene

tcel gene ENSG00000105397.9 SLE

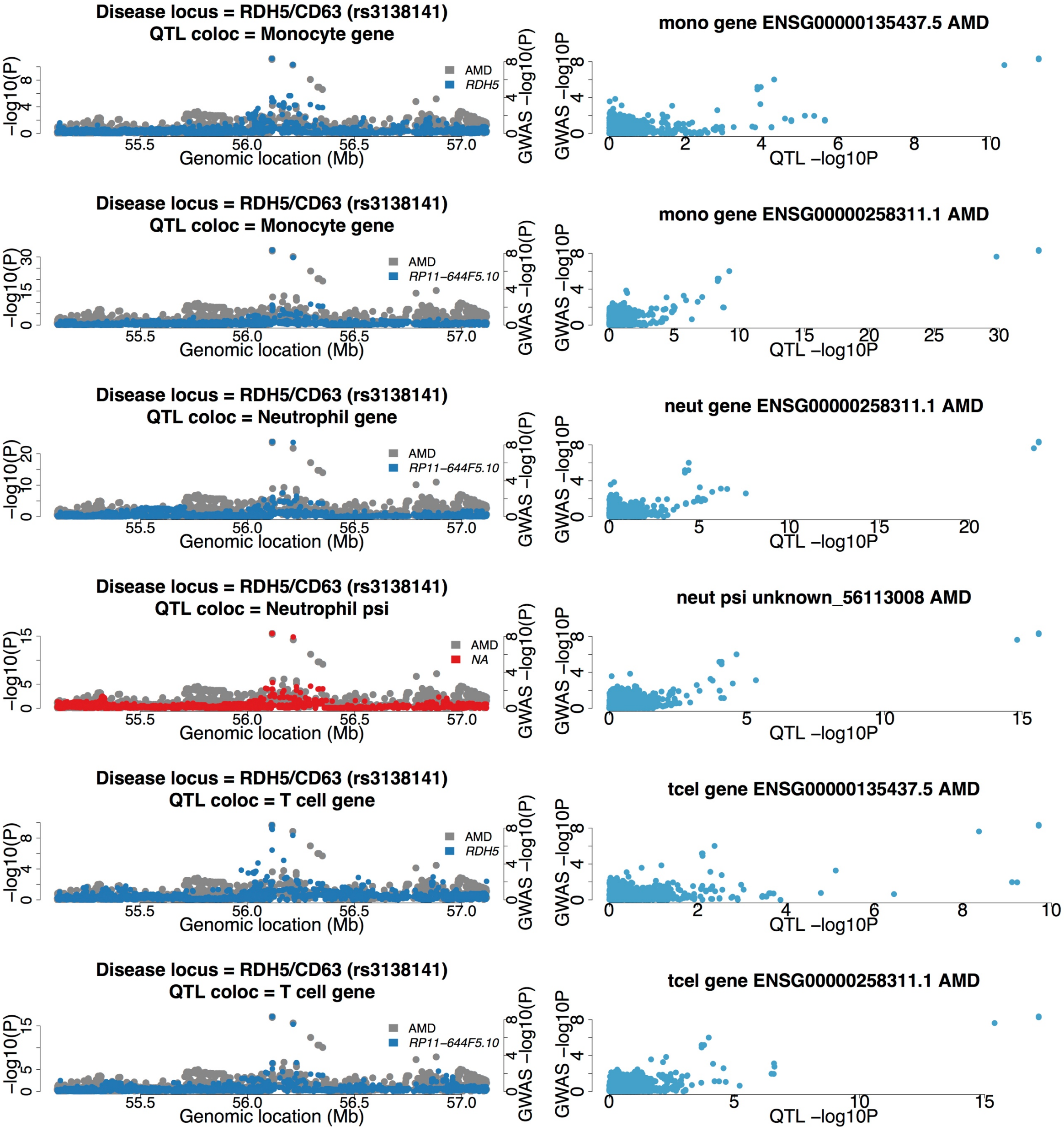
H) Coronary artery disease *SORT1* locus



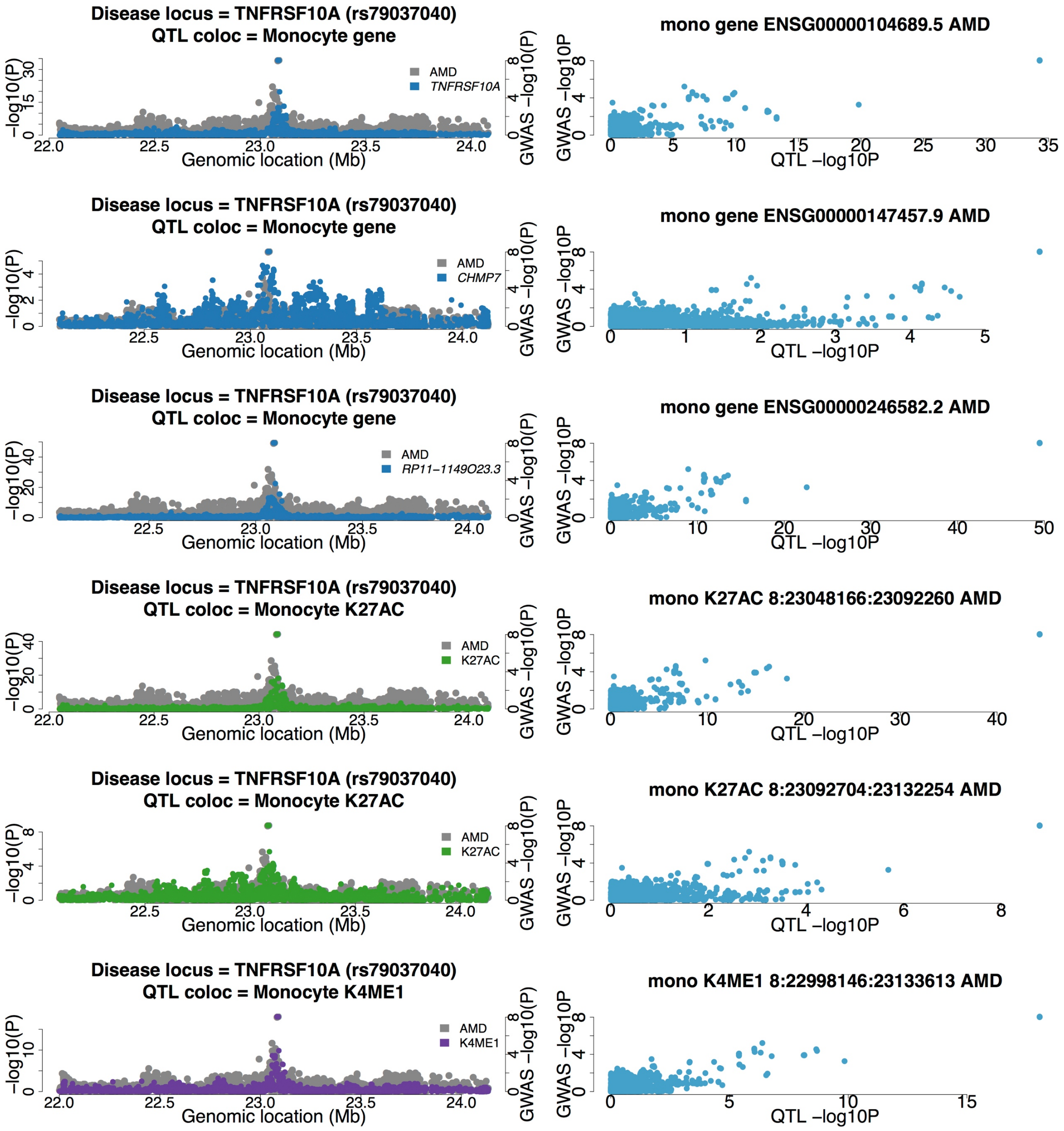
I) Extremes of FEV$_1$ lung ratio *TSEN54* locus



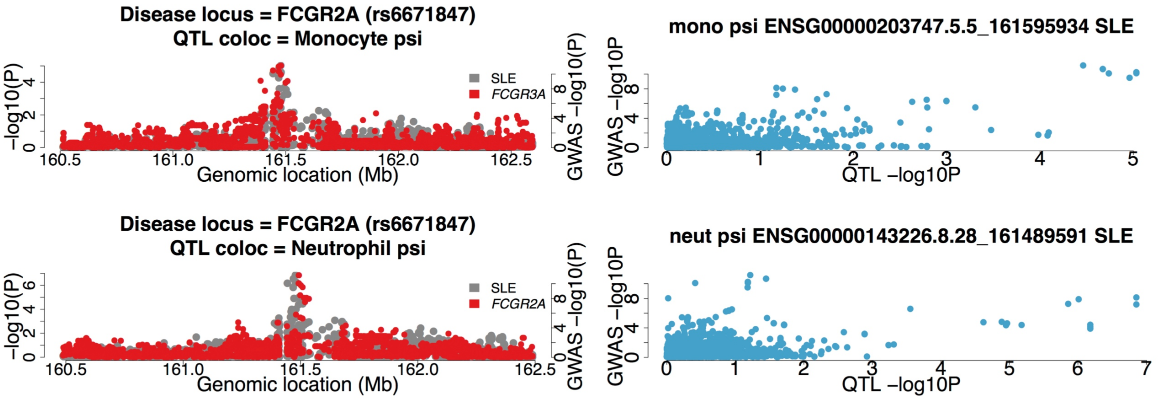
J) Advanced age-related macular degeneration *RDH5* locus



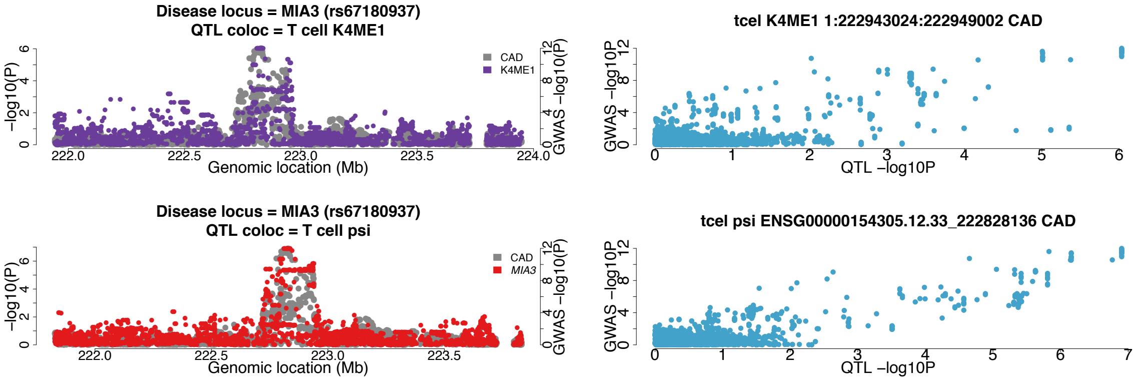
K) Advanced age-related macular degeneration *TNFRSF10A* locus



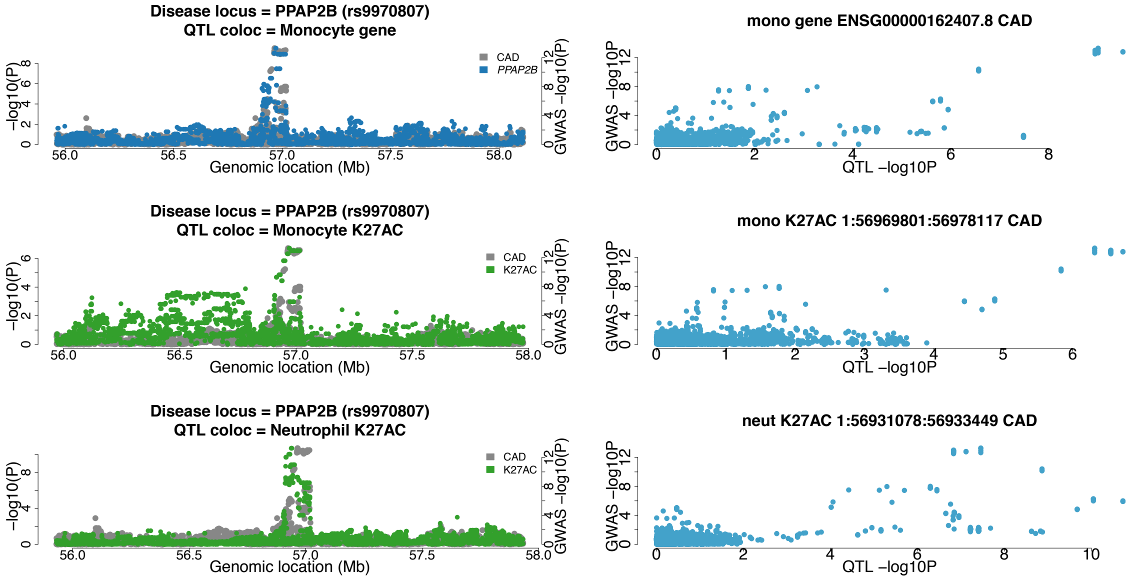
L) Systemic lupus erythematosus *FCGR2A* locus



