Investigation of Expression Quantitative Trait Loci and Regulatory Genetic Variants in Primary Human Immune Cells

Seiko Makino

St Cross College
University of Oxford
Trinity Term 2013

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy at the University of Oxford

Abstract

The post human genome sequence era has begun to explore various aspects of the functional genome in relation to disease including gene expression, genetic variation and epigenetics. The genetic determinants of common and complex phenotypes are difficult to resolve even though their heritability is recognised. Recent genome-wide association studies (GWAS) for common diseases has identified many new disease susceptibility associated loci. These loci often lie in non-coding regions of the genome and disease associated genetic variants are proposed to act by modulating gene expression.

This thesis investigated genetic variation as determinants of gene expression in the context of the immune system especially focused on the innate immune and inflammatory responses. Different primary human immune cell types were collected from healthy volunteers of European ancestry to achieve this. In order to identify genetic variants associating with gene expression, expression quantitative trait loci (eQTL) mapping was conducted in a cell type specific manner.

The primary dataset (n=288) consists of $CD19^+$ B-cells from the adaptive immune system and $CD14^+$ monocytes from the innate immune system. 78% of the total cis eQTL were found to be cell type specific and include genes relating to their roles in the immune response. Trans eQTL showed greater cell type specificity and include master regulatory eQTL on the *LYZ* locus at chromosome 12q15 in monocytes and the *KLF4* (9p31) in B-cells. The identified eQTL are implicated in association with autoimmune disease susceptibility including inflammatory bowel disease, diabetes and rheumatoid arthritis.

The second analysed dataset (n=64) consists of $CD14^+$ monocytes and macrophages differentiated *ex vivo*. Macrophages are involved in many inflammatory diseases as well as in the innate immune response. The differential gene expression and eQTL mapping analyses were conducted to investigate macrophages specific gene expression signatures and associations to genetic variants. Macrophage eQTL are involved in signal transduction for the inflammatory response (*IL1RN* and *STAT4*) and lipid metabolism (*PPARG*) with implication for metabolic disease association.

The eQTL analyses using primary immune cell types provide insights into genetic variation in association to gene expression which is involved in autoimmunity and disease susceptibility.

Acknowledgements

I would like to thank my supervisors Dr Julian Knight and Dr Benjamin Fairfax for constant support and encouragement as well as guidance throughout my study. I would especially like to thank Dr Fairfax for initiating and working together for the volunteer sample collection which is reflected in the largest part of my thesis. I would also like to thank Dr Vivek Naranbhai for supporting the second volunteer sample collection, the macrophages microarrays and insightful advice, Dr Jayachandran Radhakrishnan and Dr Peter Humburg for supporting me for the bioinformatics and statistical aspect of this thesis.

I would like to thank the Genomics Service in the Wellcome Sanger Institute especially Dr Chris McGee for prompt support for microarray data processing, Dr Anna Rautanen for guidance for genotyping data analysis, Dr Stephen Leslie and Dr Alexander Dilthey for HLA imputation, Dr Helen Lockstone and Dr Natasha Sahgal for guidance for microarray data analysis, Dr John Broxholme for guidance for basic computational data analysis. I would also like to thank my transfer examiners Dr Stephen Chapman and Dr Cecilia Lindgren for useful advice in the transfer viva.

I would like to thank all my colleagues, Dr. Katharine Plant, Miss Emma Davenport, Dr Tara Mills, Miss Evelyn Lau, Dr Esther Ng, Dr Daniel Wong and Dr Wanseon Lee for support. I would also like to thank Mrs Phyllis Fairfax and Miss Ruth Porter for their friendship and friends at OUCKC and Falcon for being nice company on the rivers.

Finally I would like to thank my friends and family even though Oxford was far away from wherever you live today.

Declarations

I declare that all work presented in this thesis is my own unless stated otherwise. Several aspects of this work relied upon collaboration where part of the work was conducted by or with others.

The healthy volunteer cohort study was designed by Dr Benjamin Fairfax and Dr Julian Knight at the Wellcome Trust Centre for Human Genetics (WTCHG), University of Oxford. I was integral to the study, participating in volunteer recruitment, sample collection, processing of nucleic acids, database management, data preprocessing and analysis throughout which I predominantly worked with Dr Fairfax. Dr Katharine Plant and Dr Jayachandran Radhakrishnan additionally helped with aspects of sample collection and data analysis respectively. The second cohort collection which included the samples for the macrophage expression study was conducted by myself, Dr Vivek Naranbhai and Dr Fairfax.

Microarray gene expression profiling on the human primary cells was performed by the Genomics Service in the Wellcome Sanger Institute, Cambridge and the Oxford Genomics Centre in the WTCHG, Oxford. Gene expression probes were BLAST aligned on the human genome with help from Dr John Broxholme in the Bioinformatics core in the WTCHG. Gene expression probes with SNPs were identified by Dr Adam Auton in the Bioinformatics core in the WTCHG. Advice regarding microarray data analysis was provided by Dr Helen Lockstone and Dr Natasha Sahgal in the Bioinformatics core in the WTCHG. Genotyping was performed by the Oxford Genomics Centre in the WTCHG. Guidance regarding genotyping data analysis was initially provided by Dr Anna Reutanen in the WTCHG. HLA type imputation was performed by Dr Alexander Dilthey and Dr Stephen Leslie in the Bioinformatics core in the WTCHG. The eQTLpath package incorporating ReactomePA and the mePipe package incorporating matrixeQTL were provided by Dr Peter Humburg in the in the WTCHG.

Associated publication

Fairfax, B. P., S. Makino, J. Radhakrishnan, K. Plant, S. Leslie, A. Dilthey, P. Ellis, C. Langford, F. O. Vannberg, and J. C. Knight (2012). "Genetics of gene expression in primary immune cells identifies cell type-specific master regulators and roles of HLA alleles". *Nat Genet* 44, pp. 502–10

Contents

Abstract i

Acknowledgements ii

Declarations iii

Associated publication iv

Contents v

List of Tables x

List of Figures xii

1 General Introduction 1

1.1 Genetic variation . . . 1

1.1.1 History . . . 1

1.1.2 Types of genetic variation . . . 4

1.1.3 Genome-wide population based studies . . . 7

1.2 Gene expression and epigenetics . . . 16

1.2.1 Expression quantitative trait loci mapping . . . 16

1.2.2 Epigenetics . . . 23

1.2.3 Summary . . . 27

1.3 Common disease, immunity and inflammation . . . 27

1.3.1 Human immune system . . . 27

1.3.2 Major histocompatibility complex . . . 29

1.3.3 Inflammatory response . . . 31

1.3.4 Genetic determinant of inflammation . . . 35

1.4 Summary . . . 36

1.5 Aims of this thesis . . . 37

2 Materials and Methods 39

2.1 PBMCs purification . . . 39

2.2 Immune cell separation . . . 39

2.3 Genomic DNA extraction . . . 39

2.4 RNA extraction . . . 40

2.5 Cell culture . . . 40
2.6 Replication of eQTL analysis by quantitative polymerase chain reaction (qPCR) . . . 40
2.7 Flow cytometry . . . 41
2.8 Cloning . . . 42
2.9 Transfection . . . 43
2.10 Luciferase assay . . . 45
2.11 Nuclear extract . . . 45
2.12 Electrophoretic mobility shift assay (EMSA) . . . 45
2.13 Gene expression array . . . 46
2.14 Genomic DNA genotyping . . . 46
2.15 Data analysis . . . 47

3 Differential gene expression data analysis of immune cell types 48
3.1 Introduction . . . 48
3.2 Sample collection of healthy volunteer cohort . . . 49
3.2.1 Recruitment of healthy European volunteers . . . 49
3.3 PBMCs and cell purifications . . . 49
3.3.1 RNA extraction . . . 51
3.4 Quality control and normalisation of microarray data . . . 52
3.4.1 Illumina BeadChip gene expression microarray . . . 52
3.4.2 Control data . . . 52
3.4.3 Preprocessing data . . . 53
3.4.4 Quality control . . . 56
3.4.5 Biological and technical replications of microarray experiments . . . 60
3.4.6 Final dataset . . . 61
3.5 Differential gene expression analysis of two immune cell types . . . 62
3.5.1 Background . . . 62
3.5.2 Processing the data . . . 62
3.5.3 Differentially expressed genes in B-cells . . . 63
3.5.4 Differentially expressed genes in monocytes . . . 65
3.6 Pathway analysis of differentially expressed gene in two immune cell types 72
3.6.1 Background . . . 72
3.6.2 Pathway analysis . . . 72
3.6.3 Pathways enriched with B-cell gene products . . . 72
3.6.4 Pathways enriched with monocyte gene products . . . 73
3.7 Correlation of gene expression within cell types . . . 75
3.7.1 Background . . . 75
3.7.2 Correlation analysis . . . 75
3.7.3 Pathway analysis of the highly correlated genes . . . 78
3.8 Discussion . . . 78
3.8.1 Sample collection . . . 78
3.8.2 Quality control of microarray data . . . 81

3.8.3 Differential expression analysis . . . 81
3.8.4 Pathway analysis . . . 82
3.8.5 Correlation . . . 83
3.8.6 Investigation for disease . . . 83
3.8.7 Summary . . . 83

4 Investigation of eQTL in primary human immune cells 84
4.1 Introduction . . . 84
4.2 Gene expression array profiling of primary immune cells . . . 86
4.3 Genotyping of the volunteer cohort . . . 87
4.3.1 Background . . . 87
4.3.2 Genotyping platform . . . 87
4.3.3 Genotyping rate and filtering . . . 89
4.3.4 Sample filtering . . . 89
4.3.5 Population stratification . . . 90
4.3.6 Final dataset . . . 91
4.4 Cell type specific expression quantitative trait mapping . . . 91
4.4.1 cis eQTL . . . 91
4.4.2 Examples of shared eQTL in both cell types . . . 93
4.4.3 Validation of eQTL by qPCR . . . 100
4.4.4 Replication of cis eQTL in published datasets . . . 100
4.4.5 Cell type specific eQTL . . . 100
4.4.6 Location of cis eQTL in relation to transcription start site . . . 108
4.4.7 trans eQTL . . . 108
4.4.8 Monocyte specific trans eQTL involving 12q15 . . . 109
4.4.9 Expression correlation of *LYZ* and trans genes . . . 110
4.4.10 B-cell specific trans eQTL involving 9q31.2 . . . 113
4.4.11 Expression correlation of *KLF4* and trans genes . . . 113
4.4.12 Pathway analysis of trans associated genes . . . 117
4.5 Cell type specific eQTL associated with GWAS loci . . . 121
4.5.1 cis eQTL associated with GWAS loci . . . 121
4.5.2 trans eQTL associated with GWAS loci . . . 125
4.5.3 trans eQTL in 12q13.2 identified in GWAS . . . 125
4.5.4 trans eQTL in the HLA region and relevance to disease . . . 128
4.6 HLA . . . 132
4.6.1 HLA type imputation . . . 132
4.6.2 Association of HLA alleles and gene expression . . . 135
4.6.3 HLA association and trans eQTL . . . 135
4.6.4 cis HLA association . . . 141
4.7 Discussion . . . 141
4.7.1 Population stratification . . . 142
4.7.2 cis eQTL mapping in two immune cell types . . . 142
4.7.3 trans eQTL . . . 143

4.7.4 GWAS and eQTL . . . 144
4.7.5 HLA . . . 145
4.7.6 Methodology and limitation of eQTL mapping . . . 145
4.7.7 Summary . . . 146

5 Paired eQTL analysis of monocytes and differentiated macrophages 147
5.1 Introduction . . . 147
5.2 Sample collection of the paired monocytes and macrophages population . 148
5.2.1 $CD14^+$ monocyte collection of healthy European individuals . . . 148
5.2.2 *ex vivo* differentiation of monocytes into macrophages . . . 149
5.3 Population stratification analysis . . . 150
5.3.1 Genotyping . . . 150
5.3.2 Quality control of genotyping . . . 150
5.3.3 Population stratification . . . 151
5.4 Gene expression profiling . . . 151
5.4.1 Normalisation of microarray data . . . 152
5.5 Differential gene expression analysis . . . 158
5.5.1 Processing the data . . . 158
5.5.2 Differentially expressed genes in macrophages compared to monocytes 159
5.6 eQTL mapping analysis . . . 173
5.6.1 Data processing . . . 174
5.6.2 Cell type specific eQTL . . . 175
5.6.3 Examples of cis eQTL in the two cell types . . . 178
5.6.4 Macrophage specific and shared eQTL associated with GWAS loci . 192
5.7 Discussion . . . 201
5.7.1 Sample collection . . . 201
5.7.2 Differential gene expression analysis . . . 202
5.7.3 eQTL analysis . . . 203
5.7.4 Summary . . . 206

6 Functional characterisation of eQTL 207
6.1 Introduction . . . 207
6.2 eQTL signal in *LYZ* region in 12q15 . . . 208
6.2.1 Imputation of 12q15 locus . . . 209
6.2.2 Regulatory sequences in *LYZ* eQTL . . . 210
6.2.3 Transcription factor binding motifs . . . 211
6.3 Assaying enhancer activity using the luciferase gene reporter system . . . 213
6.3.1 Cloning of eQTL locus flanking LYZ in 12q15 . . . 214
6.3.2 *In vitro* luciferase reporter assay . . . 215
6.4 EMSA to identify SNP alleles bound differentially to cellular proteins . . . 218
6.4.1 Background . . . 218
6.4.2 Protein binding patterns of LYZ eSNPs . . . 220
6.4.3 Competition assay . . . 220

6.4.4 Antibody supershift assay 224
6.5 Discussion 225
6.5.1 Functional characterisation of GWAS and eQTL studies 225
6.5.2 Identification of regulatory sequence 226
6.5.3 Luciferase reporter assay 227
6.5.4 EMSA 227
6.5.5 Summary 229

7 General Discussion **230**
7.1 Findings and conclusions 230
7.1.1 Differential gene expression analysis 230
7.1.2 eQTL mapping analysis 232
7.1.3 Functional characterisation of eQTL 236
7.2 Future work 236
7.2.1 Analysis methods 237
7.2.2 HLA association 238
7.2.3 Mechanism behind trans eQTL association 239
7.2.4 Disease specific eQTL analysis 239
7.2.5 Identification of regulatory variation 240
7.3 Concluding remarks 241

Appendix **242**

Bibliography **254**

List of Tables

1.1 The HapMap project overview 11

2.1 Oligonucleotide sequences used for qPCR 41
2.2 Cloning primers used for *LYZ* luciferase assays 44
2.3 Cloning plasmids used for *LYZ* luciferase assays 44
2.4 EMSA probe sequences 46
2.5 Probe content of Illumina HumanHT-12 v4 BeadChip 47
2.6 Variant marker content of Illumina HumanOmniExpress-12v1.0 BeadChip . . 47

3.1 Demographics of the healthy volunteer cohort 50
3.2 Summary of differentially expressed probes 63
3.3 Differentially expressed genes in B-cells 67
3.4 Differentially expressed genes in monocytes 68
3.5 HLA gene expression in B-cells and monocytes 71
3.6 Summary of gene correlation analysis 75
3.7 Highly correlating genes in B-cells and monocytes 77

4.1 Summary of cis eQTL in monocytes and B-cells 93
4.2 Directional eQTL in shared eSNPs 99
4.3 Comparison of eQTL studies 102
4.4 Summary of trans eQTL in monocytes and B-cells at p-value<$1.0\text{x}10^{-11}$. 109
4.5 Trans eQTL genes correlated with *KLF4* expression 117
4.6 Reactome pathway enrichment analysis on rs10816519 trans associated genes in B-cells 120
4.7 Cell type specific eSNPs identified in GWAS 121
4.8 HLA allele associations with gene expression 135
4.9 Cell type specificity in HLA class I and II alleles 136
4.10 HLA allele association to trans eQTL 138

5.1 Demographics of the 64 healthy individuals 149
5.2 Summary of differentially expressed probes in the monocytes and macrophages 159
5.3 Reactome pathway enrichment analysis on the differentially up-regulated genes in macrophages 161
5.4 Reactome pathway enrichment analysis on the differentially down-regulated genes in macrophages 162

5.5 Up-regulated genes in macrophages . . . 166
5.6 Down-regulated genes in macrophages . . . 169
5.7 Differentially down-regulated *HLA* genes in macrophages . . . 172
5.8 cis eQTL analysis results with PCs . . . 175
5.9 cis eSNPs identified with genes expressed at a similar level in monocytes and macrophages . . . 177
5.10 Differentially expressed probes involved in eQTL . . . 177
5.11 Macrophage unique cis eQTL in comparison to monocytes . . . 186
5.12 Monocyte unique cis eQTL in comparison to macrophages . . . 188
5.13 Cis eQTL shared between macrophages and monocytes . . . 190
5.14 Macrophage specific and shared eQTL associated with GWAS loci . . . 192
5.15 Macrophage specific eQTL associated with GWAS loci . . . 198
5.16 Shared eQTL associated with GWAS loci . . . 199

6.1 Transcription factor binding motifs of *LYZ* eSNPs . . . 212
6.2 *LYZ* eSNPs . . . 215

1 Reactome pathway enrichment analysis on genes significantly differentially expressed in B-cells . . . 242
2 Reactome pathway enrichment analysis on genes significantly differentially expressed in monocytes . . . 243
3 Reactome pathway enrichment analysis on the correlated genes in B-cells and monocytes . . . 247
4 HLA imputation summary . . . 248
5 List of HLA allele association in cis in monocytes and B-cells . . . 252
6 Description of the HapMap and the 1000 Genome reference populations . . . 253

List of Figures

1.1 Classes of genomic variation 5
1.2 Model of contribution of disease associated variants to phenotypes 14
1.3 Diagram of cis and trans eQTL associations 19
1.4 Genes encoded in the human MHC 31

3.1 FACS of MACS $CD14^{+}$ separation 51
3.2 Illumina BeadChip control probe data 54
3.3 Expression of housekeeping genes before normalisation 55
3.4 The distribution of expression value densities before and after normalisation . 57
3.5 The clusters of the samples before and after normalisation 57
3.6 Boxplots of expression values before and after normalisation 58
3.7 Pairwise MA plot of the expression data 59
3.8 PCA plots of the expression data 60
3.9 The correlation between two replicated samples 61
3.10 Volcano plot of differentially expressed genes in B-cells and monocytes 63
3.11 MA plot of differentially expressed genes in B-cell and monocyte 64
3.12 Pathway enriched with B-cells differentially expressed genes. 73
3.13 Pathways enriched with monocytes differentially expressed genes. 74
3.14 Frequency of probe correlation 76
3.15 Pathway enriched with monocytes highly correlated genes 79
3.16 Pathway enriched with B-cell highly correlated genes 80

4.1 Infinium SNP genplot 89
4.2 MDS plot of population stratification 92
4.3 Venn diagrams of monocytes and B-cells cis and trans eQTL associations in a cell type specific manner 92
4.4 Circos plots of monocyte and B-cell eQTL association 94
4.5 cis eSNPs at different p-value thresholds 95
4.6 Box and association plots for shared cis eQTL between monocytes and B-cells 96
4.7 Shared cis eQTL between monocytes and B-cells 97
4.8 Examples of directional eQTL in shared eSNPs 98
4.9 qPCR validation of eQTL 101
4.10 Comparison of observed eSNPs in this study with previously published eQTL studies 103

4.11 Gene expression levels of cell type specific eQTL . . . 105
4.12 B-cell specific eQTL association plot . . . 106
4.12 Monocyte specific eQTL association plot . . . 107
4.13 Location of cis eQTL in relation to distance from transcription start sites . . . 108
4.14 Manhattan plot of rs10784774 trans association in monocytes . . . 110
4.15 rs10784774 trans associated genes in monocytes. (A) Regional association plots. The X and Y axes indicate chromosome location and p-values of probes. 111
4.15 rs10784774 trans associated genes in monocytes . . . 112
4.16 Gene expression correlation of *LYZ* and trans associated genes in monocytes . 114
4.17 rs10816519 trans associated genes in B-cells . . . 115
4.17 rs10816519 trans associated genes in B-cells . . . 116
4.18 Gene expression correlation of *KLF4* and trans associated genes in B-cells . . . 118
4.19 cis eQTL identified in disease GWAS . . . 124
4.20 Haplotype analysis of disease susceptible loci in 12q13.2 . . . 126
4.21 B-cell trans association in 12q13.2 . . . 127
4.22 Haplotype analysis of disease susceptible loci in 6p21.32 . . . 130
4.23 Monocyte trans eQTL in 6p21.32 with associated genes . . . 131
4.24 Validation of HLA 2 digit alleles imputation . . . 134
4.25 *HLA-DQA1* and *HLA-DRB1* associations with trans eQTL gene expression . . . 140

5.1 qPCR of the CD markers in differentiated macrophages . . . 149
5.2 MDS plot of population stratification for the 63 individuals . . . 152
5.3 MDS plot of population stratification within Europeans for the 63 individuals 153
5.4 Density plot of the signal intensity before and after normalisation. . . . 154
5.5 The clusters of the monocyte and macrophage expression data before and after normalisation . . . 154
5.6 Boxplots of the monocytes and macrophages expression values before and after normalisation . . . 155
5.7 Pairwise MA plot of the expression data in the monocytes and macrophages . 156
5.8 PCA plots of the expression data in the monocytes and macrophages . . . 157
5.9 The number of significant probes in macrophages and monocytes differential gene expression analysis . . . 159
5.10 Volcano plot of differentially expressed genes in macrophages and monocytes 160
5.11 MA plot of differentially expressed genes in monocytes and macrophages . . . 161
5.12 Differentially expressed genes in macrophages and monocytes in relation to a pathway for cytokines and their receptors . . . 165
5.13 eQTL analysis using PCs as covariates . . . 176
5.14 Venn diagram of the cell type specific cis eQTL in macrophages and monocytes 177
5.15 Distance from TSS to eQTL in monocytes, macrophages and shared datasets . 178
5.16 Manhattan plots of macrophage and monocyte specific cis eQTL . . . 179
5.17 Examples of macrophage specific eQTL in comparison to monocytes . . . 181
5.18 Examples of monocyte specific cis eQTL in a comparison to macrophages . . . 182
5.19 Manhattan plot of shared cis eQTL between monocytes and macrophages . . . 183

5.20 Shared cis eQTL between monocytes and macrophages 185
5.21 Examples of macrophage eQTL associated with GWAS loci 195
5.22 Examples of macrophage eQTL associated with GWAS loci 197

6.1 A diagram of the 12q15 LYZ promoter region 209
6.2 Local association plots of *LYZ* expression with rs10878967 and rs12812860 . . 210
6.3 Luciferase reporter construct 214
6.4 Luciferase activity of a region including rs10784774 216
6.5 Luciferase activity of p300 TFBS 217
6.6 Luciferase activity of a region including rs10878967 217
6.7 Luciferase activity of a region including rs1384 218
6.8 Position effect of enhancers in pGL3-promoter 219
6.9 EMSA with LYZ eSNPs with K562 221
6.10 EMSA competition assay on rs1384 with K562 222
6.11 EMSA competition assay on rs10878967 using K562 nuclear extract 222
6.12 EMSA competition assay on rs10878967 with THP-1 223
6.13 EMSA competition assay on rs10878967 with LCL 223
6.14 EMSA supershift assay on rs10878967 using K562 nuclear extract 224
6.15 EMSA supershift assay on rs1384 on K562 nuclear extract 225
6.16 Summary of *LYZ* locus functional characterisation 226

1 *HLA-DQA1* and *HLA-DRB1* association to trans eQTL gene expression 249
2 *HLA-DQB1* association to trans eQTL gene expression 250
3 *HLA-C* and *-B* alleles associated with *PSORS1C3* expression 251

Abbreviations

E.coli	*Escherichia coli*
ABF-1	activated B-cell factor-1
AIDS	Acquired immunodeficiency syndrome
ALPS	autoimmune lymphoproliferative syndrome
APC	antigen presenting cell
APS	Ammonium per sulphate
AS	ankylosing spondilitis
ATM	adipose tissue macrophage
BCR	B-cell receptor
BD	bipolar disorder
BMI	body mass index
BSA	bovine serum albumin
CAD	coronary artery disease
cAMP	cyclic AMP
CD	cluster of differentiation
CDCV	common disease, common variant hypothesis
cDNA	complementary DNA
CDRV	common disease, rare variant hypothesis
CHD	coronary heart disease
ChIP	chromatin immunoprecipitation
chr	chromosome
CNP	copy number polymorphism
CNS	central nervous system
CNV	copy number variation
cRNA	complementary RNA
Cy3	Cyanine 3
DC	dendritic cell
DMSO	dimethyl sulfoxide
DNA	deoxyribonucleic acid

DTT	Dithiothreitol
EBV	Epstein Barr virus
EDTA	ethylene diamine tetraacetic acid
EGTA	ethylene glycol tetraacetic acid
EMSA	electrophoretic mobility shift assay
eQTL	expression quantitative trait loci
eSNP	expression associated SNP
FDR	false discovery rate
GFP	green fluorescent protein
GM-CSF	granulocyte-macrophage colony stimulating factor
GPCR	G protein coupled receptor
GWAS	Genome-wide association study
HIV	Human immunodeficiency virus
HLA	human leukocyte antigen
HUVEC	human umbilical vein endothelial cells
HWE	Hardy-Weinberg equilibrium
IBD	Identity by descent
IBD	inflammatory bowel disease
ICA	independent component analysis
Ig	immunoglobulin
Indels	insertions and deletions
IQR	Interquartile range
KEGG	Kyoto Encyclopedia of Genes and Genomes
LCL	lymphoblastoid cell line
LD	linkage disequilibrium
LDL	low cholesterol lipoprotein
LOD	logarithm of odds
logFC	$\log_2$ fold change
LPS	lipopolysaccharide
MACS	magnetic activating cell sorting
MAF	minor allele frequency
MDS	multidimensional scaling
MHC	major histocompatibility complex
MI	myocardial infarction
miRNA	microRNA
MS	multiple sclerosis

NHGRI the National Human Genome Research Institute

NLR NOD-like receptor

PAMP pathogen associated molecular pattern

PBMC peripheral mononuclear blood cell

PBS phosphate buffered saline

PC principal component

PCA principal component analysis

PCR polymerase chain reaction

PPV positive predicted value

PRR pattern recognition receptor

qPCR quantitative polymerase chain reaction

RA rheumatoid arthritis

RFLP restriction fragment length polymorphism

RSN robust spline normalisation

SNP single nucleotide polymorphism

SNV single nucleotide variant

SVD singular value decomposition

T1D type 1 diabetes

T2D type 2 diabetes

TBE Tris/borate/EDTA

TEMED N,N,N',N'-Tetramethylethylenediamine

TF transcription factor

TFBS transcription factor binding site

The ENCODE project The ENCyclopaedia Of DNA Elements project

TLR Toll-like receptor

TNF tumour necrosis factor

TSS transcription start site

UC ulcerative colitis

VLDL very low cholesterol lipoprotein

VNTR variable number tandem repeats

w/v weight per volume

WGS whole genome sequencing

WTCCC Wellcome Trust Case Control Consortium

1 General Introduction

Human genetic variation is a key contributor to diverse and complex phenotypes. Since the completion of the first human genome sequence in 2001, an extended effort has been made to elucidate the diversity of genetic variation and its functional consequences. This introduction provides a historical overview of the association between genetic variation and disease phenotypes, focusing in particular on common multifactorial traits, the impact of genome-wide association studies and current approaches to defining regulatory variants through expression quantitative trait mapping and other approaches. This is complemented by an overview of the biology of inflammation and immunity, showing the relevance of genomic approaches. The key questions to be addressed in this thesis are introduced; namely how to identify and functionally characterise regulatory genetic variants modulating the immune system which may determine disease risk.

1.1 Genetic variation

1.1.1 History

Discovery of heredity

Genetics underlies the heredity of traits and transmission of information from a parent to its offspring in living organisms. The inheritance of traits was first observed in plants by Gregor Mendel in 1865, when he described the inheritance of certain traits such as the colour and coat of pea plants as a dominant or recessive phenotype (Mendel, 1866). Archibald Garrod in 1902 then described recessive inheritance in humans from observation of alkaptonuria, a rare congenital disorder caused by a defect in tyrosine metabolism (Garrod, 1902). It remained unclear how these traits were carried from one generation to the next even though patterns of inheritance had been described as phenotypes, including human diseases and traits, and in other organisms such as fruits flies.

Genetic linkage and discovery of DNA

Prior to discovery of DNA, genetic linkage was first explained by William Bateson and his colleagues in 1905 (Bateson, 1905). They described alleles where some phenotypes cosegregated. Genetic linkage is caused by recombination during meiosis. Cosegregation of a marker gene carrying a trait in the next generation manifests as physical proximity on the same chromosome. Linkage analysis of traits has been used to map genetic loci since their observation. In 1911 Morgan and his student proved that genes are on chromosomes and demonstrated the phenomenon of genetic linkage using a sex-linked white eye colour mutation in fruit flies (Morgan, 1910, Sturtevant, 1913). DNA, residing in cell nuclei, was shown to be the substance of inheritance by the Hershey-Chase experiment in 1952; bacteria that had been transformed with bacteriophage DNA were found to pass on the DNA, rather than its protein product, to its progeny (Hershey and Chase, 1952). The structure of the double-helix DNA was proposed by James Watson and Francis Crick in 1953, after viewing a single X-ray diffraction image of DNA taken by Rosalind Franklin and Taymond Gosling in 1952 (Watson and Crick, 1953, Franklin and Gosling, 1953). Crick and his colleagues later showed that three bases identify each amino acid, a codon, leading to the deciphering of the genetic code (Matthaei *et al.*, 1962). The central dogma of molecular biology stated by Crick in 1958 (restated in 1970) explained the DNA-RNA-protein flow of genetic information (Crick, 1970). This work on molecular genetics in the early 20th century has promoted the gene from an abstract to a physical unit on the DNA sequence.

Genetic mapping

Mapping of disease associated genes, along with the DNA related discoveries described above, has enabled advances in molecular biological techniques. Polymorphic genetic markers have been used as tools to map associations with phenotypes long before the human genome was sequenced. Discovery of restriction enzymes which cut DNA at or near specific sequences was a breakthrough for detecting genetic polymorphisms and their use in mapping (Smith and Wilcox, 1970). Restriction fragment length polymorphism (RFLP) uses a restriction enzyme to digest homologous DNA fragments to determine whether polymorphisms are present and whether they are specific to a trait or disease. Botstein and colleagues developed RFLPs to construct a genetic linkage map using restriction enzymes in the 1980s (Botstein *et al.*, 1980). The genetic diagnosis of the Mendelian disease sickle cell anemia, which is caused by a mutation on the β haemoglobin

gene, was made possible by RFLP mapping (Saiki *et al.*, 1985). The causal trinucleotide repeat of CAG on the Huntington gene (*HTT*) was determined by the same strategy (Gusella *et al.*, 1983).

Molecular biological techniques

The long history of genetic mapping has highlighted the great diversity in genetic information across individuals, within and between populations. It has become possible to identify variation at single nucleotide resolution since the development of DNA sequencing. This is achieved by the incorporation of chain-terminating dideoxynucleotides *in vitro* DNA polymerisation, which was developed by Frederick Sanger in 1977 (Sanger *et al.*, 1977). Technological advances in molecular biology include a discovery of the reverse transcription reaction in 1970. This reaction reverses the central dogma by making complementary DNA (cDNA) from RNA and is catalysed by reverse transcriptase (Baltimore, 1970, Temin and Mizutani, 1970). Cloning of transcribed RNA was made possible by converting it to DNA (Summers and Mason, 1982, Mann *et al.*, 1983). The other essential invention was the polymerase chain reaction (PCR) in 1987 for DNA amplification (Mullis and Faloona, 1987). These technologies, together with recent implementation of massively parallel sequencing (also called whole genome sequencing, WGS) by synthesis, have promoted sequencing of DNA a a genomic scale (Metzker, 2010). The human genome project was completed in 2001 as the first reference human genome sequence using the shot-gun sequencing strategy (Lander *et al.*, 2001, Venter *et al.*, 2001). Over the last 5 years it has been possible to sequence thousands of genomes as the results of rapid falls in cost and increases in throughput. The 1000 Genomes Project has been launched to investigate genome sequences in a population level (Durbin *et al.*, 2010).

Human genome sequence and genetic disease association

The human genome is comprised of approximately 1.1% protein coding sequence, with approximately 25000 genes in the 2.9 billion base pairs (Venter *et al.*, 2001). About three quarters of the genome does not encode genes or introns. Mendelian traits and diseases have been mapped onto the genome, mostly at protein coding regions. These include 2985 genes with phenotype-causing mutations and 4902 phenotypes with a gene product basis description (OMIM database, accessed 27th March 2013, Amberger *et al.*, 2009). By contrast, for complex traits such as common or multifactorial diseases, only a small number of disease genes were identified by classic linkage analysis. Complex traits show

a heritable component although they tend not to follow Mendel's rules of inheritance and are jointly affected by environmental factors. Genetic factors may be clinically significant. This is notable in infectious disease as genetic mutation, deletion and variation have been identified as being associated with increased susceptibility or resistance to specific infectious disease phenotypes (Chapman and Hill, 2012). For example, the disease onset and the progression by infection of human immunodeficiency virus (HIV) is delayed in carriers of a chemokine receptor 5 (*CCR5*) with a 32bp deletion (Smith *et al.*, 1997). Homozygous individuals with a 32bp deletion are completely protected against HIV infection. CCR5 is an essential coreceptor for HIV infection and fails to be expressed on the cell surface in the presence of the deletion (Benkirane *et al.*, 1997). In addition, resistance to malaria pathogens *Plasmodium vivax* and *P. knowlesi* is given by mutations to a gene that encodes a Duffy antigen chemokine receptor (DARC) which functions as a receptor for *P. vivax* in the surface of erythrocytes (Miller *et al.*, 1975, Miller *et al.*, 1976). The FY*B allele of *DARC*, found in the Duffy negative blood group, with a single T to C substitution in the promoter region disrupts a binding site for GATA1 transcription factor and represses the gene expression in erythrocytes (Tournamille *et al.*, 1995). As complex diseases or traits cannot be explained in a monogenic manner by Mendel's rules of inheritance, new approaches were needed to address these traits. Genome-wide association studies (GWAS) were developed using high throughput genotyping to search association, comparing thousands of cases and controls (Section 1.1.3).

1.1.2 Types of genetic variation

Naturally occurring genetic variation has been used as a marker for diseases based on cosegregation with the causal loci in linkage studies as mentioned above. In some instances, genetic variants are functionally associated with disease susceptibility and development. Scherer and colleagues described classification of genetic variation based on size (Scherer *et al.*, 2007, Figure 1.1). Sequence variation consists of single nucleotide variation, in contrast structural variation includes variation larger than 2bp (Feuk *et al.*, 2006, Scherer *et al.*, 2007). This section describes the main types of genetic variation and relevance to disease susceptibility.

Variable number tandem repeat

Variable number tandem repeats (VNTRs) are a type of variation where a short nucleotide sequence is repeated a number of times. This type of repetitive DNA sequence makes

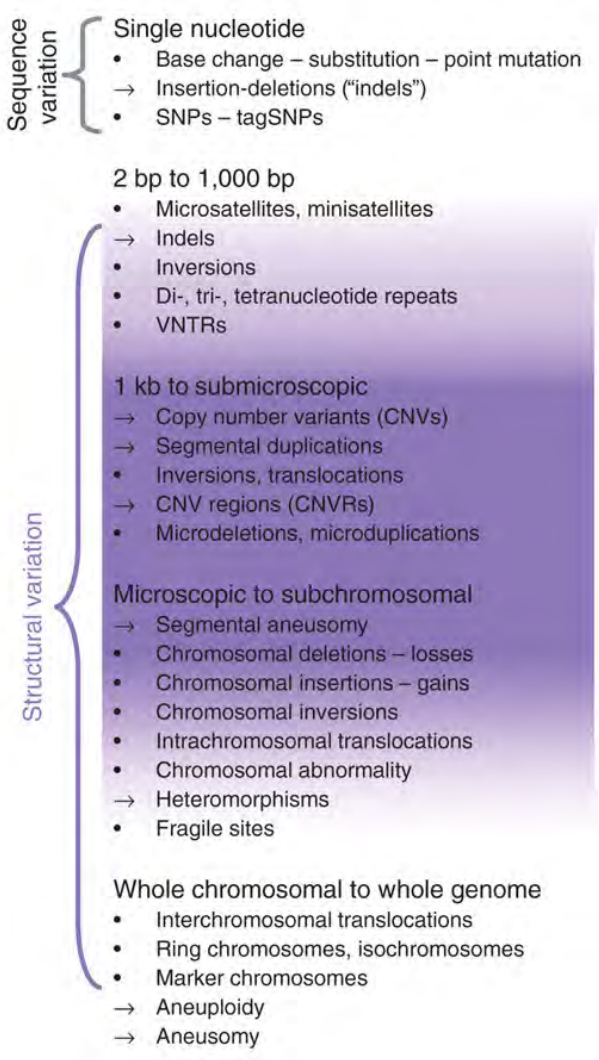


Figure 1.1: Classes of genomic variation. Modified and adapted by permission from Macmillan Publishers Lts: Nature Review Genetics, Scherer *et al.*, 2007, copyright 2007.

up 35% of the human genome sequence (Venter *et al.*, 2001). VNTRs were essential as genetic markers for linkage mapping by RFLP before the whole human genome sequence became available (Nakamura *et al.*, 1987). A genetic map of the human genome based on microsatellites, a type of VNTR with repeating short sequences of 2-6bp, was created for linkage analysis (Dib *et al.*, 1996). Some diseases are shown to associate with VNTRs. For example, a 14bp VNTR found in 5'UTR of an insulin gene (*INS*) is associated with Type 1 diabetes (T1D) (Lucassen *et al.*, 1993). The allele variation of the VNTR is positively correlated with the gene expression levels of *INS* in thymus. This suggests that the longer VNTR provides higher expression of *INS* and protects against T1D susceptibility (Pugliese *et al.*, 1997).

Copy number variation

Copy number variation (CNV) is defined as a DNA segment of one kb or more which has been found to vary in number when compared to a reference genome (Stankiewicz and Lupski, 2010). CNVs result from genomic rearrangement events such as deletions, insertions, duplications and translocations. The effect of CNVs on phenotypes can result in changing the dosage of a gene, altering the distance of regulatory DNA sequences, or unmasking recessive mutations or functional polymorphisms in a remaining allele. Genome-wide studies have revealed that CNVs occupy 12% of the genome (Redon *et al.*, 2006). For example, a lower copy number with *CCL3L1*, encoding a potent HIV-1 suppressive chemokine, is associated with significantly increased susceptibility to HIV infection (Gonzalez *et al.*, 2005). This locus has a variable copy number of a segmental duplication encompassing *CCL3L1*.

Single nucleotide variants

In contrast to VNTRs and CNVs, single nucleotide variants (SNVs) typically consist of one nucleotide with the substitution for another in the sequence. Two alternate bases, biallelic SNVs, are often observed. Single nucleotide polymorphisms (SNPs) are defined as having a minor allele frequency (MAF) greater than 1% to distinguish them from rare variants (Brookes, 1999). SNPs are abundant compared with other types of DNA variation and common SNPs (often defined as MAF>5%) are found an average every one or two kb (Sachidanandam *et al.*, 2001).

SNVs located in gene coding sequences may alter amino acid codons; these are called non-synonymous SNVs. SNVs that are associated with diseases identified by linkage studies are mainly coding SNVs because the functional causes of disease susceptibility at a transcriptional or translational level can be determined. For example, an allele of *ApoE4*, encoding Apolipoprotein E, is associated with increased risk of Alzheimer disease due to two amino acid changes caused by SNVs (Corder *et al.*, 1994). The isoform's protein conformation is affected by these variants, resulting in instability and less binding affinity to receptors for low cholesterol lipoprotein (LDL)(Hatters *et al.*, 2006). However, SNVs in coding regions account for a very small fraction of total SNPs (approximately 0.4%, Durbin *et al.*, 2010). The majority of SNPs are located in gene regulatory regions or in gene deserts.

Summary

Various types of genetic variation have been found in the genome by linkage analysis to help identify loci associated with diseases or traits. The completion of the human genome sequence has helped to understand the structure of the variation in the genome. Using known variation over the whole genome, the search for phenotype associated loci is possible on a genome-wide scale.

1.1.3 Genome-wide population based studies

Linkage analysis initially had few genetic markers available in humans to systematically map inheritance of traits. Phenotypic markers were instead used to define associations with disease. For example, blood types were associated with stomach cancer and alterations in the haemoglobin β protein were associated with sickle-cell anaemia (Aird *et al.*, 1953, Ingram, 1956).

In the 1970s, molecular biological techniques were developed to use naturally occurring DNA polymorphisms as genetic linkage markers including RFLPs as described earlier. DNA cloning made it possible to map a genomic position associated with a Mendelian trait in model organisms by crosses (Clarke and Carbon, 1980, Bender *et al.*, 1983). RFLP analysis could trace familial transmission of chromosomal regions in humans. Huntington's disease was the first case to demonstrate linkage of RFLP markers to disease (Gusella *et al.*, 1983). RFLP markers have been systematically developed to cover the human genome, starting with 400 DNA markers in 1987 and increasing to more than 5000 (Donis-Keller *et al.*, 1987, Nakamura *et al.*, 1987, Dib *et al.*, 1996).