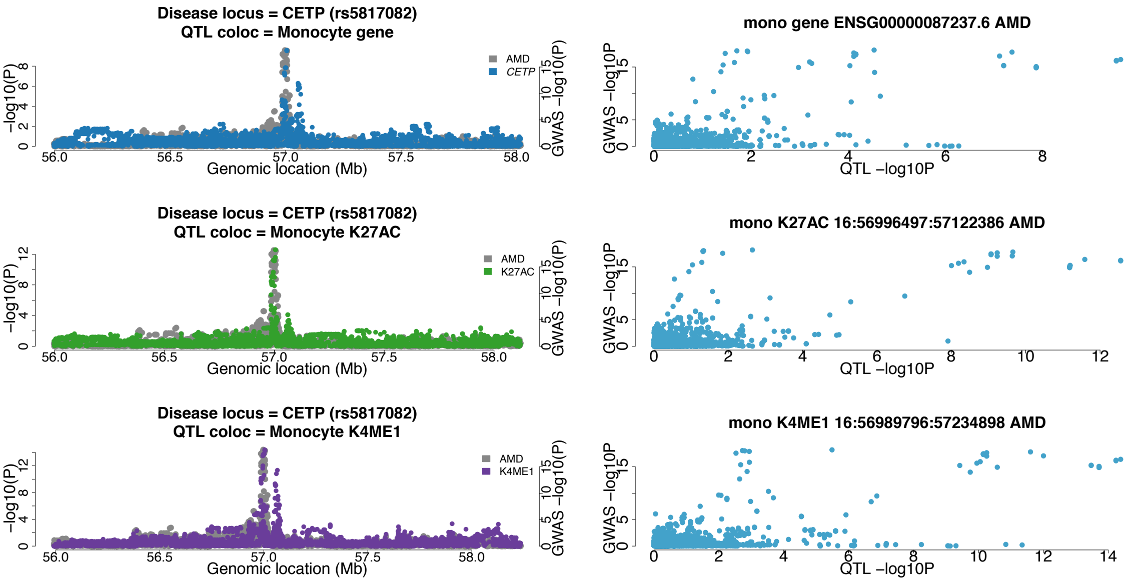
M) Coronary artery disease *MIA3* locus



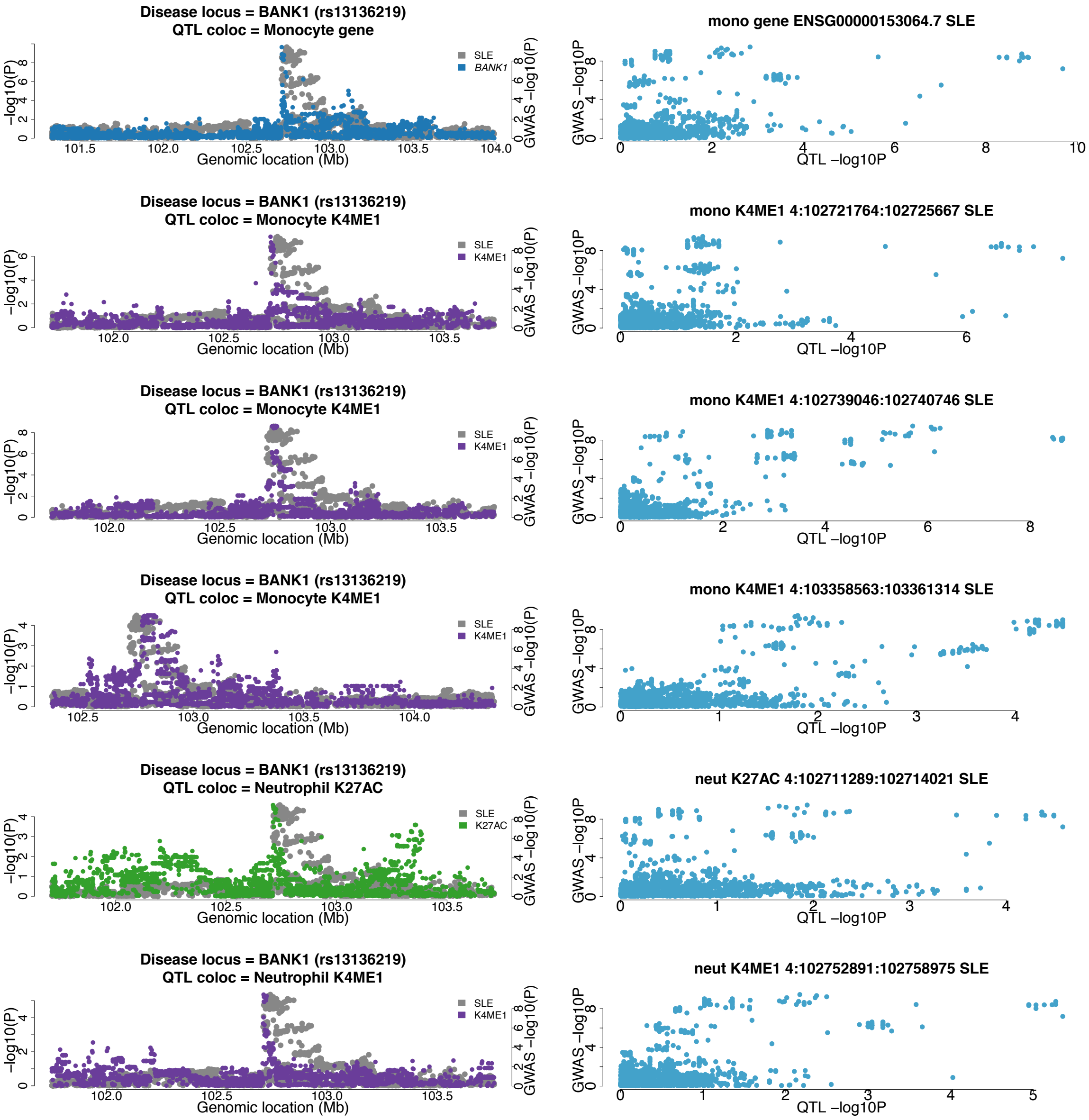
N) Coronary artery disease *PPAP2B* locus



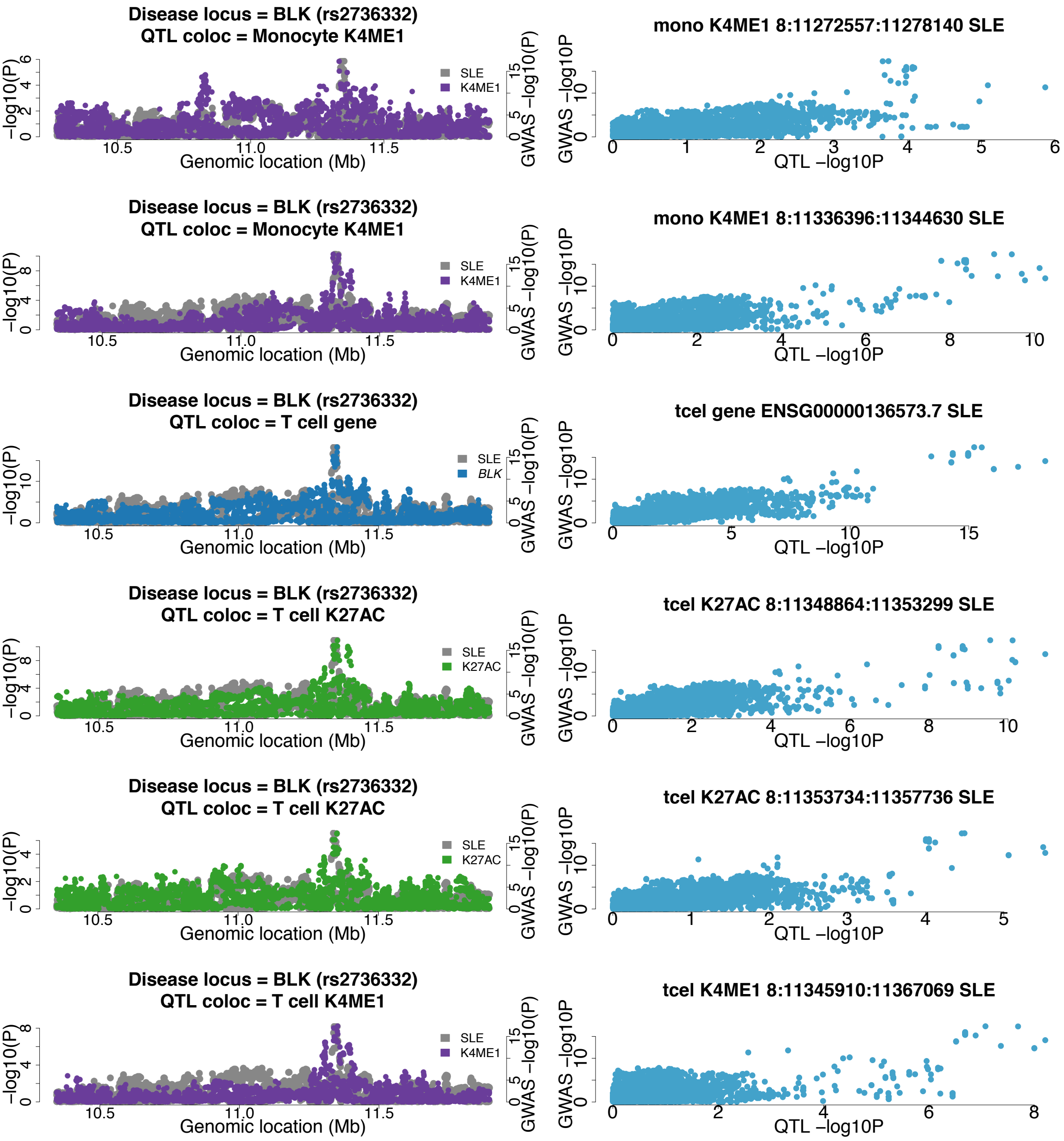
O) Advanced age-related macular degeneration *CETP* locus



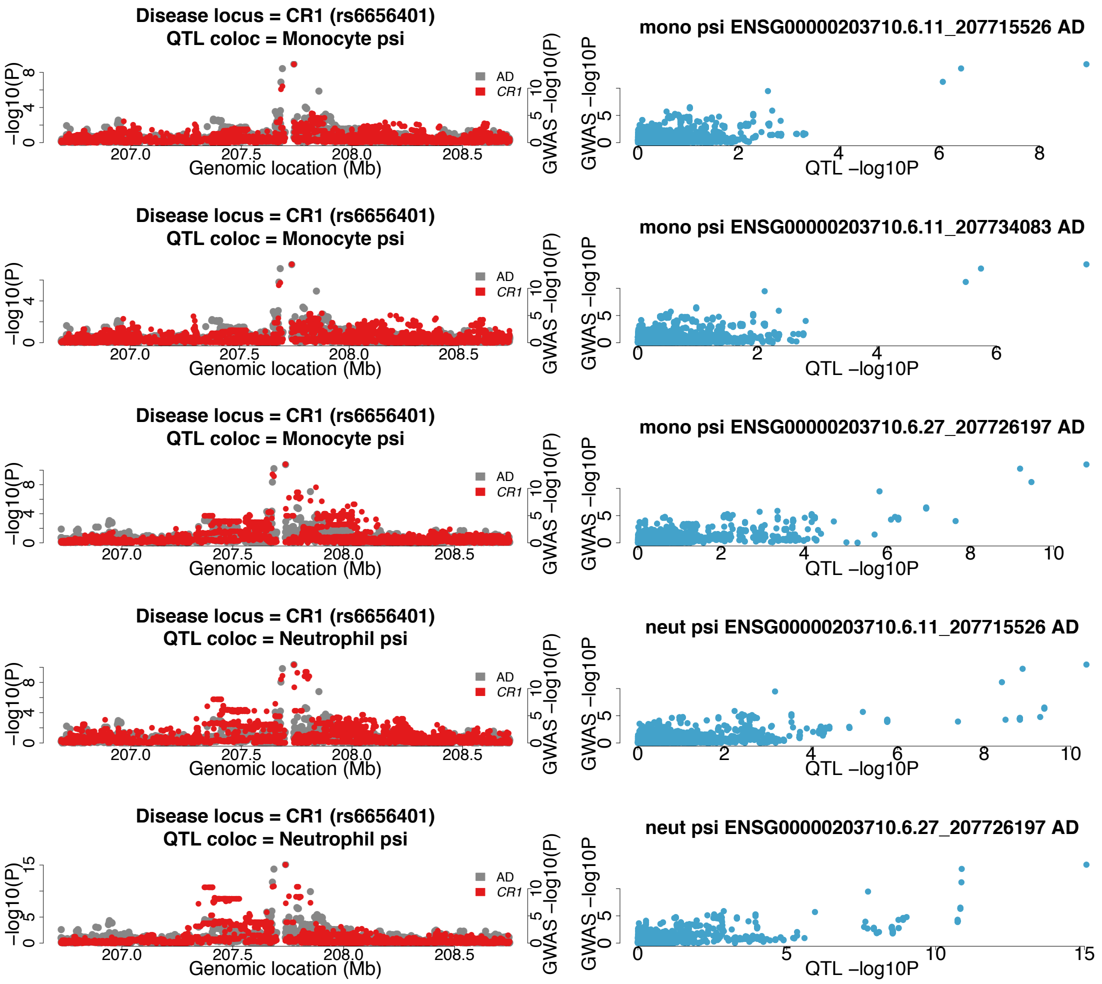
P) Systemic lupus erythematosus *BANK1* locus



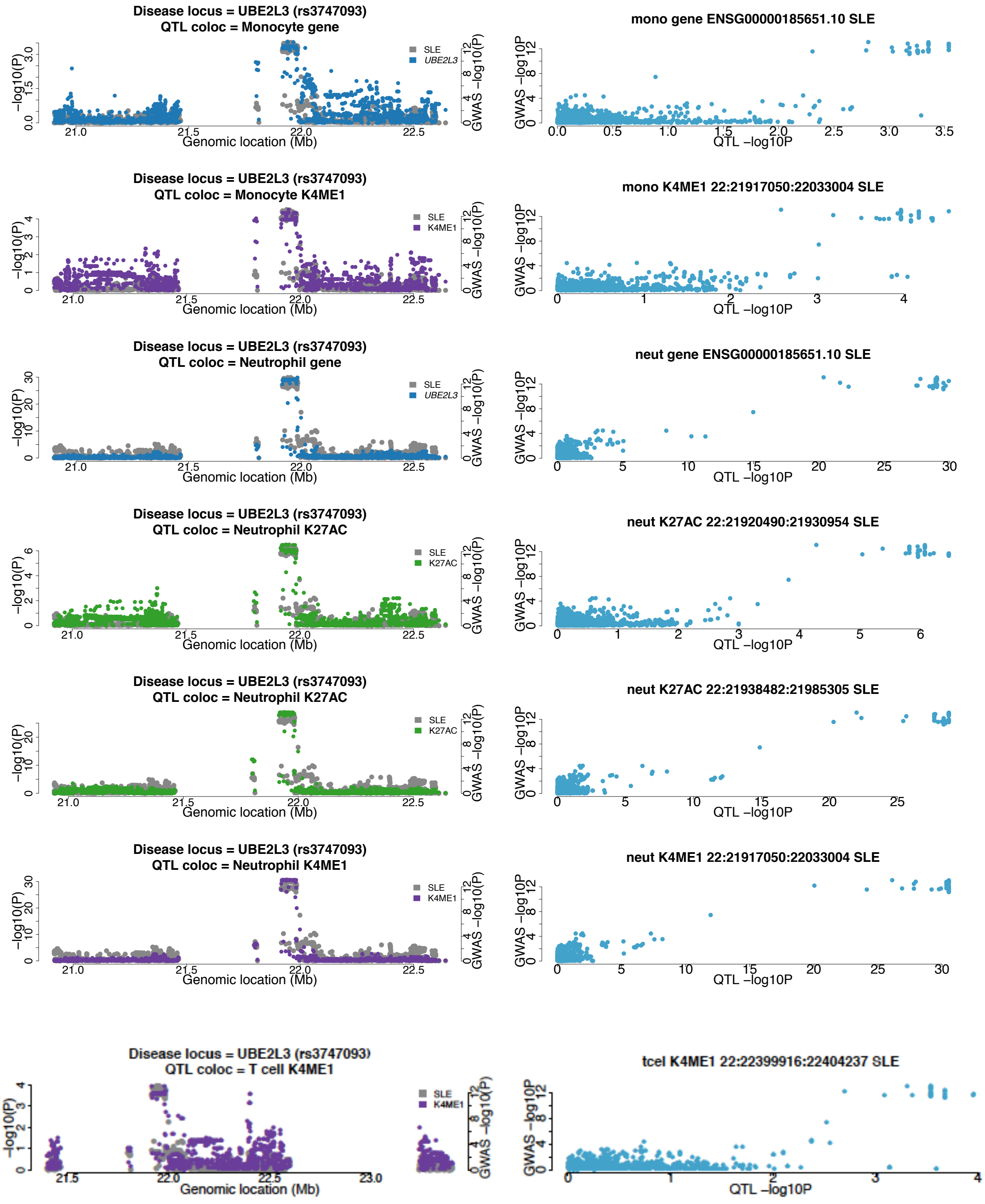
Q) Systemic lupus erythematosus *BLK* locus



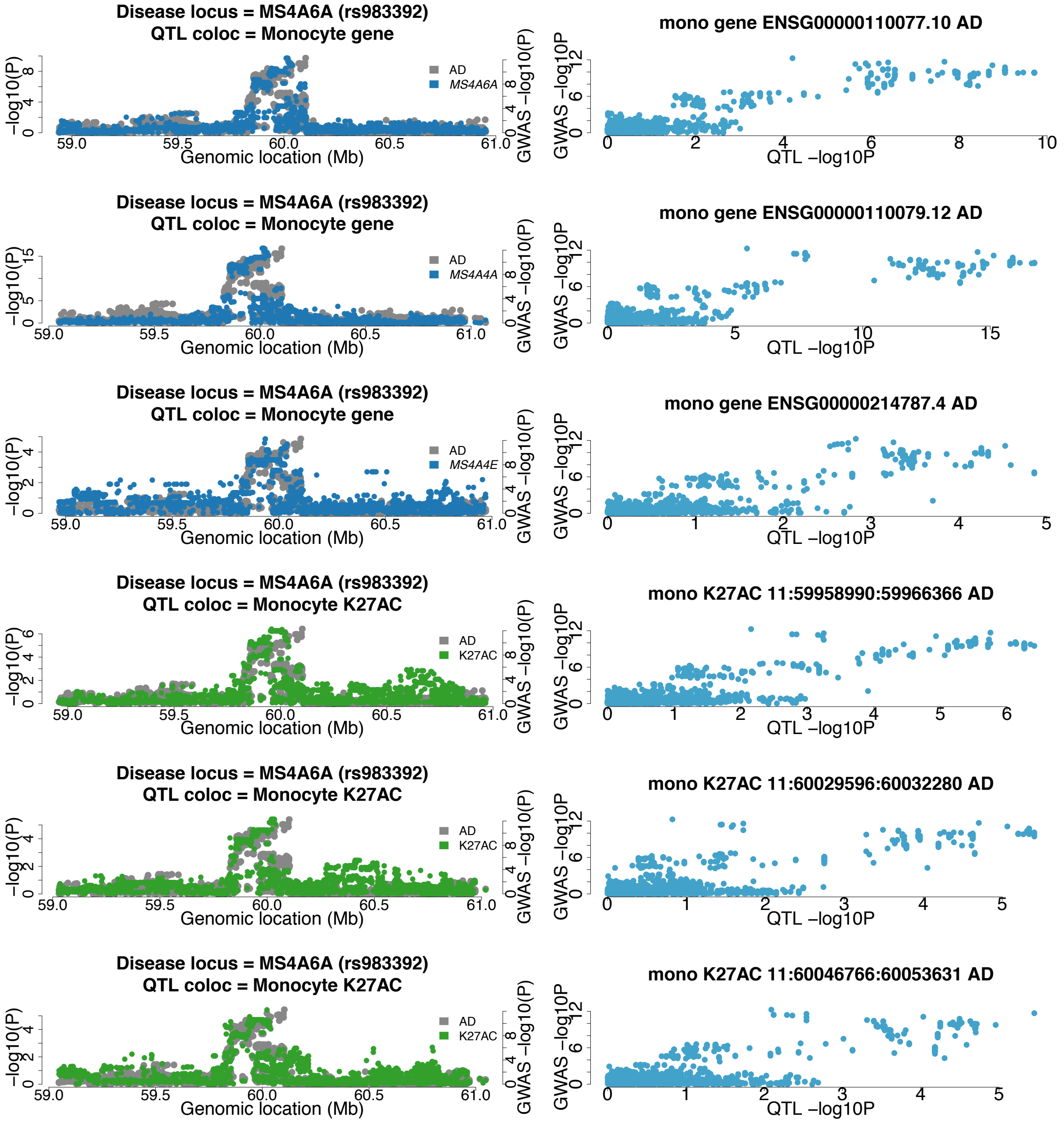
R) Alzheimer’s disease *CR1* locus

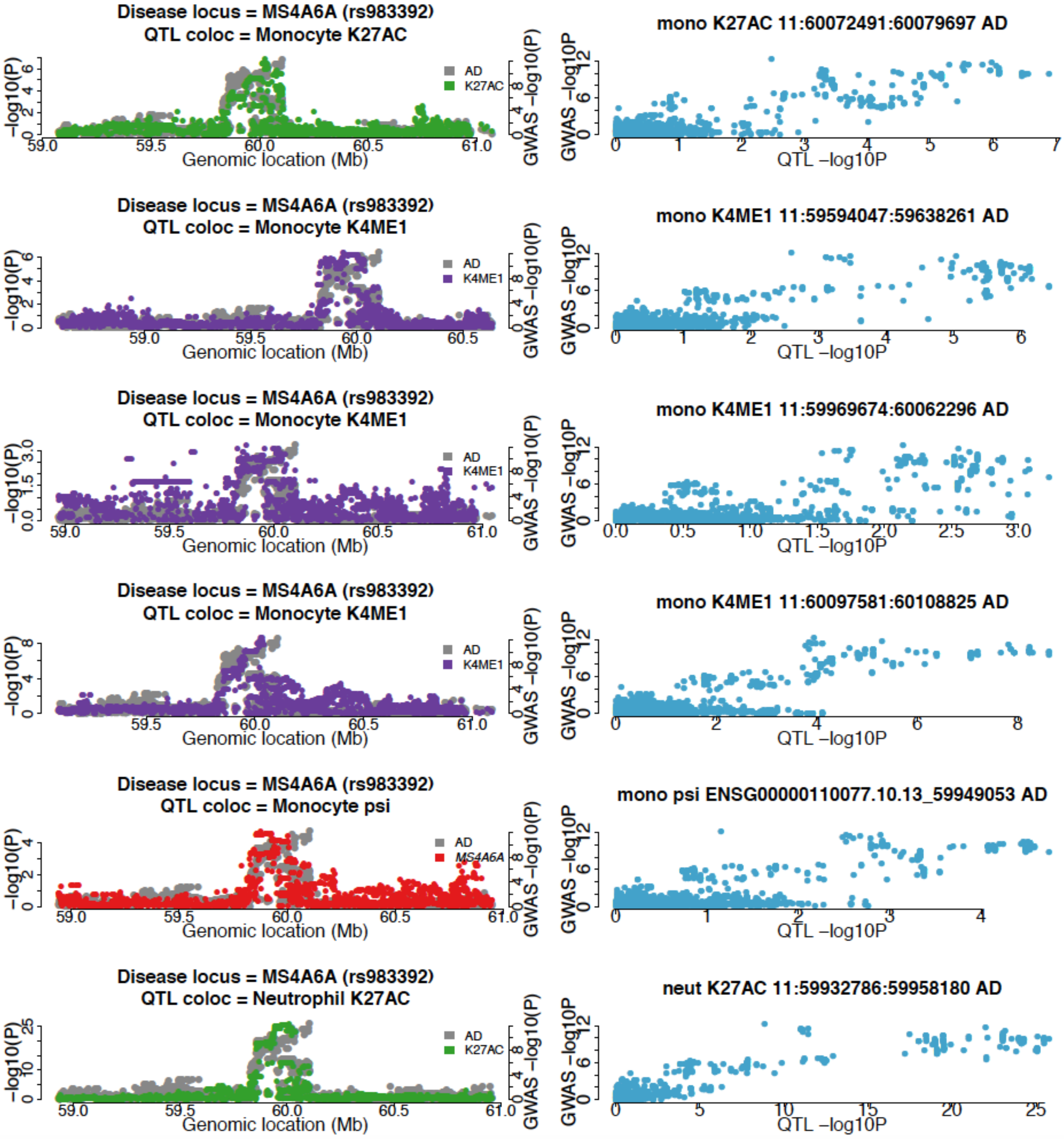


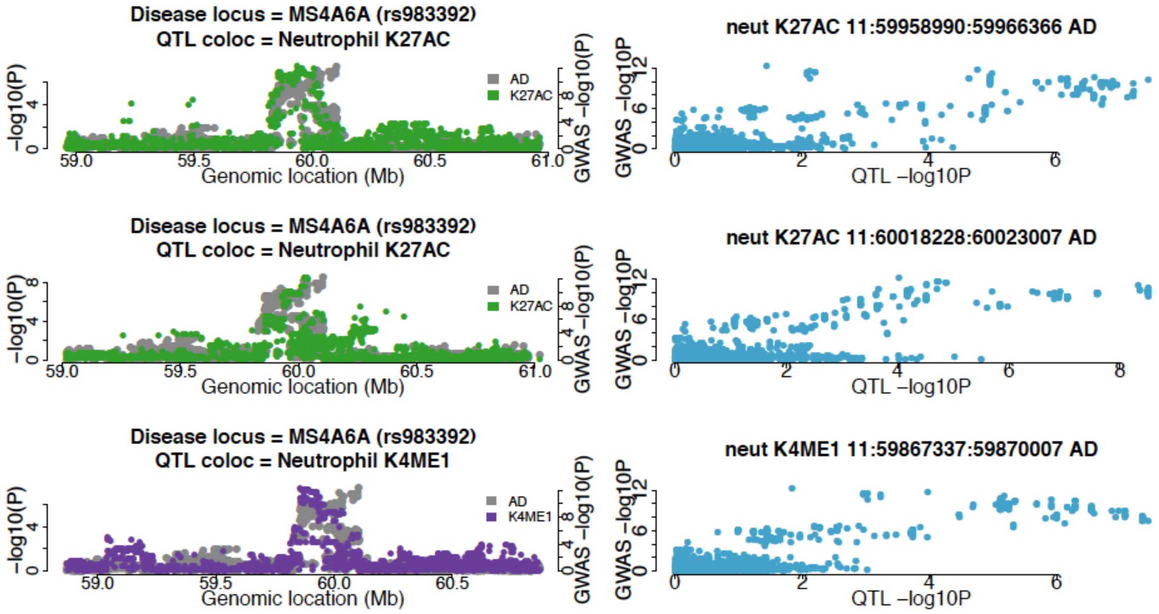
S) Systemic lupus erythematosus *UBE2L3* locus