By contrast, common human disease including cancer, heart disease, hypertension and diabetes show evidence of heritability based on family and twin studies but typically do not show a Mendelian pattern of inheritance and genetic factors have a modest magnitude of effect making the identification of specific casual variants difficult (Fuchs *et al.*, 1994, Antoniou *et al.*, 2003, Longini *et al.*, 1984, Bell and Polonsky, 2001). Genetic association studies have been used, initially based on a candidate gene approach. Type 2 diabetes (T2D) was one of the small number of instances where this led to the successful identification of an associated gene (Deeb *et al.*, 1998). A missense common variant Pro12Ala in *PPARG*, encoding a transcription factor peroxisome proliferator-activated receptor-γ, was associated with reduced risk of T2D. This association was the only one to be validated in a different cohort (Altshuler *et al.*, 2000). Another early example was the susceptibility locus to Crohn's disease that was identified by positional cloning to

NOD2 (also known as *CARD15*). *NOD2* encodes the nucleotide-binding oligomerisation domain-containing protein 2, and its product belongs to a signalling pathway important in the innate immunity and inflammation (Hugot *et al.*, 2001).

It is known that both genetic and environmental factors affect the development of common diseases. Disease susceptibility is unpredictable even when the influence of environmental factors is known. For example, tobacco smoking has been shown to influence the development of lung cancer by increasing the risk more than twenty-fold. However, not all smokers will develop the disease despite this increased risk, suggesting that other unidentified factors may be involved (Peto *et al.*, 2000, Khuder, 2001).

Linkage disequilibrium and haplotypes

Linkage disequilibrium (LD) describes non-random association of alleles at two or more loci. The term LD was first introduced by Lewontin and Kojima in 1960, when they described the relationship of two gametes by how frequently they nonrandomly cosegregate (Lewontin and Kojima, 1960). LD occurs due to non-random recombination of homologous chromosomes during meiosis. The likelihood of linkage between two genetic markers can be determined using the logarithm of odds (LOD) scores which are statistical test scores of the likelihood of two loci cosegregating whether they are actually linked or not (Morton, 1955).

The statistical measure of LD shows a relationship between two alleles. r^2, the square of the correlation coefficient between two alleles, is commonly used for measurement of LD between two alleles (Devlin and Risch, 1995). When r^2 is 1, they are perfectly correlated and are inherited together. Current conventional genotyping arrays cannot capture all variants on the genome. Many variants are found in strong linkage with others, especially common polymorphisms. Therefore, SNP genotypes analysed in the context of LD can be used to estimate (impute) correlating SNPs based on the r^2 statistic. Selecting tagging SNPs which can identify many other SNPs due to strong LD is one of the essential steps when designing a genetic association study (Carlson *et al.*, 2004).

A haplotype is a combination of loci on the same chromosome that are inherited together during homologous recombination. Haplotypes give insights into linkage as they are determined by LD causing the coinheritance of alleles. Thus, a genome-wide map of haplotypes was proposed to be an useful resource for genetic analysis. To map human haplotypes over a genome-wide scale, appropriate genetic markers and a template of the human genome sequence were required. The two complete human genome sequences

were therefore used as a template (Lander *et al.*, 2001, Venter *et al.*, 2001). SNPs were chosen as genetic markers because they are abundant throughout the genome and most are biallelic, thus easy to genotype in a high throughput manner. Strong LD was thought not to extend further than 3kb, thus approximately 500K SNPs were estimated to cover the variation over a whole genome (Kruglyak, 1999).

The first genome wide SNP map containing 1.42 million SNPs was generated in 2001 by the International SNP Map Working Group using the infrastructure of the human genome sequencing project (Sachidanandam *et al.*, 2001). The density of SNPs in the map was one every 1.9kb. The identified SNPs were deposited into a public database (dbSNP, *http://www.ncbi.nlm.nih.gov/SNP/*, Sherry *et al.*, 2001).

Few initial association studies had been successful prior to 2001, with many failing to be replicated in other populations. The low success rate for identifying true associations is thought to have been due to a small sample sizes, population admixture and few available genetic markers (Risch and Merikangas, 1996, Cardon and Bell, 2001). Association studies have become feasible through the genome wide SNP map and high throughput low cost genotyping since 2001.

The International HapMap Project

The International HapMap Project was launched with the aim of creating a haplotype catalog at a population level.

The phase I HapMap project started by genotyping common SNPs (MAF>5%) in a 500kb range and demonstrated extensive haplotype diversity over four geographically diverse populations (Table 1.1, International HapMap Consortium, 2005). The project set up the infrastructure for large scale SNP genotyping as well as data analysis. The project facilitated genome-wide SNP genotyping and has identified recombination hotspots and block-like haplotype structures of LD. It is estimated that approximately 500K SNPs with MAF>5% are able to test >90% of common SNPs for non-African populations (International HapMap Consortium, 2005). African populations require more SNPs to be genotyped in order to cover the diverse and fragmented LD structures seen.

The populations that were used for the SNP analysis were also used to investigate association between haplotypes and the major histocompatibility complex (MHC) (Bakker *et al.*, 2006). The MHC locus on human chromosome 6 encodes cell surface molecules essential for the functionality of the immune system and is strongly associated with many immune-related diseases (Trowsdale, 2011).

The phase 2 HapMap project increased the number of genotyped SNPs to 3.1 million, an average density of one SNP per 1kb (Frazer *et al.*, 2007). The SNP coverage in the phase 2 dataset included common SNPs with a lower MAF compared to phase 1. The data was assumed to cover 25-30% of common SNPs (the total common SNPs were predicted to be 11 million at the time of the publication). Currently the number of identified common SNPs is approximately 53 million on the database, of which 38 millions have been validated.

Phase III of the HapMap project included rare SNPs with low MAF (<5%) and copy number polymorphisms (CNPs) (Altshuler *et al.*, 2010). Phase III also expanded the population diversity with seven additional populations sampled (Table 1.1). The HapMap project used the sequencing data from the ENCODE project which focused on functional characterisation of non-coding DNA elements and is described in the following section (ENCODE Project Consortium, 2004).Sanger sequencing on the ENCODE 500kb regions improved the resolution of SNPs with MAF between 0.5-5%. Investigation of rare variants and copy number polymorphisms revealed less sharing of these types of variation across populations.

SNP genotypes are not phased; in other words they do not represent haplotypes. However, haplotypes can be estimated based on SNP genotypes. The HapMap project and the 1000 Genomes project which is described in the next section have provided reference panels for haplotype estimation. Statistical programs to estimate haplotypes have been developed for this purpose (Marchini and Howie, 2010). The haplotype information helps to infer other SNPs with strong LD.

Table 1.1: The HapMap project overview. CNPs - copy number polymorphisms.

Population	Description	Size	Total size	The number of markers	MAF	Range	Publication
Phase 1							
YRI	African Yorban (30 parent-offspring trios) from Ibadan, Nigeria	90	269	1 million	>5%	500kb	International HapMap Consortium, 2003
CEU	western European descent (30 trios) from Utah, USA	90					International HapMap Consortium, 2005
CHB	Han Chinese from Beijing, China	45					
JPT	Japanese from Tokyo, Japan	44					
Phase 2							
Same as Phase 1 populations			270	3.1 million	>5%	1kb	Frazer *et al.*, 2007
Phase 3							
Phase 1 populations (4 populations)			1184	1.6 million	>5%		Altshuler *et al.*, 2010
YRI		167		common			
CEU		165		SNPs			
CHB		84					
JPT		86	692	ten 100kb	0.5-5%		
Additional 7 populations			(unrelated)	regional			
ASW	African ancestry (trios) from the southwestern USA	83		sequencing			
CHD	Chinese from metropolitan Denver, Colorado, USA	85		856 CNPs	1%		
GIH	Gujarati Indians from Houston, Texas, USA	88					
LWK	Luhya from Webuye, Kenya	90					
MKK	Maasai from Kinyawa, Kenya	171					
MXL	Mexican ancestry (trios) from Los Angeles, California, USA	77					
TSI	Tuscans from Italy	88					

The 1000 Genomes Project

The 1000 Genomes Project was launched to investigate the genetic diversity of individuals from diverse populations based on the analysis of whole genome sequences (Durbin *et al.*, 2010, Abecasis *et al.*, 2012). The improvement of sequencing technology in the past decade, in particular the development of WGS, has enabled whole genome sequencing in a much short time period than previously possible (Metzker, 2010). The data from the 1000 Genomes Project has identified 38 million SNPs, 1.4 million short insertions and deletions, and more than 14,000 larger deletions. A large fraction of rare variants was also described in this project (38% of MAF 0.5-5% and 87% of MAF<0.5%). The direct sequencing of the genomes revealed that rare variants form a signature of population differentiation.

Genome wide association studies

GWAS look for an association of genetic alleles to a trait in a population. This is achieved by comparing a population with a phenotypic trait (cases) to a population without the trait (controls) and identifying alleles that occur more or less frequently in the case population. GWAS of common or complex diseases have been greatly facilitated by the abundance of genetic markers to identify loci that are associated with disease susceptibility. Association studies have been helped by both the HapMap and the 1000 Genomes Projects because they have provided hundreds of thousands of DNA variants as markers as well as an appropriate analysis of population stratification which can help remove admixed samples to avoid false positives.

One of the first and largest GWAS was led by the Wellcome Trust Case Control Consortium (WTCCC) and used shared controls for seven common multifactorial diseases; bipolar disorder (BD), coronary artery disease (CAD), Crohn's disease, hypertension, rheumatoid arthritis (RA), T1D and T2D (Wellcome Trust Case Control Consortium, 2007). The study identified at least one allele per disease above a genome-wide association p-value threshold p-value<$5x10^{-7}$. Association is statistically tested for the null hypothesis of no association. A p-value threshold for GWAS, genome-wide significance, is generally agreed to be less than $5x10^{-8}$, which is based on a conservative Bonferroni correction (Risch and Merikangas, 1996, Wellcome Trust Case Control Consortium, 2007, Stranger *et al.*, 2011). The WTCCC expected to identify more risk associated alleles based on their power calculation (Wellcome Trust Case Control Consortium, 2007). This highlights

sample size as an issue affecting the success of GWAS especially when common SNPs are a subject of disease association (Risch and Merikangas, 1996).

Meta-analysis, where several GWAS are combined to increase the sample size, has since been carried out. For Crohn's disease, only two loci identified were previously annotated in the past 40 years by linkage and candidate gene association studies. As a result of the meta-analysis, the GWAS and meta-analysis for Crohn's disease have identified more than 163 risk associated loci to date (Jostins *et al.*, 2012).

Imputation of SNP genotypes may increase the likelihood of identifying specific disease associated loci. Imputation can be used in meta-analysis where different genotyping platforms have been used. Association studies using the WTCCC data with further imputation have identified few novel loci (Huang *et al.*, 2012). This is not surprising because tagging SNPs used in GWAS have been carefully selected to define at least one haplotype.

Since the first large study, GWAS have identified many loci associated with disease susceptibility. A catalogue of GWAS by the National Human Genome Research Institute (NHGRI) is updated weekly with newly identified loci associated with traits (Hindorff *et al.*, 2009, *http://www.genome.gov/gwastudies/*). As of January 2013, 1486 publications describing 772 phenotypes have been reported. Some loci associated with particular traits have been replicated in multiple study cohorts. For example, *NOD2* and *IL23R* loci associated with susceptibility for Crohn's disease have been replicated in several studies (Jostins *et al.*, 2012, Franke *et al.*, 2010, Barrett *et al.*, 2008). Listing the GWAS data from various traits helped to identify loci that are associated with multiple autoimmune disease traits, suggesting that they share underlying causal mechanisms. For example, the *CARD9* locus is associated with susceptibility to inflammatory bowel disease (IBD) including Crohn's disease and ulcerative colitis (UC) and ankylosing spondylitis (AS) (Barrett *et al.*, 2009, Franke *et al.*, 2010, McGovern *et al.*, 2010, Anderson *et al.*, 2011, Jostins *et al.*, 2012, Evans *et al.*, 2011). Another example, *CD40* locus is associated with susceptibility to RA, multiple sclerosis (MS) and IBD (Raychaudhuri *et al.*, 2008, Stahl *et al.*, 2010, Bahlo *et al.*, 2009, Jostins *et al.*, 2012). This locus was identified as a susceptibility locus to Kawasaki disease, a paediatric vasculitis autoimmune syndrome, in Japanese and Han Chinese populations (Onouchi *et al.*, 2012, Lee *et al.*, 2012). Most of the GWAS have used European ancestry populations, however some trait associations have been replicated in non-European populations. RA has been well studied prior to GWAS, where the susceptibility locus in the MHC *HLA-DRB1* allele, presenting antigens

including cyclic citrullinated peptide, is identified in approximately 70% of affected patients (Gregersen *et al.*, 1987). The association of *HLA-DRB1* allele to RA has been replicated in not only European but also Japanese and Korean populations (Wellcome Trust Case Control Consortium, 2007, Raychaudhuri *et al.*, 2008, Stahl *et al.*, 2010, Kochi *et al.*, 2010, Freudenberg *et al.*, 2011).

Challenges in GWAS

The genetic basis of common diseases, in terms of the relative impact of common and rare variants, has been debated. Two different hypothesis have been proposed: common disease, common variant (CDCV) and common disease, rare variant (CDRV) (Bodmer and Bonilla, 2008, Schork *et al.*, 2009). The CDCV hypothesis has been a premise of GWAS as common variants have been used as genetic markers to identify haplotypes responsible for disease susceptibility. GWAS are predicated on the hypothesis that disease associated alleles are frequent in a population. The deleterious effect caused by individual common variants would thus be very small. On the other hand, the CDRV hypothesis suggests that rare variants could be highly deleterious and occur through a higher mutation rate (Pritchard, 2001). SNPs with MAF<5% have continued to be reported since the phase 3 HapMap and the 1000 Genomes Projects. Although rare variants may have a larger effect size than common variants, a much larger population size may be required for genetic studies to have enough statistical power to detect these variants (Burton *et al.*, 2009). The model of rare and common variants that contribute disease susceptibility is summarised in Figure 1.2 (McCarthy *et al.*, 2008, Manolio *et al.*, 2009).

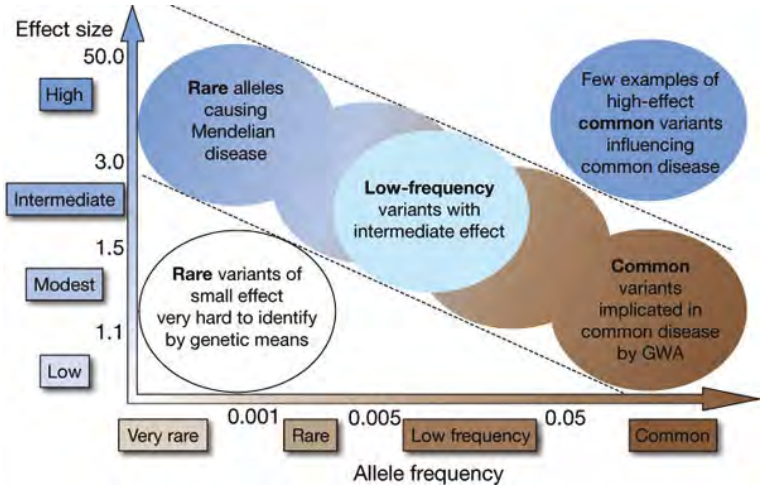


Figure 1.2: Model of contribution of disease associated variants to phenotypes. Adapted by permission from Macmillan Publishers Lts: Nature, Manolio *et al.*, 2009, copyright 2009.

Common SNPs are often a subject of GWAS and the effect size of SNP alleles discovered is normally modest. The exception is found in the MHC regions. Genetic variants in the MHC regions tend to have a larger effect size because many immune-related genes are encoded in this gene dense region. Outside of the MHC regions, many SNPs with disease association are located in non-coding gene deserts which makes functional characterisation difficult. One of the few successful GWAS associations to be functionally characterised was that of the LDL level and a regulatory SNP that modifies hepatic expression of *SORT1* (Musunuru *et al.*, 2010). SNPs identified by GWAS may have a role in the function of disease mechanism, but may also just be a marker to tag the regulatory sequence that is involved in a functional mechanism.

Findings from family-based linkage and genome-wide analysis of traits do not explain all heritability. This shows the complex aetiology of common disease. As mentioned above, common SNPs show modest effect size and the proportion of heritability explained by identified loci is relatively small (Figure 1.2, Manolio *et al.*, 2009). The reasons for missing heritability not being detected in current study settings may be speculated about in several ways. Rare variants are difficult to detect due to a limited sample size and are likely to have a larger effect size. Meta-analysis, combining several cohorts for a same phenotype, is one way to overcome a sample size issue. A large scale whole genome sequencing in a population level also helps to provide detailed reference panels of genomic variation, which is described in the following section. Most GWAS have used SNPs. Structural variants including CNVs, insertions and deletions (Indels), VNTRs are now used as alternative genetic markers. However, the effect size and frequency of these markers are similarly inversely correlated, as with SNPs. For example, rare microdeletions and micrpduplications can explain large effect sizes in psychiatric diseases, autism and schizophrenia (effect sizes of 15-100, Weiss *et al.*, 2008, Stefansson *et al.*, 2008). In contrast, common structural variants with associations to phenotypes have much smaller effect sizes. A CNP (MAF>1-5%) which was associated with susceptibility to Crohn's disease has an effect size of 1.5 (McCarroll *et al.*, 2008). Non-DNA sequence modification in the genome, known as epigenetics, is another target for searching missing heritability. This is described in the following section.

The aetiology of common diseases includes not only genetic but also environmental factors as mentioned previously (Hunter, 2005). Non-genetic factors have been known to affect disease susceptibility. One of the most recognised factors is diet which affects risk of many diseases including obesity, diabetes and heart disease by changing cholesterol levels,

body fat mass and other disease associated elements. A study model which takes non-genetic factors into account in a genetic association study has been considered (Thomas, 2010). However there are challenges; environmental factors are difficult to quantify because of their multidimensionality. For example, air pollution has mixed chemical substances which make measurement of effects by each substance difficult.

Missing heritability of common diseases may not only be explained by DNA sequences. Different approaches to identify functional interpretation of genetic variants have been developed in parallel to GWAS.

1.2 Gene expression and epigenetics

Many genetic susceptibility loci for complex diseases have been identified in non-coding regions. Approximately 3% of disease- or trait- associated variants are in non-synonymous coding variants whereas the majority of GWAS mapped loci are located in non-coding genomic regions (40% intronic and >54% intergenic, Hindorff *et al.*, 2009, Freedman *et al.*, 2011). Variants in non-coding regions are often technically difficult to functionally characterise, however this is important in order to understand which genes are being modulated and the mechanisms of action in diseases.

DNA variants in non-coding regions may be involved in modulation of gene expression in several ways. For example, the variants may be involved by changing the binding affinity for specific transcription factors, altering susceptibility to DNA methylation, causing variation in DNase I hypersensitivity sites or leading to the modification of chromatin structure (Kerkel *et al.*, 2008, Schalkwyk *et al.*, 2010, Thurman *et al.*, 2012, Neph *et al.*, 2012). Following GWAS, functional approaches have been emerging in an effort to characterise genetic variation. Such approaches include mapping genetic determinants of gene expression by treating gene expression as a quantitative trait. This section reviews several functional approaches to explore the link between genetic variation and gene expression, and the effect on phenotypes.

1.2.1 Expression quantitative trait loci mapping

Heritability of gene expression at a genome-wide scale has been observed across different tissues and among populations (Goring *et al.*, 2007, Spielman *et al.*, 2007, Emilsson *et al.*, 2008, Price *et al.*, 2011). This evidence encourages the investigation of the association of genetic variation to gene expression on a genome wide scale.

Expression quantitative trait loci (eQTL) mapping is an approach that identifies genetic signals influencing gene expression locally in cis (intrachromosomal) and at a distance in trans (which may be interchromosomal). The genetic markers and gene expression are typically measured in hundreds of individuals from a given population with the same genetic background. An eQTL explains the fraction of the genetic effect on gene expression which is treated as a quantitative phenotype. Initial studies in model organisms have constructed a framework for eQTL mapping, predominantly using yeast and the fruit fly (Brem *et al.*, 2002, Jin *et al.*, 2001). Studies using recombinant inbred model organisms in yeast and rats have also demonstrated heritability and tissue specificity of gene expression (Schadt *et al.*, 2003, Hubner *et al.*, 2005, Petretto *et al.*, 2006). Linkage analysis in these studies has determined that the heritability of gene expression is relatively similar across tissues but it does not explain all variation of gene expression (Hubner *et al.*, 2005, Petretto *et al.*, 2006, Price *et al.*, 2011). The effect size of each genetic variant is relatively small thus genetic control of gene expression is likely be polygenic. Initial eQTL mapping in humans has taken advantage of the HapMap populations because of their defined population stratification, the availability of genetic variant data, and the gene expression profiles obtained from lymphoblastoid cell lines (Cheung *et al.*, 2003, Monks *et al.*, 2004, Morley *et al.*, 2004, Stranger *et al.*, 2007a, Veyrieras *et al.*, 2008). The power of these initial studies was relatively small as the population sample size was less than 60.

cis and trans

The physical distance between genetic variants and a gene (typically a transcription start site or probe location) defines cis and trans eQTL. Local eQTL are likely cis acting and distant eQTL are likely trans acting, however distance alone cannot define a mechanism. A cis interaction is a genetic variant found close to the regulated gene (Figure 1.3A). Cis eQTL are enriched near transcription start site (TSS), in regulatory sequences including promoters and enhancers (Gaffney *et al.*, 2012). Both using model organisms and in human studies, cis eQTL have been more readily detectable than trans eQTL (Petretto *et al.*, 2006).

Comparison of eQTL between tissue types or populations finds that cis eQTL are shared more than trans associations (Price *et al.*, 2011, Stranger *et al.*, 2012). The effect size of cis eQTL is often larger than that of trans eQTL, which may be why detection of trans eQTL is difficult (Petretto *et al.*, 2006). Larger sample sizes may help to detect trans

eQTL with a small effect which may otherwise be hidden as noise in small cohorts. Not all detected eQTL explain variation of gene expression, thus many undetectable eQTL with small effect sizes are likely present but cannot be found due to sample size, the number of genetic variants tested and tissue heterogeneity.

A trans association is found at a greater distance to the regulated gene, or on a different chromosome entirely (Figure 1.3B). One explanation for trans eQTL is that a local association of a gene involved with transcription regulation such as a transcription factor may modulate gene expression of a network of genes including on different chromosomes.

Detection of trans eQTL is more difficult than cis eQTL because of the more modest effect size of trans eQTL, and the heterogeneous gene expression profile of tissues. Trans eQTL may also only appear when cells are activated by environmental stimuli according to their cellular functions (Smirnov *et al.*, 2009, Gaffney, 2013).

A trans eQTL that has an association with multiple genes may be proposed as a master regulatory locus (Figure 1.3C). The *KLF14* locus (rs4731702), which is also associated to metabolic phenotypes, is associated with hundreds of genes expressed in adipose tissues (Small *et al.*, 2011). *KLF14* encodes a family of transcription activators and repressors, and this trans master regulatory eQTL has been identified in two independent cohorts (Small *et al.*, 2011).

Genetic markers for eQTL mapping

SNPs are commonly used as genetic markers for eQTL mapping in a similar manner to GWAS. Tagging SNPs can be used to locate primary eQTL signals. Imputation of SNPs in strong LD to the associated variants may enable fine mapping of eQTL signals and permit combining genotyping data from different platforms. However, the chance of finding new eQTL signals by imputation is low because tagging SNPs used for genotyping represent most of common haplotypes. Common SNPs (MAF>5%) have been used for the majority of eQTL analyses due to the small sample sizes involved in order to avoid detection of false positives. CNVs were shown to associate with variation of gene expression (Stranger *et al.*, 2007b).The eQTL signals detected from SNPs and CNVs in the study overlap little, suggesting that different mechanisms may cause genetic variation in gene expression. Polymorphic microsatellites have also been used to map cis eQTL in association with genes with high heritability (Goring *et al.*, 2007).

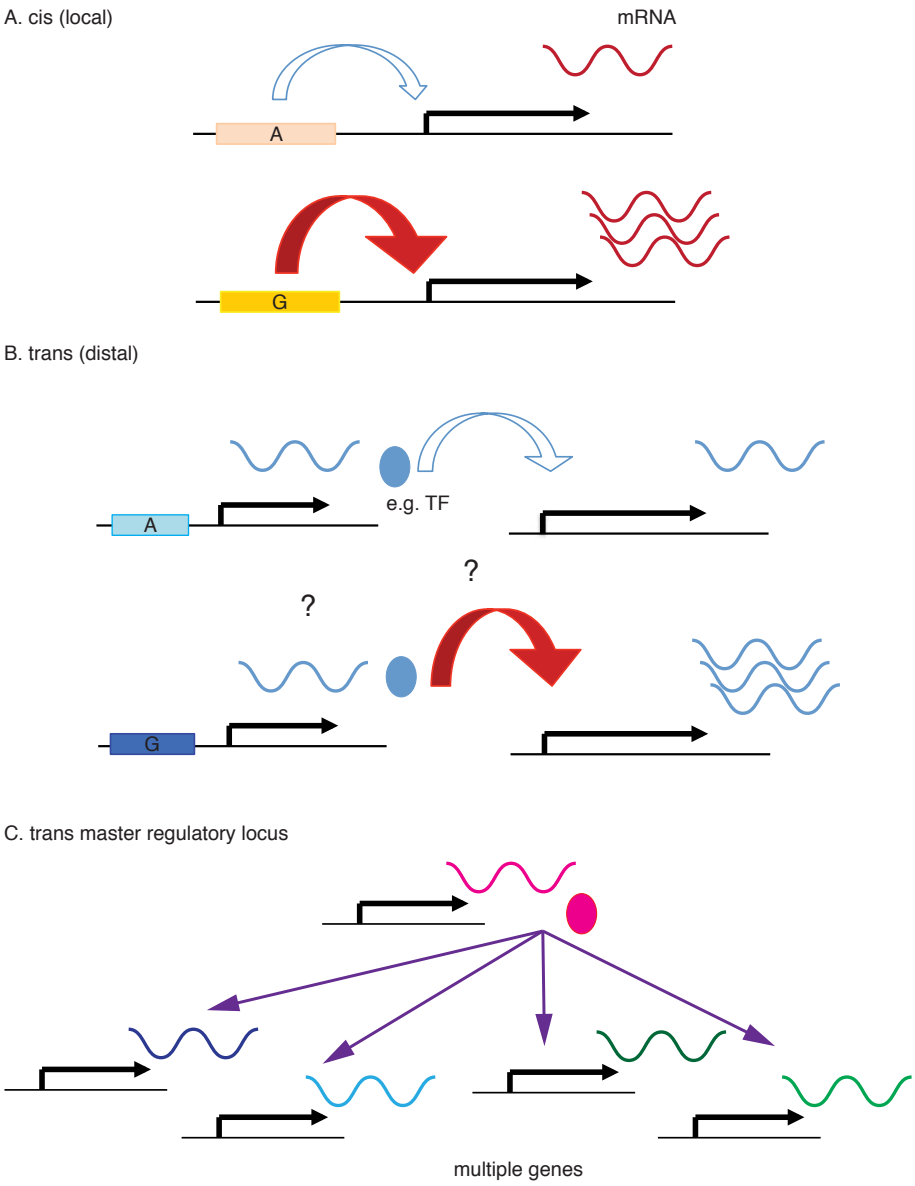


Figure 1.3: Diagram of cis and trans eQTL associations. A. cis eQTL showing association with different alleles found close to a gene transcription start site (TSS). B. trans eQTL association found distally from a gene TSS. C. trans master regulatory locus shown distal association to multiple genes.

LCLs for eQTL analysis

Many eQTL mapping studies have used accessible tissues, mainly from circulating blood cells. Lymphoblastoid cell lines (LCLs) immortalised with Eprtein-Barr virus (EBV) were initially widely used for eQTL mapping in humans due to the large established cell collections from diverse populations, ease of culture and monogeneity together with the availability of genotyping data. LCLs can be informative for gene expression of the B-cell lineage and shared eQTL between LCLs and primary tissues have been found (Bullaughey *et al.*, 2009). Many eQTL are also found across all the cell and tissue types that have been tested, highlighting the value of LCLs.

However, some studies have demonstrated that cell lines are not reliable for eQTL mapping. *In vitro* stimulation of LCLs results in various responses in gene expression because of non-genetic confounders including EBV load and the cell growth rate, which decrease the power to detect true associations between genetic variants and gene expression (Choy *et al.*, 2008). Immortalisation by EBV leads to modification of some innate gene expression profiles (Caliskan *et al.*, 2011). Established LCLs are latently infected and immortally replicate due to prevention of apoptosis by viral proteins (Gregory *et al.*, 1991). Even though the magnitude of gene expression modification of B-lymphocyte by EBV transformation is limited, EBV has been demonstrated to utilise host MHC class II genes as cofactors (Caliskan *et al.*, 2011, Li *et al.*, 1997). Therefore, the extreme monoclonality of cell lines has been shown to be unreliable for allelic specific expression analysis (Plagnol *et al.*, 2008). This evidence implies that studies on LCLs, especially in immune related fields, may not be able to detect many associations with genetic variation.

Cell type specificity in eQTL

eQTL mapping in humans has been conducted using primary tissues and single cell types including whole blood, monocytes, lymphocytes, adipose tissues, osteoblasts, skin, liver, and brain cortex (Fehrmann *et al.*, 2011, Goring *et al.*, 2007, Murphy *et al.*, 2010, Emilsson *et al.*, 2008, Grundberg *et al.*, 2009, Small *et al.*, 2011, Schadt *et al.*, 2008, Innocenti *et al.*, 2011, Myers *et al.*, 2007, Heinzen *et al.*, 2008). Some studies have used multiple tissues from same individuals to map eQTL in a tissue specific manner (Dimas *et al.*, 2009, Bullaughey *et al.*, 2009, Price *et al.*, 2011, Grundberg *et al.*, 2012, Fu *et al.*, 2012, Nica *et al.*, 2011). Sample sizes are generally small due to sample availability. However,

power of these studies may be higher than GWAS because they analyse traits (transcripts) mechanistically closer to genotype than whole organism disease traits.

Tissue or cell specific eQTL have been mapped by comparing eQTL data between different tissue or cell types. A model organism study using rat tissues showed a small proportion of tissue specific eQTL (15%, Hubner *et al.*, 2005). Tissues heterogeneously consist of many different cell types. Therefore, eQTL analyses using tissues have difficulty to find gene expression unique to a single cell type. The purity of cell fractions may also be essential to detect true eQTL associations. A study comparing three human tissue types, fibroblasts, LCLs and T cells, used samples higher purity than tissues and reported that up to 78% of eQTL was tissue specific (Dimas *et al.*, 2009). However, gene expression on LCLs cannot avoid modification by immobilisation as described above. The study of cell differentiation from stem cells to erythroid and myeloid cell lineages has shown the largest number of cell type specific eQTL associations compared to shared eQTL, suggesting dynamic change in gene expression may be associated with tissue independent eQTL (Gerrits *et al.*, 2009). Tissue specific eQTL analysis in a large scale is relatively new and statistical frameworks may be still under development. Taking into account differential gene expression in different cell types, identification of cell type specific eQTL is important to associate with common diseases where tissues are locally affected tissues.

Non-genetic influences on eQTL

The influence of environmental factors make the replication of eQTL mapping results difficult. An eQTL study investigating expression of leukocytes from a Moroccan population living in three different environments, in a city, a mountain and a desert, has estimated that >30% of the variance of a particular transcript can be explained by environmental effect (Idaghdour *et al.*, 2010). eQTL mapping in LCLs exposed to radiation found that more genes associating in trans remained significant (13%) after stimulation compared to genes associating in cis (less than 1%, Smirnov *et al.*, 2009). Appropriate measurement and control of environmental factors may be required to identify all eQTL.

eQTL and disease associated loci

One anticipated result of eQTL mapping studies is the assignment of functions to genetic variants and a greater understanding of the mechanism underlying susceptibility to common disease (Cookson *et al.*, 2009). Annotation of GWAS associated variants with eQTL revealed that trait-associated SNPs are likely to also be eQTL (Nica *et al.*, 2010,

Nicolae *et al.*, 2010). eQTL mapping in disease cohorts has been conducted in an effort to identify eQTL signals related to disease susceptibility, including studies in asthma and obesity (Dixon *et al.*, 2007, Emilsson *et al.*, 2008). Investigating tissues which are directly involved in disease mechanisms may be essential to detect disease associated eQTL as many regulatory variants are highly context specific in their actions. Tissue availability may be limiting but in some cases tractable, for example, adipose and cutaneous tissues are relatively available from individuals with metabolic disorders and psoriasis (Emilsson *et al.*, 2008, Fu *et al.*, 2012, Ding *et al.*, 2010).

There are several cases where disease associated alleles have been shown to associate with gene expression. The functional demonstration for the asthma associated locus of a cis eQTL involving *ORMDL3* showed that the increased expression of *ORMDL3* is associated with induction of inflammation by altering Ca^{+} levels within the endoplasmic reticulum and protein folding (Moffatt *et al.*, 2007, Ferreira *et al.*, 2011, Wan *et al.*, 2012, Cantero-Recasens *et al.*, 2010). High levels of LDL cholesterol are associated with the risk of myocardial infarction (MI). A liver tissue specific cis eQTL to *SORT1* found on the risk allele for MI and LDL cholesterol level has shown that an allelic change in the transcription factor binding site (TFBS) in the promoter of *SORT1* causes hepatic expression differences (Musunuru *et al.*, 2010). *SORT1* encodes Sortilin 1, a protein that helps to degrade very low cholesterol lipoprotein (VLDL) before hepatic VLDL secretion.

Following these successful investigations, eQTL analysis was applied to other GWAS including psoriasis, T2D, and schizophrenia (Ding *et al.*, 2010, Zhong *et al.*, 2010, Jong *et al.*, 2012). Risk alleles for common disease are found in many cases to be eQTL, altering levels and timing of gene expression rather than changing protein functions as seen in Mendelian diseases.

Summary

eQTL mapping has become a common tool to determine the regulatory function of genetic variants. GWAS alleles overlapping with cis eQTL are often enriched in regulatory DNA sequences and are thus relatively easy to associate with gene function. It is often difficult to predict a functional mechanism for the association of trans eQTL to GWAS loci. Environmental factors may need to be included in study designs to assess eQTL more accurately or sensitively, especially in the search for trans eQTL.

1.2.2 Epigenetics

The study of epigenetics, which focuses on modification to DNA structure rather than DNA sequence, has shown that chromatin state and DNA methylation are affected by genetic variation. The term ”epigenetics” was invented by C.H. Waddington to describe the mechanism of the DNA modifications which is heritable through programmed changes in differentiated cell types (Waddington, 1942). Therefore, developmental biology has been focusing on epigenetics in order to investigate cell fate during cell division. Epigenetic changes are also involved in disease, including both monogenic and common diseases (Feinberg, 2007). For example, the epigenetics of cancer has been studied well; tumour cells are often distinguishable based on dysregulated methylation compared to normal cells. The frequency of epigenetic changes are shown to be associated with genetic variation. This section describes how genetic variation influence epigenetic changes.

Regulatory genetic variants

Gene transcription is mainly driven by cis-regulatory sequences, gene promoters, enhancers and insulators. As eQTL analysis has determined, modification of these DNA sequences often results in a change to gene expression. Gene transcription is mainly governed by RNA polymerase II and its co-factors which together form a large molecular complex. Genetic variation in gene regulatory elements often modulates the accessibility of proteins involved in transcription. TFBS often have conserved DNA binding motifs and can be computationally predicted (Sandelin *et al.*, 2004, Tompa *et al.*, 2005).

The *in vivo* interaction of proteins and DNA sequences can be determined by chromatin immunoprecipitation (ChIP)(Solomon *et al.*, 1988). The technique is limited by the quality of the antibody that is used for protein detection and the large quantity of DNA required to carry out only immunoprecipitation for a rare transcription factor (TF) with relatively few DNA binding sites.

Determination of TF binding to a regulatory sequence and how this is modulated by genetic variants through ChIP experiments can help to understand the effect of genetic variants in gene regulation. Maynard and colleagues describe a method to detect allelic imbalance of protein-DNA interactions using heterozygous SNPs as markers at a genome-wide scale (Maynard *et al.*, 2008). Their method determines the heterogeneity of a protein binding to genetic variants using a comparison of SNP genotyping for genomic and ChIP output DNA. ChIP output can be further examined by WGS, called ChIP-seq (Park, 2009). ChIP-seq can be used to detect not only DNA-protein interactions but also the allelic

specificity of the interaction (Mikkelsen *et al.*, 2007, Zheng *et al.*, 2010, Kasowski *et al.*, 2010, Reddy *et al.*, 2012). McDaniell and colleagues have shown that chromatin signatures vary by 10% between individuals and are heritable using DNaseI sensitivity mapping and ChIP-seq for CTCF, a transcriptional and chromatin regulator (McDaniell *et al.*, 2010).

Histone modification

Chromatin structure defines regulation of gene expression (Li *et al.*, 2007). The accessibility of proteins including TFs to DNA regulatory sequences, such as promoters and enhancers, is essential for initiation of gene transcription, which was initially shown by a DNaseI hypersensitive site assay (Scott and Wigmore, 1978, Keene *et al.*, 1981). N-terminal tails of histone proteins are also subject to epigenetic modification including methylation and acetylation (Strahl and Allis, 2000, Nightingale *et al.*, 2006).

Investigation of chromatin signatures at a genome-wide scale using ChIP-seq and DNase-seq, a combined technique of DNaseI hypersensitivity assay and WGS, revealed their plasticity and cell type specificity (Thurman *et al.*, 2012). Loss of histone modifications in tumour cells has been observed as global modification of cells during tumourgenesis, suggesting disruption of normal gene expression (Esteller, 2007). Allele specific chromatin signatures were demonstrated using DNase-seq and CTCF ChIP-seq with LCL trios, of which about 10% of the active chromatin were individual or allele-specific (McDaniell *et al.*, 2010). This study also showed heritability of chromatin states.

DNA methylation

DNA methylation is frequently involved in silencing gene transcription by preventing binding of TFs or leading to formation of heterochromatin (Bird, 2002). Genetic variants on regulatory DNA sequences may modify the susceptibility to DNA methylation. Allele-specific DNA methylation is created by SNPs and leads to allele-specific mRNA expression (Kerkel *et al.*, 2008). Inter-individual differences in DNA methylation have been shown by methylation QTL mapping in human brain tissues and are directly associated with gene expression (Zhang *et al.*, 2010). Heritability of DNA methylation patterns has been demonstrated in mouse models which determined differential allele-specificity and showed that the pattern is important for genetic imprinting and gene regulation (Xie *et al.*, 2012, Schilling *et al.*, 2009).

DNA methylation is shown to be involved in disease susceptibility. Allele-specific mRNA expression in *VNN1* and *CYP2A7* implies to associate with smoking behaviour

and cholesterol levels respectively (Thorgeirsson *et al.*, 2010, Tobacco and Consortium, 2010, Kumasaka *et al.*, 2012, Schadt *et al.*, 2005, Goring *et al.*, 2007). Tumour cells are characterised by both global and gene specific hypomethylation and hypermethylation, which causes the silencing of tumour suppressor genes and activation of oncogenes, leading to higher frequency of tumour progression (Feinberg *et al.*, 2006). A high frequency of promoter hypermethylation of the tumour suppressor gene *BRCA1* is shown to associate with breast carcinoma due to the suppression of *BRCA1* gene expression (Esteller *et al.*, 2000). Some types of cancer risk may be associated with genetic variants which create or disrupt methylation sites in regulatory regions of tumour suppressor genes and oncogenes.

Post-transcriptional regulation

Recent evidence suggests that post-transcriptional regulation is influenced by genetic variation.

Alternative splicing, where different patterns of splicing occur within the same gene, is commonly observed (Modrek *et al.*, 2001, Johnson *et al.*, 2003). Additionally, alternative splicing has been commonly observed with tissue specificity (Wang *et al.*, 2008). Splicing occurs nonrandomly and is an heritable event involved in post-transcriptional regulation (Kwan *et al.*, 2007). Splicing QTL (sQTL) mapping analyses the association of genetic variants and gene expression levels at an isoform level which is measured by an exon tiling array or RNA-seq (Kwan *et al.*, 2008, Lalonde *et al.*, 2011). The assumption of sQTL mapping analysis is that alternative splicing events are modulated by splicing SNP variants by creating splicing junctions and/or modifying binding of a splicing complex. These studies have shown that alternative splicing of isoforms is associated with genetic variants. RNA-seq can quantify the types of isoforms being expressed as well as gene expression levels, therefore the technique can help to detect sQTL associations (Lalonde *et al.*, 2011, Majewski and Pastinen, 2011). sQTL may be functionally explained by proteins involved in splicing (trans association), regulatory motifs in cis, and different TSS.

The mechanism of gene expression inhibition by microRNAs (miRNAs) has been demonstrated to involve miRNAs targeting the 3'UTR on mRNA and facilitating instability of mRNA leading to decay through binding (Winter *et al.*, 2009, Chen and Rajewsky, 2007). miRNA has been implicated in many human diseases especially cancer (Sethupathy and Collins, 2008, Calin and Croce, 2006). Aberrant expression of miRNA in tumour

cells has been shown (Iorio *et al.*, 2005, Volinia *et al.*, 2006). Genetic variants in the 3'UTR as well as the miRNA coding sequence can potentially modulate mRNA decay rate or stability. eQTL analysis using miRNA expression levels has determined that genetic variants are associated with miRNA expression in cis and trans (Borel *et al.*, 2011). eQTL analysis in human adipose tissue has shown that miRNA expression levels were associated with genetic variants in cis but no negative correlation was found between miRNA and its target mRNA level at a genome-wide scale (Parts *et al.*, 2012). A study that measured mRNA decay rates in LCLs has shown that 10% of the genes tested were significantly associated with genetic variants (Pai *et al.*, 2012). The mechanism by which miRNA regulates gene transcripts remains unclear.