T) Alzheimer’s disease *MS4A6A* locus




Disease locus = MS4A6A (rs983392)
QTL coloc = Monocyte K27AC
AD
K27AC
−log10(P)
GWAS −log10(P)
Genomic location (Mb)
mono K27AC 11:60072491:60079697 AD
GWAS −log10P
QTL −log10P
Disease locus = MS4A6A (rs983392)
QTL coloc = Monocyte K4ME1
AD
K4ME1
mono K4ME1 11:59594047:59638261 AD
Disease locus = MS4A6A (rs983392)
QTL coloc = Monocyte K4ME1
mono K4ME1 11:59969674:60062296 AD
Disease locus = MS4A6A (rs983392)
QTL coloc = Monocyte K4ME1
mono K4ME1 11:60097581:60108825 AD
Disease locus = MS4A6A (rs983392)
QTL coloc = Monocyte psi
AD
MS4A6A
mono psi ENSG00000110077.10.13_59949053 AD
Disease locus = MS4A6A (rs983392)
QTL coloc = Neutrophil K27AC
AD
K27AC
neut K27AC 11:59932786:59958180 AD


Disease locus = MS4A6A (rs983392)
QTL coloc = Neutrophil K27AC
−log10(P)
0
4
59.0
59.5
60.0
60.5
61.0
Genomic location (Mb)
AD
K27AC
GWAS −log10(P)
0
4
8
neut K27AC 11:59958990:59966366 AD
GWAS −log10P
0
6
12
0
2
4
6
QTL −log10P
Disease locus = MS4A6A (rs983392)
QTL coloc = Neutrophil K27AC
−log10(P)
0
4
8
59.0
59.5
60.0
60.5
61.0
Genomic location (Mb)
AD
K27AC
GWAS −log10(P)
0
4
8
neut K27AC 11:60018228:60023007 AD
GWAS −log10P
0
6
12
0
2
4
6
8
QTL −log10P
Disease locus = MS4A6A (rs983392)
QTL coloc = Neutrophil K4ME1
−log10(P)
0
4
59.0
59.5
60.0
60.5
Genomic location (Mb)
AD
K4ME1
GWAS −log10(P)
0
4
8
neut K4ME1 11:59867337:59870007 AD
GWAS −log10P
0
6
12
0
2
4
6
QTL −log10P

Chapter 2 Supplementary Tables

Feature	SNP (EA/OA) (EAF)	Beta	SE
TNFRSF10A (rs13255394)	Univariate	-0.266 (0.017)	9.725 x 10^{-35}
	Conditional (*RP11-114O23.3* expression)	-0.075 (0.028)	8.638 x 10^{-03}
	Conditional (*RP11-114O23.3* expression + H3K27ac signal)	-0.045 (0.031)	1.355 x 10^{-01}
RP11-114O23.3 (rs13255394)	Univariate	-0.880 (0.040)	5.433 x 10^{-54}
	Conditional (*TNFRSF10A* expression)	-0.576 (0.052)	2.312 x 10^{-22}
	Conditional (*TNFRSF10A* expression + H3K27ac signal)	-0.565 (0.061)	2.325 x 10^{-16}

Supplementary Table 2.1: Conditional causality analysis in the *TNFRSF10A* locus using Blueprint phase 1 genetic data Association results (beta, SE, and p value) from a simple linear regression model. The univariate approach tests for association of the respective gene expression with the genotype of rs13255394 (lead monocyte SNP). Conditional analysis then tests for association of the gene with genotype whilst conditioning on the expression of the alternative gene. The reduction in p value is greatest when testing for association between the SNP and *TNFRSF10A* expression whilst conditioning on *RP11-114O23.3* expression, which suggests the RNA may be causal for variation in expression of *TNFRSF10A*. The further approach conditions on the gene expression and H3K27ac level.

	Disease SNP	Study SNP	LD	Locus	Feature SNP	Cell	Mark	Feature.ID
AD	rs11771145	rs11771145	NA	EPHA1	rs11771145	M	G	ENSG00000185899.1
AD	rs11771145	rs11771145	NA	EPHA1	rs112524998	M	G	ENSG00000229153.1
AD	rs11771145	rs11771145	NA	EPHA1	rs11771145	M	G	ENSG00000221855.1
AD	rs11771145	rs11771145	NA	EPHA1	rs10265814	M	K27	7:143173365:143179545
AD	rs11771145	rs11771145	NA	EPHA1	rs10237465	M	K27	7:143118095:143120257
AD	rs11771145	rs11771145	NA	EPHA1	rs11771145	M	K27	7:143067879:143115356
AD	rs11771145	rs11771145	NA	EPHA1	rs12540656	M	K27	7:143115439:143117746
AD	rs11771145	rs11771145	NA	EPHA1	rs10265814	M	K27	7:143158610:143161683
AD	rs11771145	rs11771145	NA	EPHA1	rs10237465	M	K27	7:143122978:143124643
AD	rs11771145	rs11771145	NA	EPHA1	rs112524998	M	K27	7:143133149:143136359
AD	rs11771145	rs11771145	NA	EPHA1	rs112524998	M	K27	7:143154161:143157641
AD	rs11771145	rs11771145	NA	EPHA1	rs10265814	M	K27	7:143200449:143202615
AD	rs11771145	rs11771145	NA	EPHA1	rs10265814	M	K27	7:143180293:143187266
AD	rs11771145	rs11771145	NA	EPHA1	rs11771145	M	K27	7:143195718:143199833
AD	rs11771145	rs11771145	NA	EPHA1	rs11771145	M	K4	7:143052447:143144656
AD	rs11771145	rs11771145	NA	EPHA1	rs11771145	M	K4	7:143147624:143164247
AD	rs11771145	rs11771145	NA	EPHA1	rs6966814	N	G	ENSG00000229153.1
AD	rs11771145	rs11771145	NA	EPHA1	rs11771145	N	G	ENSG00000185899.1
AD	rs11771145	rs11771145	NA	EPHA1	rs6966814	N	G	ENSG00000234066.1
AD	rs11771145	rs11771145	NA	EPHA1	rs11771145	N	G	ENSG00000221855.1
AD	rs11771145	rs11771145	NA	EPHA1	rs10265814	N	K4	7:143052447:143144656
AD	rs11771145	rs11771145	NA	EPHA1	rs112524998	T	S	7:143104980:143112160:1:3:1, 7:143104984:143112160:1:1:1, 7:143111539:143112160:1:1:0
AD	rs6656401	rs6656401	NA	CR1	rs7515905	M	S	1:207685000:207715526:1:1:1, 1:207713412:207715526:1:1:1
AD	rs6656401	rs6656401	NA	CR1	rs7515905	M	S	1:207726197:207731882:1:1:1, 1:207726197:207748939:1:1:1
AD	rs6656401	rs6656401	NA	CR1	rs7515905	M	S	1:207685000:207734083:1:1:1, 1:207713412:207734083:1:1:1, 1:207731969:207734083:1:1:1
AD	rs6656401	rs6656401	NA	CR1	rs7515905	N	S	1:207726197:207731882:1:1:1, 1:207726197:207748939:1:1:1
AD	rs6656401	rs6656401	NA	CR1	rs7515905	N	S	1:207685000:207715526:1:1:1, 1:207713412:207715526:1:1:1
AD	rs983392	rs983392	NA	MS4A6A	rs611418	M	G	ENSG00000110079.12
AD	rs983392	rs983392	NA	MS4A6A	rs611418	M	G	ENSG00000110077.10
AD	rs983392	rs983392	NA	MS4A6A	rs617135	M	G	ENSG00000214787.4
AD	rs983392	rs983392	NA	MS4A6A	rs4938931	M	K27	11:60029596:60032280
AD	rs983392	rs983392	NA	MS4A6A	rs2081545	M	K27	11:59958990:59966366
AD	rs983392	rs983392	NA	MS4A6A	rs1562990	M	K27	11:60072491:60079697
AD	rs983392	rs983392	NA	MS4A6A	rs1582763	M	K27	11:60046766:60053631
AD	rs983392	rs983392	NA	MS4A6A	rs580064	M	K4	11:59969674:60062296
AD	rs983392	rs983392	NA	MS4A6A	rs4938931	M	K4	11:60097581:60108825
AD	rs983392	rs983392	NA	MS4A6A	rs617135	M	K4	11:59594047:59638261

AD	rs983392	rs983392	NA	MS4A6A	rs4939311	M	S	11:59943085:59949053:2:2:1, 11:59945790:59949053:2:2:1, 11:59947439:59949053:2:2:1
AD	rs983392	rs983392	NA	MS4A6A	rs7107627	N	K27	11:60018228:60023007
AD	rs983392	rs983392	NA	MS4A6A	rs7933202	N	K27	11:59958990:59966366
AD	rs983392	rs983392	NA	MS4A6A	rs1019671	N	K27	11:59932786:59958180
AD	rs983392	rs983392	NA	MS4A6A	rs1441586	N	K4	11:59867337:59870007
AMD	rs10033900	rs10033900	NA	CFI	rs3181191	T	G	ENSG00000248785.1
AMD	rs11080055	rs11080055	NA	VTN	rs2027993	M	G	ENSG00000244045.5
AMD	rs11080055	rs11080055	NA	VTN	rs11080055	N	G	ENSG00000004139.9
AMD	rs11080055	rs11080055	NA	VTN	rs241777	N	G	ENSG00000244045.5
AMD	rs11080055	rs11080055	NA	VTN	rs241777	T	K27	17:27592619:27623928
AMD	rs1142	rs1142	NA	SRPK2	rs12534381	M	G	ENSG00000135250.12
AMD	rs1142	rs1142	NA	SRPK2	rs10263499	M	K27	7:104849291:104858777
AMD	rs1142	rs1142	NA	SRPK2	rs55671517	M	K27	7:104840590:104849222
AMD	rs1142	rs1142	NA	SRPK2	rs2299304	M	K27	7:104578443:104588833
AMD	rs1142	rs1142	NA	SRPK2	rs2385558	M	K4	7:104817678:105045953
AMD	rs1142	rs1142	NA	SRPK2	rs1204061	N	G	ENSG00000135250.12
AMD	rs1142	rs1142	NA	SRPK2	rs2074753	N	K27	7:104982791:105001752
AMD	rs1142	rs1142	NA	SRPK2	rs3823752	N	K27	7:104828844:104835525
AMD	rs140647181	rs140647181	NA	COL8A1	rs6791887	M	G	ENSG00000036054.8
AMD	rs140647181	rs140647181	NA	COL8A1	rs7611566	N	S	3:100029387:100034942:1:1:1, 3:100030722:100034942:1:1:1
AMD	rs140647181	rs140647181	NA	COL8A1	rs6794668	T	K27	3:99927877:99930243
AMD	rs3138141	rs3138141	NA	RDH5	rs3138142	M	G	ENSG00000135437.5
AMD	rs3138141	rs3138141	NA	RDH5	rs3138142	M	G	ENSG00000258311.1
AMD	rs3138141	rs3138141	NA	RDH5	rs3138142	N	G	ENSG00000258311.1
AMD	rs3138141	rs3138141	NA	RDH5	rs3138142	N	S	12:56113008:56113282:1:1:1, 12:56113008:56115472:1:1:1
AMD	rs3138141	rs3138141	NA	RDH5	rs3138142	T	G	ENSG00000135437.5
AMD	rs3138141	rs3138141	NA	RDH5	rs3138142	T	G	ENSG00000258311.1
AMD	rs5817082	rs5817082	NA	CETP	rs7205804	M	G	ENSG00000087237.6
AMD	rs5817082	rs5817082	NA	CETP	rs1532625	M	K27	16:56996497:57122386
AMD	rs5817082	rs5817082	NA	CETP	rs7205804	M	K4	16:56989796:57234898
AMD	rs6565597	rs6565597	NA	NPLOC4	rs112612275	M	K27	17:79585940:79590489
AMD	rs6565597	rs6565597	NA	NPLOC4	rs35816741	M	K4	17:79594768:79608902
AMD	rs6565597	rs6565597	NA	NPLOC4	rs67050149	M	S	17:79526443:79532530:2:2:1, 17:79531006:79532530:2:2:1
AMD	rs6565597	rs6565597	NA	NPLOC4	rs112612275	N	K27	17:79585940:79590489
AMD	rs6565597	rs6565597	NA	NPLOC4	rs11655377	T	K27	17:79596351:79605193
AMD	rs6565597	rs6565597	NA	NPLOC4	rs11150803	T	K27	17:79619627:79621546
AMD	rs6565597	rs6565597	NA	NPLOC4	rs35816741	T	K27	17:79622135:79624634
AMD	rs6565597	rs6565597	NA	NPLOC4	rs11150803	T	K27	17:79578768:79583302
AMD	rs6565597	rs6565597	NA	NPLOC4	rs35816741	T	K27	17:80244653:80295172
AMD	rs6565597	rs6565597	NA	NPLOC4	rs9912071	T	K4	17:79594768:79608902

AMD	rs6565597	rs6565597	NA	NPLOC4	rs8070929	T	S	17:79526443:79532530:2:2:1, 17:79531006:79532530:2:2:1
AMD	rs6565597	rs6565597	NA	NPLOC4	rs8070929	T	S	17:79526443:79530946:2:2:1, 17:79526443:79532530:2:2:1
AMD	rs6565597	rs6565597	NA	NPLOC4	rs8070929	T	S	17:79526443:79527746:2:2:1, 17:79526443:79530946:2:2:1, 17:79526443:79532530:2:2:1
AMD	rs67538026	rs67538026	NA	CNN2	rs10419072	M	G	ENSG00000064666.9
AMD	rs67538026	rs67538026	NA	CNN2	rs62131196	M	K27	19:1024681:1033920
AMD	rs67538026	rs67538026	NA	CNN2	rs10417845	N	G	ENSG00000261526.1
AMD	rs67538026	rs67538026	NA	CNN2	rs62131196	T	G	ENSG00000064687.8
AMD	rs67538026	rs67538026	NA	CNN2	rs62131196	T	K27	19:1024681:1033920
AMD	rs7803454	rs7803454	NA	PILRB/A	rs111312383	M	G	ENSG00000121716.12
AMD	rs7803454	rs7803454	NA	PILRB/A	rs111312383	M	G	ENSG00000066923.12
AMD	rs7803454	rs7803454	NA	PILRB/A	rs7803454	M	G	ENSG00000146834.8
AMD	rs7803454	rs7803454	NA	PILRB/A	rs111312383	M	K4	7:99906073:99912427
AMD	rs7803454	rs7803454	NA	PILRB/A	rs113387325	M	K4	7:99801401:99888538
AMD	rs7803454	rs7803454	NA	PILRB/A	rs111312383	M	S	7:99954456:99955842:0:0:1, 7:99954507:99955842:1:1:1
AMD	rs7803454	rs7803454	NA	PILRB/A	rs111312383	M	S	7:99943592:99947339:1:1:1, 7:99943592:99947353:1:1:1, 7:99943592:99947421:1:1:1
AMD	rs7803454	rs7803454	NA	PILRB/A	rs111312383	M	S	7:99954456:99955842:0:0:1, 7:99954507:99955842:1:1:1, 7:99954562:99955842:1:1:1
AMD	rs7803454	rs7803454	NA	PILRB/A	rs111312383	M	S	7:99933844:99935801:1:1:1, 7:99933961:99935801:1:1:1
AMD	rs7803454	rs7803454	NA	PILRB/A	rs111312383	M	S	7:99943592:99947339:1:1:1, 7:99943592:99947421:1:1:1
AMD	rs7803454	rs7803454	NA	PILRB/A	rs111312383	M	S	7:99943592:99947339:1:1:1, 7:99943592:99947353:1:1:1
AMD	rs7803454	rs7803454	NA	PILRB/A	rs67471932	N	G	ENSG00000233389.2
AMD	rs7803454	rs7803454	NA	PILRB/A	rs111312383	N	G	ENSG00000066923.12
AMD	rs7803454	rs7803454	NA	PILRB/A	rs7787825	N	G	ENSG00000146834.8
AMD	rs7803454	rs7803454	NA	PILRB/A	rs111312383	N	K4	7:99906073:99912427
AMD	rs7803454	rs7803454	NA	PILRB/A	rs111312383	N	K4	7:99801401:99888538
AMD	rs7803454	rs7803454	NA	PILRB/A	rs67471932	N	K4	7:99912739:99917535
AMD	rs7803454	rs7803454	NA	PILRB/A	rs111312383	N	S	7:99954456:99955842:0:0:1, 7:99954507:99955842:1:1:1, 7:99954562:99955842:1:1:1
AMD	rs7803454	rs7803454	NA	PILRB/A	rs111312383	N	S	7:99933844:99935801:1:1:1, 7:99933961:99935801:1:1:1
AMD	rs7803454	rs7803454	NA	PILRB/A	rs111312383	N	S	7:99954456:99955842:0:0:1, 7:99954507:99955842:1:1:1
AMD	rs7803454	rs7803454	NA	PILRB/A	rs11769886	N	S	7:99943592:99947339:1:1:1, 7:99943592:99947421:1:1:1
AMD	rs7803454	rs7803454	NA	PILRB/A	rs11769886	N	S	7:99943592:99947339:1:1:1, 7:99943592:99947353:1:1:1, 7:99943592:99947421:1:1:1
AMD	rs7803454	rs7803454	NA	PILRB/A	rs11769886	N	S	7:99943592:99947339:1:1:1, 7:99943592:99947353:1:1:1
AMD	rs7803454	rs7803454	NA	PILRB/A	rs111312383	T	G	ENSG00000085514.10
AMD	rs7803454	rs7803454	NA	PILRB/A	rs111312383	T	G	ENSG00000078487.13
AMD	rs7803454	rs7803454	NA	PILRB/A	rs113387325	T	G	ENSG00000241357.1
AMD	rs7803454	rs7803454	NA	PILRB/A	rs113387325	T	G	ENSG00000121716.12
AMD	rs7803454	rs7803454	NA	PILRB/A	rs7787825	T	G	ENSG00000146834.8

AMD	rs7803454	rs7803454	NA	PILRB/A	rs111312383	T	S	7:99951107:99951517:1:1:1, 7:99951107:99952765:1:1:1
AMD	rs7803454	rs7803454	NA	PILRB/A	rs113387325	T	S	7:99933844:99935801:1:1:1, 7:99933961:99935801:1:1:1
AMD	rs7803454	rs7803454	NA	PILRB/A	rs113387325	T	S	7:99947511:99949785:1:1:0, 7:99949034:99949785:1:1:1
AMD	rs7803454	rs7803454	NA	PILRB/A	rs111312383	T	S	7:99943592:99947339:1:1:1, 7:99943592:99947353:1:1:1
AMD	rs7803454	rs7803454	NA	PILRB/A	rs113387325	T	S	7:99947511:99948874:1:1:1, 7:99947511:99949785:1:1:0
AMD	rs7803454	rs7803454	NA	PILRB/A	rs111312383	T	S	7:99943592:99947339:1:1:1, 7:99943592:99947421:1:1:1
AMD	rs7803454	rs7803454	NA	PILRB/A	rs111312383	T	S	7:99943592:99947339:1:1:1, 7:99943592:99947353:1:1:1, 7:99943592:99947421:1:1:1
AMD	rs79037040	rs79037040	NA	TNFRSF10A	rs13255394	M	G	ENSG00000246582.2
AMD	rs79037040	rs79037040	NA	TNFRSF10A	rs13255394	M	G	ENSG00000147457.9
AMD	rs79037040	rs79037040	NA	TNFRSF10A	rs13255394	M	G	ENSG00000104689.5
AMD	rs79037040	rs79037040	NA	TNFRSF10A	rs13255394	M	K27	8:23048166:23092260
AMD	rs79037040	rs79037040	NA	TNFRSF10A	rs13255394	M	K27	8:23092704:23132254
AMD	rs79037040	rs79037040	NA	TNFRSF10A	rs13255394	M	K4	8:22998146:23133613
CAD	chr2:203828796:I	chr2:203828796:I	NA	WDR12	rs148513392	M	G	ENSG00000163596.12
CAD	chr2:203828796:I	chr2:203828796:I	NA	WDR12	rs78488377	M	G	ENSG00000144426.14
CAD	chr2:203828796:I	chr2:203828796:I	NA	WDR12	rs140201293	M	G	ENSG00000138380.12
CAD	chr2:203828796:I	chr2:203828796:I	NA	WDR12	rs16839813	M	K4	2:204391511:204403180
CAD	chr2:203828796:I	chr2:203828796:I	NA	WDR12	rs4675290	N	K27	2:204364327:204367436
CAD	chr2:203828796:I	chr2:203828796:I	NA	WDR12	rs72934535	N	K27	2:203772046:203778875
CAD	chr2:203828796:I	chr2:203828796:I	NA	WDR12	rs72932553	T	G	ENSG00000236047.1
CAD	rs11191416	rs11191416	NA	NT5C2	rs79780963	M	S	10:104934740:104940941:2:2:1, 10:104934740:104940987:2:2:0
CAD	rs11191416	rs11191416	NA	NT5C2	rs111374294	N	G	ENSG00000237827.1
CAD	rs11191416	rs11191416	NA	NT5C2	rs111374294	N	G	ENSG00000076685.14
CAD	rs11191416	rs11191416	NA	NT5C2	rs3740390	T	K27	10:104811999:104815290
CAD	rs11191416	rs11191416	NA	NT5C2	rs79780963	T	S	10:104934740:104936242:2:2:0, 10:104934740:104937274:2:2:0, 10:104934740:104940941:2:2:1, 10:104934740:104940987:2:2:0, 10:104934740:104952992:2:2:1

CAD	rs11191416	rs11191416	NA	NT5C2	rs79780963	T	S	10:104934740:104940941:2:2:1, 10:104934740:104940987:2:2:0, 10:104934740:104952992:2:2:1
CAD	rs1412444	rs1412444	NA	LIPA	rs1412444	M	G	ENSG00000107798.12
CAD	rs1412444	rs1412444	NA	LIPA	rs1320496	M	K27	10:90993615:91006217
CAD	rs1412444	rs1412444	NA	LIPA	rs1412444	M	K27	10:90976768:90986051
CAD	rs1412444	rs1412444	NA	LIPA	rs1412444	M	K27	10:91010098:91013357
CAD	rs1412444	rs1412444	NA	LIPA	rs1320496	M	K27	10:91013477:91017078
CAD	rs1412444	rs1412444	NA	LIPA	rs1332326	M	K4	10:91050031:91073249
CAD	rs1412444	rs1412444	NA	LIPA	rs1332326	M	K4	10:91306722:91311404
CAD	rs1412444	rs1412444	NA	LIPA	rs1412444	M	K4	10:91131855:91136597
CAD	rs1412444	rs1412444	NA	LIPA	rs1320496	M	K4	10:90987967:91024823
CAD	rs1412444	rs1412444	NA	LIPA	rs1320496	M	K4	10:91042400:91044298
CAD	rs1412444	rs1412444	NA	LIPA	rs1332326	M	K4	10:90962723:90987091
CAD	rs1412444	rs1412444	NA	LIPA	rs1412444	M	K4	10:91027656:91032000
CAD	rs1412444	rs1412444	NA	LIPA	rs1332328	N	K27	10:90993615:91006217
CAD	rs1412444	rs1412444	NA	LIPA	rs1412444	N	K4	10:90987967:91024823
CAD	rs1412444	rs1412444	NA	LIPA	rs1332327	T	G	ENSG00000107798.12
CAD	rs1412444	rs1412444	NA	LIPA	rs1320496	T	K27	10:90248309:90252291
CAD	rs17087335	rs17087335	NA	REST	rs12645070	M	G	ENSG00000084093.11
CAD	rs17087335	rs17087335	NA	REST	rs12645070	M	K27	4:57771837:57788561
CAD	rs17087335	rs17087335	NA	REST	rs12645070	M	K27	4:57823529:57826313
CAD	rs17087335	rs17087335	NA	REST	rs12645070	M	K4	4:57770788:57806908
CAD	rs17087335	rs17087335	NA	REST	rs12645070	M	K4	4:57820927:57828891
CAD	rs17087335	rs17087335	NA	REST	rs12645070	N	K27	4:57823529:57826313
CAD	rs17087335	rs17087335	NA	REST	rs6554401	N	K4	4:57820927:57828891
CAD	rs6689306	rs6689306	NA	IL6R	rs7549250	N	K27	1:154372031:154419908
CAD	rs6689306	rs6689306	NA	IL6R	rs7549338	N	K4	1:153669232:153673360
CAD	rs6689306	rs6689306	NA	IL6R	rs11265611	N	K4	1:154342399:154479953
CAD	rs6689306	rs6689306	NA	IL6R	rs4845625	T	S	1:154422457:154437609:1:1:1, 1:154427058:154437609:1:1:1
CAD	rs6689306	rs6689306	NA	IL6R	rs4845625	T	S	1:154422457:154426963:1:1:1, 1:154422457:154437609:1:1:1
CAD	rs6689306	rs6689306	NA	IL6R	rs4845625	T	S	1:154422457:154426548:1:1:0, 1:154422457:154426963:1:1:1, 1:154422457:154437609:1:1:1
CAD	rs67180937	rs67180937	NA	MIA3	rs35700460	T	K4	1:222943024:222949002
CAD	rs67180937	rs67180937	NA	MIA3	rs35700460	T	S	1:222828136:222831263:1:1:0, 1:222828136:222832063:1:1:1
CAD	rs7528419	rs7528419	NA	SORT1	rs12740374	M	G	ENSG00000134222.11
CAD	rs7528419	rs7528419	NA	SORT1	rs660240	M	K27	1:109812607:109818851
CAD	rs7528419	rs7528419	NA	SORT1	rs12740374	M	K4	1:109779241:109861456
CAD	rs7528419	rs7528419	NA	SORT1	rs12740374	N	G	ENSG00000134222.11
CAD	rs7528419	rs7528419	NA	SORT1	rs660240	N	K27	1:109812607:109818851
CAD	rs7528419	rs7528419	NA	SORT1	rs660240	N	K4	1:109779241:109861456
CAD	rs7528419	rs7528419	NA	SORT1	rs660240	T	K27	1:109109862:109115257