The ENCODE project

The ENCyclopaedia Of DNA Elements (ENCODE) project systematically catalogued the functional annotation of non-coding DNA elements at a genome-wide scale using distinct TFs and diverse human cell lines (Birney *et al.*, 2007, The ENCODE Project Consortium; Dunham *et al.*, 2012). The study used more than 300 types of cell lines, primary cells and tissues in order to cover diverse profiles of gene expression. These are classified into three tiers; Tier 1 includes the highest priority set consisting of three cell lines, K562 erythroleukaemia cells, GM12878 LCL and H1 embryonic stem cell (H1 hESC) line. Tier 2 includes the second priority set consisting of HeLa-S3 cervical carcinoma cells, HepG2 hepatoblastoma cells and primary (non-transformed) human umbilical vein endothelial cells (HUVECs) and 12 additional cell types. Tier 3 includes 338 other types. The project gives an insight into the function of non-coding regions which may be involved in the modulation of gene expression by changing the binding affinity for specific TFs, susceptibility to DNA methylation, DNaseI hypersensitive sites, modification of chromatin structure, and non-coding RNA (Thurman *et al.*, 2012, Neph *et al.*, 2012, Djebali *et al.*, 2012, Gerstein *et al.*, 2012, Sanyal *et al.*, 2012). The ChIP-seq experiments for the ENCODE project used 12 histones and 189 TFs (The ENCODE Project Consortium; Dunham *et al.*, 2012). The coverage of the genome sequence reached 80% which was covered by at least one experiment. Their conservative assessment showed 8.5% of the genome sequence was covered by either a TFBS motif (4.6%) or a DNase hypersensitive site (5.7%). The extensive data obtained from the ENCODE project may help assign functions to disease associated loci, although the study cautiously advised that the data may be skewed by the limited number of cell types and TFs used.

1.2.3 Summary

Regulation of gene expression is heritable and influenced by genetic variants, especially involving regulatory elements. eQTL and epigenetic analysis at a genome-wide scale aims to assign regulatory functions to these genetic variants. Genetic variants that associate with gene expression are likely to overlap with disease associated variants. This evidence may contribute to greater understanding of the mechanism of common disease susceptibility.

1.3 Common disease, immunity and inflammation

Inflammation and the immune response are key pathological processes underlying many common diseases including diabetes, IBD, RA, MS, asthma, and cancer. This section describes the biology of these responses to provide context for the subsequent work in this thesis seeking to define genetic modulators of inflammation and the immune response.

1.3.1 Human immune system

The mammalian immune system has been developed to protect host organisms from microbial and viral infection. It is classified into non-specific and specific components. The non-specific component, innate immunity, is equipped with physical barriers, a surveillance system involving circulating immune cells in the blood stream and an antibody independent innate immune system associated with inflammation. The specific component, adaptive immunity, is based on acquired immunity and is antibody dependent. Both immune systems are tightly linked to each other to eliminate microbes and viruses efficiently (Fearon and Locksley, 1996). Immune defects arising from non-synonymous mutations in genes which are involved in the immune system have been highly informative in defining the biology of the immune response.

Innate immune system

Activation of inflammation in the innate immune system is best characterised for microbial infections (Medzhitov, 2007). The recognition of distinctive microbial structures, pathogen associated molecular patterns (PAMPs), is mediated by pattern recognition receptors (PRRs) that are expressed in innate immune cells, especially monocytes and macrophages (Takeuchi and Akira, 2010).

Toll-like receptors (TLRs) are well characterised PRRs of acute inflammation that are expressed on the cell surface or endosomes of peripheral immune cells. For example, TLR4 along with CD14 on monocytes recognises endotoxin lipopolysaccharide (LPS) from Gram-negative bacteria and rapidly induces a signal transduction network to activate an inflammatory response (Kumar *et al.*, 2009, Wright *et al.*, 1990). *TLR4* mutations in humans and mice are associated with endotoxin hyporesponsiveness and failure to activate the signalling pathways (Arbour *et al.*, 2000, Poltorak *et al.*, 1998), indicating that TLR4 is an essential receptor of LPS to activate the innate immune system.

Another PRR family, NOD-like receptors (NLRs), intracellularly recognise bacterial peptideglycan and activate downstream signalling pathways to produce the pro-inflammatory cytokine IL1β (Chen *et al.*, 2009). This signalling pathway activates an inflammasome, a protein complex consisting of caspase 1 and NALP to promote maturation of IL1β (Rathinam *et al.*, 2012). Susceptibility to Crohn's disease is associated with *NOD2* mutations (Ogura *et al.*, 2001a, Hugot *et al.*, 2001). Studies in mice have shown that *NOD2* expression in the intestine is essential to recognise bacterial infection and to regulate the adaptive immunity (Kobayashi *et al.*, 2005)

These PRRs are vital to activate downstream signalling pathways that are involved in production of the pro-inflammatory cytokines and chemokines, and in antigen-presentation to lymphocytes resulting in antibody production.

Adaptive immune system

Adaptive immunity is acquired after birth, based on the history of microbial infection. Antibody production is triggered by the innate immune system which recognises microbial antigens and presents them to lymphocytes. This immunological memory is to prevent further infection by the same pathogens. The system is highly adaptable because of hypermutation in the genetic loci which govern production of antibodies. B-cells, a type of lymphocyte, are vital to the adaptive immune system because they differentiate into different types of B-cells and recognise the antigens of invaders presented by antigen presenting cells (APCs) then produce antibodies. APCs largely express major histocompatibility complex (MHC) class II molecules to present antigens to T cells and B-cells.

1.3.2 Major histocompatibility complex

The MHC region in mammals encodes many genes essential for the immune system and has been investigated for susceptibility to autoimmune and infectious diseases (Trowsdale, 2011). The MHC region was initially studied in inbred mice when investigating immune tolerance and rejection in tissue grafting and transplantation (Klein, 1986). A genetic linkage study in inbred mice demonstrated that the histocompatibility-2 locus determined susceptibility or resistance to tumour transplant (Snell and Higgins, 1951). In humans, the MHC molecules were discovered to be expressed on the surface of leukocytes (Payne and Rolfs, 1958, Bodmer, 1972). Human leukocyte antigen (HLA) genes within the MHC are classified by their roles in immune responses. The major role of molecules encoded by these genes is antigen presentation in leukocytes and transmits information of infectious agents into T cells.

MHC encoded molecules

HLA corresponding to the MHC class I and II genes encodes protein receptors with distinctive roles in the immune systems in different cell types; the class I genes are expressed in most nucleated cells and the class II genes are expressed in professional immune cells including monocytes and dendritic cells (DCs) to form heterodimeric protein receptors. These two classes are encoded in genes clusters on chromosome 6p21 (class I; telomeric, class II; centromeric, Figure 1.4). Genes on the clusters are often co-expressed because they form dimers or complexes in order to function properly (Horton *et al.*, 2004). The classical class I genes include *HLA-A*, *-B* and *-C*, which encode hyperpolymorphic polypeptide chains to present intracellular antigens to $CD8^+$ T cytotoxic cells (Bjorkman *et al.*, 1987). Activation of $CD8^+$ cytotoxic T cells is only triggered where the class I molecules on the surface of APCs present peptides derived from foreign proteins such as viral proteins.

The classical class II genes include *HLA-DP*, *-DQ* and *-DR*. These are expressed on the surface of APCs to present exogenous antigens to $CD4^+$ helper T cells (Jones *et al.*, 2006).The activation of $CD4^+$ helper T cells leads to acquired immunity through several stages of differentiation of helper T cells to effector, memory and regulatory T cells. The non-classical class II genes, *HLA-DM* and *-DO*, are less variable in sequence compared to the classical genes. They are expressed in the cytosol to form heterotetrameric complexes and are involved in antigen processing and loading onto the classical class II molecules (Alfonso and Karlsson, 2000).

The MHC class III genes are encoded in the region between the class I and II gene clusters (Carroll *et al.*, 1984). The gene products include complement proteins (C2, C4 and factor B) and inflammatory cytokines such as tumour necrosis factor (TNF) family, lymphotoxin α and β, together with heat shock proteins and other immune mediators.

Genomic structure of the human MHC

The MHC region in humans is highly polymorphic due to somatic recombination to adapt to numerous microbial infections present in human populations. One of the first physical maps of the human MHC was constructed based on DNA cloning and sequence data from research studies by Campbell and Trowsdale in 1993, which included class I, II and III genes (Campbell and Trowsdale, 1993). In 1999, the complete sequencing of the human MHC showed that the region is 3.6Mbp long and encodes 224 genes including the extended MHC class I and II genes (Consortium, 1999, reviewed in Horton *et al.*, 2004, Figure 1.4). The study also showed that the region has complex haplotypes and high density of genes. The association of the MHC with susceptibility to autoimmune diseases was shown first by serological typing in the 1970s and confirmed by GWAS including the WTCCC (Crohn's disease; Kobayashi *et al.*, 1990, T1D; Cudworth and Woodrow, 1975, and RA; Stastny, 1978, Wellcome Trust Case Control Consortium, 2007). The extreme polymorphism of the MHC region makes GWAS challenging because it is difficult to identify true causal variants for disease phenotypes. To facilitate this, haplotypic analyses of the MHC were conducted to map LD structure and to construct reference haplotype sequences using direct sequencing in HLA homozygous cell lines (Miretti *et al.*, 2005, Bakker *et al.*, 2006, Horton *et al.*, 2008). The results of these studies provided a resource to help with the identification of loci associated with common diseases.

MHC associated diseases

The MHC is associated with many autoimmune and infectious diseases and some disease susceptibility loci are shared by several diseases (Horton *et al.*, 2004, Trowsdale, 2011). For example, class I genes *HLA-A* and *-B* are associated with AS, T1D, celiac disease, psoriasis, MS and HIV (Brewerton *et al.*, 1973, Nejentsev *et al.*, 2007, Stokes *et al.*, 1972, Nair *et al.*, 2006, Jersild *et al.*, 1973, Carrington *et al.*, 1999). The viral load of HIV infection is associated with *HLA* class I alleles; antiviral cytotoxic T cells with specific alleles fail to recognise viral peptides presented by infected cells (Moore *et al.*, 2002). On the other hand, restriction of HIV replication is highly associated with *HLA-B*5701*

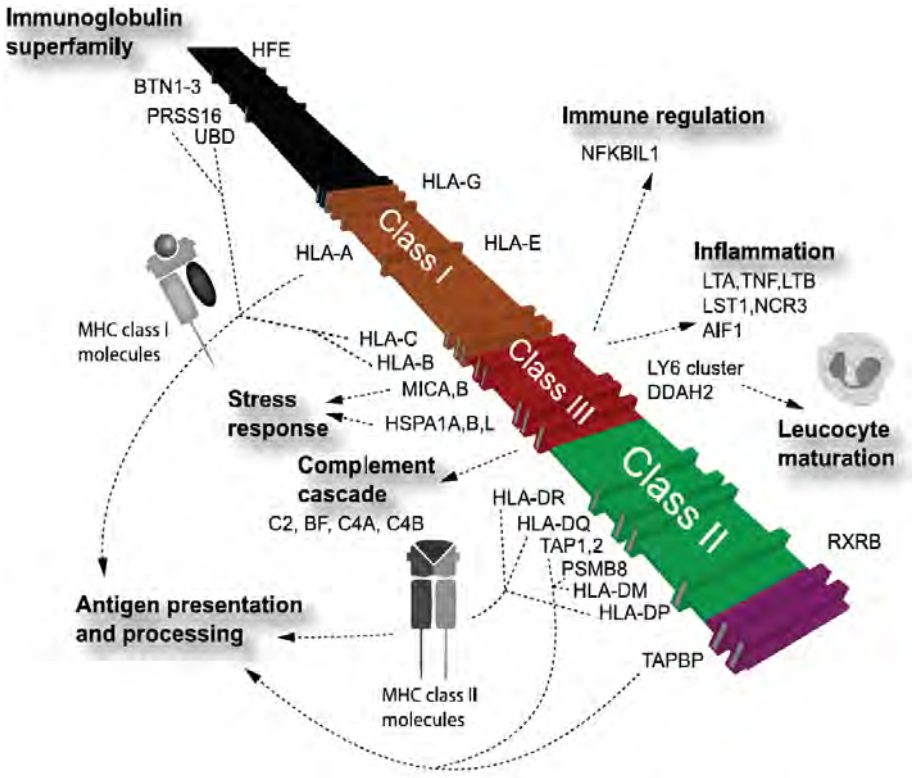


Figure 1.4: Genes encoded in the human MHC. Reproduced from Knight, 2009, Human Genetic Diversity: Functional Consequences of Health and Disease, page 310, by permission of Oxford University Press.

because $CD8^+$ T cells with HLA-B*57 specifically respond to the HIV derived peptide (Migueles *et al.*, 2000). The class II loci, *HLA-DR* and *-DQ* are associated with autoimmune diseases including RA, celiac disease, MS, T1D and IBD (Stastny, 1978, Sollid *et al.*, 1989, Bugawan *et al.*, 1989, Kolstad *et al.*, 1989, Todd *et al.*, 1987, Kobayashi *et al.*, 1990). The class III genes, C2 and C4 complements, are associated with susceptibility to SLE (Rahman and Isenberg, 2008). Functional mechanisms of these associations have yet to be fully understood. However, further understanding of the MHC region by sequencing and haplotypic analysis may help discovery of new loci associated with common diseases.

1.3.3 Inflammatory response

The role of inflammation in normal immune response

Inflammation is a physical response to a cellular insult, most commonly in the form of infection. Inflammation is triggered where tissue is damaged by infection, autoimmunity or physical injury. The classic inflammatory response is defined as a defence to protect the body from infected and/or damaged tissues in an acute manner (Medzhitov, 2008).

Inflammatory pathway

The inflammatory response can be dissected into 3 parts (inducers, sensors and mediators) in order to understand types of inflammatory response (Medzhitov, 2008). Exogenous inducers are typically derived from microbial (PAMPs, such as LPS, peptidogylcan) or non-microbial (e.g. allergens, irritants) sources. Endogenous inducers are less well defined but are produced by signals from damaged or stressed, necrotic tissues which release cellular constituents such as ATP and potassium ion efflux. A gene product of *NALP3* responds to both exo- and endogenous inducers which include pore-forming exotoxin from Gram-positive bacteria and ATP released by disruption of the plasma membrane. The protein then activates the inflammasome, a cytosolic protein complex, to produce the pro-inflammatory cytokines IL-1β and IL-18 (Mariathasan *et al.*, 2006).

Sensors, mainly PRRs, recognise inducers and activate the production of mediators such as the pro-inflammatory cytokines TNFα, IL6 and IL1β (Nathan, 2002). Acute inflammation, including local vasodilation of the blood endothelial cells, is initiated by these inflammatory mediators for extravasation of neutrophils. Inducing inflammation in the affected tissues by pro-inflammatory mediators alters the functionality of the tissues and downstream signalling pathways. Leukocytes are then recruited into the inflamed tissue. Inflamed tissue may recover to a normal state, or may undergo apoptosis in order to maintain tissue homeostasis.

Resolution of inflammation

Resolution of inflammation is often more important than the initiation of inflammation because unresolved inflammation may lead to tissue damage and disrupt tissue homeostasis (Serhan and Savill, 2005). The inflammatory response is a tightly regulated mechanism since overproduction of pro-inflammatory cytokines would cause tissue damage and a pathological state where sepsis, autoimmune disorders, chronic inflammation or cancer can occur in the absence of infection (Nathan and Ding, 2010). The fine balance between triggering and resolving inflammation is therefore an important mechanism for our survival. Once inflammation has ceased, activated macrophages clear invading agents and the infiltrated neutrophils in the inflamed tissue. Anti-inflammatory cytokines such as TGFβ and IL10 are produced and tissue repair is facilitated by macrophages.

Macrophages during inflammation

Leukocytes circulate in the peripheral blood stream normally for surveillance, and migrate to the site of inflammation once sensors transmit signals of infection via mediators. Neutrophil granulocytes are one of the first responders of acute inflammation to migrate into the site of inflammation. Residential and circulating monocytes are attracted to the peripheral site by pro-inflammatory stimuli and then migrate transendothelially, where they differentiate into macrophages or myeloid DCs (Ingersoll *et al.*, 2011).

Monocytes have a heterogeneous cell lineage and take active roles in both innate and adaptive immunities. They are involved in the initiation of acute inflammation, clearance of pathogens and dead cells, tissue repair, and initiation of antibody based immune responses. The heterogeneous population of mature monocytes constitute 5-10% of peripheral mononuclear blood cells (PBMCs). The subsets of monocytes are typically classified by their expression of particular cluster of differentiation (CD) markers, CD14 and CD16 (Passlick *et al.*, 1989). Mouse models together with *in vitro* and *in vivo* studies have attempted to determine monocyte differentiation but it is difficult due to the heterogeneity of the monocyte population (Geissmann *et al.*, 2010). Classical monocytes ($CD14^{++}CD16^{-}$ also expressing chemokine receptor 2 (*CCR2*)) typically migrate into inflamed tissues. $CD14^{+}CD16^{+}$ monocytes have features that resemble tissue macrophages (Ziegler-Heitbrock *et al.*, 1993). Monocytes expressing $CD16^{+}$ have been shown to preferentially become DCs (Randolph *et al.*, 2002).

Inflammatory macrophages have pleiotropic roles during acute inflammation. They sense non-self through PRRs, express chemotaxis and pro-inflammatory cytokines, become phagocytes to clear invading agents and cellular debris, repair damaged tissues, and present antigens as APCs with the expression of the MHC class II molecules to transmit the antigen information to T lymphocytes in order to activate the adaptive immunity (Murray and Wynn, 2011).

Macrophages reside in specific tissues as subpopulations such as osteoclasts (bone), alveolar macrophages (lung), kupffer cells (liver), microglia (central nervous system, CNS) and adipose tissue macrophages (adipose tissue) (Gordon and Taylor, 2005, Pollard, 2009). The role of these mature macrophages is to maintain tissue homeostasis and their dysfunction has been associated with disease (Pollard, 2009, Murray and Wynn, 2011). For example, adipose tissue macrophages which are altered tissue macrophage to classically activated state and cause inflammation are involved in obesity and insulin resistance in T2D (Olefsky and Glass, 2010). Tissue macrophages are involved in tumourigenesis

(Grivennikov *et al.*, 2010). While inflammatory macrophages activate tumour-killing mechanisms, alternatively activated macrophages suppress such mechanisms and promote tumour growth (Ruffell *et al.*, 2012).

Nonresolving inflammation

Innate immunity underlies common diseases as a result of the pathological status of inflammation. Normally anti-inflammatory molecules promote resolution of inflammation after successful elimination of invading agents (Serhan and Savill, 2005). However, nonresolving inflammation persists as chronic inflammation which is distinct from acute inflammation. The mechanism of chronic inflammation, which appears in autoimmune and chronic inflammatory diseases in the absence of inflammation inducers, is incompletely understood compared to infection triggered inflammation. Restoration of homeostatic status can be disrupted by many factors including dysregulation of pro-inflammatory cytokines, persistent presence of self-antigens, abnormal proliferation of lymphocytes and failure of tissue-repair response (Nathan and Ding, 2010). Switching the inflammatory phenotype of macrophages to tissue repair or an anti-inflammatory state is shown to be important in resolving inflammation (Arnold *et al.*, 2007).

Diseases caused by dysfunction of inflammation

In many common diseases, there is strong evidence that inflammation can contribute to the disease phenotype. Persistent inflammation is observed in many autoimmune diseases. Beyond inflammatory myeloid cells, some lymphocyte lineages have also been implicated in regulation and inflammation of the innate immune response (Walker *et al.*, 2013). Nonresolving inflammation in the artery, which causes atherosclerosis, often results in cardiac related death including either MI or coronary heart disease (CHD)(Ross, 1993). Atherosclerosis development is associated with accumulation of LDL which is retained by macrophages and other inflammatory cells, T cells and mast cells causing pro-inflammatory phenotypes in artery (Tabas, 2010). RA is associated with production of pro-inflammatory cytokines including IL-17 by Th17 (Treg), and production of autoantibodies by B-cells (Choy, 2012). The pathogenesis of IBD, including Crohn's disease and UC, is associated with defective responses in both innate and adaptive immunity (Khor *et al.*, 2011, Geremia *et al.*, 2013). Chronic inflammation in the gut may result from defective innate microbial recognition, autophagy, dysregulation of T cell response and T cell differentiation (Ogura *et al.*, 2001a, Glocker *et al.*, 2009, Hsu *et al.*, 2007, Rioux *et al.*,

2007, McCarroll *et al.*, 2008, Khor *et al.*, 2011). T1D is characterised by autoimmune mediated destruction of insulin producing β-cells in the pancreas and failure to produce insulin (Bottazzo *et al.*, 1985, Belle *et al.*, 2011). T2D, on the other hand, develops due to insulin resistance in the cells for glucose intake. The majority of T2D cases are related with environmental factors such as obesity and diet. Recent studies in T2D have started to understand the role of inflammation, observed as elevated levels of pro-inflammatory cytokines, in circulating blood and adipose tissues (Donath and Shoelson, 2011). As mentioned previously inflammatory macrophages are involved in adipose tissue inflammation.

B-cell involvement in inflammation

B-cells are vital for humoral immunity. B-cell involvement in autoimmune disease and inflammation has been demonstrated not only in antibody production and antigen presentation to T cells but also in inflammation suppression. Autoantibodies produced by B-cells have often been predicted as a cause of pathogenic autoimmune responses (Edwards *et al.*, 1999). Regulatory B-cells (B_{regs}) are a subset of B-cells which have shown suppressive function on Th1 differentiation and activation of macrophages during the inflammatory response (Yang *et al.*, 2013). Differentiation of B_{regs} is speculated to occur via stimuli from activated $CD4^+$ T cells (Mauri, 2010). The association of inflammation suppression by B_{regs} with autoimmune disease has been shown in mice models for RA, diabetes and autoimmune encephalomyelitis (Yang *et al.*, 2013).

1.3.4 Genetic determinant of inflammation

How our genetic makeup may influence inflammation remains unclear. Many multifactorial diseases share inflammatory and autoimmune features (Cho and Gregersen, 2011). Adaptive immunity has been the focus of research in autoimmune diseases, however, recent studies have shown that innate immunity is also involved in disease development. The evidence for macrophage involvement in autoimmune disease is a good example as previously mentioned. Obesity promotes activation of inflammatory macrophages and production of proinflammatory cytokines, which leads to insulin resistance of the targeted tissue (Olefsky and Glass, 2010). There are few autoimmune diseases whose causal genetic mutations have been identified. Autoimmune lymphoproliferative syndrome (ALPS) has been shown to have functional mutations in the TNF receptor superfamily *FAS* gene which encodes a cell-surface receptor to trigger apoptosis (Holzelova *et al.*,

2004). Somatic mutation of the *FAS* gene identified in patients with ALPS were shown resistance to normal apoptosis of precursor T cells which have no T cell receptors. Several mutations on the genes that are associated with *NOD2* signalling pathway have been identified to contribute to Crohn's disease (Geremia *et al.*, 2013). Monogenic mutations are informative for rare diseases but not common diseases associated with the inflammatory response. Here the MHC loci plays a key role, notably in autoimmunity but non-MHC loci are also important, implying polygenic aetiology of common diseases. For example, AS is associated with *HLA-B27* locus as higher than 90% of the patients carry the locus (Schlosstein *et al.*, 1973, Brewerton *et al.*, 1973). GWAS for AS confirmed this MHC association and identified non-MHC loci including *ERAP1* coding variants. *ERAP1* encodes endoplasmic reticulum amino peptidase 1 to trim HLA class I binding peptides in endoplasmic reticulum for antigen presentation (Hammer *et al.*, 2006). Synergistic effects between HLA-B27 and ERAP1 have been proposed as a cause of AS (Keidel *et al.*, 2013).

Autoimmunity and inflammation are complex quantitative traits. GWAS has identified variants that are associated with susceptibility to inflammatory and infectious diseases including IBD, psoriasis, SLE, celiac disease, tuberculosis and leprosy (Barrett *et al.*, 2008, Strange *et al.*, 2010, Hom *et al.*, 2008, Dubois *et al.*, 2010, Thye *et al.*, 2010, Zhang *et al.*, 2009). Loci identified by GWAS may help to partially explain missing heritable components of autoimmunity and inflammatory disease (Manolio *et al.*, 2009). Although the specific functional variants in most diseases are yet to be determined, the evidence that the inflammatory response is controlled at not only signalling pathways but also gene transcriptional levels by regulatory variants supports the hypothesis that genetic variants may contribute to susceptibility to common inflammatory diseases (Foster *et al.*, 2007, Medzhitov and Horng, 2009).

1.4 Summary

Since the first complete human genome sequence in 2001, the shift from candidate gene studies to genome-wide scale association studies in common disease traits has identified many loci associated with disease susceptibility. These loci include not only coding regions but also, most commonly, non-coding DNA sequences. Analysis of genetic modulators of gene regulation and functional genomic profiling have become essential to follow up GWAS and understand the functional consequence of genetic variation.

1.5 Aims of this thesis

The objective of this thesis is to understand how genetic variation may influence gene expression in the context of specific immune cell lineages and human immune responses, and ultimately determine the basis of inflammatory disease phenotypes. In order to achieve this, a cohort of healthy volunteers of European ancestry was established to map genetic modulators of gene expression in human primary immune cell types by adopting an eQTL approach.

Specific objectives and overview

1. **To characterise the transcriptional response for two primary immune cell populations, B-cells and monocytes.** Chapter 3 describes differential gene expression analysis between human primary immune cells, $CD19^+$ B-cells and $CD14^+$ monocytes. This includes a comparison of genes differentially expressed in B-cells and monocytes due to their cell type specific roles in human immune systems.

2. **To determine cell type specificity in the genetic determinants of gene expression for primary immune cell populations.** Chapter 4 describes one of the first large scale eQTL mapping studies using highly purified primary cells, $CD19^+$ B-cells and $CD14^+$ monocytes. Comparison of cis and trans eQTL associations between two cell types shows a high proportion of cell type specific associations and the presence of master regulatory loci in trans.

3. **To determine the influence of cell differentiation on eQTL observed in human monocytes/macrophages.** Chapter 5 describes paired eQTL analysis of primary human $CD14^+$ monocytes and differentiated macrophages. The analysis of this dataset shows that a large proportion of genes are highly differentially expressed. The eQTL analysis shows a relatively large proportion of the unique associations to each cell type in both cis and trans and that a master trans regulatory locus is retained after macrophage differentiation.

4. **To functionally characterise a master regulatory locus identified through trans association mapping.** Chapter 6 describes the functional characterisation of the monocyte eQTL involving *LYZ* that was detected as a master regulatory locus in both cis and trans. The detailed analysis of this locus defines potential regulatory elements and the enhancer activity of the locus was experimentally determined in an allelic specific manner.

5. **Discussion and future work** Finally the results from the proceeding chapters will be summarised and discussed. The impact of the results on our understanding of genetic variants and their roles in modulating gene expression for immune cells is assessed and future directions for further work are outlined.

2 Materials and Methods

2.1 PBMCs purification

50 mL whole blood was collected into anticoagulant EDTA-containing blood collection tubes (Vacutainer system, Becton Dickinson). Peripheral blood mononuclear cells (PBMCs) were purified by Ficoll gradient using Ficoll-paque (cat# 17-1440-02, GE Healthcare, Piscataway, NJ, USA) and 50 mL of whole blood (Boyum 1976). PBMCs were washed three times in Hanks buffered saline solution (HBSS) without Ca^{2+} and Mg^{2+} (cat# 14175-129, Invitrogen, Carlsbad, CA, USA) and the number of cells was counted by haemocytometer.

2.2 Immune cell separation

Magnetic activating cell sorting (MACS, Miltenyi Biotech) was used for all separations. Monocytes ($CD14^+$) and B-cells ($CD19^+$) were positively separated using anti-CD14 and anti-CD19 conjugated microbeads respectively (CD14 MicroBeads monocyte positive separation kit; cat# DS130-050-201, CD19 MicroBeads B-cell positive separation kit; cat# DS130-050-301, Miltenyi). 20 million PBMCs were used in each cell type separation where available. $CD19^+$ B-cells were subsequently resuspended in RLT reagent supplemented with β-mercaptoethanol (Qiagen) for storage and RNA extraction. $CD14^+$ monocytes were counted by the automated cell counter (Countess, cat# C10227, Invitrogen) and $5x10^5$ cells were resuspended in RLT for RNA extraction. The remaining cells were used for stimulation experiments.

2.3 Genomic DNA extraction

Genomic DNA was extracted from 1-2 mL whole blood using the Gentra PureGene Blood kit following the manufacturers instruction (cat# 158467, Qiagen). The protocol was based on a salting out protocol for human DNA extraction. Briefly, red blood cells from 1-2 mL whole blood were lysed in order to pellet leukocytes. Leukocytes were lysed with cell lysis solution with proteinase K at 100 µg/mL for at least 2 hours. Then proteins were precipitated with protein precipitation solution and DNA was precipitated with isopropanol. The DNA was quantified on the Nanodrop (Nanodrop 1000, Thermo Scientific) and PicoGreen dsDNA quantification assay (cat# 11496, Invitrogen).

2.4 RNA extraction

Total RNA was extracted using the RNeasy mini kit from cells stored in RLT (range approximately 1x10^5 to 5x10^5 cells, cat# 74106, Qiagen). Cells in 350 µL RLT reagent were lysed and total RNA was captured in a RNeasy mini spin column. After several washes the total RNA was eluted in 30 µL nuclease free H_2O. Total RNA was quantified by Nanodrop or Bioanalyzer following the manufacturers instruction (Bioanalyzer RNA 6000 Nano kit, cat# 5067-1511, Agilent Technologies). RNA quality was scored based on peaks of 28S and 18S ribosomal RNA and a trace of electrophoresis (Imbeaud *et al.*, 2005).

2.5 Cell culture

Monocytes were cultured at 4x10^5 cells/mL in 5 mL polypropylene cell culture tubes (cat# 734-0447, Becton Dickinson) in a 400 µL volume for the gene expression study, and at 2.5x10^4 cells/mL for *ex vivo* ligand stimulation in round bottom 96-well polystyrene tissue culture plates (cat# 734-0027, Becton Dickinson) for 200 µL volume in RPMI 1640 growth medium (cat# R-0883, Sigma-Aldrich) supplemented with 2 µM L-glutamine (cat# G-7513, Invitrogen), 100 U/mL penicillin, 0.2 mg/mL streptomycin (cat# P-0781, Invitrogen) and 20% v/v Gold standard fetal calf serum (PAA Laboratories GmbH, Pasching, Austria). Monocytes were cultured at 37 °C in 5% humidified CO^2 overnight prior to cell stimulation experiments. Macrophages were differentiated by culturing 4x10^5 $CD14^+$ monocytes as described with the addition of TNF-α at 0.5 ng/mL (TNF recombinant human protein, PHC3015L, Invitrogen) and granulocyte-macrophage colony stimulating factor (GM-CSF) at 0.5 ng/mL (GM-CSF recombinant human protein, PHC2015, Invitrogen) at 4x10^5 cells/mL for 18 days, replacing the culture media every 3-4 days.

2.6 Replication of eQTL analysis by quantitative polymerase chain reaction (qPCR)

Single strand complementary DNA (cDNA) was synthesised by reverse transcription reaction using the SuperScript III First-Strand Synthesis System following the manufacturer's instructions (cat# 18080-051, Invitrogen). Briefly, a 10 µL volume reaction with approximately 100-300 ng of total RNA, 50 µg of random hexamers, 10 µmol of dNTPs, and ionised water up to 10 µL was incubated at 65 °C for 5 minutes and chilled on ice at least 1 minutes. The 10 µL enzyme reaction mix consisting of 200U of SuperScript III reverse transcriptase, 2 µL of 10x RT buffer, 4 µL of 25 mM $MgCl_2$, 2 µL of 0.1 M DTT, 40U of RNAseOUT enzyme was added to the cDNA mix and it was incubated at 25 °C for 10 minutes, at 50 °C for 50 minutes, at 85 °C for 5 minutes, and then chilled on ice at least for 1 minute. 1U of RNase H was added in the reaction mix and incubated at 37 °C for 20 minutes.

qPCR was performed using 1 µL of 1:5 diluted cDNA in a 12.5 µL reaction with SYBR Green Supermix using CFX96 Real-Time PCR Detection System following manufacturer's instructions (iQ SYBR Green Supermix; cat# 170-8885 and CFX96; cat# 185-5096, BioRad

Laboratories). Oligonucleotide primers used for PCR are listed in Table 2.1. Briefly, a 12.5 µL volume reaction with 6.25 µL of 2x SYBR Green Supermix, 0.5 µL of each 6.25 µM forward and reverse primers, 1 µL of diluted cDNA and 4.75 µL of ionised water was run qPCR using CFX96 Real-Time PCR Detection System (BioRad). PCR cycling used is initial denature at 95 °C for 3 minutes, 40-50 cycles of 95 °C for 10 seconds, 60 °C for 20 seconds, 72 °C for 20 seconds. Relative gene transcript levels were determined by the $\delta\delta$-Ct method with *ACTB* expression used as the reference base line (Livak and Schmittgen, 2001).

Name	Sequence	Tm (°C)	Length
BCL2L13_fwd	TTAAGACTGAGGCTGCTCTGC	59.9	21
BCL2L13_rvs	GGATCAGCCAGAGGAAAGAA	59.4	20
BLK_3UTR_fwd2	GAGGTCATCCGCAACCTG	60.2	18
BLK_3UTR_rvs2	ACTGCAGGAACTCGAAGGTG	60.4	20
CD40_3UTR_fwd	GCTCCAGTGCAGGAGACTTT	59.6	20
CD40_3UTR_rvs	CTCTGGCCAACTGCCTGT	59.9	18
CD52_fwd	CTCAGGACAAAACGACACCA	59.7	20
CD52_rvs	GCTGAGACGTGTCACCTCAA	60	20
CD14_fwd	TCTCTGTCCCCACAAGTTCC	60.1	20
CD14_rvs	CTTGGCTGGCAGTCCTTTAG	60.0	20
CD68_fwd	CAGGGGTCTTTGGGCAAAG	62.9	19
CD68_rvs	CCCGGCTTTCTTTTTGTTCT	60.6	20
DFNA5_fwd	TGACTGAAAATGAACTGTAAATGC	58.4	24
DFNA5_rvs	GGCAAGGTCACATAAAATGC	58.1	20
LITAF_3UTR_fwd	AAACACCTTCATCATGGAACTC	58.0	22
LITAF_3UTR_rvs	GTTCTGGTTCCCAAGCATCC	61.8	20
LYZ_fwd	TCGATCTCCTGACCTTGTGA	59.3	20
LYZ_rvs	TGGGCATAATCGTATCATTCA	58.9	21
SELL_3UTR_fwd	AAATGGAAGCTGAAGGAACTG	58.4	21
SELL_3UTR_rvs	TAAATCGCCCTTGGTGAAAG	60.1	20
VWA5A_fwd	GCCTGAAAGACTTGGGTTGA	60.2	20
VWA5A_rvs	AGGGCCTGACTTAGAGGACA	58.9	20

Table 2.1: Oligonucleotide sequences used for qPCR

2.7 Flow cytometry

Primary cells were resuspended in phosphate buffered saline (PBS) with 10% fetal bovine serum (FBS) at 4 °C. Primary antibodies labelled with fluorophores (anti-CD14 PE-Cy7 conjugated; cat# 25-0149-41, anti-CD16 FITC conjugated; cat# 12-0569-41, eBioscience) were incubated with the primary cells at 4 °C for 30 minutes. The cells were washed with PBS with 10% FBS for three times and resuspended again in the buffer. Flow cytometry was performed in the flow cytometry facility in the Old Road Research Building using the Dako CyAn flow cytometer (Dako).

2.8 Cloning

Cloning plasmids used for in this study are shown in Table 2.3.

PCR condition

Target DNA sequences were amplified from approximately 40-80ng human genomic DNA using the polymerase chain reaction (PCR) (1x Phusion HF buffer, 200 µM dNTPs, 0.5 µM forward and reverse primers, 4% DMSO, 0.4U Phusion HighFidelity DNA polymerase (all NEB)) using PCR cycling of initial denaturation 98 °C for 30 seconds, followed by 30-40 cycles of 98 °C for 10 seconds, 63 °C for 15 seconds, 72 °C for 45 seconds, and final extension 72 °C for 10 minutes. The oligonucleotide primer sequences were shown in Table 2.2.

TOPO cloning

The PCR products were purified using a gel extraction kit according to the manufacturer's instruction (Qiagen). Briefly, excised gel fragments were solubilised in a DNA-binding buffer and DNA was isolated on a column and washed, then eluted in 30 µL pH 8.0 Tris-HCl elution buffer. 2 µL of the purified PCR products were cloned into the pCR-Blunt II TOPO cloning vector (Invitrogen) following the manufacturer's instruction. Briefly, a 3 µL reaction of 2 µL purified PCR product, 0.5 µL Salt solution (1.2M NaCl and 0.06M $MgCl_2$), 0.5 µL pCR-Blunt TOPO vector master mix was incubated for 5 minutes at room temperature. The ligation reaction was incubated with 50 µL of Alpha-Select Bronze Efficiency chemically competent *E.coli* (cat# BIO-85025, Bioline) on ice for 30 minutes. Following this incubation the cells were heat shocked at 42 °C for 45 seconds and immediately returned on ice for at least 1 minute. 250 µL S.O.C. medium (2% w/v tryptone, 0.5% w/v Yeast extract, 8.65 mM NaCl, 2.5 mM KCl, 10 mM $MgCl_2$, 20 mM glucose) was added and the cells were incubated with shaking for 1 hour at 37 °C. After this incubation they were plated onto LB agar plates with appropriate antibiotics (kanamycin at 50 µg/mL and carbenicillin at 100 µg/mL) and incubated overnight at 37 °C.

Miniprep

The following day, single colonies were picked and grown in 3 mL of LB medium with appropriate antibiotics overnight with shaking at 37 °C. The following day, transformed plasmids were isolated from 2 mL using a miniprep kit (Qiagen) following the manufacturer's instructions. Briefly, *E.coli* cells were pelleted and resuspended in a resuspended solution and lysed in a lysis buffer. After neutralising by adding a neutralising buffer the cell debris were pelleted and plasmid DNA was captured using a DNA binding column and was washed with an ethanol containing wash buffer and eluted in pH8.0 Tris-HCl elution buffer. Miniprepped plasmids were analysed by restriction digestion using EcoRI enzyme to determine the size of the DNA insert.

Sequencing

The sequences of the DNA inserts were confirmed by a conventional dye-terminator sequencing method (BigDye Terminator V3.1, Applied Biosystems). Briefly, a 10 µL reaction mix containing 0.5 µL of Ready Reaction Premix, 1.75 µL of 5x BigDye Sequencing buffer, 3.2 pmol of a primer and approximately 50-100ng of a plasmid DNA was run using a conventional thermal cycler for initial denaturation at 96 °C for 1 minute, 25 cycles of 96 °C for 10 seconds, 50 °C for 5 seconds and 60 °C for 4 minutes. A sequencing reaction was stopped by adding 2.5 µL of 125 mM EDTA followed by ethanol precipitation of DNA for 15 minutes incubation at room temperature. DNA was pelleted by centrifugation at 3000x*g* for 30 minutes at 4 °C and was washed in 70% ethanol. After pelleting by centrifugation at 1650x*g* for 15 minutes at 4 °C, DNA was air dried for 10-15 minutes at room temperature and run electrophoresis on an ABI PRISM 3700 DNA analyser (Applied BioSystems). Sequence chromatograms were analysed on a VectorNTI program (Invitrogen).

Subcloning

The pGL3-promoter and the TOPO plasmids with the inserts were digested by restriction digestion using KpnI and NheI for 5' and 3' ends respectively in a 40 µL reaction volume of 10U KpnI, 10U NheI, 100 µg/mL BSA, 1x NEB reaction buffer 1, 0.5-1µg plasmid DNA and incubated for 1-2 hours at 37 °C (NEB). The digested DNA was gel purified as described above. The ligation reaction was carried out using T4 ligase in a 10 µL reaction volume with 0.5 µL Quick T4 DNA Ligase, 1x ligation buffer and 25 µg of 1:3 molar ratio of a vector and an insert DNA (Quick Ligation kit, cat# M2200, NEB). It was incubated at room temperature for 5 minutes and was transformed to *E.coli* competent cells as described above. The inserts were determined by minipreps followed by restriction digestion as described.

Midiprep

After confirmation of the inserts in pGL3-promoter, all plasmids were purified using an endotoxin-free midiprep kit following the manufacturer's instructions (cat#12243, Qiagen). Briefly, 50 mL of overnight LB *E.coli* culture was pelleted, resuspended, lysed and neutralised. The supernatant of the cell lysate was added an endotoxin removal reagent and was incubated on ice for 30 minutes. DNA was precipitated in isopropanol, was washed in 70% ethanol and eluted in pH8.0 10 mM Tris- 1 mM EDTA buffer. Their yields were quantified using the Nanodrop.

2.9 Transfection

Lipofectamine

Transfection of luciferase assay plasmids to K562 human chronic myelogenous leukaemia cell line (ATCC Number CCL-243), Jurkat acute T cell leukaemia cell line (ATCC

Table 2.2: Cloning primers used for *LYZ* luciferase assays.

eSNP	Name	Size (bp)	Sequence (5' to 3', forward and reverse)
rs10784774	pLYZ 2kb	1961	TCAGGCGGTACCAGAGCTGGGTGACTGGCTTA
			GATAAAGCTAGCAGCATGATTTGAGGGCAGAG
	p300 TFBS	367	GAGGTACCTCTGGGGCTTAACTGTTCCA
			AGCTAGCATGGGCACACAGGTTTCTCT
rs1384 and	3'UTR LYZ	1463	TCAGGCGGTACCGGTGGAGGTTGCAGAGGAT
rs12812860			GATAAAGCTAGCACAAGGCATGTGTTGCAAAG
rs10878967	pLYZ 1.5kb	1545	TCAGGCGGTACCAATTATAGGCGCACCACCAC
			GATAAAGCTAGCACAAGCCCCAGAACAATGAG

Table 2.3: Cloning plasmids used for *LYZ* luciferase assays. Kan - kanamycin, Amp - ampicillin.

Name	Size (bp)	Selection marker	Description
pCR II Blunt	3519	Kan	TOPO cloning vector (Invitrogen)
pGL3-promoter	5010	Amp	Firefly luciferase assay vector (Promega)
pRL-TK	4045	Amp	Renilla luciferase assay vector, internal control (Promega)

number TIB-152, Schneider *et al.*, 1977) was performed using Lipofectamine LTX as the transfection reagent following the manufacturer's instruction (cat# 15338-100, Invitrogen). Briefly, K562 cells were replaced in a fresh medium the day before transfection. 400 ng of the pGL3-promoter plasmid and 100 ng of the reference plasmid pRL-TK were diluted in 100 µL of Opti-MEM I Reduced Serum medium (cat# 31985-062, Invitrogen) and incubated with 0.5 µL of PLUS reagent for 5 minutes at room temperature followed by incubation with 1.5 µL of Lipofectamine LTX reagent for 25 minutes at room temperature. The DNA-lipofectamine LTX complexes were directly added to 500 µL of cells at a density $5x10^5$ cells/mL in 24-well polystyrene plates. Replicates were performed for each transfection. The transfected cells were incubated for 24 hours before harvesting for the luciferase assay.

Nucleofection

Transfection of luciferase assay plasmids to THP-1 human acute monocytic leukaemia cell line (ATCC number TIB-202, Tsuchiya *et al.*, 1980) and GM18057 the HapMap CEU LCLs (Dausset *et al.*, 1990) was performed using a Nucleofector device (cat# AAB-1001, Lonza) and a Cell Line Nucleofector kit V following the manufacturer's instructions (cat# VVCA-1003, Lonza). The Nucleofector technology uses an electrical pulse to create small pores in cell membranes to deliver nucleic acids to the cytoplasm. Briefly, cells were replace in a fresh medium the day before transfection. $1x10^6$ cells were resuspended in 100 µL Buffer V and added with 1.6 µg of the pGL3-promoter plasmid and 0.4 µg of the reference plasmid pRL-TK. The cell-DNA mixture was immediately transferred to a microcuvette with 2 mm gap and was electric pulse shocked using a V-001 program for THP-1 and X-001 program for GM18057. The cells were added with 500 µL of a supplemented medium and transferred to 12-well polystyrene plates supplemented with 1 mL of a supplemented

medium. The transfected cells were incubated for 16 hours before harvesting for the luciferase assay.

2.10 Luciferase assay

The transfected cells were washed twice using 500 µL PBS pH 8.0. They were incubated with 100 µL 1x passive lysis buffer (PLB, Promega) for 15 min with rocking at room temperature. The supernatants were harvested for the luciferase assay. 30 µL of the PLB supernatants were mixed with 100 µL Luciferase Assay Reagent II (LAR II, Promega) and immediately measured for 10 seconds for pGL3 Firefly luciferase reporter expression activity using a luminometer (Safire2 microplate reader, Tecan). Subsequently 100 µL of Stop&Glo reagent was mixed to measure the pRL-TK Renilla luciferase reporter activity to determine the transfection baseline. Two technical replicates of the luciferase assay were performed for each transfection well. The readings of the pGL3-promoter plasmids were normalised by those of the pRL-TK plasmid.

2.11 Nuclear extract

Nuclear extracts were prepared from cultured cell lines as previously described (Schreiber *et al.*, 1989). Briefly, 5-7x10^6 cells were washed in cold PBS twice. The cells were resuspended in 1 mL cold Buffer A (10 mM HEPES pH 7.9, 10 mM KCl, 0.1 mM EDTA, 0.1 mM EGTA, 1X Complete Protease Inhibitors cocktail (Roche) and 0.1 mM DTT) and incubated for 15 min on ice. Soluble proteins were extracted by cell lysis and vortexing with 67 µL 10% Nonidet NP-40 (Fluka). The homogenate was washed once with Buffer A and resuspended in cold Buffer B (20 mM HEPES pH 7.9, 0.4 M KCl, 1 mM EDTA, 1 mM EGTA, 1X Complete Protease Inhibitors cocktail (Roche) and 0.1 mM DTT) at approximately 1 mg in 30 µL concentration.The supernatant was snap-freeze on dry ice and stored at -80°C.

2.12 Electrophoretic mobility shift assay (EMSA)

25-nt oligonucleotide probes were designed with 5'-AGCT-3' flanking sequences to produce a 5' overhanging end for ^{32}P radioisotope labelling (Table 2.4). Forward and reverse probes at 1.5 µM in 100 µL with 1X New England Biolaboratory buffer 2 (NEB) were denatured at 100°C for 5 min and cooled at a rate of 1°C per minute to room temperature (approximately 20°C). An oligoduplex was radiorabelled by Klenow fragment to incorporate α^{32}P dCTP to 5' overhang at room temperature for 45 min followed by heat inactivation at 65°C for 10 min. Unincorporated ^{32}P dCTP and nucleotides were captured in a Sephadex G-25 column (GE). The binding reaction of oligoduplex probes with crude nuclear extracts, competitors or antibodies was incubated at room temperature for 10 to 20 min in binding buffer (10 mM HEPES pH 7.8, 50 mM KCl, 1 mM EDTA, 1 mM EGTA,

12.5% glycerol, 1 µg polydIdC). Antibodies used for super shift assay were anti-P300 (sc-585), anti-CREB1 (sc-186X), anti-P65 (sc-372), and rabbit IgG (sc-2027, all Santa Cruz Biotechinologies) at 1 µg per lane. The reaction complex underwent electrophoreses in a polyacrilamide gel (5% acrylamide, 1X TBE buffer, 0.25% w/v APS, 1.5% TEMED) at 4°C (200V) for 2.5 hours. The gel was dried on 3MM chromatography paper (Whatmann) at 80°C and exposed to an autoradiography film for 1 to 3 days. The film was developed on the Kodak X-Omat film processor (Kodak).

Table 2.4: EMSA probe sequences. Forward sequences without the flanking 5′-AGCT-3′ are shown.

SNP	Forward sequence (5' to 3')
rs10784774	TTCCAGGATCT (A or G)GGGCACAAACT
rs1384	GAATTTTGATTT (C or T) TTAAAGTGTGT
rs12812860	ATAAAAAATAAA (T or C) GTGGAATTCACT
rs10878967	CCCTATTTATTA (T or C) TGAAGAATCAAA

2.13 Gene expression array

$CD14^+$ monocytes, $CD19^+$ B-cells and macrophages total RNA (200 ng) from each of the 288 healthy volunteers was analysed using the Illumina HumanHT-12v4 BeadChip gene expression array platform which included 47231 probes (26135 genes) based of RefSeq genes (Release 38, Table 2.5). The experiment was performed by the Wellcome Trust Sanger Institute Microarray Services and the Wellcome Trust Centre for Human Genetics. Briefly, 200 µg or less of total RNA was reverse transcribed, amplified and labelled with biotin. An oligo-dT oligonucleotide with the sequence of a viral T7 RNA polymerase promoter was hybridised to the polyA tail of the total RNA. Complementary DNA was reverse-transcribed with template mRNA to synthesise double strand DNA template for RNA amplification. The RNA was linearly amplified by T7 RNA polymerase by incorporating nucleotides and biotin-labeled dUTP. Biotinylated complementary RNA (cRNA) was hybridised with the BeadChips at 58 °C overnight (16-18 hours). After washing to remove excess unhybridised cRNA, the hybridised BeadChips were stained with Cyanine 3 (Cy3) conjugated streptavidin and were scanned using an iScan (Illumina). The background was subtracted by using the R packages lumi and limma. The raw data was transformed by $\log_2$. Both cell type datasets were normalised together using the robust spline normalisation method (RSN, Du *et al.*, 2008).

2.14 Genomic DNA genotyping

SNPs and other genetics variants were determined using the Illumina Infinium high-density genotyping bead arrays (Illumina HumanOmniExpress-12v1.0 Beadchips, NCBI36 Build). Genotyping was performed by the Wellcome Trust Centre for Human Genetics Core Genomics service. The calling of genotypes was performed using the GenomeStudio

Probes	Description	Number of probes
NM	Coding transcript, well-established annotation	28,688
XM	Coding transcript, provisional annotation	11,121
NR	Non-coding transcript, well-established annotation	1,752
XR	Non-coding transcript, provisional annotation	2,209
UniGene	Experimentally confirmed mRNA sequences that align to EST clusters	3,461
Total		47,231

Table 2.5: Probe content of Illumina HumanHT-12 v4 BeadChip

program (Illumina). The numbers of genetic markers on the Beadchip are shown in Table 2.6.

Marker categories	Number of markers
Number of SNPs within a 10kb range of RefSeq	381,329
Nonsynonymous SNPs (NCBI annotated)	12,134
MHC[1]/ADME[2]	7,566/16,680
Sex Chromosome(X/Y/PAR[3] Loci)	18,239/1,697/540
Total	733,202

Table 2.6: Variant marker content of Illumina HumanOmniExpress-12v1.0 BeadChip. 1: MHC major histocompatibility complex, 2: ADME Absorption, Disruption, Metabolism, and Excretion, 3: PAR pseudoautosomal region.

2.15 Data analysis

The quality control of the genotyping data and association analysis were performed using PLINK, the whole genome data analysis toolset (*http://pngu.mgh.harvard.edu/~purcell/plink/*, Purcell *et al.*, 2007). Linkage disequilibrium (LD) was analysed using Haploview with the 1000 Genomes pilot 1 CEU reference panel (*http://www.broadinstitute.org/scientific-community/science/programs/medical-and-population-genetics/haploview/haploview*, Barrett *et al.*, 2005). Imputation of genotyping data was performed using Impute2 (*http://mathgen.stats.ox.ac.uk/impute/impute_v2.html*, Howie *et al.*, 2009). Imputation of HLA alleles was performed using HLA*IMP (*https://oxfordhla.well.ox.ac.uk*, Dilthey *et al.*, 2011). Statistics were analysed using R and appropriate packages (*http://www.r-project.org/*). Parameters used for analysis were indicated in the results chapters.

3 Differential gene expression data analysis of immune cell types

3.1 Introduction

Cells differentiate along particular lineages to perform specialised functions throughout the body. The genome-wide gene expression profile, defined as a transcriptome, of a particular cell undergoes dynamic changes on a temporal basis dependent upon the state of differentiation. Transcriptome profiling on a single cell type or tissue using a microarray takes a snap shot of these changes at a gene expression level at the moment of sample collection. Microarrays have been used to identify a molecular signature from one cell type to another. For example, it was used to detect a set of differentially expressed genes in tumours in comparison to normal cells, classifying different types of tumours, prediction of clinical outcomes (Alizadeh *et al.*, 2000, Ross *et al.*, 2000, Veer *et al.*, 2002). Investigation of transcriptomes for fully differentiated normal cell types are also of interest to determine genes involved in their specific biological roles. Monocytes and B-cells were chosen in this study because they both have important roles in innate and adaptive immune responses and are easily isolated from peripheral blood. Monocytes are circulating leukocytes that, in the case of infection or tissue damage, migrate into the site of inflammation to mediate the primary immune response. Monocytes can additionally differentiate into macrophages and dendritic cells under the influence of specific cytokine stimuli for subsequent immune responses (Shi and Pamer, 2011). B-cells are a type of lymphocyte that form a vital component of the adaptive immune system. The main role of B-cells is the production of antibodies, when fully activated by the presentation of antigen from APCs to B-cell receptor (BCR). Previous research has also attributed immune suppressive function to a B-cell subtype (Mauri and Bosma, 2012).

The aim of the work presented in this chapter was to investigate cell type specific gene expression for primary human monocytes and B-cells isolated from the same individuals and resolve differences that may be associated with the function of these cell types in immune responses. Microarrays were used as a robust and well established tool for genome wide gene expression profiling. The gene expression data generated was also used for subsequent mapping of the genetic basis of inter-individual differences in gene expression described in Chapter 4.

Objectives

The aim of the work described in this chapter is to determine the profiles of cell type specific gene expression which may be associated with the functionality of the cell types in immune responses. Specific objectives are:

1. To collect primary immune cell samples from healthy volunteers of European descent
2. To obtain gene expression data from two immune cell types, $CD14^+$ monocytes and $CD19^+$ B-cells
3. To analyse differential gene expression between these two cell types

3.2 Sample collection of healthy volunteer cohort

3.2.1 Recruitment of healthy European volunteers

This study was approved by the Oxfordshire Research Ethics Committee (COREC reference 06/Q1605/55). A total of 288 volunteers, on average 10 volunteers per day, were recruited in the Oxfordshire area between March 2010 to July 2010 following written informed consent at the time of sample collection. All volunteers reported no history of autoimmune or other relevant diseases including infection at the time of sample collection. The median age of the population was 33.1 years and their self-reported ethnicities are shown in Table 3.1.

3.3 PBMCs and cell purifications

PBMCs were prepared from 50 mL whole blood using a Ficoll gradient (median cell count $73.4x10^7$ cell/mL, interquartile range (IQR) +/-30.18). The immune cell subpopulations

		Total	Male	Female
	Sample number (%)	288	125 (43.4)	163 (56.6)
Age (years)	Range	18-62	18-54	18-62
	Median	29	27	31
Race (%)	British Isles	243 (84.4)		
	Northern European /Scandinavian	29 (10.1)		
	Southern European	9 (3.1)		
	Eastern European	4(1.4)		
	Mixed heritage	3 (1.0)		

Table 3.1: Demographics of the healthy volunteer cohort. Race is based on self-reported geographical origin of volunteers.

were separated by Miltenyi magnetic activated cell sorting (MACS). The MACS method separates cells using magnetic nanoparticles coated with antibodies. Positive cell separation uses an antibody against a surface antigen of the cell type of interest to directly pull out the single cell population by magnetic beads binding to a column in a magnetic field. Negative separation uses a cocktail of antibodies that correspond to cell surface antigens of all cell types except the cell type of interest. This intentionally avoids touching a cell population by antibody binding. It has been demonstrated that there is no significant consistent difference due to activation by binding of antibodies and the positive separation method showed consistently more reproducible results in microarray replication due to less contamination of other cell types (Lyons *et al.*, 2007). In order to compare positive and negative selection for separating monocyte and B-cell populations from PBMCs, cells purified by both methods were analysed using flow cytometry.

$CD14^+$ monocyte population was selected positively using anti-CD14 beads and negatively using the MACS monocyte isolation kit II. $CD19^+$ B-cell population was selected positively using anti-CD19 beads and negatively using the MACS B-cell isolation kit II. The isolated cells were subsequently immunostained with anti-CD14 conjugated with PE-Cy7 and anti-CD16 conjugated with FITC for $CD14^+$ monocytes, and anti-CD19 conjugated FITC for $CD19^+$ B-cells. The comparison between the methods showed higher purity in the positively separated cell populations (only monocyte results shown, Figure 3.1). Thus the positive separation method was chosen for the rest of the collection.

Monocytes (288 samples) and B-cells (287 samples) were freshly separated from PBMCs using the MACS method using $CD14^+$ and $CD19^+$ magnetic beads respectively (median counts $3.2x10^6$ cell/mL, IQR +/-2 for monocytes). The purity of the separated

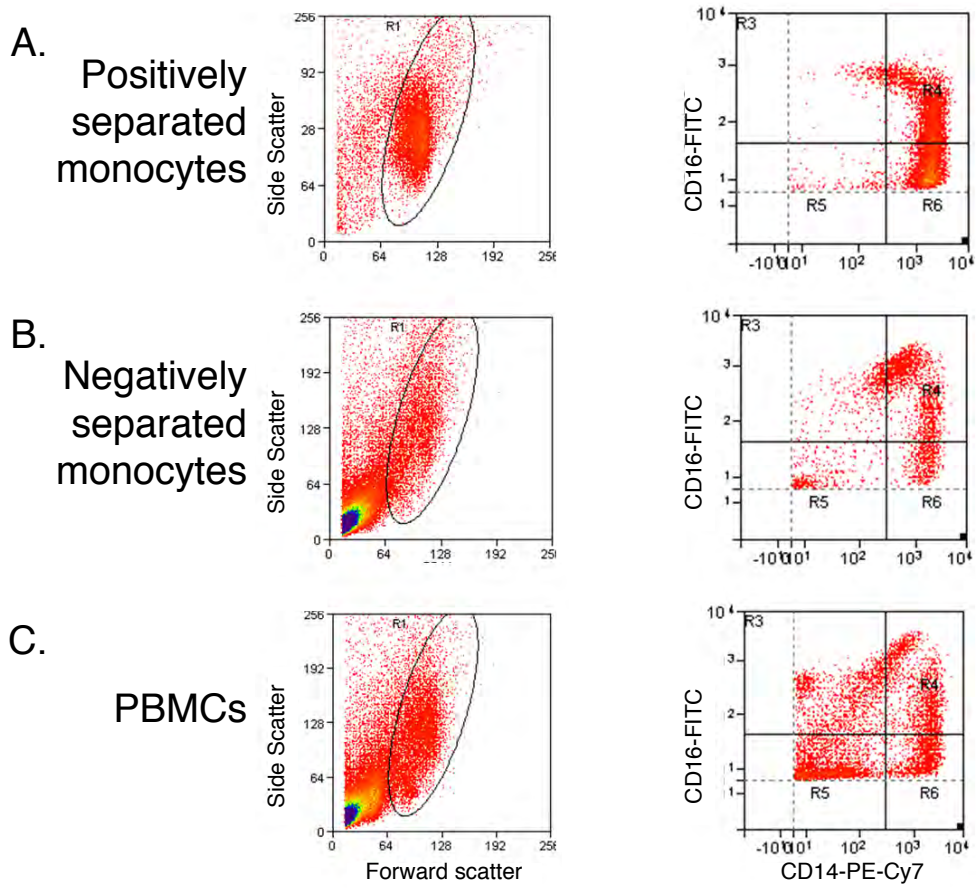


Figure 3.1: Monocytes separated by positive and negative selection were analysed by flow cytometry by immunostaining using anti-CD14 conjugated with PE-Cy7 and anti-CD16 conjugated with FITC. (A) Positive selection showing the enriched $CD14^+$ and $CD16^+$ population, contrasting with (B) negative selection and (C) unsorted PBMCs control.

cells was determined by flow cytometry and was 99% for monocytes and 90-95% for B-cells. Separated cells were resuspended in RLT reagent and stored at −80 °C.

3.3.1 RNA extraction

Total RNA was extracted from the stored RLT samples and was quantified by Nanodrop. A subset of the RNA samples were analysed using a Bioanalyzer (Agilent). The median RNA yields from B-cells and monocytes were 38.2 and 25.6 ng/µL (IQR +/-23.8 and +/-14.8 respectively). The median of RNA integrity number (RIN) from Bioanalyzer analysis was 8.75 (IQR +/-1.3, tested in 4.5% of the population for each cell type).

3.4 Quality control and normalisation of microarray data

3.4.1 Illumina BeadChip gene expression microarray

In total 588 RNA samples were assayed on Illumina HumanHT-12v4 BeadChip arrays at the Wellcome Trust Sanger Centre. This included 288 monocytes, 287 B-cells, and 13 replicates to allow quality control. The replicate samples consisted of 5 B-cell biological replicates, 4 B-cell technical replicates and 4 monocyte biological replicates.

Illumina microarrays use BeadArray technology in which 3-micron silica beads are attached to 29-mer address and 50-mer transcript specific oligonucleotides (Kuhn *et al.*, 2004). Beads were randomly deposited into wells of BeadChips that were created by lithographic techniques on glass slides. Each bead position was decoded by 29-mer address sequences in the Illumina factory (Gunderson *et al.*, 2004). Biotin-labeled cRNA was hybridised to 50-mer probe sequences of BeadChips. The hybridised BeadChips were stained with Cy3 conjugated streptavidin and were scanned on Illumina iScan system to determine the Cy3 signal intensity of each bead. BeadArray chips were scanned on Illumina iScan system. The GX (Gene Expression) module of Illumina GenomeStudio software analysed the gene expression data based on the Cy3 signal intensity to determine data integrity and perform basic quality control checks. The raw expression data was exported in a text file format from software after the background subtraction based on the control probe set. The format of the raw data has a matrix of one line per probe with 8 columns per sample that include minimal, average and maximum signals for the probe, the number of beads for the probe and its standard deviation, and detection score.

3.4.2 Control data

The GX module separately outputs the results of the internal control probe set, which include signals from biotin, Cy3 hybridisation, housekeeping genes, labelling probes, low stringency hybridisation and negative probes (Figure 3.2). This information is useful to assess the quality of the data. Signals from the biotin probes determine successful secondary staining during cRNA amplification. Signals from the Cy3 labelled probes determine gradient hybridisation responses independent of RNA quality. The housekeeping genes consist of genes ubiquitously expressed in any cellular samples and determine the intactness of RNA samples. The labelling signals that were provided by the external RNA controls confirm a successful reverse transcription reaction. The low stringency hybridisation controls determine adequacy of signal stringency. The

negative controls, which consist of random sequences with no correspondence to the genomes, defines the system background that is caused by non-specific hybridisation in the dataset. The background correction was performed using the GX module of the GenomeStudio software before the data export. The expression of 7 housekeeping genes before normalisation is shown in Figure 3.3.

3.4.3 Preprocessing data

The signal intensity of each probe reflects the abundance of a gene transcript. The signal is a relative measure in each BeadChip with a unique baseline. In order to compare directly all gene expression data across chips without bias from difference baselines, normalisation is required. Normalisation adjusts the balance of the individual hybridisation intensities and is important for meaningful biological comparisons because biases of microarray data are not only caused by hybridisation on different chips but also variable starting RNA quantities, amplification and labelling of cRNA. The data was processed in 3 steps: background subtraction, transformation and normalisation. The data for both B-cell and monocyte gene expression was processed in the R statistics program with the lumi package (The R project for statistical computing, *http://www.r-project.org/*, Du *et al.*, 2008). The lumi package is specially designed to process Illumina BeadChip microarray data.

Filtering probes

The detection scores in the raw data outputs from GenomeStudio indicate the quality of probe hybridisation. Signals from poor hybridisation may cause false positive results in further analysis. Probes with detection scores >0.99 in 5% of the sample population (n=30) were remained for further preprocessing (18252 probes out of 47231).

Log_2 transformation

A simple $\log_2$ transformation was applied for this dataset. The $\log_2$ transformation converts exponentially distributed raw expression values to a normal scale and preserves the ranks of expression of each probe.

Normalisation

For this dataset, Robust Spline Normalisation method (RSN) was chosen based on normalisation method comparisons by Schmid et al (Schmid *et al.*, 2010). RSN combines features of the quantile and LOESS normalisations. A quantile normalisation is first

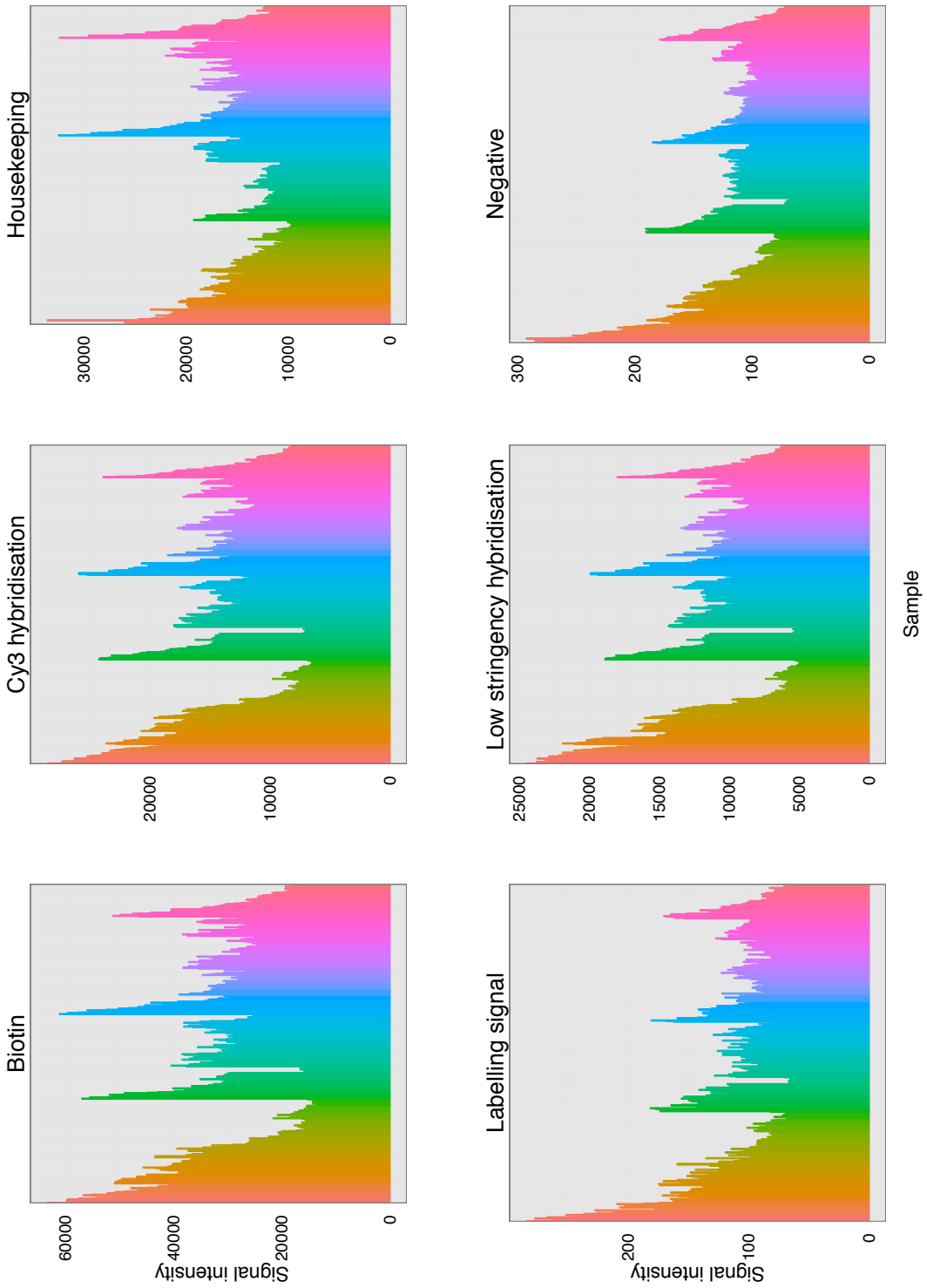


Figure 3.2: Illumina BeadChip control probe data. Distribution of control probe data across all the samples before normalisation. Each panel presents relative signals from biotin, Cy3 hybridisation, housekeeping genes, labelling signal, low stringency hybridisation and negative probe. Samples were coloured based on hybridised BeadChips.

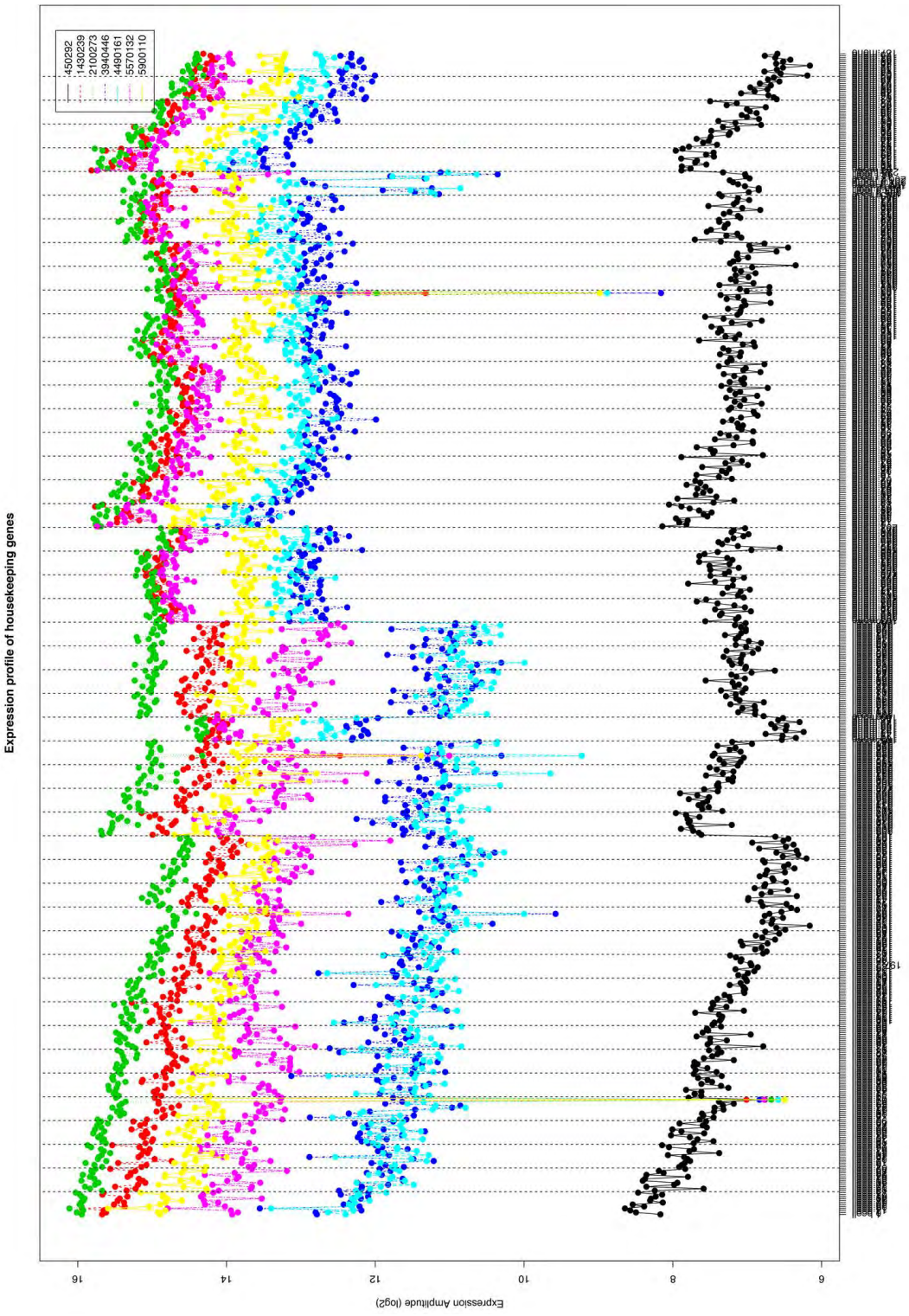


Figure 3.3: Expression of housekeeping genes before normalisation. Log_2 transformed expression values (X-axis) of 7 housekeeping probes were plotted across all samples (Y-axis). Two samples had fluctuating signals in the majority of the housekeeping controls. This was likely due to failed hybridisation and the samples were excluded from further analysis. 450292 - *CNR2* (black), 1430239 - *UBC* (red dash), 2100273 - *EEF1A1* (green dot), 3940446 - *TXN* (blue dash), 4490161 - *GAPDH* (light blue dash), 5570132 - *ACTB* (pink dash), 5900110 - *RPS9* (yellow).

performed to normalise quantiles of each chip and preserve the distribution of expression values. LOESS normalisation is then performed to fit to a local regression (monotonic) curve. This method preserves both the rank order of probes and the effect sizes of the differentially expressed genes.

RSN was applied to the $\log_2$ transformed data iteratively until no outlier samples were identified at each normalisation. The distance from the mean of the samples after removing the extreme 10% samples was used to detect outliers. The outliers were detected when its distance to the mean was larger than two times the median distances to the mean.

For this dataset, two iterative procedures were required. Two samples were detected as outliers (sample ID 79 in B-cells and 191 in monocytes) and removed from further analysis. These are the two samples presented in Figure 3.3. The RNA concentrations of the two samples were 47.81 and 35.43 ng/µL which were above the averages.

3.4.4 Quality control

The quality control of the normalisation process was verified by examining the distribution of expression values. The distribution of expression value densities before and after normalisation clearly shows a shift of the density curves towards a common curve (Figure 3.4). Cluster analysis was used to reveal relationships between the samples and identify potential outliers. Notably, clustering of the samples became tighter following normalisation (Figure 3.5). Two large clusters are composed of the B-cell and monocyte datasets. The box and whisker plots of the expression values in each sample show the effect of normalisation as the mean of the raw data was shifted to the same baseline (Figure 3.6).

The efficacy of normalisation was assessed using an MA plot. It consists of the ratio of $\log_2$ signal intensity in each probe (M) and the average (A) between two samples (Dudoit *et al.*, 2002). The ratio of each probe between the two samples indicates intensity dependent biases. Normalisation is expected to correct this bias close to zero throughout low to high intensities. Randomly selected samples were plotted before and after normalisation (Figure 3.7). Following normalisation, the median values were approaching zero thus showed the normalisation was adequate.

Assessing variation in the data helps to identify experimental artefacts that cause non-biological variation. Principal component analysis (PCA) reduces dimensionality of the data while retaining most variation in the data (Ringner, 2008). The first component explains the largest variation of the dataset. This is often due to experimental artefacts

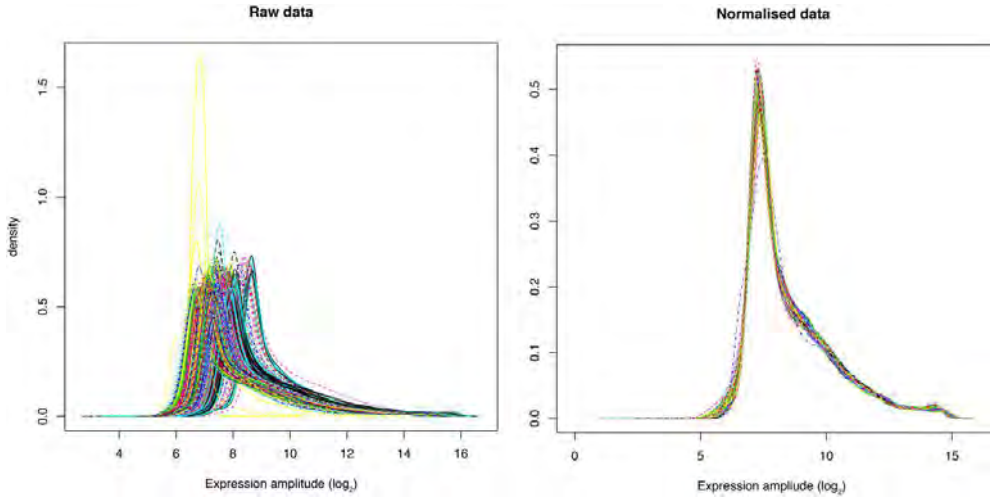


Figure 3.4: The distribution of the expression value densities before (left panel) and after (right panel) normalisation.

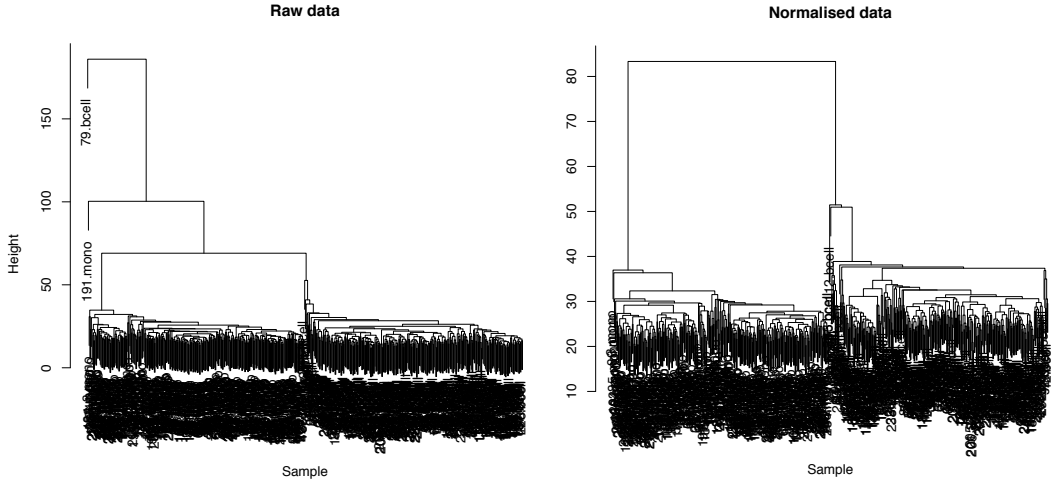


Figure 3.5: The clusters of the samples based on expression values before (left panel) and after (right panel) normalisation. The two large clusters were formed by the B-cell and monocyte datasets.

such as array chips and quality of RNA. The biplots of principal component 1 (PC1) and PC2 before normalisation showed that PC1 explained sample outliers and array chips (Figure 3.8). In contrast the PC1 after normalisation clearly grouped the samples into two groups; B-cells and monocytes (Figure 3.8).

What element in gene expression data is explained by individual PCs can be identified by correlating PCs and covariates such as array chips and RNA concentration. Briefly what PC2 explains was visually examined by colouring each sample with different covariates addition to array chips (Figure 3.8). RNA concentration and sample collection dates were examined as a possible major source of noise. None of these covariates matched distribution of PC2 in both raw and normalised data (data not shown).

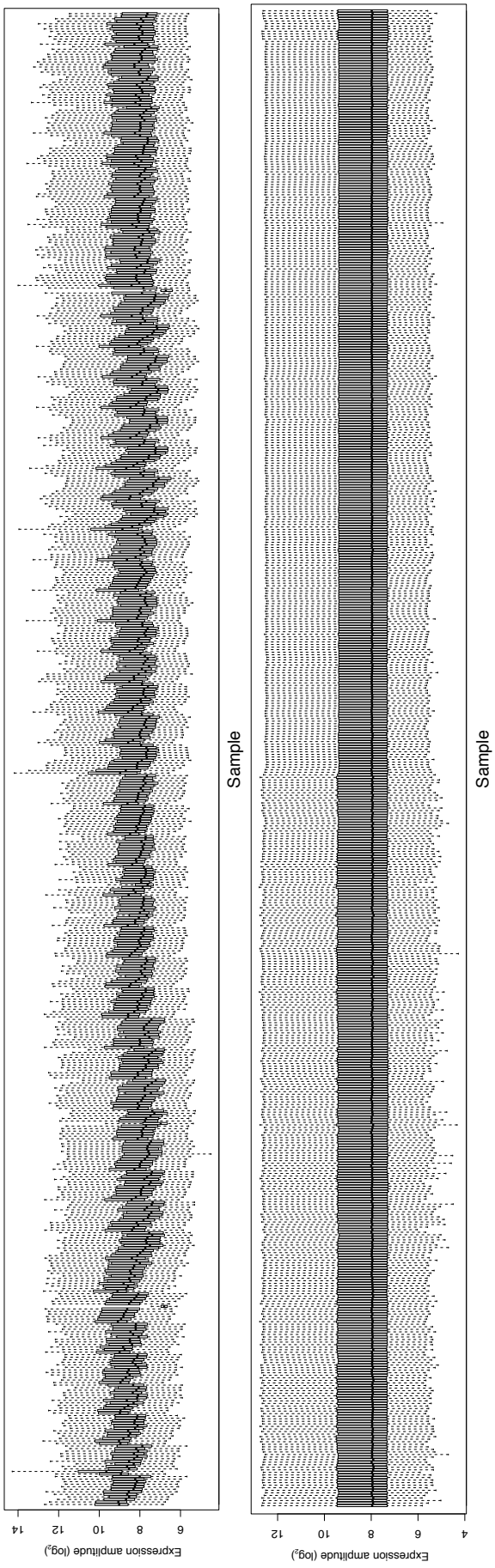


Figure 3.6: The distribution of expression values before (top panel) and after (bottom panel) normalisation.

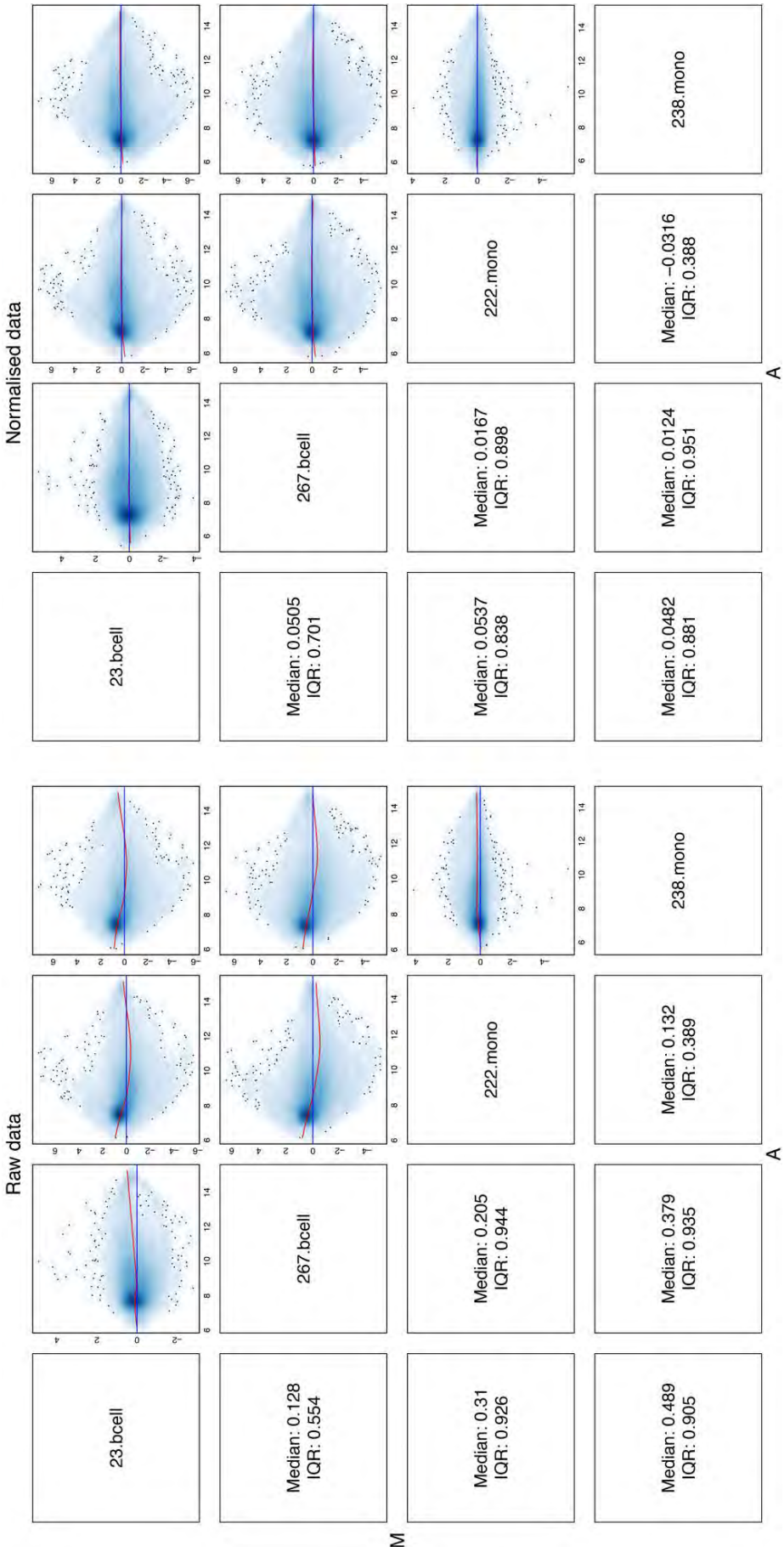


Figure 3.7: Pairwise MA plot of the expression data from randomly selected samples before (left panel) and after (right panel) normalisation. The median values of each plot (red lines) show the efficacy of normalisation. The Y and X axes show the ratio of $\log_2$ signal intensity (M) and the average between the two samples (A). bcell; B-cells, mono; monocytes. Medians and interquartile ranges (IQR) for each pair are shown.