CAD	rs7568458	rs7568458	NA	GGCX	rs12714145	M	G	ENSG00000115486.6
CAD	rs7568458	rs7568458	NA	GGCX	rs1078004	N	G	ENSG00000115486.6
CAD	rs7568458	rs7568458	NA	GGCX	rs56819945	N	K27	2:85760296:85771243
CAD	rs7568458	rs7568458	NA	GGCX	rs10176176	N	K4	2:85523177:85561159
CAD	rs7568458	rs7568458	NA	GGCX	rs11891260	T	G	ENSG00000115486.6
CAD	rs7568458	rs7568458	NA	GGCX	rs1561198	T	G	ENSG00000118640.6
CAD	rs9970807	rs9970807	NA	PPAP2B	rs56186267	M	G	ENSG00000162407.8
CAD	rs9970807	rs9970807	NA	PPAP2B	rs56186267	M	K27	1:56969801:56978117
CAD	rs9970807	rs9970807	NA	PPAP2B	rs6588634	N	K27	1:56931078:56933449
FEV1	rs7218675	rs7218675	NA	TSEN54	rs35643020	M	G	ENSG00000182173.8
FEV1	rs78420228	rs78420228	NA	CDC123	rs12241367	M	K27	10:12310277:12315701
FEV1	rs78420228	rs78420228	NA	CDC123	rs61848342	M	K27	10:12374604:12376790
FEV1	rs78420228	rs78420228	NA	CDC123	rs11599700	M	K4	10:12273289:12320006
FEV1	rs78420228	rs78420228	NA	CDC123	rs10795944	N	G	ENSG00000228302.2
FEV1	rs7842022	rs78420228	NA	CDC123	rs12241367	N	K4	10:12273289:12320006
SLE	rs1143679	rs34572943	0.94	ITGAM	rs34082782	M	G	ENSG00000140688.12
SLE	rs1143679	rs34572943	0.94	ITGAM	rs34550882	N	G	ENSG00000261385.1
SLE	rs1143679	rs34572943	0.94	ITGAM	rs34550882	N	G	ENSG00000260219.1
SLE	rs1143679	rs34572943	0.94	ITGAM	rs9673398	N	G	ENSG00000169896.11
SLE	rs1143679	rs34572943	0.94	ITGAM	rs9673404	T	K4	16:31355247:31421179
SLE	rs13136219	rs10028805	0.98	BANK1	rs7683892	M	G	ENSG00000153064.7
SLE	rs13136219	rs10028805	0.98	BANK1	rs34029191	M	K4	4:102739046:102740746
SLE	rs13136219	rs10028805	0.98	BANK1	rs4270588	M	K4	4:103358563:103361314
SLE	rs13136219	rs10028805	0.98	BANK1	rs7683892	M	K4	4:102721764:102725667
SLE	rs13136219	rs10028805	0.98	BANK1	rs7683892	N	K27	4:102711289:102714021
SLE	rs13136219	rs10028805	0.98	BANK1	rs7683892	N	K4	4:102752891:102758975
SLE	rs2736332	rs2736340	0.9	BLK	rs13257831	M	K4	8:11336396:11344630
SLE	rs2736332	rs2736340	0.9	BLK	rs12680762	M	K4	8:11272557:11278140
SLE	rs2736332	rs2736340	0.9	BLK	rs922483	T	G	ENSG00000136573.7
SLE	rs2736332	rs2736340	0.9	BLK	rs2736345	T	K27	8:11353734:11357736
SLE	rs2736332	rs2736340	0.9	BLK	rs922483	T	K27	8:11348864:11353299
SLE	rs2736332	rs2736340	0.9	BLK	rs922483	T	K4	8:11345910:11367069
SLE	rs35251378	rs2304256	0.95	TYK2	rs11085725	N	G	ENSG00000105397.9
SLE	rs35251378	rs2304256	0.95	TYK2	rs280497	T	G	ENSG00000105397.9
SLE	rs3747093	rs7444	0.88	UBE2L3	rs5749485	M	G	ENSG00000185651.10
SLE	rs3747093	rs7444	0.88	UBE2L3	rs2070512	M	K4	22:21917050:22033004
SLE	rs3747093	rs7444	0.88	UBE2L3	rs2298429	N	G	ENSG00000185651.10
SLE	rs3747093	rs7444	0.88	UBE2L3	rs11089620	N	K27	22:21920490:21930954
SLE	rs3747093	rs7444	0.88	UBE2L3	rs140488	N	K27	22:21938482:21985305
SLE	rs3747093	rs7444	0.88	UBE2L3	rs140488	N	K4	22:21917050:22033004
SLE	rs3747093	rs7444	0.88	UBE2L3	rs5754102	T	K4	22:22399916:22404237

SLE	rs58688157	rs12802200	0.48	IRF7	rs7120313	M	G	ENSG00000070047.6
SLE	rs58688157	rs12802200	0.48	IRF7	rs2246614	M	G	ENSG00000161328.10
SLE	rs58688157	rs12802200	0.48	IRF7	rs11246177	M	G	ENSG00000185507.14
SLE	rs58688157	rs12802200	0.48	IRF7	rs936472	M	G	ENSG00000185522.4
SLE	rs58688157	rs12802200	0.48	IRF7	rs936472	M	K27	11:600961:621989
SLE	rs58688157	rs12802200	0.48	IRF7	rs386614207	M	K4	11:601613:633623
SLE	rs58688157	rs12802200	0.48	IRF7	rs2246614	N	K27	11:600961:621989
SLE	rs58688157	rs12802200	0.48	IRF7	rs11246217	N	S	11:614038:614173:2:2:1, 11:614038:614475:2:2:1
SLE	rs58688157	rs12802200	0.48	IRF7	rs12290989	N	S	11:614038:614475:2:2:1, 11:614400:614475:2:2:1
SLE	rs58688157	rs12802200	0.48	IRF7	rs4963128	T	G	ENSG00000185507.14
SLE	rs58688157	rs12802200	0.48	IRF7	rs12419618	T	K27	11:600961:621989
SLE	rs58688157	rs12802200	0.48	IRF7	rs12803048	T	K4	11:601613:633623
SLE	rs6671847	rs1801274	0.89	FCGR2A	rs4657041	M	S	1:161512990:161595934:2:2:0, 1:161594430:161595934:2:2:1
SLE	rs6671847	rs1801274	0.89	FCGR2A	rs12129787	N	S	1:161487928:161489591:1:1:0, 1:161488906:161489591:1:1:0, 1:161489451:161489591:1:1:1

Supplementary Table 2.2: Summary of all features colocalised with disease loci that colocalised with at least one gene or splicing QTL
All disease loci that colocalised with at least one gene or splicing QTL is summarised here. Loci colocalised with only histone features are not listed. The Disease SNP columns describes the SNP assigned from the GWAS summary statistics (Materials and Methods). The Study SNP is the disease lead SNP listed in the study, and the LD is the 1000G LD between the Study SNP and Disease SNP if these differ. Differences may occur if the summary statistics available were a subset of a full study meta-analysis. The Study SNP was used to assign the locus, designated in the study. The Feature SNP is the molecular feature lead SNP. Cell is the cell-type of the corresponding feature (M = monocyte, N = neutrophil, T = T cell). Mark describes the feature type (G = Gene, S = splicing/PSI, K4 = H3K4me1, K27ac, H3K27ac). The feature ID describes the Ensembl gene ID or the histone signal peak defined as chr:start:end. The splicing ID describes the splicing junctions defined by Chen *et al* (2016a).

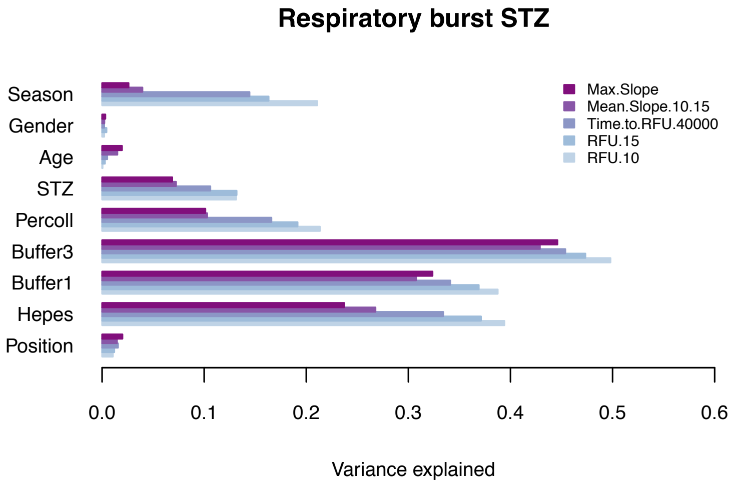
BLUEPRINT summary statistics for lead QTLs can be found here: http://blueprint-dev.bioinfo.cnio.es/WP10/
Lead study SNP summaries can be found here: AMD PMID: 26691988, AD PMID: 24162737, CAD PMID: 26343387, FEV_1 PMID: 26423011, SLE PMID: 26502338

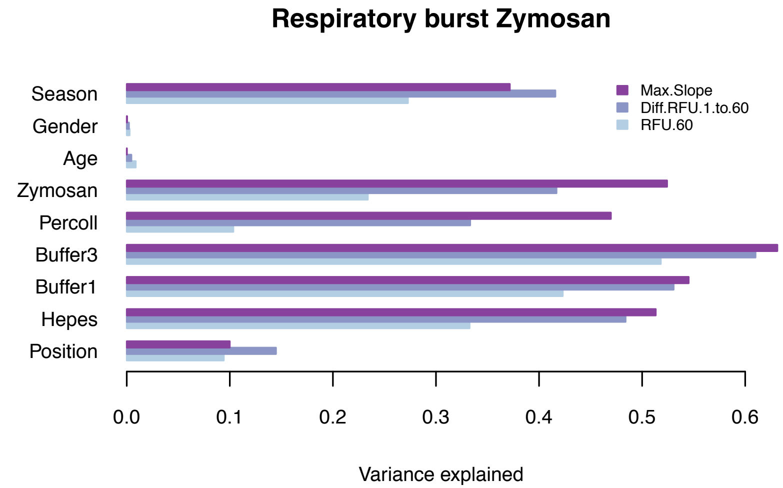
Freely downloadable Summary statistics: Alzheimer’s Disease, Lambert *et al* (2013): http://web.pasteur-lille.fr/en/recherche/u744/igap/igap_download.php

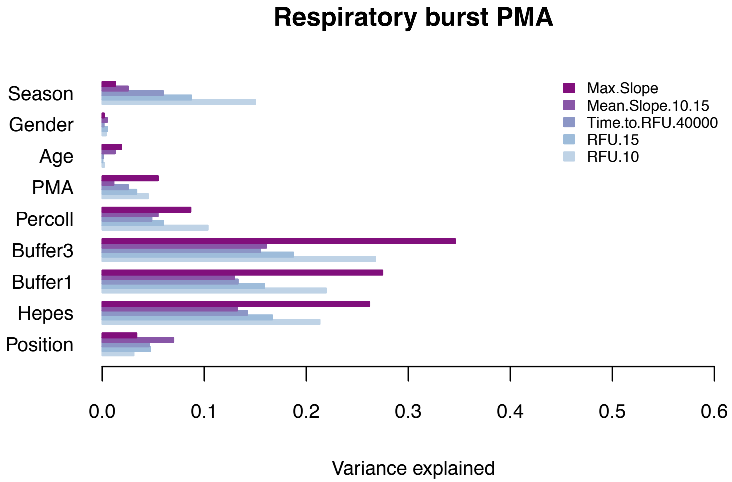
Coronary artery disease, Nikpay *et al.* (2015): http://www.cardiogramplusc4d.org/data-downloads/ Systemic lupus erythematosus, Bentham *et al.* (2015):

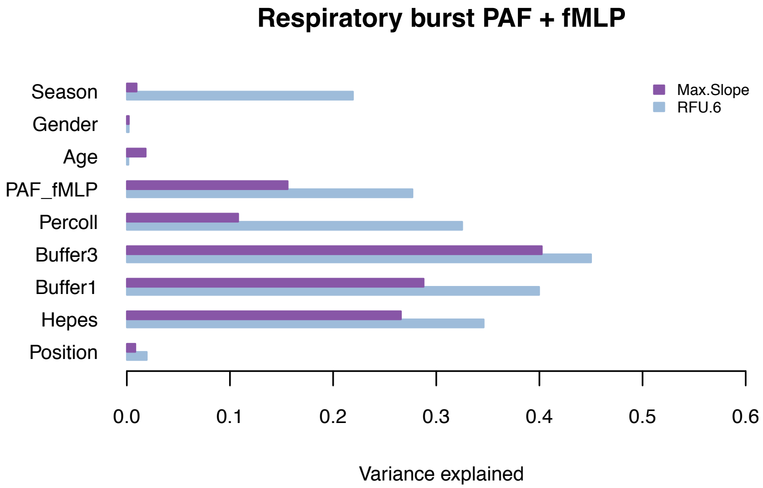
https://www.immunobase.org/downloads/protected_data/GWAS_Data/