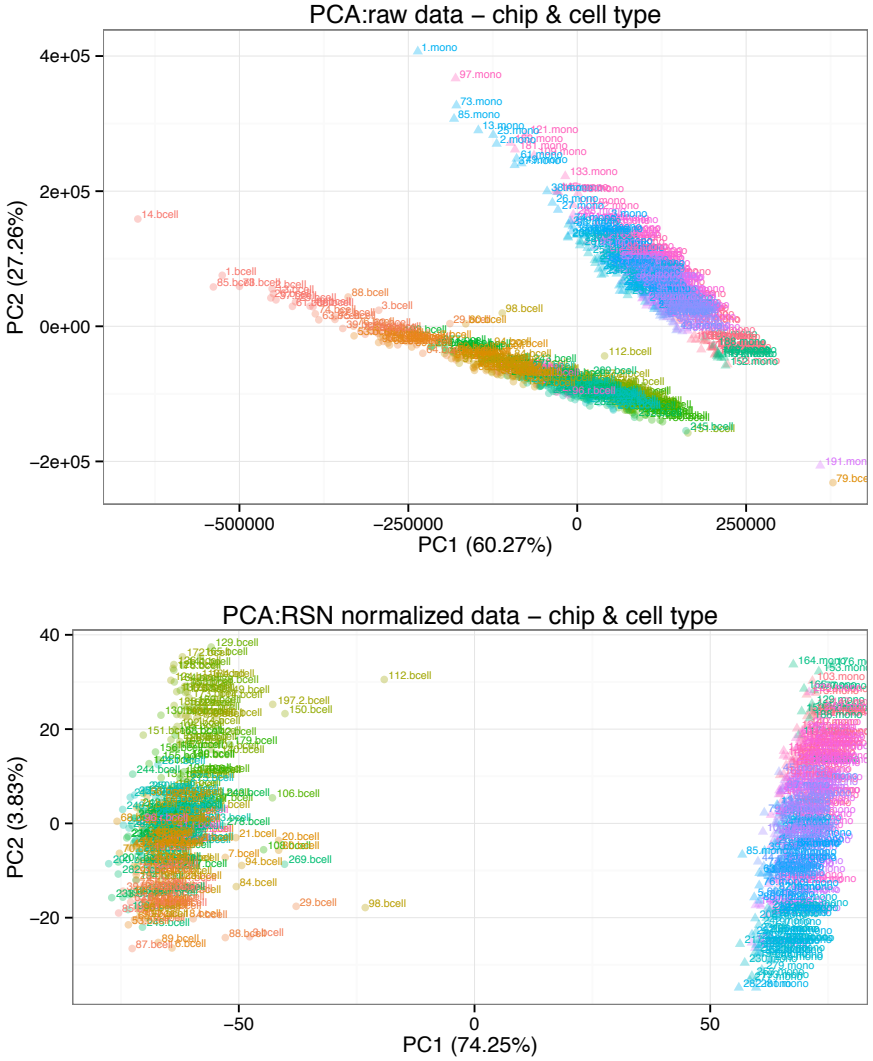


Figure 3.8: PCA plots of the expression data before (top panel) and after (bottom panel) normalisation. The samples are coloured by microarray chips (n=49) and are pointed by cell types (circles; B-cells, triangles; monocytes). The cell types of B-cell and monocyte colours indicate chip groups. Percentages of variance explained by PCs before and after normalisation are shown: PC1 - 60.27% and 74.25%, PC2 - 27.26% and 3.83%.

3.4.5 Biological and technical replications of microarray experiments

In order to examine reproducibility of the data obtained from Illumina BeadChips, the experiment included 13 replicated samples (5 from B-cells and 4 from monocytes as biological replicates, 4 from B-cells as technical replicates). The biological replicates were generated using blood samples taken from the same individuals several days after the initial blood collection. The technical replicates were hybridised twice on different BeadArray chips. Analysis of the biological and technical replicated samples showed high correlation (an average coefficient of correlation 0.94 and 0.97 respectively) ensuring

that the sample collection and preparations were uniformly performed. The correlation between biological replicated samples was also very high (Figure 3.9).

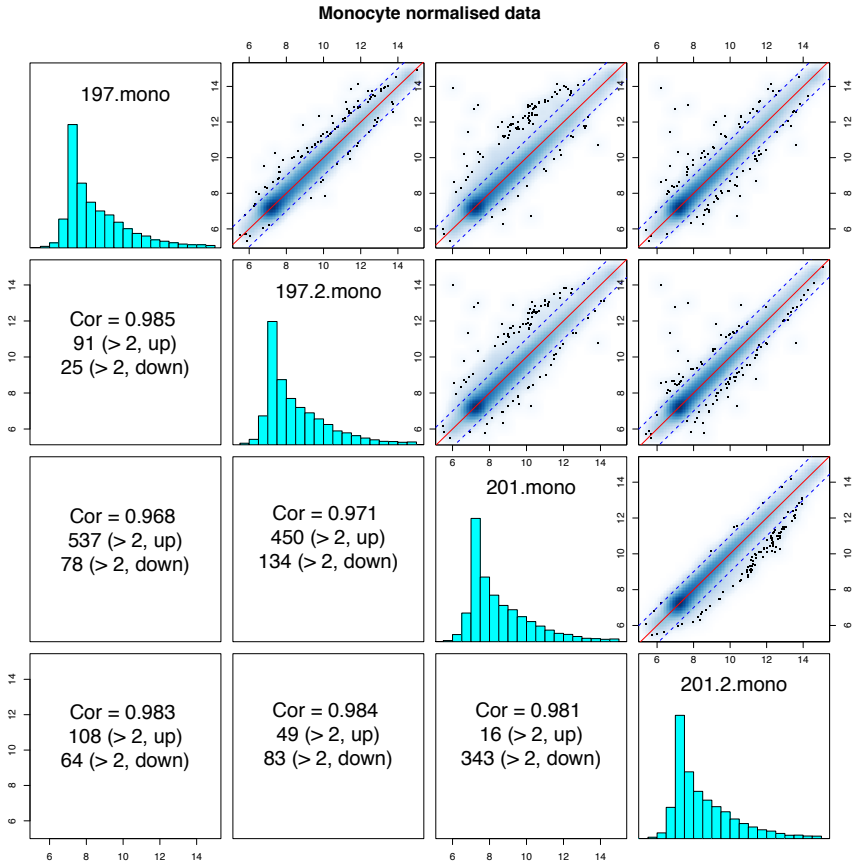


Figure 3.9: The correlation between monocyte biological replicates. Two monocyte examples (mono) with biological replicates are plotted in a pairwise manner for correlation (Cor). Density plots of gene expression are shown for each sample.

3.4.6 Final dataset

The final dataset consists of samples that have been normalised and filtered. After removing the two outliers and 13 replicates, a total of 573 samples were retained in the final dataset. This included 286 B-cells and 287 monocytes samples. Gene expression data was available from 18252 probes.

3.5 Differential gene expression analysis of two immune cell types

3.5.1 Background

The R package limma is designed to analyse gene expression by fitting a linear model to each gene and estimate the log-ratios between two or more samples. This method is able to accommodate different variances as opposed to fitting a single model to an entire experiment. Limma differential expression analysis outputs moderated t-statistics (t), p-values based on t, and log-odds of differential expression (B) as well as gene expression log-fold changes. Moderated t-statistics were calculated from the ratio of $\log_2$-fold change (M) to their standard errors as residual standard deviations. The B-statistic (B) is the log-odds that the gene is differentially expressed.

3.5.2 Processing the data

First each probe of the expression data was fitted to a linear model separately in B-cells and monocytes using a generalised least square method that estimates parameters of unequal variances. A generalised least square method is efficient when variances of observations are not equal. The coefficients of each linear model describes the difference between the two cell types. The contrasts of each probe were computed based on the estimated coefficient of a linear model. The empirical Bayes statistics method was applied to compute moderated t-statistics (t) with false discovery rate (FDR) (Benjamini and Hochberg, 1995) and log-odds (B) of differential expression. FDR explains the expected proportion of false discoveries for multiple comparisons.

In summary, 81.2% of probes tested (14827 probes out of 18252) were significant at an adjusted p-value cutoff 0.001. The details of the summary statistics with different thresholds are shown in Table 3.2. At lower p-values, a greater number of probes were significantly up-regulated in monocytes compared to B-cells (1721 probes in monocytes compared to 1255 in B-cells at $p<1x10^{-200}$). The volcano plot shows that a significant number of probes were differentially expressed between two cell types (Figure 3.10). At similar levels of $\log_2$ fold changes (>4), the log-odds (B) in B-cells were larger than that of monocytes (54 probes in B-cells have B>1000 as opposed to 13 probes in monocytes). The $\log_2$ fold changes (M) between B-cells and monocytes were larger when average expression

was around $\log_2 10$ (Figure 3.11). The highly significant genes are described in the next sections based on the characteristics of the cell types (Tables 3.3 and 3.4).

Table 3.2: Summary of differentially expressed probes between B-cells and monocytes. Thresholds for cutoff are the adjusted p-value with FDR. Up-regulated probes found in only one cell type are shown for B-cells and monocytes.

Threshold	Number of probes (%)			Collective (%)		
p (<)	total	B-cells (%)	monocytes (%)	total (%)	B-cells (%)	monocytes (%)
total	18252	9514 (52.13)	8738 (47.87)	100.00		
0.001	14827	8016 (54.06)	6811 (45.94)	81.20	43.92	37.32
$1x10^{-50}$	9109	4795 (52.64)	4314 (47.36)	49.91	26.27	23.64
$1x10^{-200}$	2976	1255 (42.17)	1721 (57.83)	16.30	6.88	9.43
0	998	352 (35.27)	646 (64.73)	5.50	1.93	3.54

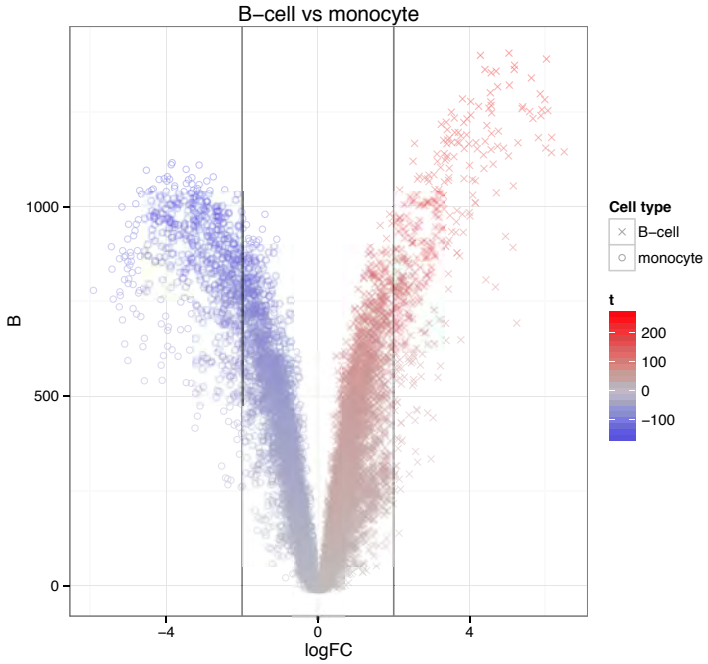


Figure 3.10: Volcano plot of differentially expressed genes in B-cells and monocytes. Probes up-regulated in B-cells (red cross) and monocytes (blue circle) are plotted based on $\log_2$ fold change (x-axis, B-cells>0 and monocytes<0) and B log-odds (y-axis). The moderated t-statistics are shown in two colours (red is B-cell and blue is monocyte up-regulation). Vertical lines indicate $\log_2$ fold change of +/-2.

3.5.3 Differentially expressed genes in B-cells

B-cells have a vital role in immune systems especially humoral immunity in the adaptive immune system. This include antigen presentation and antibody production. Many genes highly expressed in B-cells in comparison to monocytes are related to the activation of B-cells and their functions are described below (Table 3.3).

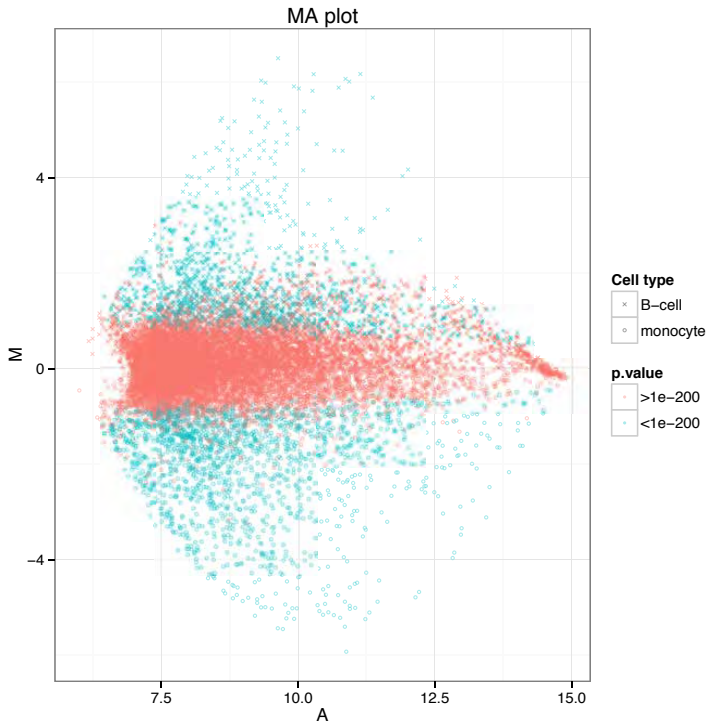


Figure 3.11: MA plot of differentially expressed genes in B-cells and monocytes. Probes that are up-regulated in B-cells (cross) and monocytes (circle) are plotted based on average expression values across the two cell types (x-axis) and $\log_2$ fold changes (y-axis). Adjusted p-values (blue, $p<1x10^{-200}$, pink, $p>1x10^{-200}$) indicate distribution of the most significant probes.

Among differentially expressed genes identified, many are cell surface receptors. These include CD19, the antigen by which B-cell were purified ($\log_2$ fold change (logFC) 5.91, p=0). BCR activation requires CD19 as a co-receptor with CD21 and CD81 for subsequent signal transduction and antigen presentation to helper T cells. A heterodimer transmembrane protein CD79, encoded by *CD79A* and *CD79B*, forms a complex with BCR and generates a signal following activation of BCR upon antigen binding (logFC 5.59 and 4.56 respectively, p=0 for both). Both *CD79A* and *CD79B* were noted to be more highly expressed in B-cells. Higher expression was also seen for *CD40*, encoding a member of the TNF receptor superfamily (logFC 2.0, $p=5.32x10^{-283}$). It is expressed in antigen presenting cells (APCs) and its recognition of a ligand CD40L (also known as CD154) in T cells leads to maturation of B-cells to antibody-producing B-cells. *CD24* gene encodes a glycoprotein that modulates signal transduction for B-cell activation following co-stimulation for T cell proliferation (logFC 5.99, p=0). B-cells differentiate into regulatory B-cells that help control immune tolerance. Anti-inflammatory cytokine IL10 is produced by regulatory B-cells and inhibits production of pro-inflammatory cytokines to facilitate differentiation of regulatory T cells (logFC 1.65, $p=5.5x10^{-171}$). B-cell lymphoid tyrosine kinase, encoded by *BLK* gene, transmits signalling from the surface immunoglobulins

downstream in the BCR signalling pathway promoting B-cell maturation (logFC 5.18, p=0).

Pathogen associated molecular patterns (PAMPs) are distinctive microbial structures that are recognised by pathogen recognition receptors (PRRs), an essential component of innate immunity to trigger inflammation. Toll-like receptors (TLRs) are well characterised PRRs (Kumar *et al.*, 2009). *TLR9* and *TLR10* were up-regulated in B-cells (logFC 1.07 and 4.0, p=9.61x10^{-154} and 0 respectively). Higher expression of *TLR9* and *TLR10* are induced via BCR activation of naive B-cells (Bernasconi *et al.*, 2003). Activation of B-cells via TLR9 binding CpG (double strand DNA, a TLR9 ligand) can also up regulate the production of IL10 and B-cell proliferation (Viglianti *et al.*, 2003).

3.5.4 Differentially expressed genes in monocytes

The primary roles of monocytes involve the innate immune response and activation of inflammation. The genes identified as differentially expressed in monocytes are involved in those processes (Figure 3.4). They include *CD14*, a component of the endotoxin LPS receptor complex with TLR4 (logFC 5.92 and 2.95, p=0 for both *CD14* and *TLR4* respectively). In order to sense invasion of microbes, monocytes express *TLR* genes excluding *TLR9* and *TLR10*. *MyD88* encodes an adaptor protein that is essential for the TLR pathway to activate downstream signalling for induction of the NFκB transcription factor (logFC 2.16, p=0). Another type of PRR, *NOD2* (Nucleotide-binding oligomerisation domain-containing protein 2) that is independent from MyD88, is also significantly expressed in monocytes (logFC 4, p=0).

Monocytes migrate through blood vessels into a site of inflammation (Ley *et al.*, 2007). Chemokine receptors (CCR2 and CX3CR1) recognise chemokines that guide migration. Transendothelial migration of monocytes is mediated by adhesion receptors for cell migration (PECAM1 and ICAM1, logFC 2.38 and 0.88, p=0 and 1.54x10^{-163} respectively).

Pro-inflammatory cytokines are secreted by activated monocytes and regulate activation of inflammation. Genes involved in inflammation responses are differentially expressed. *TNFRSF1A* and *TNFRSF1B* encode tumour necrosis factor (TNF) receptors that mediates the TNF-α signal to downstream pathways and promotes the cytokine cascade (logFC 4.55, p=0).

The differential gene expression analysis demonstrates that the majority of genes expressed in B-cells and monocytes are significantly differentially expressed when these two cell types are compared. In order to investigate those genes systematically, pathway

analysis was conducted in the following section to determine enrichment of these genes with functional pathways in each cell type.

Table 3.3: Differentially expressed genes in B-cells. logFD: $\log_2$ fold changes, t: moderated t-statistics, B: B statistics (log-odds), adj.P.Val: adjusted p-values by FDR.

B-cell up-regulated genes								Average expression		
Gene	Gene product	Refseq_ID	Location	logFC	t	B	adj.P.Val	B-cells	monocytes	Both
B-cell activation										
BLK	B lymphoid tyrosine kinase	NM_001715	8p23	5.18	257.62	1361.73	0	11.43	6.25	8.84
CD19	CD19 molecule	NM_001770	16p11.2	5.91	213.77	1255.50	0	12.88	6.97	9.92
CD24	CD24 molecule	NM_013230	Yq11.2	5.99	223.97	1282.02	0	12.40	6.42	9.40
CD40	CD40 molecule, TNF receptor superfamily member 5	NM_001250	20q12	2.00	70.05	640.48	$5.32\text{x}10^{-283}$	9.80	7.80	8.80
CD79A	CD79a molecule, immunoglobulin-associated alpha	NM_001783	19q13.2	5.59	211.85	1250.36	0	12.14	6.55	9.34
CD79B	CD79b molecule, immunoglobulin-associated beta	NM_021602	17q23	4.56	221.62	1276.00	0	13.10	8.54	10.81
CD81	CD81 molecule	NM_004356	11p15.5	1.24	80.99	716.49	$7.49\text{x}10^{-316}$	13.87	12.63	13.25
CD82	CD82 molecule	NM_001024844	11p11.2	1.51	71.16	648.66	$1.55\text{x}10^{-286}$	10.03	8.51	9.27
IRF8	interferon regulatory factor 8	NM_002163	16q24.1	1.05	79.62	707.47	$5.90\text{x}10^{-312}$	13.22	12.17	12.70
Anti-inflammatory cytokine										
IL10	interleukin 10	NM_000572	1q32.1	1.65	40.71	381.30	$5.55\text{x}10^{-171}$	12.30	10.65	11.48
Inflammatory response										
LTB	lymphotoxin beta (TNF superfamily, member 3)	NM_002341	6p21.3	4.70	255.54	1357.11	0	12.69	7.99	10.34
TLR pathway										
TLR10	toll-like receptor 10	NM_001017388	4p14	4.00	164.93	1108.51	0	11.58	7.58	9.57
TLR9	toll-like receptor 9	NM_017442	3p21.3	1.07	37.00	341.44	$9.61\text{x}10^{-154}$	7.15	6.08	6.61
TRAF5	TNF receptor-associated factor 5	NM_004619	1q32	4.24	217.14	1264.39	0	10.86	6.62	8.74

Table 3.4: Differentially expressed genes in monocytes. logFD: $\log_2$ fold changes, t: moderated t-statistics, B: B statistics (log-odds), adj.P.Val: adjusted p-values by FDR.

Monocyte up-regulated genes								Average expression		
Gene	Gene product	Refseq_ID	Location	logFC	t	B	adj.P.Val	B-cells	monocytes	Both
Inflammatory response										
AOAH	acyloxyacyl hydrolase	NM_001637	7p14-p12	-4.10	-127.84	965.57	0	6.81	10.91	8.86
CCR2	chemokine (C-C motif) receptor 2	NM_000647	3p21	-3.10	-82.82	728.38	$5.32\text{x}10^{-321}$	7.87	10.97	9.42
CD163	CD163 molecule	NM_203416	12p13	-4.19	-154.39	1071.27	0	7.35	11.53	9.44
CLEC4A	C-type lectin domain family 4, member A	NM_194447	12p13	-3.02	-134.25	992.87	0	8.72	11.74	10.24
CX3CR1	chemokine (C-X3-C motif) receptor 1	NM_001337	3p21.3	-2.21	-76.67	687.56	$2.35\text{x}10^{-303}$	7.28	9.49	8.39
ICAM1	intercellular adhesion molecule 1	NM_000201	19p13.2	-0.88	-39.09	364.09	$1.54\text{x}10^{-163}$	6.63	7.51	7.07
LTBR	lymphotoxin beta receptor (TNFR superfamily, member 3)	NM_002342	12p13	-3.91	-154.58	1071.98	0	6.88	10.79	8.84
PECAM1	platelet/endothelial cell adhesion molecule 1	NM_000442	17q23.3	-2.38	-116.31	913.09	0	9.63	12.01	10.82
TNFRSF1A	tumor necrosis factor receptor superfamily, member 1A	NM_001065	12p13.2	-4.55	-128.10	966.69	0	6.54	11.09	8.82
Monocyte differentiation										
KLF4	Kruppel-like factor 4 (gut)	NM_004235	9q31	-3.56	-96.83	812.52	0	6.22	9.78	8.00
Pathogen recognition receptors										
NOD2	nucleotide-binding oligomerization domain containing 2	NM_022162	16q12	-4.00	-145.96	1039.70	0	6.45	10.45	8.45
TLR pathway										
CD14	CD14 molecule	NM_000591	5q31.3	-5.92	-90.99	778.80	0	7.91	13.84	10.88
MYD88	myeloid differentiation primary response 88	NM_002468.3	3p22	-2.16	-111.00	887.27	0	8.72	10.89	9.81
TLR1	toll-like receptor 1	NM_003263	4p14	-0.96	-19.01	130.79	$1.40\text{x}10^{-62}$	6.24	7.20	6.72
TLR2	toll-like receptor 2	NM_003264	4q32	-1.64	-82.22	724.53	$2.55\text{x}10^{-319}$	5.86	7.51	6.69

TLR4	toll-like receptor 4	NM_138554	9q33.1	-2.95	-128.59	968.82	0	6.20	9.14	7.67
TLR5	toll-like receptor 5	NM_003268	1q32.3-q42	-2.95	-120.86	934.34	0	7.45	10.40	8.93
TLR6	toll-like receptor 6	NM_006068	4p16.1	-1.38	-63.82	592.92	$1.89x10^{-262}$	8.37	9.75	9.06
TLR7	toll-like receptor 7	NM_016562	Xp22.3	-0.59	-15.35	89.19	$1.52x10^{-44}$	8.77	9.35	9.06
TLR8	toll-like receptor 8	NM_016610	Xp22	-4.39	-146.79	1042.91	0	6.27	10.66	8.47

Major histocompatibility complex genes

HLA genes located on chromosome 6 (chr6, 6p21.3) encode antigen presenting proteins that are vital to the immune system. Major histocompatibility complex (MHC) class I genes encode molecules that present foreign derived small peptides following internal processing in immune nucleated cells and platelets. MHC class II genes encode heterodimeric protein molecules to present antigens to a T cell receptor in T helper cells. Overall expression levels of the MHC genes are high in both cell types ($\log_2$ expression>10, Table 3.5). B-cells were found to express higher levels of class II genes compared to monocytes, and monocytes expressed higher levels of class I genes.

Table 3.5: HLA gene expression in B-cells and monocytes. logFD: $\log_2$ fold changes, t: moderated t-statistics, B: B statistics (log-odds), adj.P.Val: adjusted p-values by FDR.

						Average expression		
SYMBOL	REFSEQ_ID	logFC	t	B	adj.P.Val	B-cells	monocytes	Both
B-cell up-regulated *HLA* genes								
HLA-DOB	NM_002120	4.90	178.73	1153.93	0	12.26	7.36	9.80
HLA-DOA	NM_002119	3.02	114.50	904.40	0	11.46	8.45	9.95
HLA-DPB1	NM_002121	1.04	27.87	236.59	$2.16\text{x}10^{-108}$	10.61	9.58	10.09
HLA-DMA	NM_006120	0.30	20.57	149.21	$1.47\text{x}10^{-70}$	13.85	13.55	13.70
HLA-DQA1	XM_936128	1.20	19.69	138.81	$4.68\text{x}10^{-66}$	13.51	12.32	12.91
HLA-DPB2	NR_001435	0.38	18.58	125.76	$2.12\text{x}10^{-60}$	7.36	6.99	7.18
HLA-DQB1	NM_002123	1.28	10.36	39.54	$5.82\text{x}10^{-23}$	10.98	9.69	10.33
HLA-DPA1	NM_033554	0.12	7.16	14.90	$3.42\text{x}10^{-12}$	13.98	13.86	13.92
Monocyte up-regulated *HLA* genes								
HLA-A	NM_002116	-0.51	-38.90	362.02	$1.21\text{x}10^{-162}$	14.12	14.63	14.37
HLA-B	NM_005514	-0.54	-35.31	322.69	$1.24\text{x}10^{-145}$	13.55	14.09	13.82
HLA-H	NR_001434	-0.52	-28.76	247.15	$5.84\text{x}10^{-113}$	13.60	14.13	13.87
HLA-E	NM_005516	-0.19	-10.83	43.79	$8.20\text{x}10^{-25}$	13.27	13.46	13.37
HLA-DRA	NM_019111	-0.11	-10.21	38.24	$2.14\text{x}10^{-22}$	14.53	14.64	14.58
HLA-F	NM_001098479	-0.24	-9.52	32.47	$7.00\text{x}10^{-20}$	11.37	11.60	11.48
HLA-C	NM_002117	-0.47	-7.40	16.48	$6.96\text{x}10^{-13}$	9.74	10.21	9.98

3.6 Pathway analysis of differentially expressed gene in two immune cell types

3.6.1 Background

The data obtained from the differential gene expression analysis in B-cells and monocytes provided examples of specific differentially expressed genes that related to their cellular specificities and roles in biological processes. However, it is laborious to curate thousands of genes manually for functional annotation. In order to interpret the genes expressed in two immune cell types systematically, gene enrichment analysis was performed using Reactome database to identify functional pathways up-regulated in each cell type (Joshi-Tope *et al.*, 2005, Croft *et al.*, 2011, *www.reactome.org*). The Reactome database is a curated, peer-reviewed, resource of human biological processes. A basic unit of Reactome is a reaction in which inputs convert into outputs with experimental evidence of one or more literature citations. Groups of these reactions form pathways that for each reaction can be sequential, in parallel or in a cycle.

3.6.2 Pathway analysis

Out of 7061 and 6011 probes significantly expressed in B-cells and monocytes respectively, 6817 and 5894 genes annotated with Entrez IDs were analysed for over-representation against 17248 genes as a background with p-value threshold <0.05. 5654 and 4725 genes were significantly identified as enriched in 8 and 34 pathways in B-cells and monocytes respectively (most significant pathways p_{B-cell}=7.63x10^{-12} and p_{mono}=1.29x10^{-35}, appendix Tables 1 and 2). The number of the pathways significant include up to second levels in the pathway hierarchy in the Reatome dataset. The genes not differentially expressed in B-cells and monocytes were also tested for enrichment in different pathways (4120 genes). However, no pathway related with immunity was significantly identified.

3.6.3 Pathways enriched with B-cell gene products

The pathways over-represented in B-cell genes belong to gene expression (pathway ID 74160, p=7.63x10^{-12}), translation (ID 156842, p=1.41x10^{-11}) and maintenance of the cell cycle (ID 1640170, p=3.22x10^{-8}). The most significant pathway was the gene expression with lower level pathways (Figure 3.12). Within this pathway, B-cell expressed genes are enriched in biological processes for metabolism of mRNA, nonsense-mediated decay and

translation termination. No pathway that is involved with the adaptive immune system was significantly identified in this analysis.

3.6.4 Pathways enriched with monocyte gene products

In monocytes, the most significantly enriched pathways included the immune system (ID 168256, p=2.57x10^{-25}), signal transduction (ID 162582, p=4.04x10^{-15}) and metabolism (ID 1430728, p=1.29x10^{-35}). Compared to the pathways significant in B-cell genes, the significant pathways in monocytes were much more directly related to their roles in the immune system. For example, antigen presentation is one of the primary roles of monocytes, to transfer information of a non-host to T cells (Antigen processing-cross presentation, ID 1236975, p=1.36x10^{-13}). The TLR pathways, which were also up-regulated, are vital to the innate immune system (Toll Receptor Cascades, ID 168898, p=1.23x10^{-9}). The pathways enriched with genes over-represented in monocytes are shown in Figure 3.13.

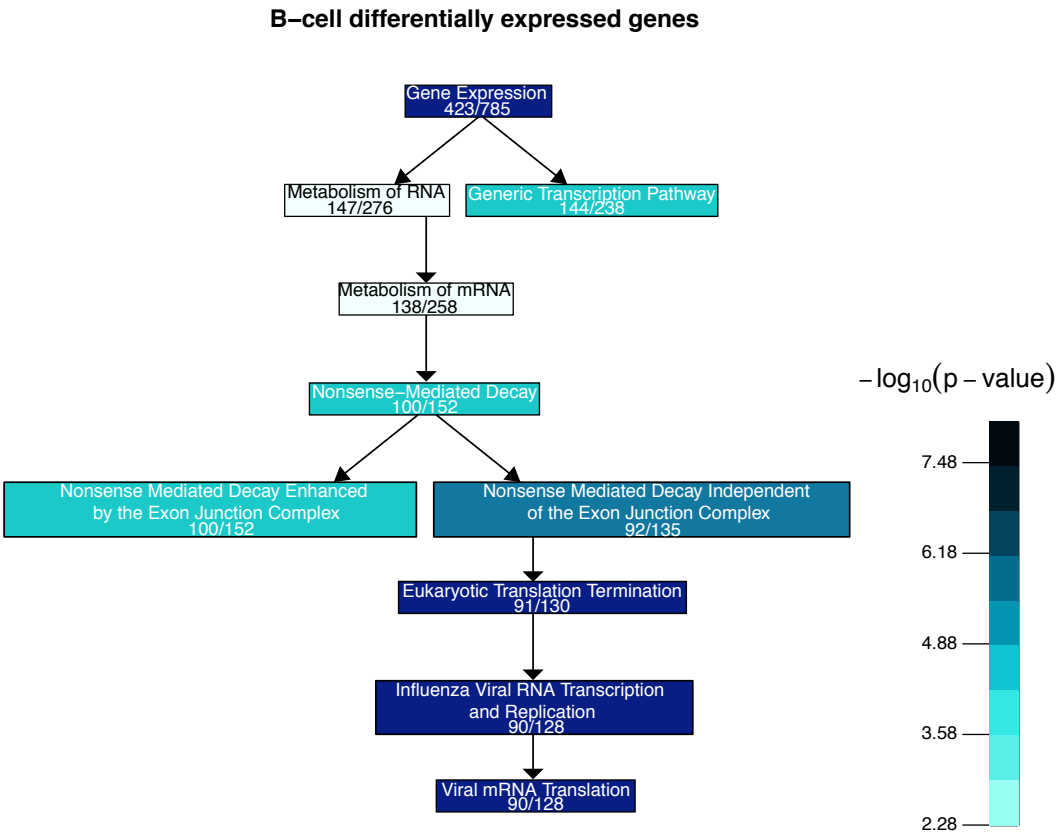


Figure 3.12: Pathways enriched with B-cells differentially expressed genes. P-values are indicated by colour gradient. The most significant pathway is Gene Expression at p=7.63x10^{-12}.

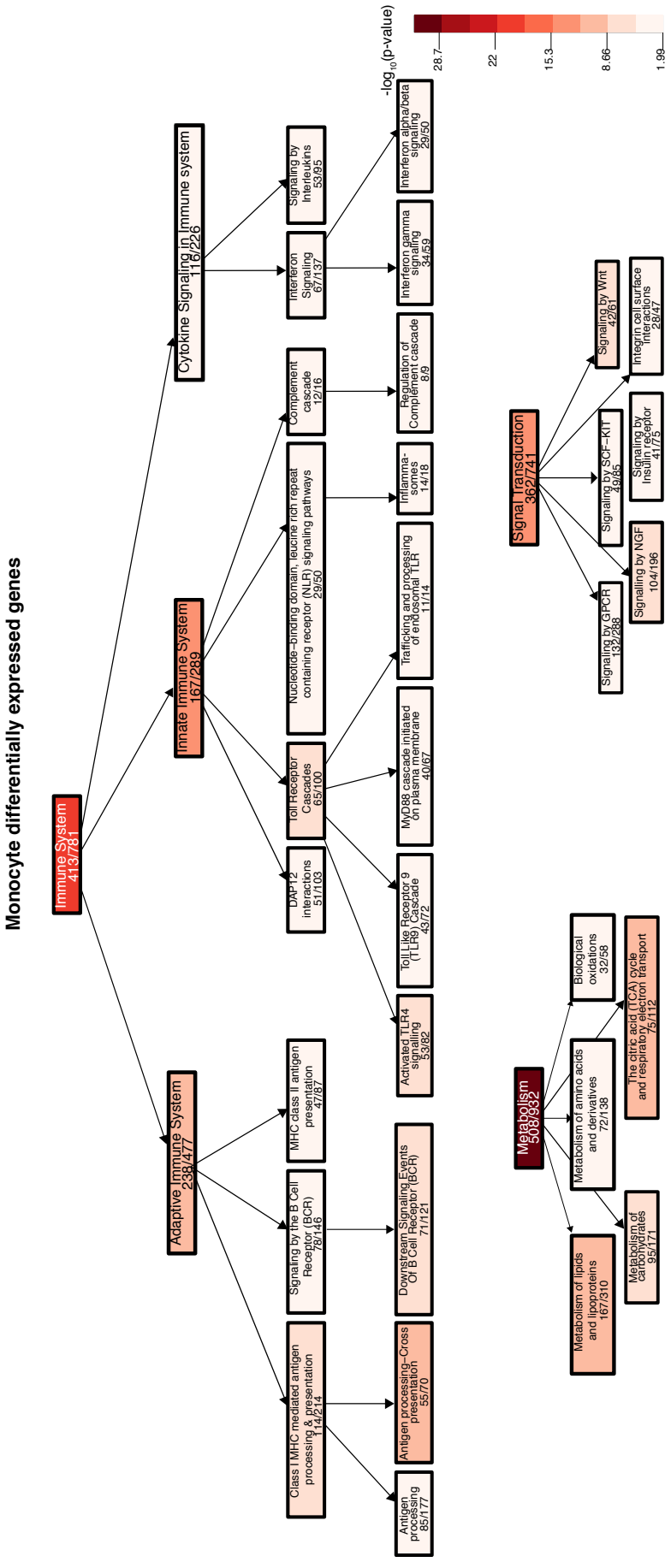


Figure 3.13: Pathways enriched with monocytes differentially expressed genes. P-values were indicated by colour gradient.

3.7 Correlation of gene expression within cell types

3.7.1 Background

Correlation of genes expression levels among cell types may provide insight into biological function. I was interested in identifying key subsets of nodal genes whose expression values were correlated with many other genes. Pathway analyses on such nodal genes may further inform as to underlying function of the cell type and additionally identify such key genes. Correlation was performed within cell types and analysed using pathway analysis.

3.7.2 Correlation analysis

Genes highly up-regulated in each cell type were used for correlation analysis. In total 6011 and 7061 probes for B-cells and monocytes which were significantly expressed in each cell type were used for correlation analysis ($p<1.5x10^{-10}$, Bonferroni correction for multiple testing for differential gene expression). Correlation between gene expression in $\log_2$ values was analysed using a linear model with p-value threshold $2.41x10^{-11}$ (Bonferroni correction). Among differentially expressed genes, 4342 and 5220 probes are significantly correlated in B-cells and monocytes respectively (Table 3.6). As a number of probes are correlated in both cell types, some probes were correlated with the large number of other probes, where frequency distribution plots are shown in Figure 3.14.

Table 3.6: Summary of gene correlation analysis.

	Probes tested	Probes correlated (%)	Number of correlations ($p<2.41x10^{-11}$)
B-cells	6011	5220 (86.8)	1139064
Monocytes	7061	4342 (61.5)	744237

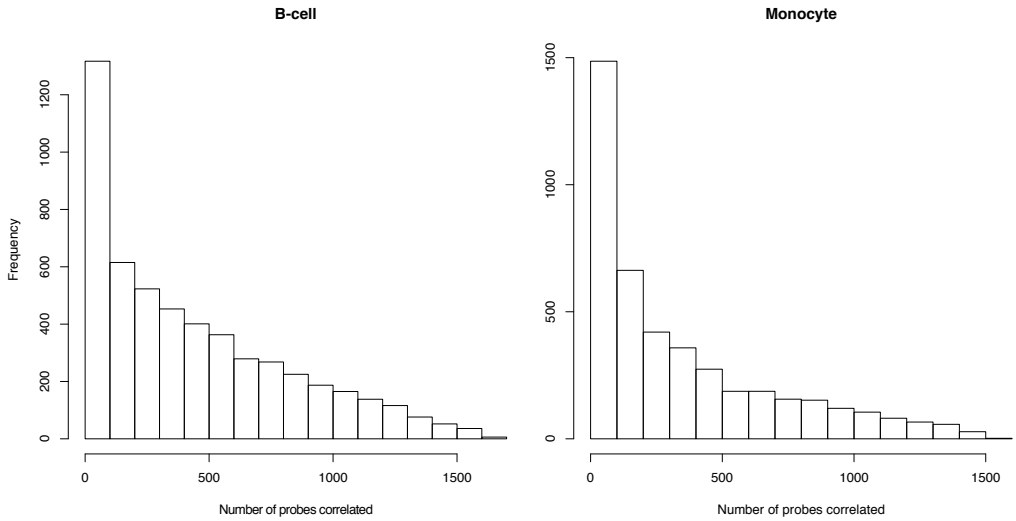


Figure 3.14: Frequency of probe correlation in differentially expressed in B-cells (left) and monocytes (right). The number of probes correlating more than 20% of significantly correlated probes (more than 1044 and 868 probes respectively) is 512 and 507 for B-cells and monocytes respectively.

Table 3.7: Highly correlating genes in B-cells and monocytes. Median of t.statistics are shown for positive and negative correlations respectively. The highly correlating probes have greater t-statistics in a positive direction.

Gene	Gene product	Number of correlations	t.statistics (median) positive	negative	Refseq_ID	Location
B-cells						
TCP1	t-complex 1	1614	12.64	-8.97	NM_030752	6q25-q27
PIK3AP1	phosphoinositide-3-kinase adaptor protein 1	1610	12.99	-9.04	NM_152309	10q24.2
RPLP1	ribosomal protein, large, P1	1599	12.65	-9.04	NM_001003	15q22
ADD3	adducin 3 (gamma)	1598	12.49	-8.79	NM_001121	10q25.2
ROD1	polypyrimidine tract binding protein 3	1586	11.62	-8.95	NM_005156	9q32
PRKRA	protein kinase, interferon-inducible double stranded RNA dependent activator	1584	12.43	-9.18	NM_003690	2q31.2
TUBGCP6	tubulin, gamma complex associated protein 6	1576	9.91	-9.75	NM_020461	22q13.31-q13.33
SUMO2	SMT3 suppressor of mif two 3 homolog 2 (S. cerevisiae)	1573	12.91	-8.84	NM_001005849	17q25
UBLCP1	ubiquitin-like domain containing CTD phosphatase 1	1568	12.4	-8.87	NM_145049	5q33.3
DEK	DEK oncogene	1563	12.5	-8.82	NM_003472	6p23
Monocytes						
ANXA2P1	annexin A2 pseudogene 1	1508	14.54	-9.77	NR_001562	4q31.3
LGALS3	lectin, galactoside-binding, soluble, 3	1506	14.29	-9.77	NM_002306	14q22.3
NDUFA11	NADH dehydrogenase (ubiquinone) 1 alpha subcomplex, 11, 14.7kDa	1486	9.74	-12.35	NM_175614	19p13.3
ATP1B3	ATPase, Na+/K+ transporting, beta 3 polypeptide	1483	13.87	-9.72	XM_001133534	3q23
NPTN	neuroplastin	1483	14.82	-9.71	NM_012428	15q24.1
TMED7	transmembrane emp24 protein transport domain containing 7	1466	14.03	-9.68	NM_181836	5q22.3
LIMS1	LIM and senescent cell antigen-like domains 1	1462	13.87	-9.73	XM_938773	2q12.3
CSNK2B	casein kinase 2, beta polypeptide	1459	9.7	-12.06	NM_001320	6p21.33
MPV17	MpV17 mitochondrial inner membrane protein	1457	9.8	-12.92	NM_002437	2p23.3
SGK3	serum/glucocorticoid regulated kinase family, member 3	1451	12.68	-9.58	NM_001033578	8q12

3.7.3 Pathway analysis of the highly correlated genes

The products of the highly correlating genes in B-cells and monocytes had variable cellular functions (Table 3.7). The functions of those genes were determined by enrichment of gene functions in biologically pathways using the Reactome database. Probes with Entrez gene IDs that are correlated with more than 20% of 5220 and 4342 probes in B-cells and monocytes were selected for an enrichment analysis (512 and 507 probes respectively). With the p-value threshold <0.05, 491 and 490 genes were identified as over-represented in pathways of each cell type.

The top correlated genes in each of the cell types are enriched in pathways involved in essential cellular functions. In monocytes, the pathways enriched belong to respiratory electron transport and ATP synthesis (most significant $p=6.95x10^{-9}$), the citric acid cycle ($p=1.06x10^{-8}$), and metabolism ($p=1.36x10^{-8}$, Figure 3.15). By contrast, the pathways enriched in B-cells are involved with RNA transcription, translational regulation and termination, and metabolism of RNA (most significant $p=3.3x10^{-6}$, Figure 3.16).

Secondly, the genes correlated with the top 5 highly correlating genes were investigated for their enrichment in biological pathways. 2058 and 1796 genes in B-cells and monocytes were tested for gene enrichment and 1867 and 1641 genes were identified as over-represented out of the total genes. Similar to the highly correlating genes, B-cell genes were over-represented in gene transcription and translation pathways and monocyte genes were over-represented in metabolism. Thus genes correlating with each other are typically specific to basal cellular mechanisms (appendix Table 3).

3.8 Discussion

The data presented in this chapter consists of the transcriptome profiles of highly purified human primary immune cells from a population of healthy volunteers of European descent. The analysis on the dataset has generated information for each cell type which can be related to their functions in the immune system.

3.8.1 Sample collection

The highly purified B-cells and monocytes were isolated from PBMCs for 288 healthy volunteers using the MACS technology. The positive separation method was chosen for this sample collection because the comparison on FACS showed higher purity than the negative method (99% and 90-95% purity in monocytes and B-cells respectively).

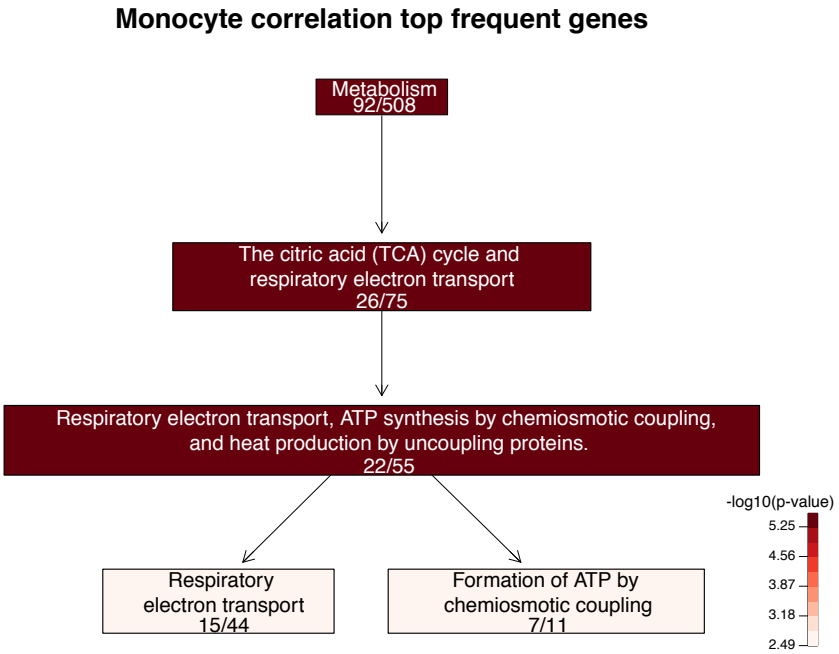


Figure 3.15: Pathway enriched with monocytes highly correlated genes. The diagram created by ReactomePA shows enriched pathways with connections to other related pathways.

Previous studies using PBMC gene expression analysis, for example to compare patients and controls, have identified descriptive/informative expression signatures (Alizadeh *et al.*, 2000, Ross *et al.*, 2000, Veer *et al.*, 2002). However, those signatures may be caused by cell type specific gene expression and/or changes in the cellular constituents of blood (Baechler *et al.*, 2003, Batliwalla *et al.*, 2005a, Batliwalla *et al.*, 2005b, Achiron *et al.*, 2004, Bennett *et al.*, 2003). The dataset described here was derived from highly purified immune cell types and has greatly reduced noise caused by cell impurity compared to many previous studies.

Positive separation by antibodies of CD cell surface markers may potentially cause activation of cell types as CD14 and CD19 are both receptors. An initial study which compared positive and negative separation methods in $CD4^+$ and $CD8^+$ T cells suggested that negative separation could avoid this, however there is no consensus over whether negative selection in fact reduces this potential cell activation (Stanciu *et al.*, 1996). Keeping cells on ice during separation can also avoid further cell activation. Lyons and colleagues demonstrated that positive separation was superior to negative in gene expression profiling of five immune cell types (Lyons *et al.*, 2007). Using FACS to determine cell purity in the samples of the healthy volunteer dataset, it was clear that positive selection was advantageous for cell purity and the differential gene expression data supported this.

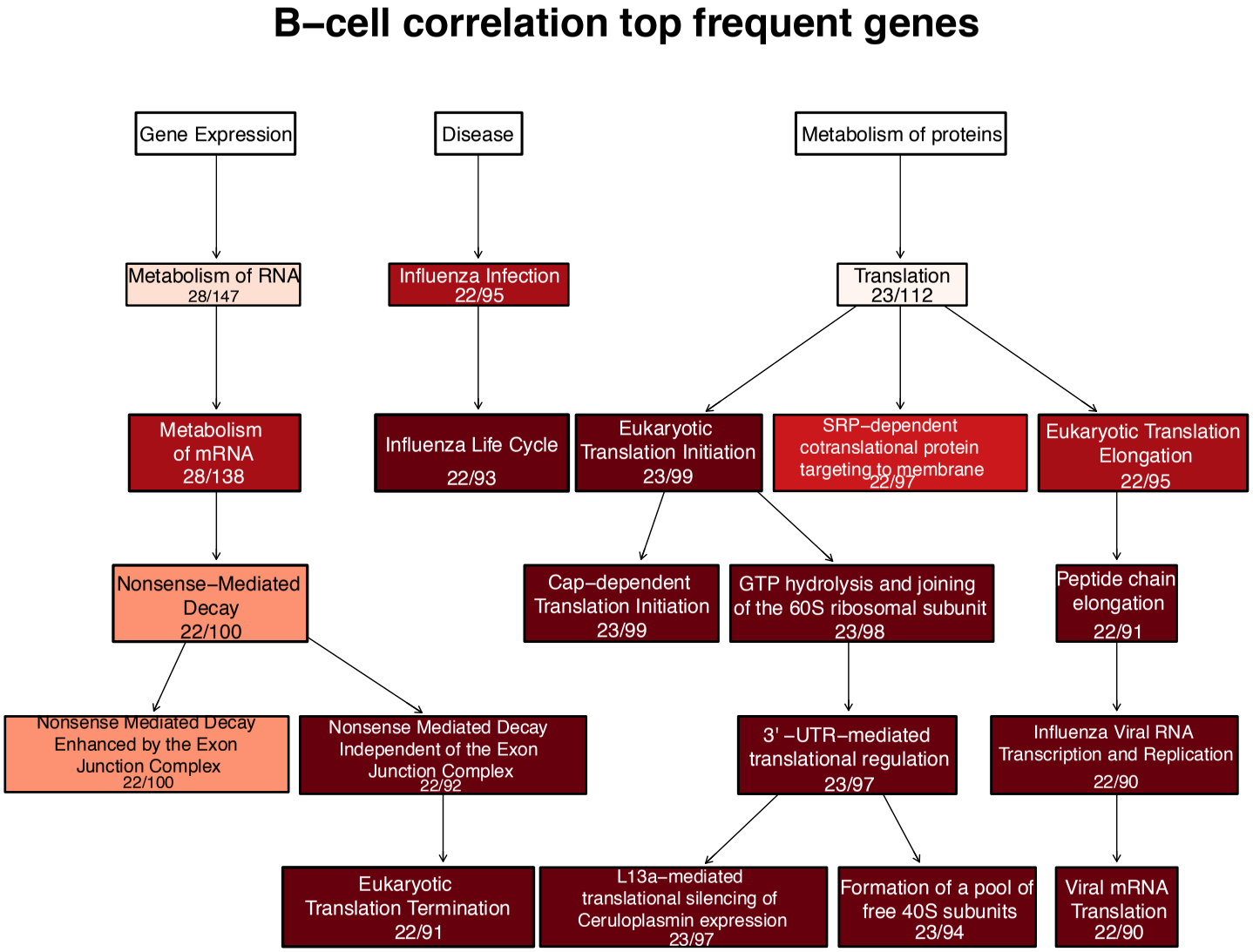


Figure 3.16: Pathways enriched with B-cells highly correlated genes.

3.8.2 Quality control of microarray data

The quality control using the control dataset has shown that each step of the RNA sample preparation was performed properly. The hybridisation of the control probes provided insights into the efficiency of the RNA hybridisation. The plot revealed two potential outliers which corresponded to the outliers detected by the normalisation step.

The normalisation of microarray data was essential in order to extract the gene expression profiles without technical biases and to study the differential gene expression between cell types. The normalisation was applied in the dataset in order to reduce biases caused by artefacts such as RNA concentration, cRNA preparation and different chips used. Using data obtained from different chips is often a significant cause of technical bias (Draghici *et al.*, 2006, The MicroArray Quality Control (MAQC) project *et al.*, 2006). Those technical biases were successfully removed from the dataset by RSN. The data uniformity after RSN was shown by the plots in the quality control section.

In total 586 samples out of 588 passed the quality check after normalisation (99.7%). This high sample pass rate implies that the samples have been processed homogeneously in each step, namely cell separation, RNA purification and microarray processing.

The biological and technical replicates of the microarray experiment showed good reproducibility and demonstrated high quality data implying the sample collection and preparation were uniformly performed. This gave confidence that comparison of the data from each sample could be made from the difference in gene expression.

It has been demonstrated that genetic background influences gene expression when compared unrelated to related individuals (Monks *et al.*, 2004, Morley *et al.*, 2004, Cheung *et al.*, 2003). The individuals in this study sample collection reported their genetic background as European descent, which makes the gene expression profile of the cohort more homogeneous. However, the precise genetic stratification of the healthy volunteer cohort is described in the Chapter 4 for eQTL mapping analysis.

3.8.3 Differential expression analysis

Comparison of the gene expression profiles from the two immune cell types was performed using the R limma package. The method involved fitting a linear model to each probe and contrasting between B-cell and monocyte values to calculate the $\log_2$ fold change. More than 80% of probes tested were differentially expressed in B-cells and monocytes. One explanation for this high proportion may be the larger number of

individuals tested, which provided very high power, as well as high purity of the cell fractions and the RNA samples.

The roles of B-cells and monocytes in the immune systems are different. Many genes significantly differentially expressed in monocytes correspond to gene products whose functions are primarily for surveillance, recognition of microbe ligands (PRRs), activation of inflammation responses and cell migration to the site of inflammation. In circulating peripheral blood, B-cells are in a steady unactivated status. $CD19^+$ is expressed in B-cells from an early stage of cell maturation and thus was chosen to be a marker of the B-cell population in this study, however this makes $CD19^+$ B-cells a mixed cell population.

The genes significantly differentially expressed in B-cells in this dataset are mainly CD markers of B-cells, components of BCR, MHC class II molecules and PRRs. The activation of B-cells can be initiated by encountering APCs, antigen processing and activation of the T helper cells by APCs following differentiation to plasma B-cells for antibody production.

The MHC class I and II genes were shown to be expressed highly in both B-cells and monocytes. The statistical comparison precisely distinguished that the class I genes were more significantly expressed in monocytes and the class II genes in B-cells. The MHC class I genes are generally expressed in almost all cell types in humans, and the MHC class I receptors present antigens derived from pathogens, which are processed by phagocytosis. $CD14^+$ monocytes can differentiate into macrophages, a professional phagocyte. This may explain higher expression of the class I genes in monocytes than B-cells. The MHC class II genes are more expressed in B-cells than monocytes. This may be because monocytes require stimulation by cytokines such as IFNγ to increase expression of the MHC class II genes upon immune response (Steimle *et al.*, 1994).

3.8.4 Pathway analysis

The pathway analysis of the differentially expressed genes in B-cells and monocytes helped to identify the position and roles of the genes in biological processes. The genes expressed in monocytes were enriched in pathways involved in the innate and adaptive immune systems as well as signal transduction and metabolism. Pathways involved in the MHC class I molecule genes, PRRs including TLRs and cytokine signalling and production were enriched in monocyte genes. They provide an immediate defence against infectious microbes and virus, and tissue damages. On the other hand, the genes expressed in B-cells were enriched in pathways of gene expression, translation and maintenance of the cell cycle. These pathways are not directly related in antibody and plasma production,

primary roles of B-cells in the adaptive immune system. This may relate to the fact that B-cells in the bloodstream are not in an activated state.

3.8.5 Correlation

Correlation analysis of gene expression in B-cells and monocytes was performed to investigate possible functional relationships and to find subsets of genes whose expression were highly correlating to many other genes. t-statistics shows direction of correlation and many highly correlating genes had greater positive relationships. Pathway analysis on those genes shows that genes highly correlating with each other belong to basic cellular metabolism and maintenance, not to immunity related genes, in both B-cells and monocytes. The gene expression profiles shown in this dataset were obtained from native, untouched immune cells from the bloodstream, which may explain why no correlation of immune related genes was identified. The analysis of gene expression in an activated status may be able to capture correlation of immune related genes.

3.8.6 Investigation for disease

A transcriptome analysis in diseases is used to find specific patterns in a comparison to healthy condition. For example, gene expression in tumour cells show distinctive expression signatures and can be used to classify types of cancers (Golub *et al.*, 1999,Ross *et al.*, 2000, Veer *et al.*, 2002, Sotiriou and Pusztai, 2009). To my knowledge, this dataset may be one of the largest cohorts collected with homogeneous genetic background. This cohort size may be large enough to define inter-individual variability of gene expression. Therefore, this dataset can be used to compare with gene expression profiles in disease to find a signature associated with the phenotype. Correlation type of analysis with a gene mutation may be possible to predict other affected genes.

3.8.7 Summary

The gene expression datasets presented in this chapter consist of highly purified B-cell and monocyte primary cells that were carefully processed for data acquisition. The differential gene expression analysis revealed the genes significantly differentially expressed in monocytes and B-cells correspond to their roles in the immune system at both the individual gene and biological pathway level. This dataset has potential to be a valuable resource for the study of cell type specific gene expression.

4 Investigation of eQTL in primary human immune cells

4.1 Introduction

The architecture of gene expression is highly heritable and varies within and between populations (Schadt *et al.*, 2003, Monks *et al.*, 2004, Stranger *et al.*, 2007a, Goring *et al.*, 2007). How genetic variants relate to differences in gene expression has been studied by associating the abundance of gene transcripts per individual to genetic polymorphisms specific to the subject (Morley *et al.*, 2004, Dixon *et al.*, 2007, Spielman *et al.*, 2007). When this is performed within a population of subjects it is known as expression quantitative trait loci (eQTL) mapping. eQTL mapping provides information to functionally characterise variants associating with gene expression. This is of particular value in GWAS where the question of which gene(s) may be affected by a SNP showing evidence of disease association is often unclear and implied based on physical proximity.

Cis-acting variants are associated with local genes, generally within 100kb. These variants may drive modulation of gene expression levels through effects in promoter and enhancer sequences. Trans-acting variants are associated with genes not in their proximity. The precise definition of trans association has been stated as a physical distance between genetic variants and genes, typically not on the same chromosome. Their mechanism of action is often unclear; but can be envisaged to be due to distal enhancer activity, or alternatively in effects of genes upstream of the target eQTL. These genes will typically be key regulating genes, for example transcription factor genes, and involve transcription activators (Small *et al.*, 2011). Therefore, this type of association may appear more in activated or stimulated states (Smirnov *et al.*, 2009, Gaffney, 2013).

It is important to understand the genetic determinants of gene expression because they may have phenotypic consequences including disease susceptibility. The goal of

eQTL mapping is to characterise functionality of genetic variation in association with gene regulation. GWAS of complex diseases and traits have facilitated identification of new susceptibility loci in more than 700 phenotypes in the past decade (more than 1400 susceptibility loci as of January 2013, Hindorff *et al.*, 2009). However, the effect sizes of these common variants are often far too small to be causal for disease. When regulatory effects are postulated based on GWAS associations in intergenic regions, the biological insight from the GWAS will depend on identifying which genes are modulated. Thus, assigning roles to DNA variants can contribute to understanding of genetic susceptibility to complex and common diseases that have a heritable component including T1D, T2D, asthma, coronary artery disease, Crohn's disease, and RA (Hirschhorn, 2003, Polychronakos and Li, 2011, Stumvoll *et al.*, 2005, Sladek *et al.*, 2007, Ober and Hoffjan, 2006, Vercelli, 2008, Watkins and Farrall, 2006, Tysk *et al.*, 1988, Franke *et al.*, 2010, MacGregor *et al.*, 2000, Jawaheer *et al.*, 2001).

A number of eQTL studies have now been reported using cultured cells or tissue samples, and have used diverse platforms for gene expression and genotyping. This has led to a number of common local associations likely acting in cis being identified. For example, cis eQTL associations to interferon regulatory factor 5 (*IRF5*) has been identified in several studies in human lymphoblastoid cell lines (LCLs) (Kwan *et al.*, 2008, Morley *et al.*, 2004, Cheung *et al.*, 2005, Dixon *et al.*, 2007). On the other hand distal associations acting in trans have smaller effect sizes and are therefore more difficult to validate. Furthermore, it is proposed that trans eQTL might be more sensitive to context (Gaffney, 2013). Studies in mice have demonstrated biological reproducibility using the same platforms (Nas *et al.*, 2010), however, studies in humans are frequently confounded by diverse methodologies and sample handling. Environmental factors and cell types also have profound effects on detecting eQTL (Idaghdour *et al.*, 2010, Dimas *et al.*, 2009).

In order to investigate context specific information from the non-immortalised transcriptome of a single cell type relevant to disease, it is important to study highly purified primary cells. The eQTL analysis on the HapMap populations is powerful due to the large and defined population size and the availability of genotyping data. However, the immortalisation performed in this study by Epstein Barr Virus (EBV) to create LCLs cannot avoid modification of their innate gene expression profiles (Caliskan *et al.*, 2011). Recent eQTL studies began to explore gene expression profiles from primary tissues or cells in order to extract the information *ex vivo* (Dimas *et al.*, 2009, Fehrmann *et al.*, 2011, Zeller *et al.*, 2010, Murphy *et al.*, 2010). An eQTL comparison among three umbilical

cord-derived cultured cells (fibroblasts, LCLs and T cells) estimated that up to 80% of regulatory variants are cell type specific (Dimas *et al.*, 2009).

The eQTL analysis presented in this chapter investigates cell type specific eQTL in two human primary immune cells, monocytes and B-cells for a cohort of 288 healthy volunteers. Monocytes, which are innately active immune cells, and B-cells from the humoral immune system, have been investigated to determine common and cell-type specific genetic determinants of the immune response. The data obtained in this analysis is further analysed to understand genetic variants identified in GWAS with their associated genes.

Objectives

The aim of the work described in this chapter is to map eQTL in human primary immune cell populations in a cell type specific manner. Specific objectives are:

1. Perform cell type specific eQTL mapping in primary B-cells and monocytes in both cis and trans
2. Interpret identified eQTL in the context of known GWAS loci and immunological pathways

4.2 Gene expression array profiling of primary immune cells

The gene expression profiles of primary B-cells and monocytes for 288 healthy volunteers were obtained using Illumina HumanHT-12v4 BeadChips. The details of data normalisation are described in Chapter 3. In brief, the expression data for 286 and 287 samples from $CD19^+$ B-cells and $CD14^+$ monocytes respectively were selected after normalisation.

Probes with SNPs in their target sequences were removed in order to avoid false positive associations. In order to identify probes with SNPs, first the probe coordinates were determined using the NCBI BLAST ftp site onto human genome assembly 36 (NCBI36, (*ftp://ftp.ncbi.nlm.nih.gov/blast/*). BLAST, Basic Local Alignment Search Tool, is a program to align two sequences to find local similarity (Altschul *et al.*, 1990, Altschul *et al.*, 1997). For exon spanning probes, only probes that matched more than 24bp in one exon were used. Probes that matched more than 2 genomic locations were removed. Then SNPs in probe target sequences were identified by overlapping probe coordinates with

SNP coordinates from the EUR population in the 1000 Genomes Project (Interim Release 23/11/2010, Durbin *et al.*, 2010). Probes containing SNPs with MAF >1% were removed from further analysis. A total of 29022 probes corresponding to 22011 genes (based on NCBI36) were retained for further analysis. This analysis was specific to autosomal genes.

4.3 Genotyping of the volunteer cohort

4.3.1 Background

High throughput genotyping has become an essential tool for GWAS of human diseases in the past decade. To date, approximately 53 million SNPs have been reported in the human genome (38 million SNPs validated, dbSNP build 137 release, 26th June 2012, Sherry *et al.*, 2001), although it had been estimated that the number of SNPs was around 11 million before efforts by the International HapMap Project and the 1000 Genomes Project. The number of novel SNPs is expected to increase because many current projects are applying whole genome sequencing (WGS) to various populations. One of the aims of the International HapMap Project was to characterise linkage disequilibrium (LD) structures across the human genome. Information about LD structures contributes to identifying SNPs with the same LD (r^2 1). These SNPs can be used to accurately impute other SNPs with the same or strong LD. In order to genotype millions of SNPs in an efficient and cost-effective manner, selection of tagging SNPs was an important step to cover SNPs across the genome. Although whole genome sequencing is attractive to identify novel single nucleotide variants (SNVs), the technology is not economical to genotype thousands of individuals yet. Thus, the high density microarray platform for genome-wide genotyping is still the favoured option to reliably genotype hundreds of thousands of SNP markers simultaneously in the most economical fashion.

4.3.2 Genotyping platform

The Illumina Infinium HD assay is based on the Illumina BeadChip technology, similar to the gene expression technology (Steemers and Gunderson, 2007). A high-density array platform, the Sentrix BeadChip, has wells created by photo lithography and plasma etching on silicon wafers. Silica beads (3 µm diameter) are randomly dispersed on the BeadChip. Oligonucleotide probes (50-mer long) attached to silica beads are designed to hybridise with genomic DNA up to one base pair before the allele of interest. After

hybridisation of fragmented genomic DNA, a single-base extension reaction incorporates one of the four dideoxynucleotides labeled with one of two-colour fluorephores (one colour for A and T and another for G and C). The advantages of this method include a PCR-free reaction which requires a small amount of DNA, and single probe design is required to detect both alleles of a SNP. The disadvantage is that only 83% of all SNPs can be genotyped since the platform uses two colour labels and 17% of SNPs those with A/T or G/C alleles can not be measured. The BeadChips hybridised with genomic DNA are scanned by the iScan System (Illumina) for image capture. The acquired images are analysed on the GenomeStudio Genotyping (GT) module for genotype calling.

The normalisation of Infinium HD raw genotyping data by the GT module adjusts for channel-dependent background and global intensity differences and scales the data (Illumina, 2006). The scanned image data is normalised by an parallel coordinate transformation in each sub-bead pool which is a set of beads manufactured together and located roughly in the same stripe on a BeadChip. The raw signal intensities of the two colours, X and Y, are rotated and rescaled to decrease dependence between two alleles. Genotyping calling is performed on the normalised signal intensities (norm R = X_{norm}+Y_{norm}) and allele frequency (norm Theta = (2/pi)*arctan2(Y_{norm}, X_{norm})) relative to canonical cluster positions for each SNP marker.

Each SNP from the normalised data is plotted on a SNP genplot with normalised Theta and R (Figure 4.1, Illumina, 2010). The GT module calls SNP genotypes by shading clusters as homozygous and heterozygous. The module also outputs a quality score called a GenCall score that indicates the reliability of each genotyping call (between 0 and 1) and is calculated based on angles, dispersion, overlap and signal intensity of each SNP genplot. The GT module sets the threshold on GenCall scores >0.15 and SNPs lower than this cutoff were excluded for further analysis.

In order to conduct eQTL analysis on purified immune cells, genome-wide genotyping on 288 healthy volunteers of European descent was performed. Genomic DNA was extracted from whole blood of 288 individuals (125 males and 163 females, see Materials and methods, Table 3.1). In total 733202 genetic variant markers, primarily consisting of SNPs, were genotyped using Illumina HumanOmniExpress-12v1.0 BeadChips (NCBI36 Build). The genotypes were called by the GT module as described above and the final report was used for further analysis. PLINK, the whole genome association analysis toolset (*http://pngu.mgh.harvard.edu/~purcell/plink/*, Purcell *et al.*, 2007), was used for quality control and population stratification analysis in the following sections.

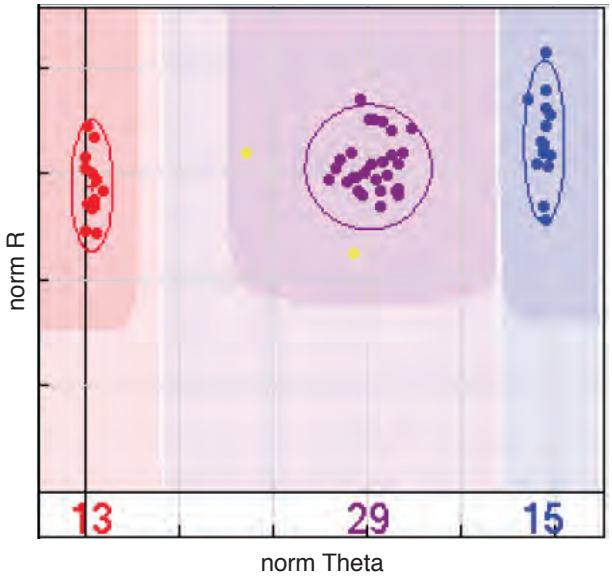


Figure 4.1: An example of the Infinium SNP genplot. The x and y axes represent normalised theta and signal intensity R respectively. Samples clustered in the blue and red shaded regions are homozygous. Samples clustered in the purple shaded region are heterozygous. Samples displayed as yellow points are outliers. Adapted from the Illumina GenomeStudio GT module guide (Illumina, 2010).

4.3.3 Genotyping rate and filtering

Due to the nature of Illumina Infinium HD genotyping system that uses probes to hybridise flanking SNP loci, not all SNPs are successfully genotyped because of failure of probe hybridisation and other factors such as poor DNA quality. The failure of genotype calling was determined by genotyping rates of each SNP as well as the GenCall scores. Minor allele frequency (MAF) is another determinant of SNP filtering because rare variants in a population may result in false positives. Although the definition for common and rare variants vary in the literature, MAF>1% cutoff was used in this dataset. Two criteria for SNP inclusion were used; for SNPs with MAF >5%, genotyping call rate at >98%; for SNPs with MAF between 1 to 5%, genotyping call rate at>99%. A total of 639719 SNPs (592775 with MAF>5% and 46944 with MAF between 1 and 5%) were included for population stratification analysis. After SNP filtering, the average SNP call rate was 99.9%.

4.3.4 Sample filtering

It is important to ensure that there is no discordance on samples in sex, DNA contamination or duplication. Discordance of sex was determined by the heterozygosity rate of SNPs on the X chromosome. The discrepancy of sex between genotyping and that recorded at

collection can be identified by calculation of the heterozygosity rate of SNPs on the X chromosome because males have only one X chromosome resulting in a rate of 1. Typically the heterozygosity rate of females is <0.2. All individuals in the healthy volunteer cohort passed sex check without discrepancy to the reported information.

Poor quality DNA causes missed calls in genotyping and results in skewed heterozygosity rates of individuals. The genotyping and heterozygosity rates per individual can measure the quality of DNA samples. Typically individuals with 3-7% genotyping missing rates are excluded (Anderson *et al.*, 2010). In this population, the average of the sample call rates was 99.72% but one sample had a lower rate (94%) and was removed. The heterozygosity rate varies in each population and genotyping platforms, therefore the distribution of heterozygosity in a population can be used to identify individuals with biased proportion of the heterozygosity rates. These are calculated by the number of homozygous SNPs (O) and the number of SNPs genotyped without missing values (N). The heterozygosity rate is given by (N-O)/N. The identified samples are suspected to arise from DNA contamination or inbreeding. In this population, the mean heterozygosity rate (33.2%) and the standard deviation (+/-2.9%) were calculated based on 288 individuals. Four individuals, including the sample with low call rate, were identified as potential outliers when 4 standard deviations were applied as threshold.

Related individuals have biased allele frequencies and will over represent genotypes and skew association with possible false positives. In order to identify duplicated or related individuals in the population, identity by descent (IBD) distance between individuals was calculated based on the average proportion of common alleles without sex chromosomes. For this purpose, only independent SNPs that were not in LD to others (r^2<0.2) were selected (130132 SNPs). The IBD calculation identified that two individuals were second-degree relatives (Pi-hat value=0.2451), with the result that one sample of the two was excluded. Thus after quality control, 283 out of 288 (98.3%) samples remained for further analysis.

4.3.5 Population stratification

Population stratification consists of systematic difference in allele frequencies due to different ancestry. It is important to study genetic association in a population of unrelated individuals that have the same genetic ancestry because biased allele frequency leads to spurious allelic association (Cardon and Palmer, 2003). Population stratification, also called population structure, of the study cohort can be assessed by LD structure as well as

allele frequencies. LD is created as a consequence of a number of factors and individuals of shared ancestry tend to inherit the same LD structure. Comparison of LD structure of individuals in a population with other defined populations reveals homogeneity of the genetic background.

Multidimensional scaling (MDS) was chosen to compare the population stratification of the healthy volunteer population with the defined HapMap populations (Altshuler *et al.*, 2010). MDS is an exploratory technique to visualise the level of similarity of individuals in a dataset in a low dimensional space (Kruskal, 1964a, Kruskal, 1964b). MDS implemented in PLINK uses a pairwise IBD matrix to determine differences in ancestry. A total of 77063 SNPs (LD pruned SNPs, r^2<0.2) were used for MDS with our population and CEU (Utah residents with Northern and Western European ancestry from the CEPH collection), TSI (Toscani in Italy), CHB (Han Chinese in Beijing), YRI (Yoruba in Ibadan, Nigeria), and GIH (Gujarati Indians in Houston, Texas) HapMap populations (Figure 4.2). Our population, consisting of 283 individuals after removal of 5 outliers, overlapped with the CEU and TSI European populations (Figure 4.2).

4.3.6 Final dataset

The QC analysis for the genotyping data identified 5 outliers and they were removed for further analysis. I decided to use a less stringent cutoff for SNP call rate for the final dataset. The stringent SNP call rate (>98%) was important for the QC step in order to identify individual outliers with a mixed genetic background. Therefore, the final dataset consists of 283 out of 288 individuals (98.2%) with a total of 651210 SNPs with MAF >1%, SNP call rate >96% and sample call rate >97%.

4.4 Cell type specific expression quantitative trait mapping

4.4.1 cis eQTL

Local association (referred to here and in subsequent text as cis associations) of gene expression with flanking SNPs was defined using a 2.5Mbp flanking window either side of the probe start position. The definition of cis distance for this analysis was kept to a relatively wide region in order to detect novel cell specific eQTL. Cis eQTL association was performed using a linear model with permutation (n=1000) and with p-values<1.0×10^{-3}

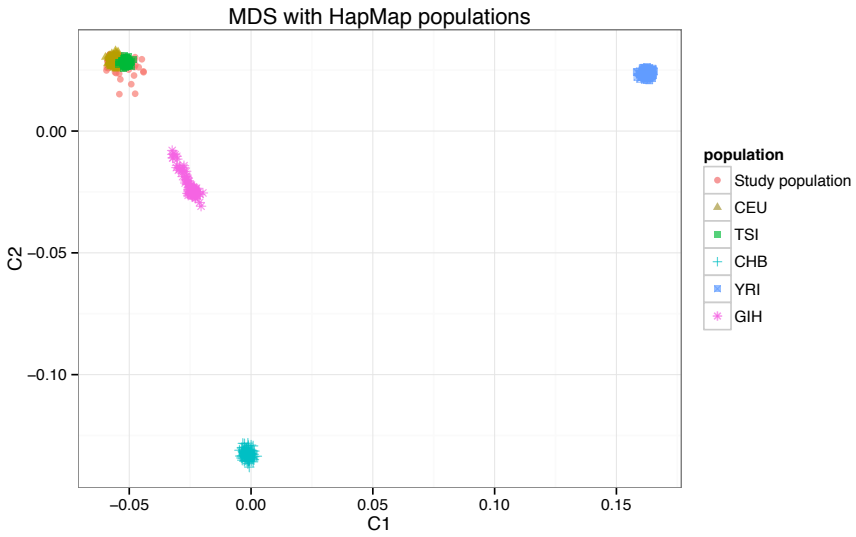


Figure 4.2: MDS analysis of our population with 283 individuals (pink circles), CEU (yellow triangles), TSI (green squares), CHB (light blue crosses), YRI (blue squares), GIH (pink stars).

defined as significant. A total of 82346 SNP-probe associations (referred to here and in subsequent text as expression associated SNPs, eSNPs) were identified in monocytes, B-cells and shared in two cell types involving 37835, 26549 and 17962 associations respectively (Figures 4.3, 4.4 and Table 4.1). 78.1% of eQTL involved cell type specific association. The number of eSNPs involving cis association at different permutation p-value thresholds showed that more associations were identified in monocytes and shared between two cell types than in B-cells (Figure 4.5).

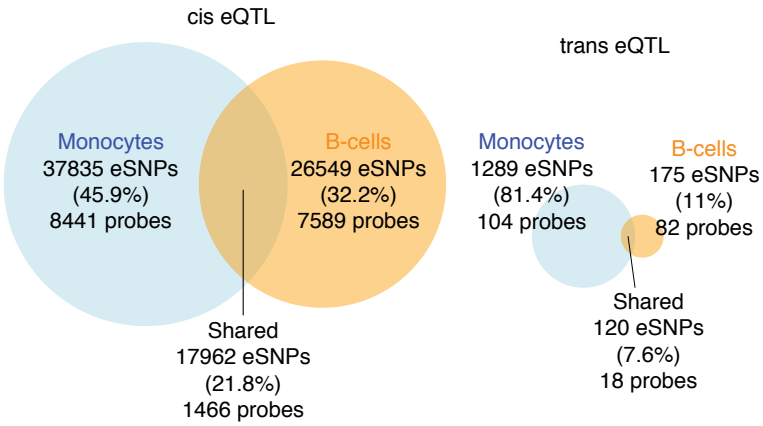


Figure 4.3: Venn diagrams of monocytes and B-cells cis (left) and trans (right) eQTL associations showing the number of unique eSNPs and probes in a cell type specific manner.

Cell type specific eQTL may arise due to differences in gene expression between cell types rather than due to functional genetic variants acting in a cell type specific manner. In order to investigate this, genes expressing similar levels across the two cell types (<0.5 difference in average $\log_2$ expression between monocytes and B-cells) were analysed. This

revealed that 70149 eSNPs were identified as significant cis eQTL. The effect size of each eSNP can emphasise how much effect each eSNP has on the level of gene expression. Although the majority of cis eQTL were cell type specific, the effect size (r^2) of the shared eQTL was larger than cell type specific eQTL (median r^2 in the peak eQTL per gene = 5.5%, 5.2% and 9.4% for monocyte specific, B-cell specific and shared eQTL respectively).

Table 4.1: Summary of cis eQTL in monocytes and B-cells at permutation p-value $1.0x10^{-3}$.

	eSNPs (%)	Probes	Genes
Monocyte	37,835 (45.9)	8,441	7,468
B-cell	26,549 (32.2)	7,589	6,831
Shared	17,962 (21.8)	1,466	1,323
Total	82,346 (100)	17496	15622

4.4.2 Examples of shared eQTL in both cell types

Although the two types of immune cells studied have distinctive roles in the immune system, shared eQTL identified included immunologically important genes. *ERAP2* on chromosome 5 (chr5) was associated in both monocytes and B-cells with rs4360063 and rs10044354 ($p_{mono}=1.05x10^{-110}$, $p_{B-cell}=2.51x10^{-115}$ for both SNPs, Figure 4.6). *ERAP2* encodes endoplasmic reticulum aminopeptidase 2 and is involved with trimming of HLA class I peptides (Saveanu *et al.*, 2005). These two SNPs are in perfect LD (r^2 1). r^2, the square of the correlation coefficient between two alleles, is commonly used for measurement of LD between SNPs (Devlin and Risch, 1995). r^2 is calculated based on a square of a maximum coefficient of LD given allele frequencies (D) over all allele frequencies observed. It ranges from 0 to 1 and where $r^2=1$, two alleles are in perfect LD. rs10239340 is associated with *IRF5* on chr7 ($p_{mono}=1.95x10^{-101}$ and $p_{B-cell}=7.67x10^{-93}$). *IRF5* encodes interferon regulatory factor 5 that has many biological roles, including roles in signal transduction of Toll-like receptors, cell cycle and apoptosis (Takaoka *et al.*, 2005, Barnes *et al.*, 2003). This locus may be important for susceptibility to autoimmune diseases, since a common haplotype of *IRF5* is associated with risk of SLE (Graham *et al.*, 2006). rs4559 in chr12 is associated with *STAT6* (($p_{mono}=1.95x10^{-85}$ and $p_{B-cell}=6.65x10^{-53}$). *STAT6* encodes a transcription factor and signal transducer in the STAT family and is required for IL4 mediated signalling and development of T helper 2 (Th2) cells (Takeda *et al.*, 1996).