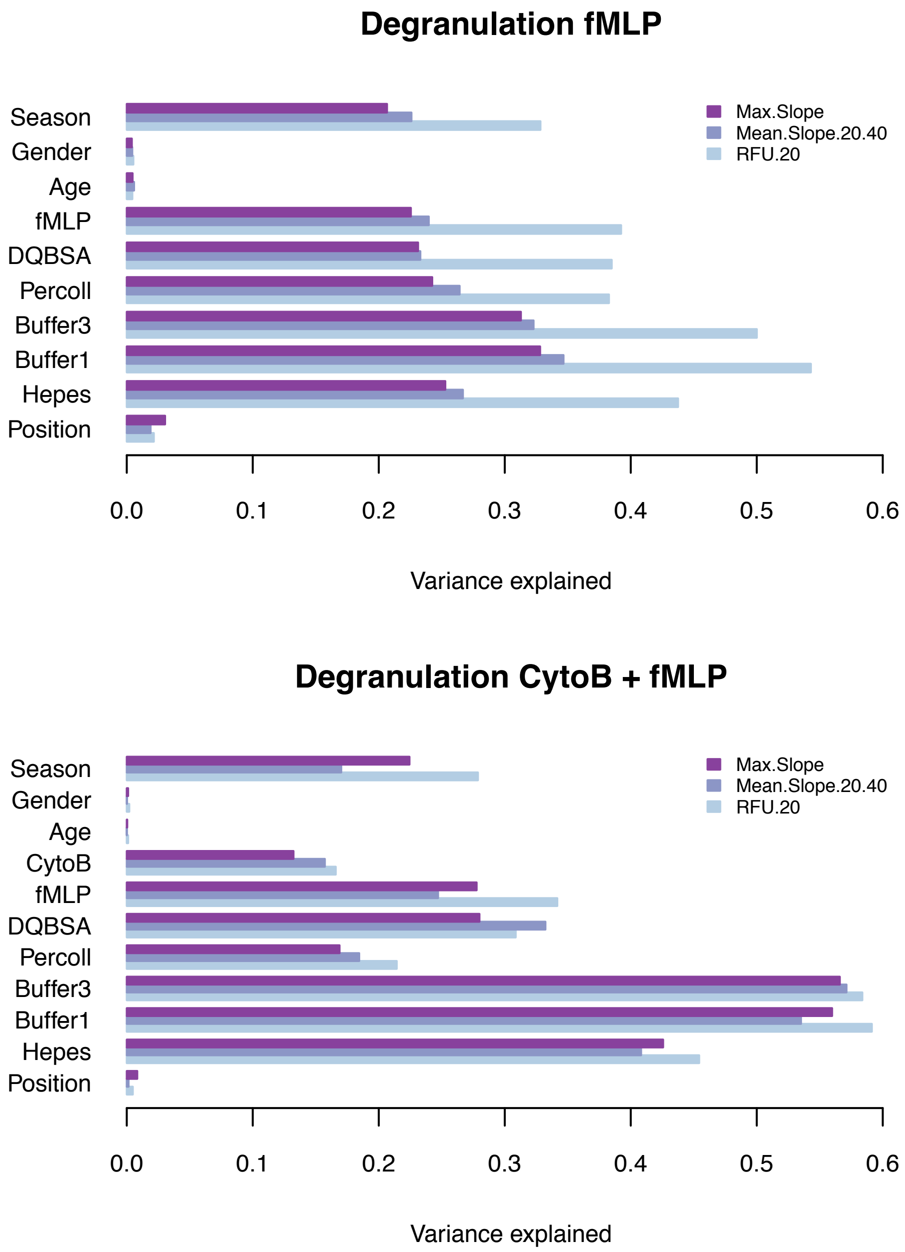
Chapter 3 Supplementary information


Respiratory burst STZ

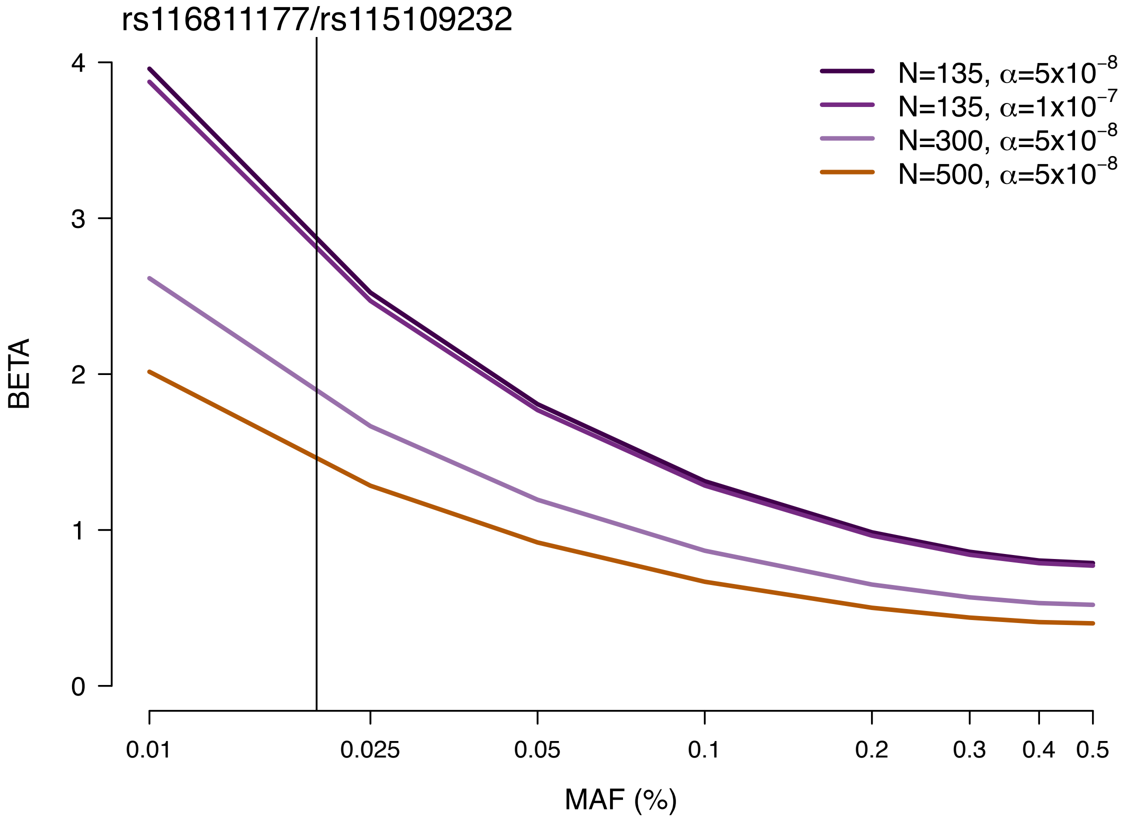

Respiratory burst Zymosan


Respiratory burst PMA


Respiratory burst PAF + fMLP

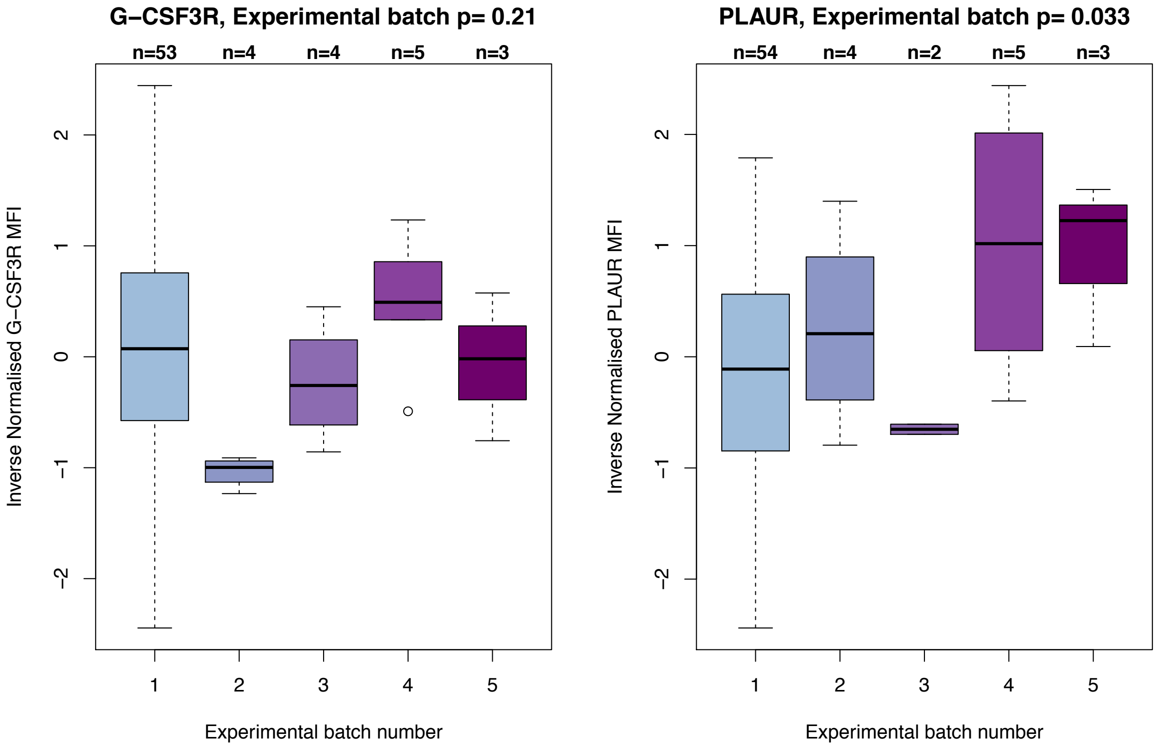


Supplementary Figure 3.1: Contribution of known sources of co-variation to respiratory burst and degranulation neutrophil responses
Barplots show R^2 values from fitting linear models of each trait with each covariate independently.



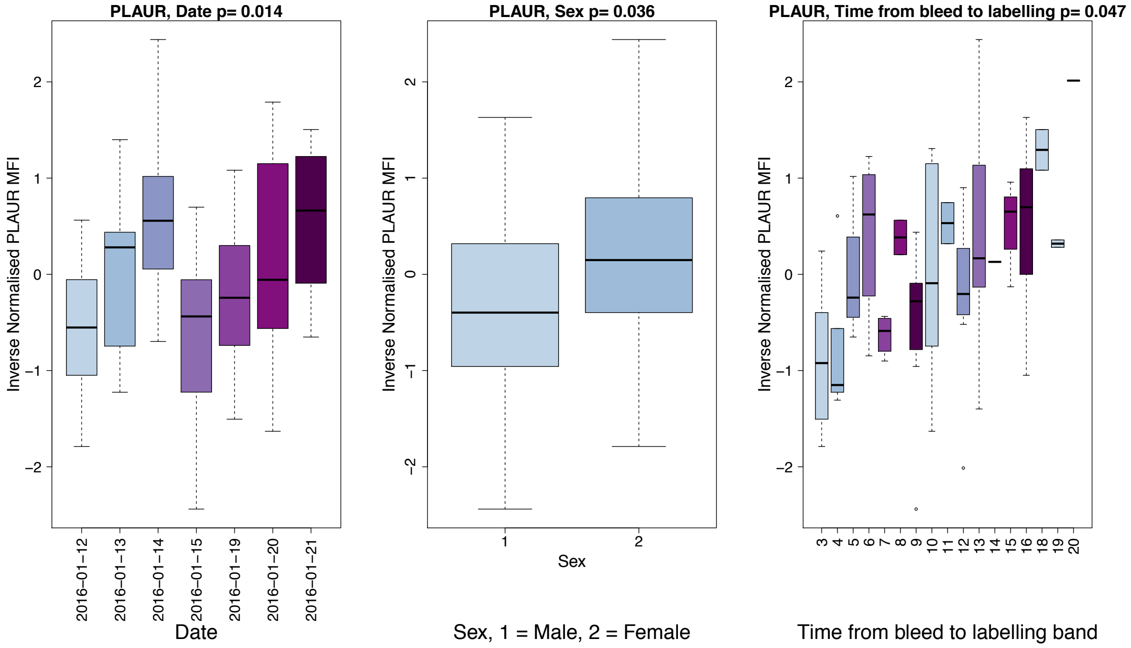
Supplementary Figure 3.2: Predicted power of the neutrophil function study
A power calculation was performed for a study of 135 individuals (approximately equal to this function cohort) and for increasing sample sizes up to 500 individuals. Variants will be likely detected by the study if the variant falls above the line which describes the relationship between beta (effect size in standard deviation of the trait value) and minor allele frequency. This study is powered to detect common variants for moderate effect size or low frequency variants with higher effect sizes of beta greater than 2. The power for the suggestive p value threshold of 1 x 10^{-07} for N = 135 is also shown. The corresponding frequency of the genome-wide significant locus (rs116811177/rs115109232) identified in this study is marked. The beta for the association for both SNPs from this study was 2.92, which is close to the predicted power of this study to detect SNPs of this low-frequency. The pwr package was used in this power calculation (Champely, 2012).

Chapter 4 Supplementary Information



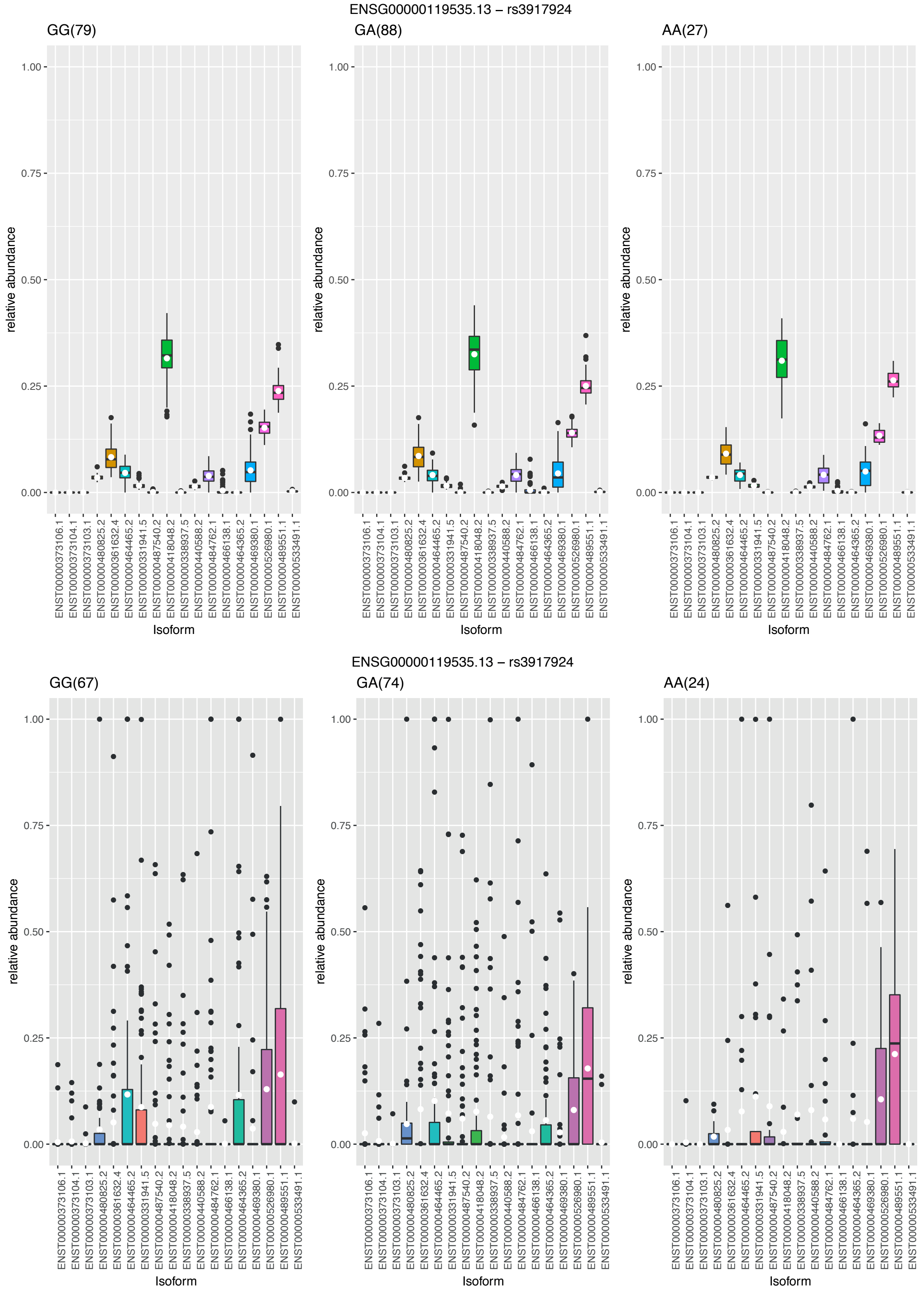
Supplementary Figure 4.1: Receptor MFI values stratified by experimental processing batch

Normalised GCSFR and PLAUR MFI stratified by experimental processing batch with outliers removed from both datasets. ANOVA testing with inverse normalised trait values found this covariate to be significant for PLAUR MFI only.



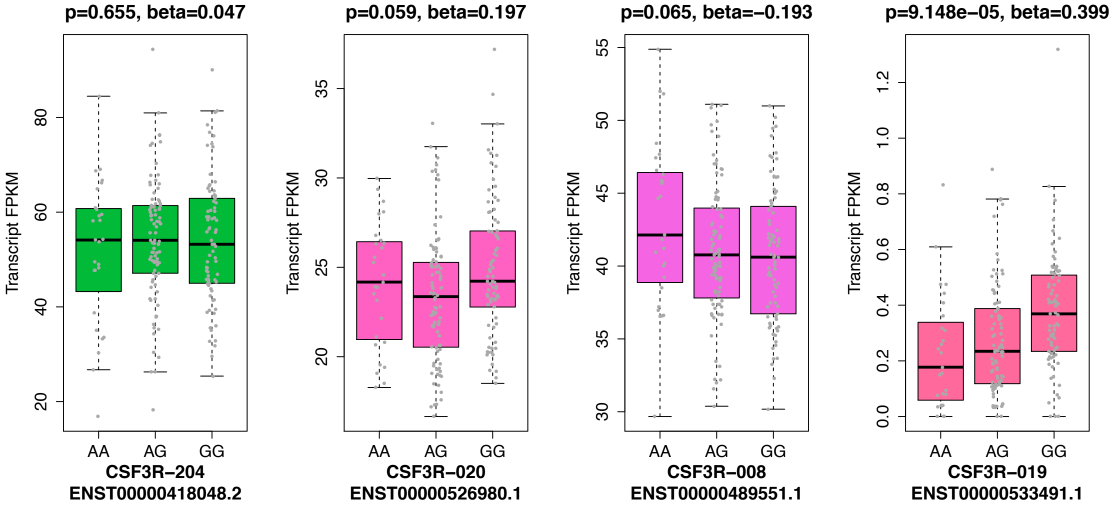
Supplementary Figure 4.2: Covariate effects on MFI values for PLAUR receptor

Normalised MFI values for the PLAUR receptor, excluding outliers, are shown stratified by date, sex and time from bleed to labelling, which were all found to be significant covariates using ANOVA. The p value for each covariate from the ANOVA testing is listed in the corresponding plots.



Supplementary Figure 4.3: Relative abundance of *CSF3R* transcripts in T cells and monocytes stratified by rs3917924 genotypes

This abundance by genotype plot shows all 18 *CSF3R* transcripts (Gencode v15) for T cells (top panel) and monocytes (bottom panel). rs3917924 is common GCSFR surface expression index SNP). The receptor-encoding transcript (green) is most abundantly expressed in monocytes but not highly expressed in T cells.



Supplementary Figure 4.4: *CSF3R* transcript expression level in monocytes
The three most abundant transcripts (left to right) including the short truncated transcript, CSF3R-019 (far right) are shown with the transcript expression level (FPKM) in monocytes stratified by the genotype of rs3917924. The p value and beta are shown after using a linear regression to test association of transcript level with genotype. Although the expression level of CSF-020 is significant (beta = 0.483, SE = 0.151, p value 0.002), the effect does not look additive. CSF3R-019 transcript expression is significantly associated with rs3917924 genotype (beta = 0.48, SE = 0.147, p value = 0.001) but the effect is much smaller than observed with neutrophils (Figure 4.12). The transcript expression levels of all four transcripts are lower than in neutrophils.

SNP	Proteins bound	Motifs altered	Histone	CEBPB/PU1	ChromHMM	HL60 TF	Differentiated HL60
rs3917912		AP-2_disc2, ELF1_disc2, Ets_known3	K27ac (1:36928496: 36957321) K4me1 (1: 36778894: 36993055)	PU1 (1: 36947299: 36949245)	Active TSS high H3K4me3+K27Ac (1: 36947600: 36948600)	P300 (1: 36947664: 36948535) CEBPE (1: 36947667: 36949236), H3K4me1, H3K4me3, H3K27ac	H3K27ac DMSO, H3K4me3 ATRA (1:36946936: 36948821), H3K27ac ATRA, H3K4me1 DMSO, H3K4me3 ATRA
rs3917914			K27ac (1:36928496: 36957321) K4me1 (1: 36778894: 36993055)	PU1 (1: 36947299: 36949245)	Active TSS high H3K4me3+K27Ac (1: 36947600: 36948600)	P300 (1: 36947664: 36948535) CEBPE (1: 36947667: 36949236), H3K4me1, H3K4me3, H3K27ac	H3K27ac DMSO, H3K4me3 ATRA (1:36946936: 36948821), H3K27ac ATRA, H3K4me1 DMSO, H3K4me3 ATRA
rs3917924	CEBPB, JUND, P300, TBP, BAF170 (Hela)	BCL_disc5, NRSF_disc3	K27ac (1:36928496: 36957321) K4me1 (1: 36778894: 36993055)	PU1 (1: 36944990: 36945853) CEBPB (1: 36944999: 36945946)	Active TSS High H3K4me3 (E10) (1: 36945600: 36945800)	CEBPE (1: 36944824: 36946074), H3K4me1, H3K4me3	H3K4me1 ATRA, H3K4me1 DMSO
rs3917932		AP-1_disc1, ATF3_disc1, ATF3_known1, ATF3_known10, ATF3_known9, ATF4, ATF6, E2F_disc1, HEY1_disc1, LXR_2, Maf_disc2, Mxi1_known1, Myc_disc1, SIRT6_disc1, SREBP_known4, T3R, TATA_disc2	K27ac (1:36928496: 36957321) K4me1 (1: 36778894: 36993055)		Active Enh (K27ac, K4ME1,E9) (1: 36943800: 36944000)	H3K4me1	

Supplementary Table 4.1: Intersection of *CSF3R* significant variants with epigenomic and molecular data

Common and rare lead SNPs for GCSFR MFI and neutrophil count are listed and annotations/ First those using HaploReg v4.1 including predicted disrupted motifs (Keradpour and Kellis, 2014). Column four onwards are epigenome-SNP overlaps (bedtools intersect). The data used was for primary neutrophils from the BLUEPRINT consortium (Chen et al, 2016, Carrillo-de-Santa-Pau et al., 2017) as well as data from Stephen Watt (manuscript in preparation). The histone peaks used for intersection were those as tested for QTL in the blueprint consortium using a 1Mb window. The final two columns list epigenome-SNP overlaps of ChIP-seq data generated in the neutrophil model cell line, HL60 and the differentiated more-mature HL60 model (ATRA and DMSO) (Chapter 2).

SNP	EA/OA	PU1 1Mb QTL	Exon	Splicing Junction	Allele-specific eQTL (WASP-ASE)
rs3917924	G/A	1:36924227:36924948, 1.518 x 10^{-06}, -0.6867 (peak falls 3' outside of CSF3R gene)	ENSG00000119535.13.46, 1.091 x 10^{-37}, 1.104 ENSG00000119535.13.41, 4.353 x10^{-04}, 0.3637	1:36945118:36945587:2:2:1 4.487 x 10^{-58}, 1.245 1:36945682:36947078:2:2:1 6.407 x 10^{-55}, 1.231 1:36943279:36945033:2:2:1 2.056 x 10^{-09}, 0.600 1:36941275:36943205:2:2:0 5.667 x 10^{-06}, 0.463	ENSG00000119535.13 5.587 x 10^{-33}, 0.103
rs3917931	C/T	1:36924227:36924948, 1.518 x 10^{-06}, -0.6867	ENSG00000119535.13.46, 1.483 x 10^{-38}, 1.122 ENSG00000119535.13.41, 2.871 x 10^{-04}, 0.378	1:36945118:36945587:2:2:1 1.687 x 10^{-59}, 1.264 1:36945682:36947078:2:2:1 9.940 x 10^{-56}, 1.247 1:36943279:36945033:2:2:1 7.018 x $10^{-10,}$ 0.622 1:36941275:36943205:2:2:0 4.664 x $10^{-06,}$ 0.471	ENSG00000119535.13 7.191 x 10^{-32}, 0.100
rs3917925	G/A	1:36924227:36924948, 1.518 x 10^{-06}, -0.6867	ENSG00000119535.13.46, 1.483 x 10^{-38}, 1.122 ENSG00000119535.13.41, 2.871 x 10^{-04}, 0.378	1:36945118:36945587:2:2:1 1.687 x 10^{-59}, 1.264 1:36945682:36947078:2:2:1 9.940 x 10^{-56}, 1.247 1:36943279:36945033:2:2:1 7.018 x 10^{-10}, 0.622 1:36941275:36943205:2:2:0 4.664 x 10^{-06}, 0.471	ENSG00000119535.13 7.191 x 10^{-32}, 0.100
rs6667127	C/T	1:36924227:36924948, 1.690 x 10^{-05}, -0.6704	ENSG00000119535.13.46, 3.067 x 10^{-22}, 0.969	1:36945118:36945587:2:2:1 1.407 x 10^{-32}, 1.119 1:36945682:36947078:2:2:1 2.881 x 10^{-30}, 1.092 1:36943279:36945033:2:2:1 4.997 x 10^{-06}, 0.500 1:36941275:36943205:2:2:0 1.349 x 10^{-04}, 0.420	ENSG00000119535.13 1.095 x 10^{-22}, 0.090
rs955115	C/A	1:36924227:36924948, 5.247 x 10^{-05}, -0.6151	ENSG00000119535.13.46, 7.116 x 10^{-08}, 0.5853	1:36945682:36947078:2:2:1 1.519 x 10^{-10}, 0.683 1:36945118:36945587:2:2:1 6.471 x 10^{-09}, 0.623 1:36935442:36937033:2:2:1 1.899 x 10^{-04}, 0.415 1:36943279:36945033:2:2:1 7.702 x 10^{-04}, 0.374	ENSG00000119535.13 3.852 x 10^{-09}, 0.068
rs3917933	G/A	1:36924227:36924948, 1.518 x 10^{-06}, -0.6867	ENSG00000119535.13.46, 1.483 x 10^{-38}, 1.122 ENSG00000119535.13.41, 2.871 x 10^{-04}, 0.378	1:36945118:36945587:2:2:1 1.687 x 10^{-59}, 1.264 1:36945682:36947078:2:2:1 9.940 x 10^{-56}, 1.247 1:36943279:36945033:2:2:1 7.018 x 10^{-10}, 0.622 1:36941275:36943205:2:2:0 4.664 x 10^{-06}, 0.471	ENSG00000119535.13 7.191 x 10^{-32}, 0.100

Supplementary Table 4.2: Summary of significant QTL associations within the GCSFR locus that were tested in the BLUEPRINT consortium
Significant QTL associations with the common SNPs from the BLUEPRINT consortium (Chen et al. 2016). Features with a qvalue of 5% or less are listed. For comparison, the uncorrected raw p value of association is listed for each feature. The phenotype for PU1 is the binding of this transcription factor assayed using ChIP-seq in primary neutrophils in a subset of the same individuals as the main Blueprint consortium (Stephen Watt, manuscript in preparation). 1Mb refers to the window size around the feature for the QTLs tested. Exon and splicing junction QTLs were identified using RNA-seq data and processed in the same way as gene expression QTLs from Chen *et al.* (2016), but the reads for exon QTLs are summed over each individual exon in a gene and over each splicing junction for the latter QTLs. These were not published as part of the Chen *et al.* (2016) paper. Allele-specific eQTLs were published as part of the main study and analysed using the WASP software.