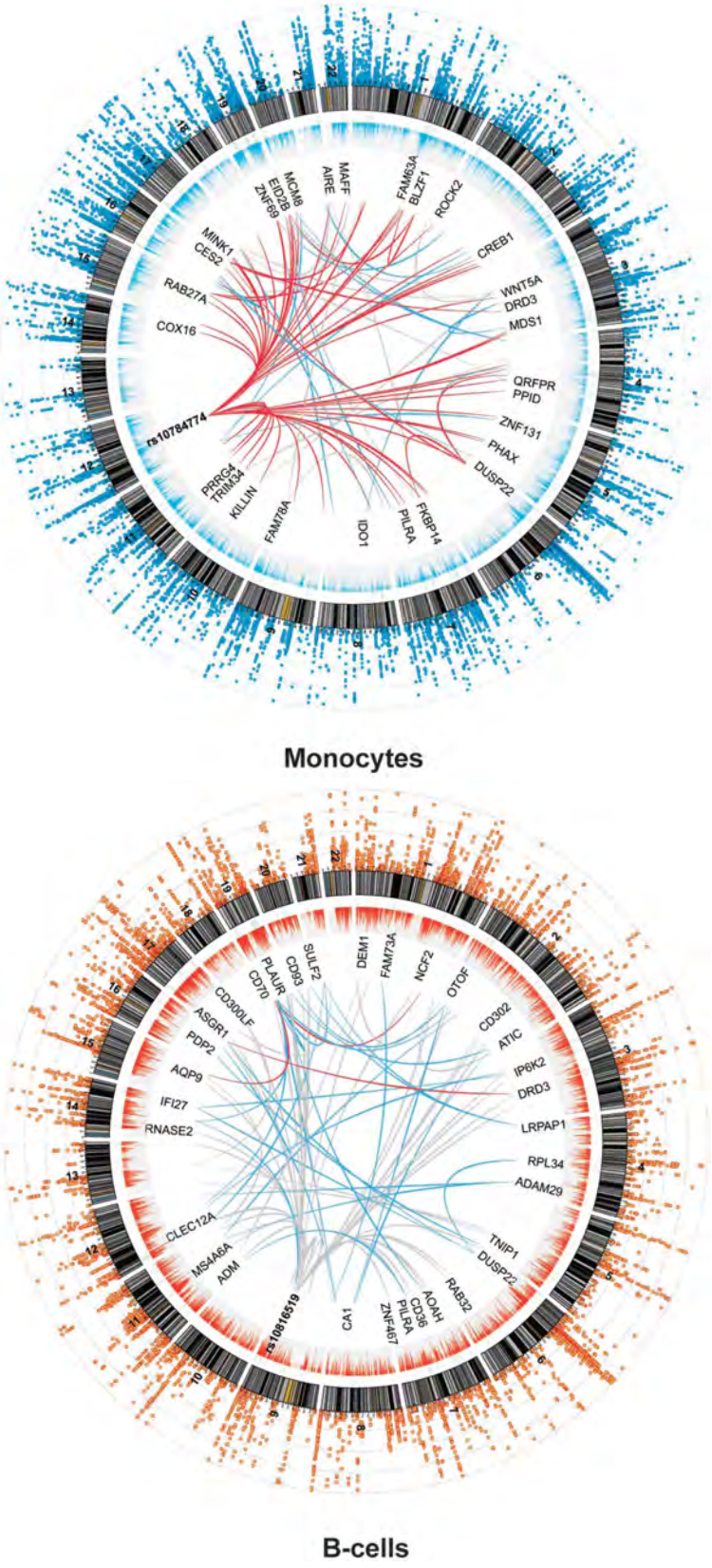


Figure 4.4: Circos plots of monocyte and B-cell eQTL association. The outermost circle indicates Manhattan plots with permutation p-values. The inner circle indicates gene expression levels. The gene names in the inner circle are selected genes that have trans eQTL at $p < 1x10^{-11}$. The lines in the inner circle indicate association of eSNP and trans associated genes. The number of eSNPs associated with each gene is shown by red lines (>10), blue lines (>1 and <10) and grey lines (1).

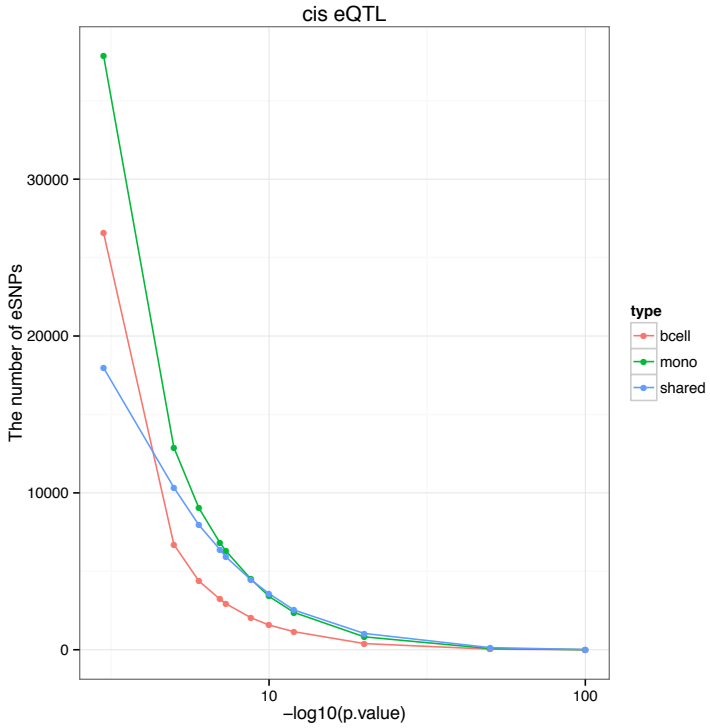


Figure 4.5: The number of cis eSNPs at different permutation p-value thresholds after permutation p-value cutoff $1x10^{-3}$. monocytes (mono, green), B-cells (bcell, pink) and shared (shared, blue).

Directional effect in shared cis eQTL

The regression coefficient (beta) resulting from linear model analysis is an estimate of how much SNP alleles affect gene expression. It also indicates the direction of effect of eSNPs. Previous eQTL analysis across cell types reported a consistent direction of effect of individual eSNPs. Consistent with this, it was observed that the majority of the shared cis eQTL were identified in the same direction in monocytes and B-cells (Figure 4.7, top right and bottom left areas).

Interestingly, among shared cis eQTL (perm $p<1.0x10^{-3}$), a small proportion of eSNPs had opposite effects in monocyte and B-cell gene expression. The opposite direction of regression coefficients from linear model analysis between monocytes and B-cells was defined as directional eQTL. A total of 59 genes and 371 eSNPs were identified (Figure 4.7, top left and bottom right areas). The median effect size of the directional eQTL was r^2=8%. Some genes with directional eQTL are important in immunological processes (Table 4.2). For example, expression of *SELL* encoding L-selectin (also known as *CD62L*) is associated in the opposite way with eSNP rs2223286 at $p_{mono}=1.05x10^{-22}$ and $p_{B-cell}=4.56x10^{-11}$ ($beta_{mono}$=-0.33 and $beta_{B-cell}$=0.14, Figure 4.8). The rs2223286 major homozygous TT genotype (genotype frequency 0.45) is associated with higher expression of *SELL* than the

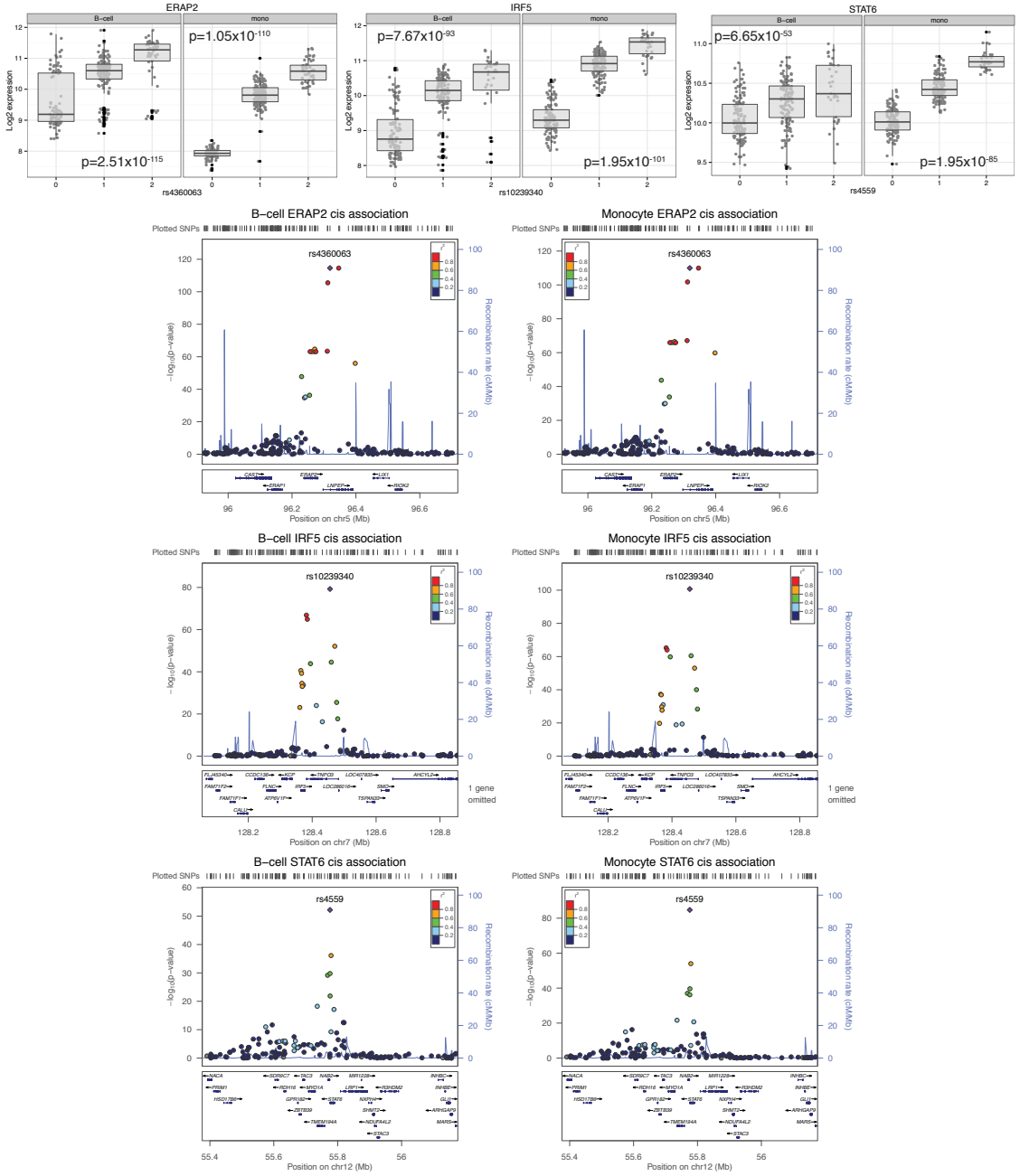


Figure 4.6: Box (top) and association (bottom) plots for shared cis eQTL between monocytes (mono) and B-cells (B-cell). P-values for the peak SNPs are indicated in the box plots (The most significant eSNPs: *ERAP2*; rs4360063, *IRF5*; rs10239340, *STAT6*; rs4559).

minor homozygous CC genotype (genotype frequency 0.11) in monocytes. On the other hand, TT genotype individuals has lower expression than CC genotype individuals in B-cells. Selectin-L is a cell surface adhesion molecule required for migration of leucocytes during inflammation (McEver *et al.*, 1995, Xu *et al.*, 2008). A second example is rs754555 associated with *DFNA5* in both cell types (p_{mono}=1.28x10^{-7} and p_{B-cell}=3.47x10^{-46}, $beta_{mono}$=0.29 and $beta_{B-cell}$=-0.84). The rs754555 major homozygous CC genotype (genotype frequency 0.70) is associated with higher expression of *DFNA5* in B-cells but lower expression in monocytes. The minor homozygous TT genotype frequency is 0.03 and is associated with *DFNA5* expression in the opposite manner. Epigenetic silencing of *DFNA5* has been reported in gastric, colorectal and breast tumours (Masuda *et al.*, 2006, Akino *et al.*, 2007, Kim *et al.*, 2008, Kim *et al.*, 2008).

In the following section, these eQTL associations were validated using quantitative PCR (qPCR).

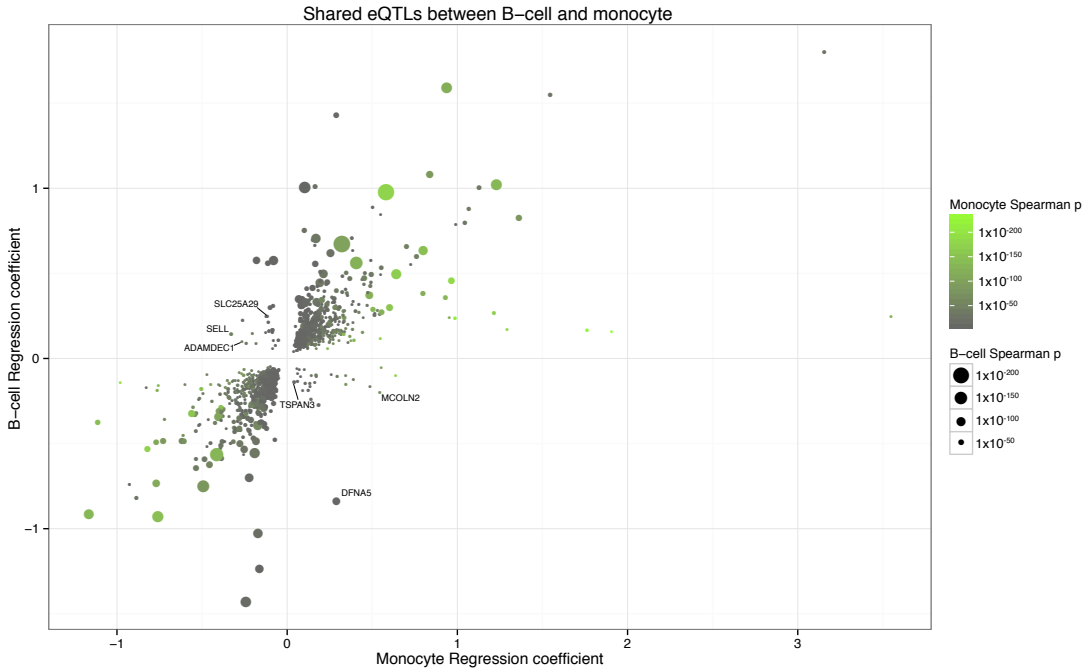


Figure 4.7: Shared cis eQTL between monocytes and B-cells. Regression coefficients from linear model analysis indicating degree of change in gene expression were used to plot and show the direction of eQTL.

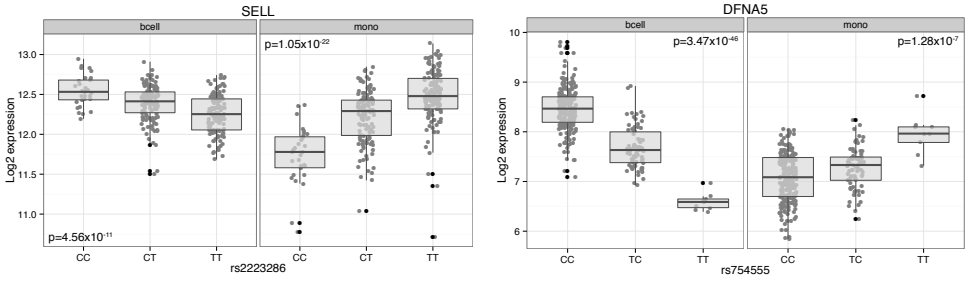


Figure 4.8: Examples of directional eQTL in shared eSNPs in monocytes (mono) and B-cells (bcell). rs2223286 is associated with *SELL* expression. rs754555 is associated with *DFNA5* encoding a protein called deafness autosomal dominant 5 (Van Laer *et al.,* 1998).

Table 4.2: Directional eQTL in shared eSNPs where genes are involved in immunological pathways. mono; monocytes, bcell; B-cells, Beta; regression coefficient in a linear model.

eSNP	p_{mono}	Beta	p_{B-cell}	Beta	Gene	Known function	Reference
rs2223286	$1.05x10^{-22}$	-0.329	$4.56x10^{-11}$	0.142	*SELL*	Selectin-L, an adhesion molecule	McEver *et al.*, 1995, Xu *et al.*, 2008
rs754555	$1.28x10^{-07}$	0.291	$3.47x10^{-46}$	-0.836	*DFNA5*	A tumour suppressor	Van Laer *et al.*, 1998, Masuda *et al.*, 2006, Akino *et al.*, 2007, Kim *et al.*, 2008
rs4559243	$6.31x10^{-20}$	-0.265	$2.56x10^{-04}$	0.1	*ADAMDEC1*	Correlated with severity of pulmonary sarcoidosis	Crouser *et al.*, 2009
rs10873515	$6.76x10^{-21}$	0.287	$1.67x10^{-04}$	-0.189	*SLC25A29*	A mitochondrial transporter	Camacho and Rioseco-Camacho, 2009
rs7169831	$2.05x10^{-09}$	0.132	$6.92x10^{-08}$	-0.126	*TSPAN3*	Tetraspanins, involved in tumour development	Wright *et al.*, 2004
rs10873682	$1.53x10^{-37}$	0.544	$4.23x10^{-06}$	-0.2	*MCOLN2*	Type I interferon response	Schoggins *et al.*, 2011

4.4.3 Validation of eQTL by qPCR

In order to validate findings of eQTL based on array profiling, qPCR with RNA from randomly selected individuals in the study cohort (n=14-29 homozygous individuals) was performed for validation for the eQTL findings. *DFNA5* and *SELL* were tested for qPCR and validated in both B-cells and monocytes in the opposite direction (*DFNA5*: p_{B-cell}=1.88x10^{-8} and p_{mono}=2.32x10^{-3}, *SELL*: p_{B-cell}=6.17x10^{-3} and p_{mono}=2.87x10^{-5}, Figure 4.9H and I). qPCR on PBMCs with *SELL* failed to show any evidence of eQTL (p_{PBMC}=0.28), demonstrating that cell specificity is an important consideration for investigating eQTL, and detecting directional eQTL will be missed in a mixed cell population.

4.4.4 Replication of cis eQTL in published datasets

In order to validate the reproducibility of eQTL identified in this study in terms of published eQTL studies, the current dataset was compared with a human primary monocyte dataset (Zeller *et al.*, 2010), human primary whole blood dataset (Fehrmann *et al.*, 2011), and human lymphoblastoid cell line datasets (Dixon *et al.*, 2007 and Stranger *et al.*, 2007a). The eSNPs with p <5x10^{-8} cutoff from this study were matched by SNP-gene relationships with the published datasets. In order to overcome different tagging SNPs used in various genotyping platforms, additional proxy SNPs (r^2>0.7) were identified for the published datasets using SNAP, an annotation and proxy SNP search tool with the 1000 Genomes Pilot 1 CEU data (*www.broadinstitute.org/mpg/snap/*). SNP-gene pairs were counted as replicated in this study dataset if they significantly associated in published datasets (adjusted p-value<0.05). The primary cell datasets, Zeller *et al* and Fehrmann *et al*, shared more than 50% of all eQTL except comparison with the B-cell dataset (Figure 4.10A). On the other hand, only 27.9% of associations were replicated in shared SNP-gene associations among human primary peripheral blood cell eQTL studies when compared to associations identified in LCLs (Figure 4.10B). This difference may be caused by a difference in genotyping and expression platforms as well as cell types and immortalisation.

4.4.5 Cell type specific eQTL

Novel cell type specific eQTL identified in this study are illustrated by plotting gene expression levels by allele and local association plots centred on the gene regions (Figures

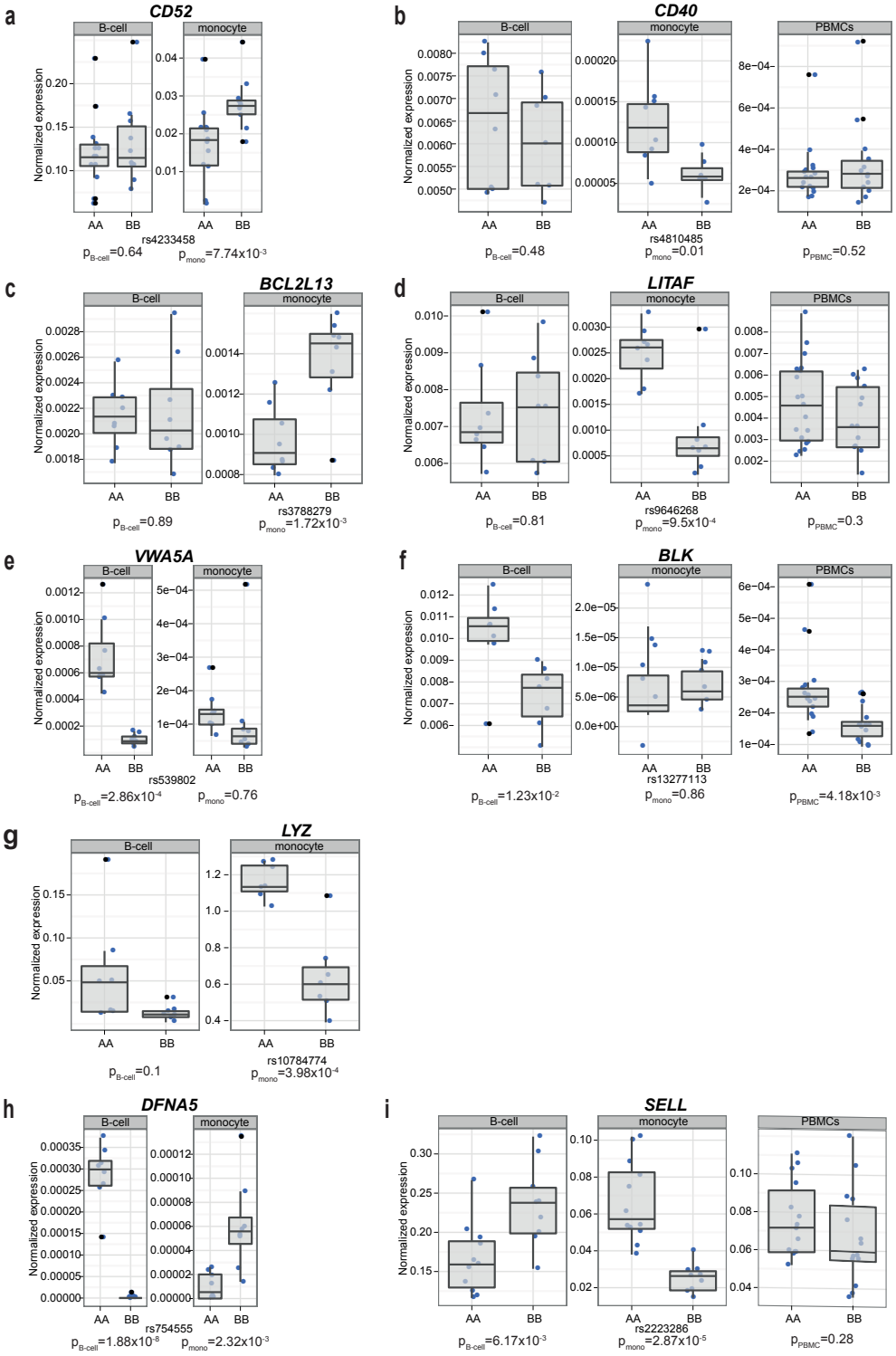


Figure 4.9: qPCR validation of eQTL. (A-D, G) Monocyte specific eQTL. (E-F) B-cell specific eQTL. (H-I) Shared directional eQTL. (A) rs4233458 with *CD52* (p_{mono}=7.74x10^{-3}). (B) rs4810485 with *CD40* (p_{mono}=0.01). (C) rs3788279 with *BCL2L13* (p_{mono}=1.72x10^{-3}). (D) rs9646268 with *LITAF* (p_{mono}=9.5x10^{-4}). (E) rs539802 with *VWA5A* (p_{B-cell}=2.86x10^{-4}). (F) rs13277113 with *BLK* (p_{B-cell}=1.23x10^{-2}). (G) rs10784774 with *LYZ* (p_{mono}=3.98x10^{-4}). (H) rs754555 with *DFNA5* (p_{B-cell}=1.88x10^{-8}, p_{mono}=2.32x10^{-3}). (I) rs2223286 with *SELL* (p_{B-cell}=6.17x10^{-3}, p_{mono}=2.87x10^{-5}).

Table 4.3: Comparison of eQTL studies using significant SNP-gene associations ($p<5x10^{-8}$). *; patients include 49 UC, 111 celiac disease, 453 chronic obstructive pulmonary disease, 856 amyotrophic lateral sclerosis. ND; no data. Reference: Zeller - Zeller *et al.*, 2010, Fehrmann - Fehrmann *et al.*, 2011, Dixon - Dixon *et al.*, 2007, Stranger - Stranger *et al.*, 2007a.

Study	Population	Size	Detail	Cell type
Present dataset	Healthy European descents	288		Monocytes & B-cells
Zeller	Gutenberg Heart Study in Mainz, Germany	1490		Monocytes
Ferhmann	Mixed population of Dutch and British patients and controls	1469	414 healthy, 1055 patients*	Whole blood
Dixon	British children with asthma	400		LCLs
Stranger	Mixed HapMap II population	270	30 CEU trios, 45 CHB, 45 JPT, 30 YRI trios	LCLs
Study	Genotyping platform	SNPs	MAF (%)	
Present dataset	Illumina Human OminiExpress v1	651210	>1	
Zeller	Affymetrix Genome-Wide Human SNP Array 6.0	675350	>1	
Ferhmann	Illumina HumanHap300, HumanHap370, 610 Quad	289044	>5	
Dixon	Illumina Sentrix Human-1, HumanHap300	408273	>5	
Stranger	Illumina WG-6 v1	~2.2 million	>5	
Study	Expression array platform	Probes	Genes	
Present dataset	Illumina HumanHT-12 v4	29022	22011	
Zeller	Illumina HumanHT-12 v3	ND	12808	
Ferhmann	Illumina HumanRef-8 v2, HumanHT12	52061	19609	
Dixon	Affymetrix U133 Plus 2.0	54675	20599	
Stranger	Illumina Sentrix Human-6 v1	14456	13643	

4.11, 4.12 and 4.12). The identified cell type specific eQTL included immunologically important genes previously implicated in disease pathogenesis. The genes that formed eQTL were tested for enrichment in biological pathways using an R package based on the Reactome curated database as described in Section 3.6. Unique Entrez gene IDs in each cell type, B-cells 5418 and monocytes 5931, and shared between the two cell types 1287 genes were interrogated against 26766 genes as the background using the eQTLpath that implemented an R package ReactomePA provided by Peter Humburg (Wellcome Trust

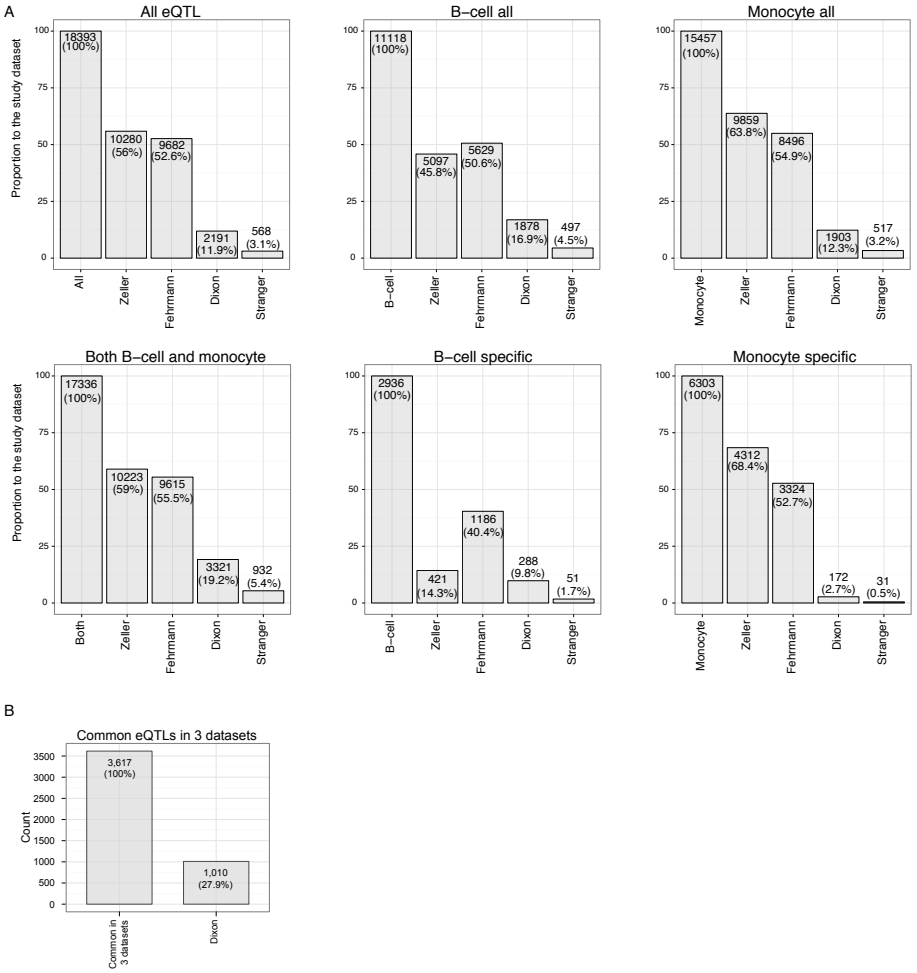


Figure 4.10: Proportion of eSNPs ($p<5x10^{-8}$) observed in this study dataset with evidence of replication in previously published eQTL studies. (A) The counts of matches in the eSNP-gene relationships were compared with eSNPs in this study dataset ($p<5x10^{-8}$). (B) eSNP-gene associations shared among three human primary peripheral blood cell eQTL studies.

Centre for Human Genetics). There was no significant enrichment in pathways when B-cell specific, monocyte specific or shared eQTL groups were analysed.

B-cell specific eQTL

B-cell specific eQTL include examples of genes involved in B-cell development, signal transduction for B-cell activation, even though these genes were not enriched in biological pathways as determined above. For example, rs13277113 on chr8 is associated with *BLK* in B-cells at $p=3.57x10^{-14}$, this gene encodes a B-lymphoid tyrosine kinase exclusively expressed in B-cells and thymocytes (Dymecki *et al.*, 1990). The ectopic expression of *BLK* is observed in a malignant T cell line from patients with cutaneous T cell lymphoma (Krejsgaard *et al.*, 2009). This locus on chr8 has been identified from a GWAS on SLE with odds ratio 1.39 ($p=1x10^{-10}$, Hom *et al.*, 2008).

B-cell specific eQTL were found involving cell surface receptors. For example, rs6773238 on chr3 is associated with *CD96* at p=1.34^{-93}. *CD96* (also known as Tactile) encodes a protein of the immunoglobulin superfamily and may play a role in the adhesive interactions of activated T and NK cells during the late phase of the immune response (Fuchs *et al.*, 2004). rs539802 on chr11 is associated with *VWA5A* at p=$7.48\text{x}10^{-55}$. *VWA5A* encodes von Willebrand factor A domain containing 5A and is a potential tumour suppressor. qPCR for *VWA5A* and *BLK* validated the array quantified eQTL was significant in B-cells but not in monocytes (Figure4.9E and F). Other examples of biologically interesting eQTL include rs734074 on chr1 which is associated with *FAM189B* at p=$1.97\text{x}10^{-29}$. *FAM189B* encodes a protein that may interact with proteins involved in apoptosis and tumour suppression. Finally, rs2168519 on chr15 is associated with *CSK* at p=$1.97\text{x}10^{-9}$. *CSK* encodes a cytoplasmic tyrosine kinase and negatively regulates lymphocyte-specific protein tyrosine kinase (Okada *et al.*, 1991).

Monocyte specific eQTL

Monocyte specific eQTL genes include genes involved in signal transduction for monocytes during immune and inflammatory responses. As determined above, there was no biological pathway that was enriched for the monocyte-specific eQTL. Interesting associations include rs745307 on chr20 which was associated with *CD40* at p=$3.25\text{x}10^{-25}$. *CD40* encodes a member of the TNF-receptor superfamily and its protein product interacts with TNF receptor-associated factors (TRAFs) to mediate the signal transduction that induces various immune and inflammatory responses (O'Sullivan and Thomas, 2003). rs4810485, another peak SNP with *CD40* (p=$5.95\text{x}10^{-15}$) is identified as a risk variant of autoimmune diseases RA and MS (p=$8.2\text{x}10^{-9}$ and $2.2.\text{x}10^{-4}$ respectively, Raychaudhuri *et al.*, 2008, Bahlo *et al.*, 2009).

rs281437 on chr19 is associated with *ICAM4* at p=$2.77\text{x}10^{-92}$. *ICAM4* encodes intercellular adhesion molecule 4 and the Landsteiner-Wiener blood group antigen that belongs to the immunoglobulin (Ig) superfamily. rs9646268 on chr16 is associated with *LITAF* at p=$3.12\text{x}10^{-38}$. *LITAF* encodes a lipopolysaccharide-induced TNF factor. The protein product can mediate *TNF* expression by directly binding to the promoter region to activate several inflammatory cytokines, TNF, IFNγ, MCP-2 and IL-10 (Myokai *et al.*, 1999, Tang *et al.*, 2005). rs3788279 on chr22 is associated with *BCL2L13* at p=$1.11\text{x}10^{-37}$. *BCL2L13* (also known as *Bcl-rambo*) encodes a mitochondrially-localised protein that

regulates apoptosis (Kataoka *et al.*, 2001). qPCR on *CD40*, *LITAF* and *BCL2L13* validated these eQTL as significant in monocytes but not in B-cells (Figure 4.9B-D).

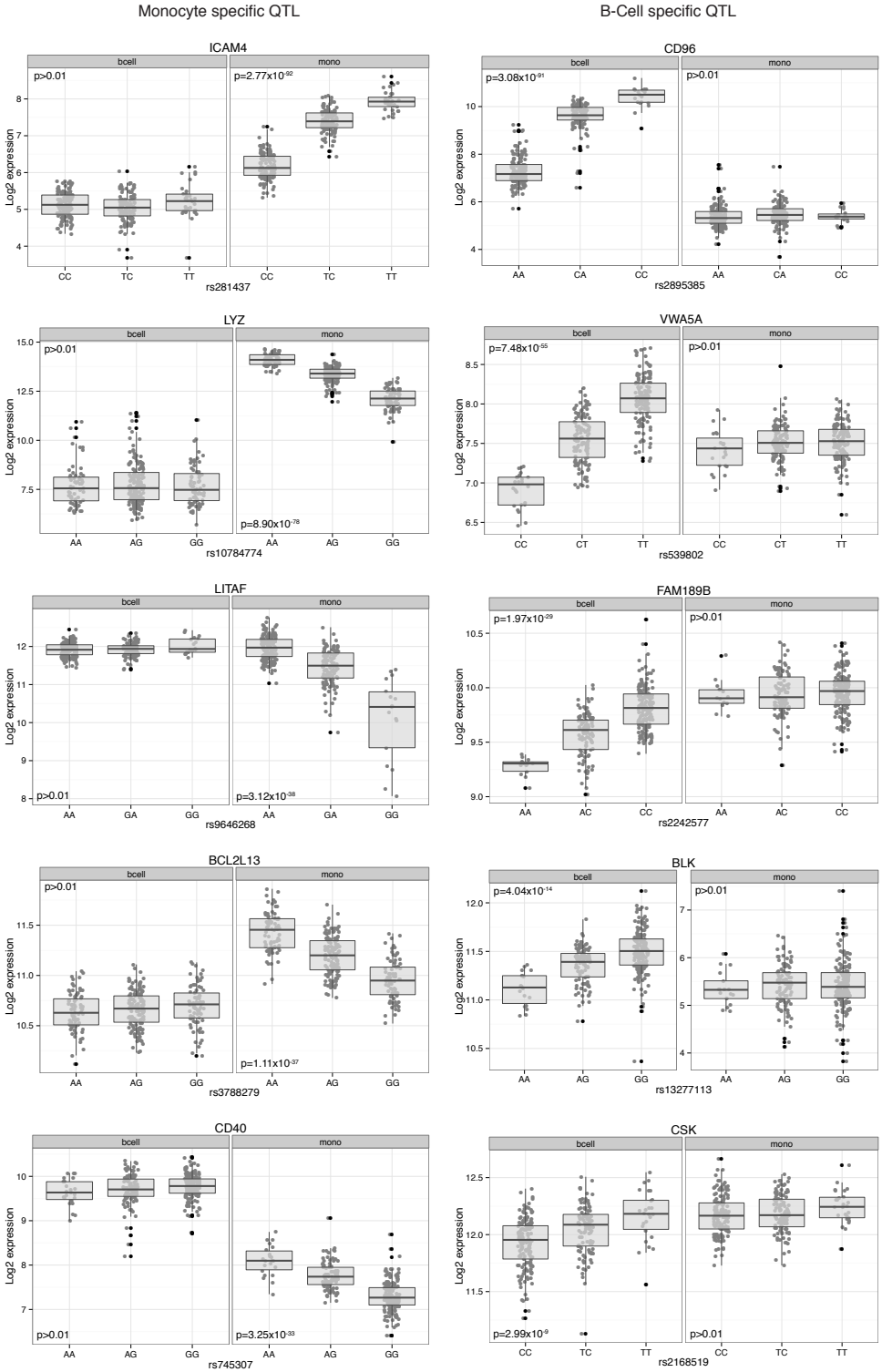


Figure 4.11: Gene expression levels of cell type specific eQTL. Left column - monocyte (mono) specific eQTL. Right column - B-cell (bcell) specific eQTL.

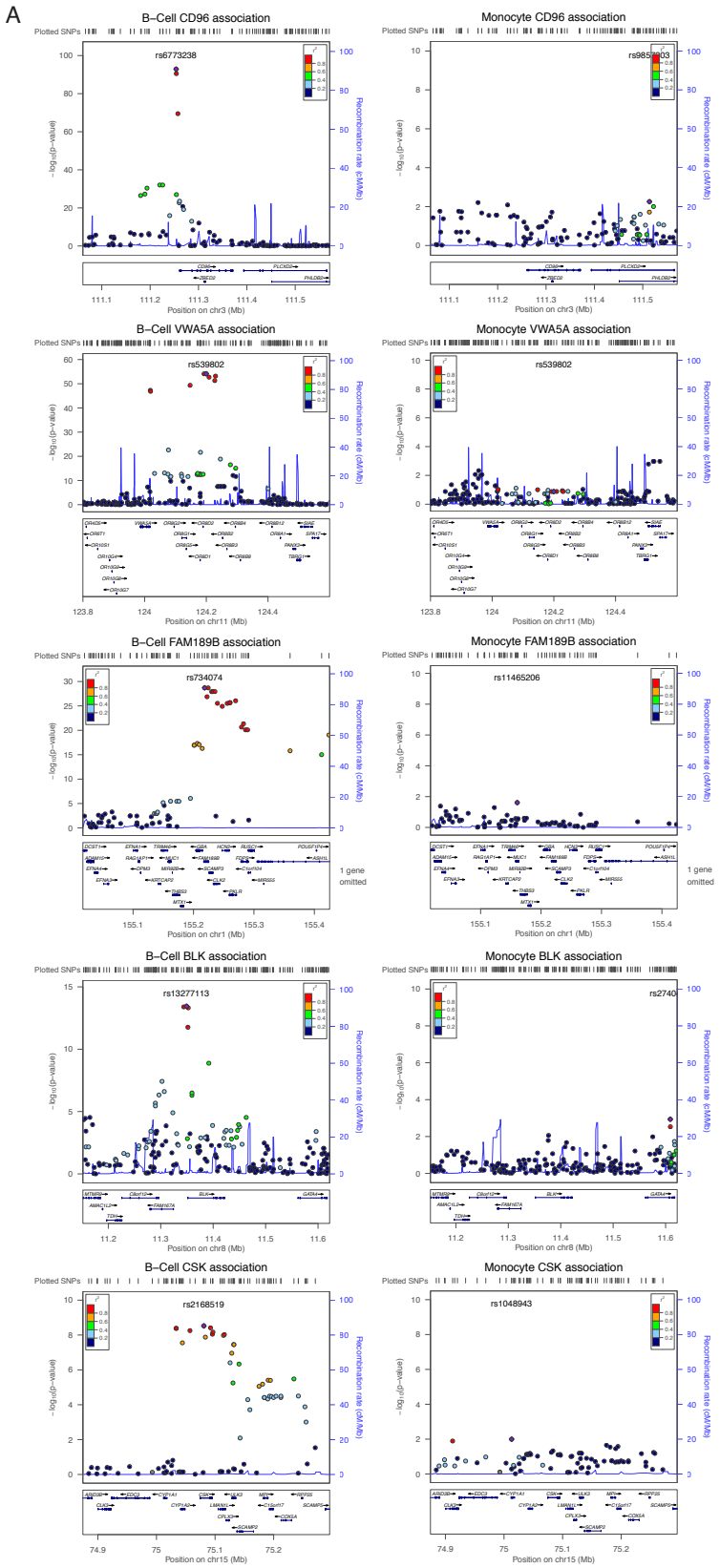


Figure 4.12: Cell type specific eQTL association plot. The x and y axes indicate chromosome position and p-values. (A) B-cell specific eQTL. rs6773238 is associated with *CD96* at p=1.34^{-93}. rs539802 is associated with *VWA5A* at p=7.48x10^{-55}. rs734074 is associated with *FAM189B* at p=1.97x10^{-29}. rs13277113 is associated with *BLK* at p=4.04x10^{-14}. rs2168519 is associated with *CSK* at p=1.97x10^{-9}.

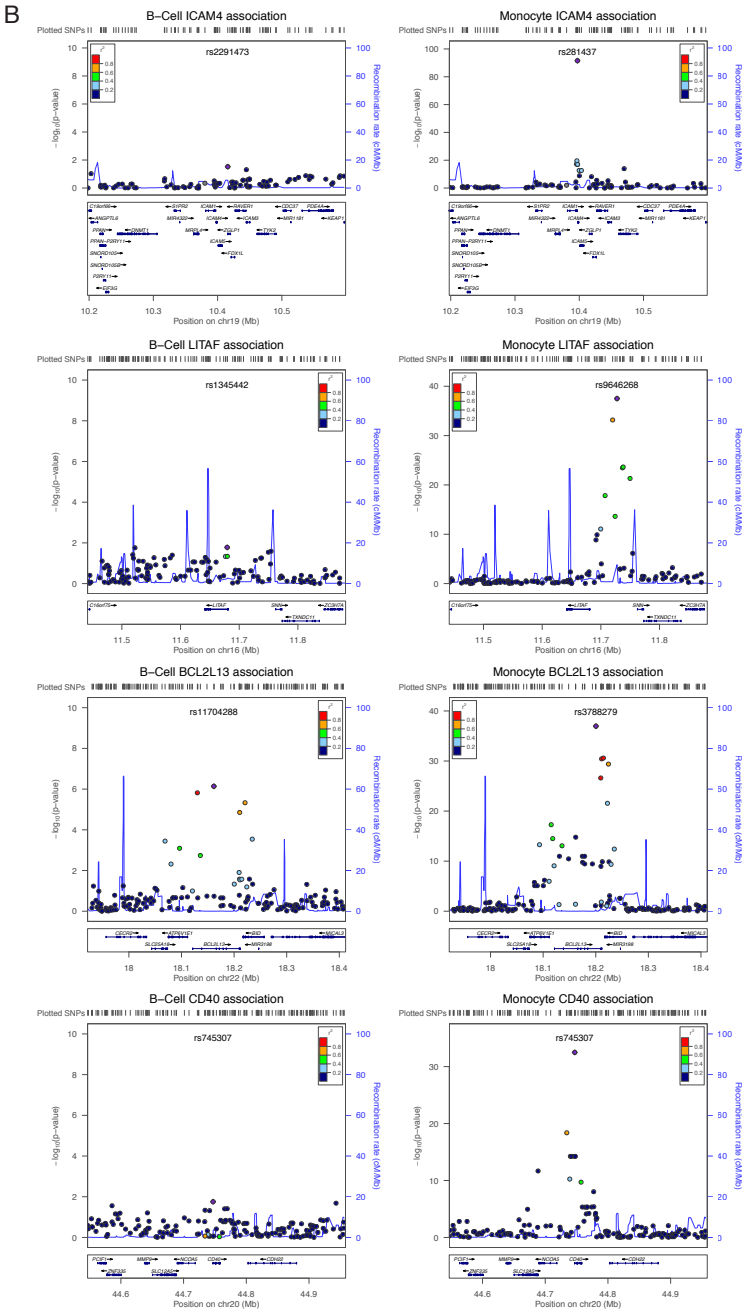


Figure 4.12: Cell type specific eQTL association plot. The x and y axes indicate chromosome position and p-values. (B) Monocyte specific eQTL. rs281437 is associated with *ICAM4* at p=2.77x10^{-92}. rs9646268 is associated with *LITAF* at p=p=3.12x10^{-38}. rs3788279 is associated with *BCL2L13* at p=1.11x10^{-37}. rs745307 is associated with *CD40* at p=3.25x10^{-25}.

4.4.6 Location of cis eQTL in relation to transcription start site

The relationship between the location of cis eQTL and transcription start sites (TSS) was analysed. This is of interest given previous reports suggesting that more distal cis eQTL regulatory effects may be critical in determining cell type specific gene expression and hence cis eQTL (Dimas *et al.*, 2009). The plot based on distance from TSS to all eQTL showed that the cell type specific eQTLs spread more distally to both sides of the TSS in comparison to the shared eQTL for the two cell types (Figure 4.13). The Kruskal-wallis test was performed to determine the distribution of the distance between eQTL and TSS. It was significantly different in the cell type specific datasets compared to when it was tested against the shared (p_{bcell}=0.01, p_{mono}=5.63x10^{-5}). There was no significant difference between the B-cell and monocyte datasets (p=0.71).

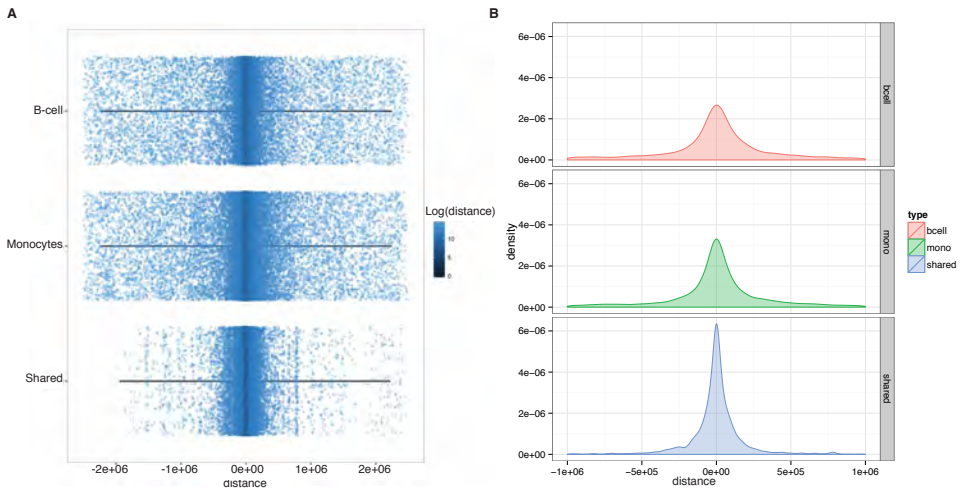


Figure 4.13: Boxplot (A) and density plot (B) of eQTL in relation to distance from TSS for cell type specific and shared eQTL. (A) +/-2.5Mb and (B) +/-1Mbp distance from TSS.

4.4.7 trans eQTL

Although the precise mechanism driving trans association is not often clear, it is thought that many result from cis eQTL driving a gene network that results in association with expression of several genes, often on another chromosome. Previous studies have shown trans eQTL in LCLs and primary cells (Dixon *et al.*, 2007, Stranger *et al.*, 2007a, Fehrmann *et al.*, 2011, Rotival *et al.*, 2011). This dataset with two highly purified cell types allows investigation of cell type specific trans eQTL. In this dataset trans eQTL was defined with association to SNPs greater than 2.5Mbp away from the probe target sequence and a p-value<1.0x10^{-11}. This threshold based on Bonferroni correction in order to avoid type I error (false positives) due to multiple testing performed in this dataset (p-value cutoff

0.05, 29022 probes and 648835 SNPs tested). In autosomal genes and SNPs, there were 1289, 175 and 120 SNP-probe associations (eSNPs) identified in monocytes, B-cells and shared in both cell types corresponding to 101, 75 and 19 genes and 204, 86 and 124 SNPs (Table 4.4, Figure 4.4). Some particular eSNPs were identified to associate with more than 20 genes in trans in a cell type specific manner (Figure 4.4, inner line networks). Of particular interest, eSNPs in 12q15 in monocytes and 9q31.2 in B-cells were identified as potential master regulators of transcription given the observed in trans eQTL. Although monocyte trans association accounts for more than 80% of the total trans association, this is driven by the very large number of associations involving the 12q15 locus in monocytes (1032 out of 1289 eSNPs).

Table 4.4: Summary of trans eQTL in monocytes and B-cells at p-value<$1.0x10^{-11}$.

	eSNPs (%)	SNP (%)	Probes	Genes
Monocyte	1289 (81.4)	204 (49.3)	104	101
B-cell	175 (11)	86 (20.7)	82	75
Shared	120 (7.6)	124 (30)	18	19
Total	1584 (100)	414 (100)	204	195

4.4.8 Monocyte specific trans eQTL involving 12q15

The most significant trans-associated SNP rs10784774 is located approximately 4kb upstream of *LYZ* at 12q15 and was shown to associate with 72 probes (62 genes) in trans (Figure 4.14). This SNP was associated in cis to *LYZ* (Figure 4.11, p=$8.9x10^{-78}$). *LYZ* encodes lysozyme, an enzyme that degrades peptidoglycans derived from bacterial cell walls (Fleming and Allison, 1922). Another SNP rs1384 located in the 3'UTR of *LYZ* associated at the same level (p=$8.9x10^{-78}$) and is in strong LD with rs10784774 (r^2 0.967).

The locus 12q15 has 23 SNPs that associate with at least 73 probes (72 genes) throughout the genome. The top associated genes to rs10784774 in trans are *CREB1* encoding a transcription factor on chr2 (p=$2.03x10^{-67}$), *DUSP19* encoding dual specificity phosphatase 19 on chr2 (p=$2.21x^{-59}$) and *RAB27A* encoding a protein in the RAS oncogene family on chr15 (p=$2.03x10^{-57}$) (Figure 4.15). This locus was also identified, although not described in detail, in previously published studies in human primary cell eQTL datasets (negatively separated monocyte population by Zeller *et al.*, 2010 and whole blood by Fehrmann *et al.*, 2011). Fehrmann *et al.*, 2011 demonstrated rs2168029, downstream of *LYZ*, was strongly associated with *LYZ* (p=$9.8x10^{-198}$) and *CREB1* (p=$3.2x10^{-67}$),

which is consistent with the finding in this monocyte dataset. rs2168029 is in strong LD with rs10784774 and rs1384 (r^2 0.868 and 0.901 respectively).

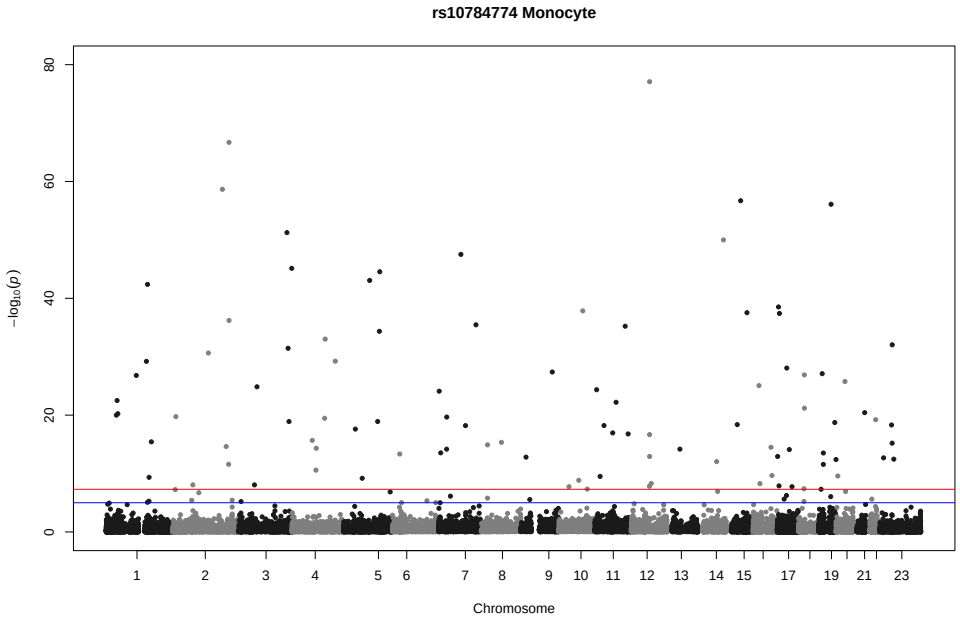


Figure 4.14: Manhattan plot of rs10784774 trans association in monocytes. X and Y axes indicate chromosome location and p-values of probes. Red and blue lines indicate p-values for Bonferroni corrected significant ($5x10^{-8}$) and suggestive ($1x10^{-5}$) thresholds for GWAS. 72 probes (62 genes) associated with the SNP are scattered around the genome.

4.4.9 Expression correlation of *LYZ* and trans genes

Although *LYZ*, in association with rs10784774 in cis, encodes an enzyme and not a TF, its expression level may be correlated with the expression levels of the trans genes associated in an allelic manner. Thus the expression levels of the trans associated genes in each genotypes (two homozygous, one heterozygous) were analysed for their correlation with *LYZ* expression in a linear model.

The analysis identified that the expression of 41 expression probes were correlated with *LYZ* expression in each genotype with at p-value<$6.85x10^{-4}$ (Bonferroni correction for multiple testing for 73 genes). *CREB1*, *DUSP19* and *RAB27A* are negatively correlated with *LYZ* expression (Figure 4.16) although the correlation of *RAB27A* expression with *LYZ* was not significant in the minor homozygous genotypes (p=0.011, Figure 4.16). *CREB1* encodes a transcription factor cAMP responsive element binding protein 1 that stimulates many viral and cellular promoters. Its co-transcription factor P300 was not identified as significant in this correlation analysis. *KLLN* encoding a p53-regulated DNA replication inhibitor and a nuclear protein was identified as significant in all

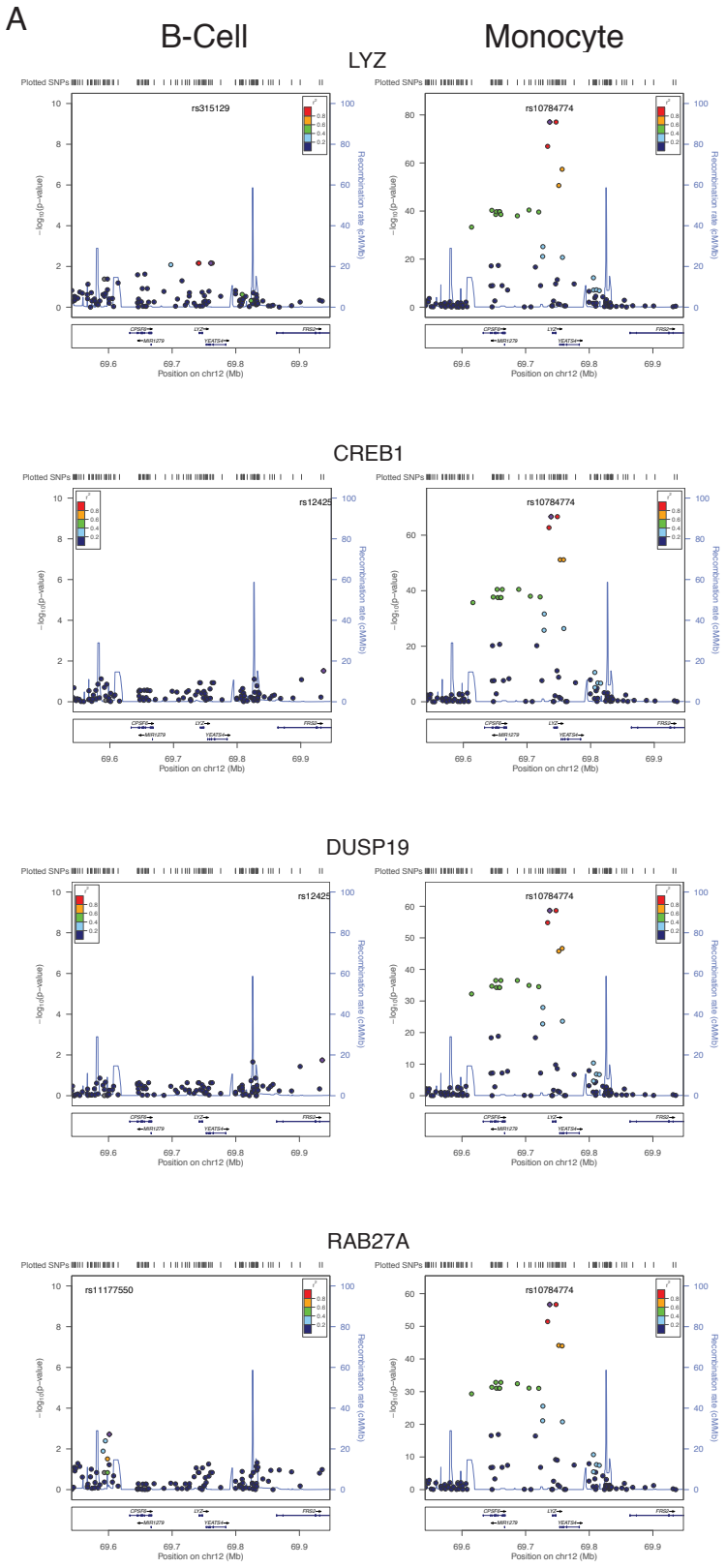


Figure 4.15: rs10784774 trans associated genes in monocytes. (A) Regional association plots. The X and Y axes indicate chromosome location and p-values of probes.

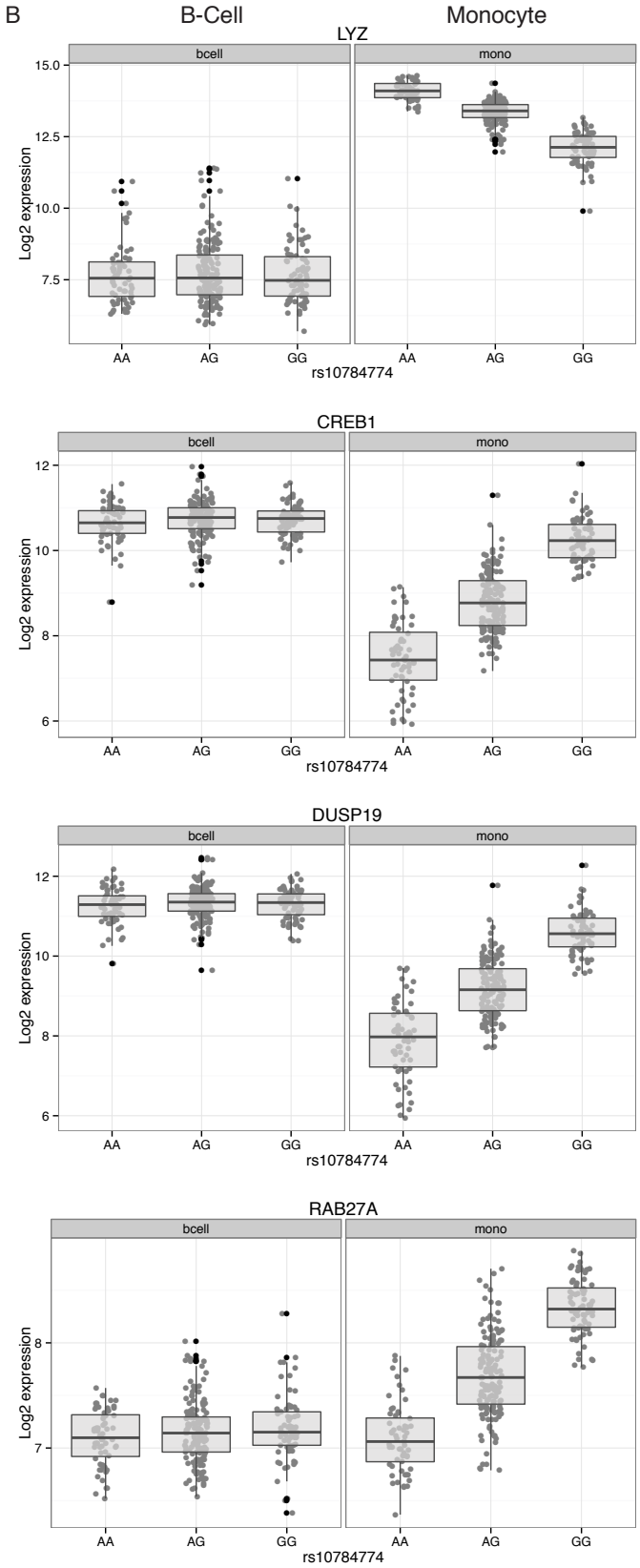


Figure 4.15: rs10784774 trans associated genes in monocytes. (B) Box plots of *CREB1* (p_{mono}=2.03x10^{-67}), *DUSP19* (p_{mono}=2.21x10^{-59}) and *RAB27A* (p_{mono}=2.03x10^{-57}). B-cells (bcell) plots shown for comparison to monocytes (mono).

three genotypes (Figure 4.16, Cho and Liang, 2008). The expression of *KLLN* is strongly associated in trans to 12q15 including rs10784774 ($p=1.59x10^{-38}$). The expression of *GSDM1* was associated with rs10784774 ($p=8.92x10^{-29}$) and correlated with *LYZ* expression in all three genotypes (Figure 4.16). *GSDM1* encodes gasdermin A which is involved with induction of apoptosis and presumably linked to carcinogenesis (Saeki *et al.*, 2007). The analysis revealed that *LYZ* expression is highly associated with the expression of the trans associated genes in an allelic manner.

rs10784774 is also associated with the adjacent TF gene *YEATS4* in cis at $p_{mono}=1.26x10^{-13}$ and $p_{B-cell}=1.55x10^{-7}$. Although *YEATS4* is a candidate TF for driving trans eQTL activity, the eQTL is additionally found in B-cells and the low correlation in gene expression with *LYZ* and *CREB1* (estimates to *LYZ* and *CREB1* are 0.29 and -0.16) suggest it is less likely.

4.4.10 B-cell specific trans eQTL involving 9q31.2

The most significant SNP rs10816519 at 9q31.2 is associated with 52 probes (46 genes) in trans and accounts for 29.7% of all B-cell trans eQTL. rs10816519 is located downstream of *KLF4*, encoding Krueppel-like factor 4, and associated in cis with *KLF4* expression at $p=2.52x10^{-8}$. *KLF4* acts as both a transcriptional suppressor and activator in cell proliferation and differentiation especially in the context of oncogenesis (Rowland *et al.*, 2005). The genes such as *PLAUR* on chr19 ($p=3.58x10^{-14}$), *TNFSF13B* on chr13 ($p=8.03x10^{-14}$), *CD302* in chr2 (C-type lectin) ($p=5.33x10^{-14}$) and *CLEC12A* on chr12 ($p=2.75x10^{-13}$) are associated in trans with rs10816519 (Figure 4.17). However, the frequency of the minor allele C is very low (0.018 in CEU population) and this dataset only has individuals homozygous for the major allele (TT) and heterozygous TC individuals (273 and 10 individuals respectively). Thus the effect of the minor genotype CC is unknown.

4.4.11 Expression correlation of *KLF4* and trans genes

The SNP rs10816519 is associated with gene expression in not only trans but also cis with *KLF4*, a gene encoding a TF which has a role as a transcription suppressor and activator. The expression levels of *KLF4* were analysed in terms of correlation with trans associated genes using a linear model. Due to the small number of the individuals with the minor allele tested (10 heterozygous individuals), only 9 probes (9 genes) out of 52 probes (46 genes) were significant at the p-value threshold $9.43x10^{-4}$ (Bonferroni correction, Table 4.5). All correlations were positive (Figure 4.18). The significant genes included

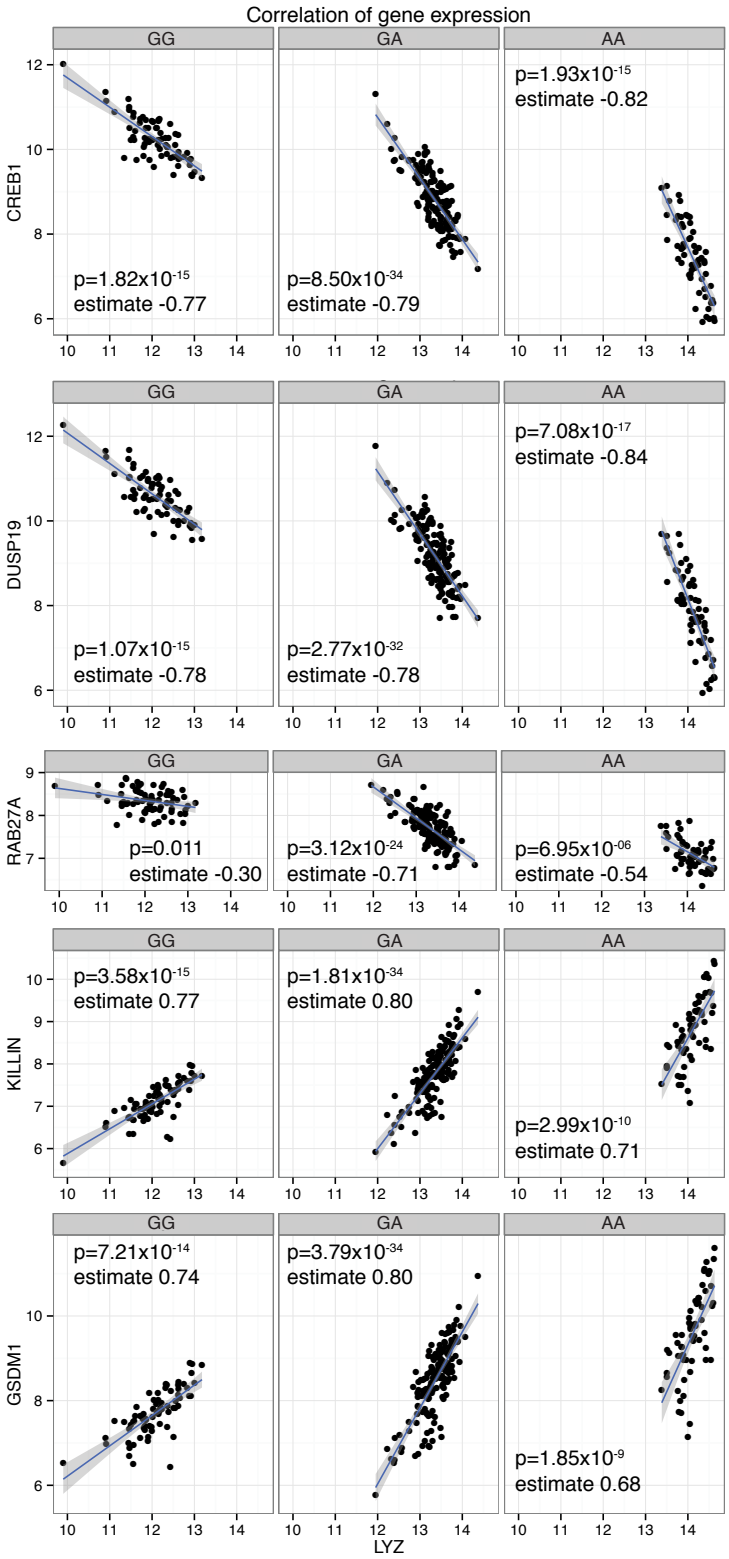


Figure 4.16: Gene expression correlation of *LYZ* and trans associated genes in monocytes for the three genotypes of rs10784774. p-values and estimates of linear correlation are shown in the plot.

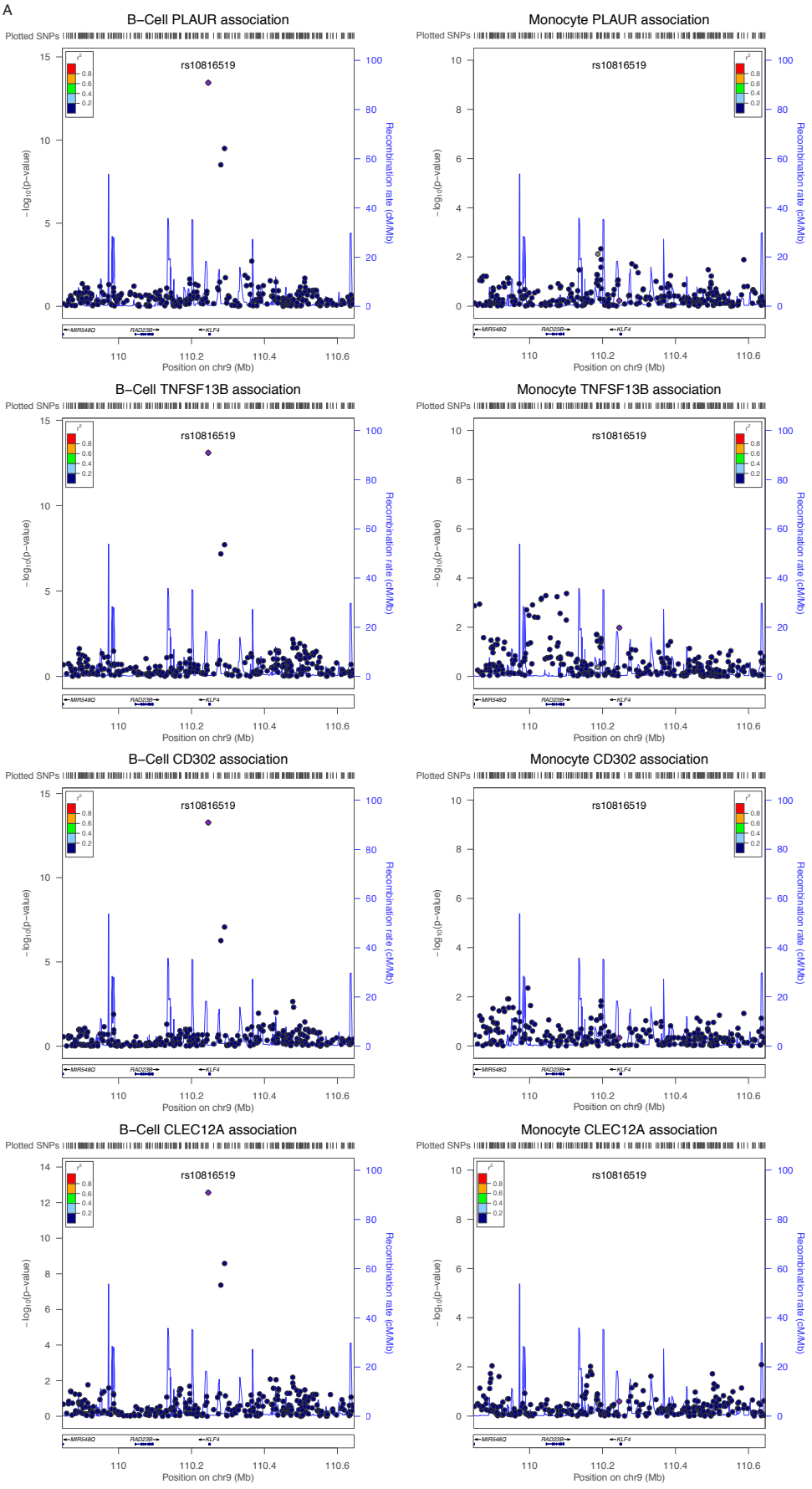


Figure 4.17: rs10816519 trans associated genes in B-cells. (A) Regional association plots. The X and Y axes indicate chromosome location and p-values of probes.

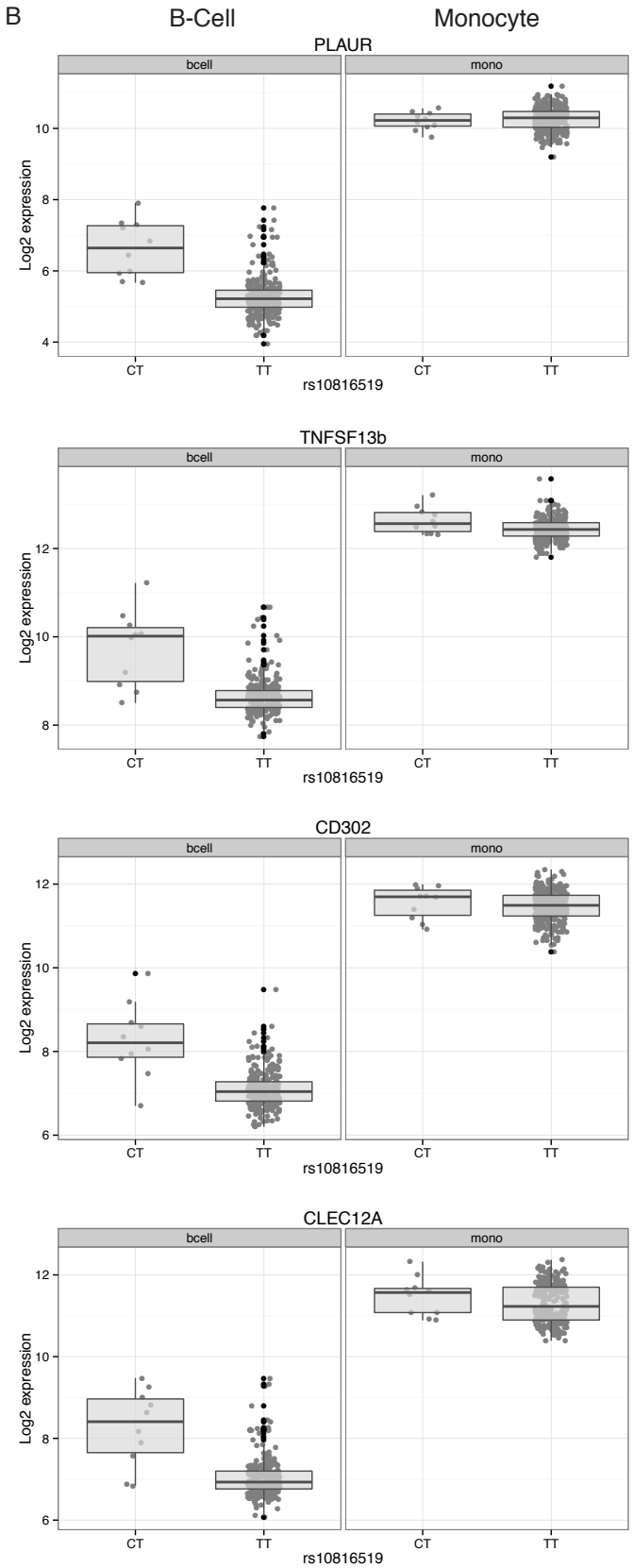


Figure 4.17: rs10816519 trans associated genes in B-cells. (B) Box plots. *PLAUR* on chr19 ($p=3.58x10^{-14}$) encodes a urokinase receptor that is involved in plasminogen activation in both normal and pathological processes (Borglum *et al.*, 1992). *TNFSF13B* on chr13 ($p=8.03x10^{-14}$) encodes a tumour necrosis factor ligand family and acts as a potent B-cell activator. *CD302* on chr2 ($p=5.33x10^{-14}$) and *CLEC12A* on chr12 ($p=2.75x10^{-13}$) encode C-type lectin receptors involved in phagocytosis and the immune response (Kato *et al.*, 2007 and Lahoud *et al.*, 2009). ~~bcell; B-cells, mono; monocytes.~~

a membrane protein *MS4A6A*, iron transporter *CYBRD1* and genes encoding enzymes as well as *PLAUR* and *TNFSF13B*. No TF has been identified as significantly correlating with *KLF4* expression. Notably, *PLAUR* has previously been identified as a target of *KLF4* activity (Wang *et al.*, 2004).

Table 4.5: Trans eQTL genes correlated with *KLF4* expression using a linear model.

		Genotype			
		TC		TT	
Gene	Gene product	estimate	p-value	estimate	p-value
CYBRD1	cytochrome b reductase 1	0.88	$8.48x10^{-4}$	0.45	$1.13x10^{-14}$
PLAUR	plasminogen activator, urokinase receptor	0.93	$8.44x10^{-5}$	0.57	$1.59x10^{-24}$
TNFSF13B	tumor necrosis factor (ligand) superfamily, member 13b	0.90	$4.12x10^{-4}$	0.58	$1.61x10^{-27}$
CD302	CD302 molecule, C-type lectin receptor	0.89	$5.35x10^{-4}$	0.50	$1.83x10^{-18}$
QPCT	glutaminyl-peptide cyclotransferase	0.91	$2.40x10^{-4}$	0.41	$1.72x10^{-12}$
PILRA	paired immunoglobin-like type 2 receptor alpha	0.93	$6.00x10^{-4}$	0.53	$9.39x10^{-29}$
ALDH1A1	aldehyde dehydrogenase 1 family, member A1	0.84	$8.07x10^{-5}$	0.51	$2.57x10^{-19}$
SULF2	sulfatase 2	0.90	$3.83x10^{-4}$	0.53	$1.29x10^{-20}$
MS4A6A	membrane-spanning 4-domains, subfamily A, member 6A	0.86	$4.92x10^{-4}$	0.57	$2.70x10^{-25}$

4.4.12 Pathway analysis of trans associated genes

In order to determine functional relationships of genes associated in trans in each cell type, pathway analysis was performed using an R package based on the Reactome curated database (*http://www.reactome.org/*). Trans gene sets were analysed and some found to be enriched in a particular biological pathways. The pathway analysis identified that 46 B-cell trans genes associated with rs10816519 belong to 9 pathways (p-value cutoff 0.05, Table 4.6)including Immune system (Reactome pathway ID 168256, $p=7.23x10^{-3}$) and Adaptive Immune system (ID 1280218, p=0.014). For example, *NLRP3* belongs in the immune system pathway and encodes a member of the inflammasome complex to activating NF-κB signalling pathway ($p=6.04x10^{-12}$ to rs10816519, Bauernfeind *et al.*, 2009). *FPR1* belongs in the adaptive immune system and encodes a protein in the G-protein coupled receptor family that mediates the response of phagocytic cells during invasion of microorganisms ($p=2.47x10^{-13}$ to rs10816519, Miettinen, 2011). Genes associated

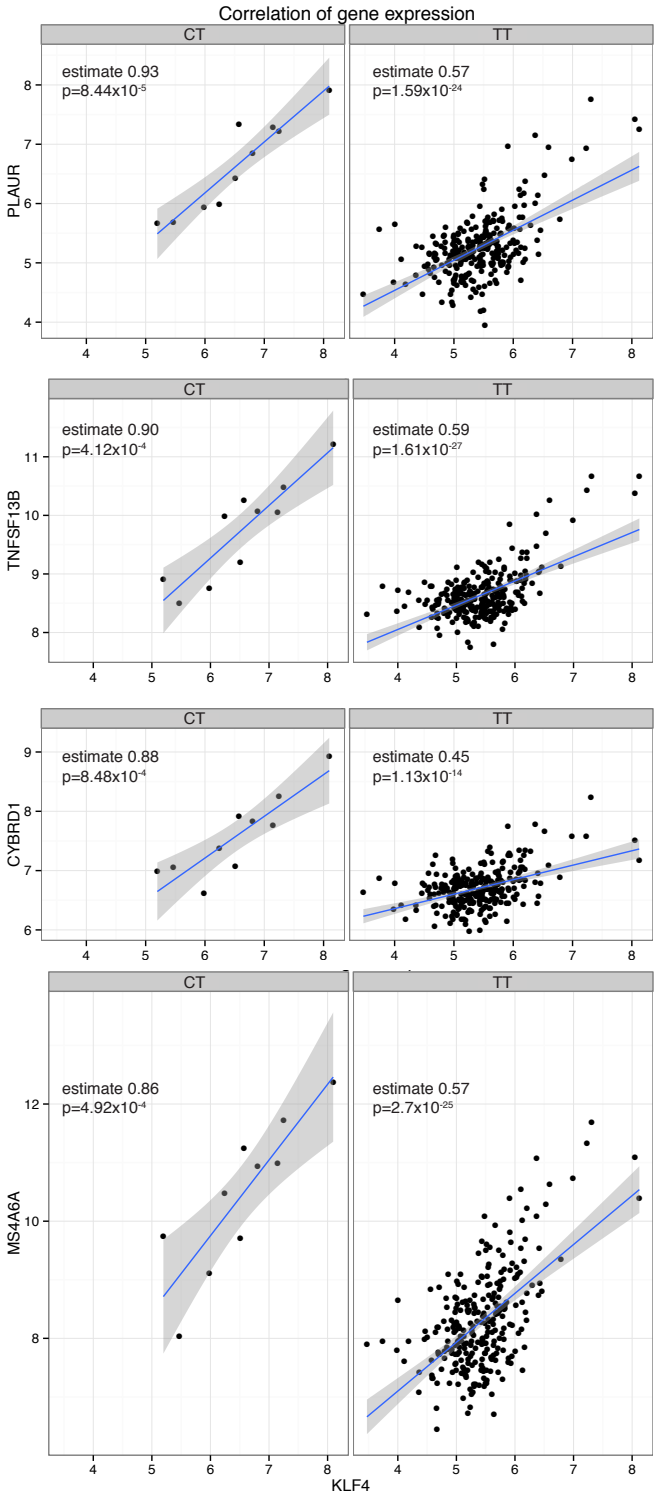


Figure 4.18: Gene expression correlation of *KLF4* and trans associated genes in B-cells in two genotypes of rs10816519. p-values and estimates of linear correlation are shown in the plot.

in trans with rs10784774 in monocytes were not significantly enriched with biological pathways in the Reactome database.

Table 4.6: Reactome pathway enrichment analysis on rs10816519 trans associated genes in B-cells. GeneRatio: over-represented genes in genes tested in each cell type. BgRatio: a ratio of genes in a pathway and background.

Pathway ID	Description	GeneRatio	BgRatio	pvalue	qvalue
211945	Phase 1 - Functionalization of compounds	3/46	54/19347	$2.87\text{x}10^{-4}$	$2.95\text{x}10^{-3}$
500792	GPCR ligand binding	5/46	365/19347	$1.69\text{x}10^{-3}$	$1.24\text{x}10^{-2}$
211859	Biological oxidations	3/46	109/19347	$2.21\text{x}10^{-3}$	$1.31\text{x}10^{-2}$
198933	Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell	2/46	48/19347	$5.82\text{x}10^{-3}$	$1.64\text{x}10^{-2}$
168256	Immune System	6/46	728/19347	$7.23\text{x}10^{-3}$	$1.64\text{x}10^{-2}$
1280218	Adaptive Immune System	4/46	391/19347	$1.37\text{x}10^{-2}$	$2.20\text{x}10^{-2}$
76005	Response to elevated platelet cytosolic Ca2+	2/46	80/19347	$1.55\text{x}10^{-2}$	$2.29\text{x}10^{-2}$
109582	Hemostasis	4/46	443/19347	$2.08\text{x}10^{-2}$	$2.57\text{x}10^{-2}$
388396	GPCR downstream signaling	5/46	677/19347	$2.17\text{x}10^{-2}$	$2.57\text{x}10^{-2}$
372790	Signaling by GPCR	5/46	763/19347	$3.40\text{x}10^{-2}$	$3.24\text{x}10^{-2}$

4.5 Cell type specific eQTL associated with GWAS loci

GWAS identified disease- or trait- associated loci are predicted to affect gene expression because the majority of associations are found in noncoding DNA where variants may disrupt or create gene regulatory elements such as enhancers (Moffatt *et al.*, 2007, Stranger *et al.*, 2007a). In particular comparing cis eQTL to GWAS loci associated with disease susceptibility may reveal gene expression associated with GWAS loci. In order to investigate the association of GWAS loci to eSNPs identified in this dataset, the GWAS catalog created by the National Human Genome Research Institute (NHGRI) was searched for overlaps with the eSNPs. The database has archived 8265 SNPs associated with 772 phenotypes with p-values less than the p-value threshold of $1x10^{-5}$ (as of 22nd January 2013, Hindorff *et al.*, 2009, *http://www.genome.gov/gwastudies/*). In order to overcome discrepancy of genotyping platforms that use different tagging SNPs, proxy SNPs (r^2 >0.8) were determined for the GWAS markers using the SNAP proxy SNP search with the CEU 1000 Genomes Project pilot 1 dataset (Johnson *et al.*, 2008, *http://www.broadinstitute.org/mpg/snap/*). This resulted in a list of 150550 SNPs with a GWAS association. In total 943 genes have significant cis eQTL related to GWAS (23.2% shared, 43.4% monocyte and 33.4% B-cell specific, Table 4.7). The majority of genes in trans eQTL related to GWAS were monocyte specific eQTL (55 monocyte specific eQTL genes out of 59 total).

Table 4.7: Cell type specific eSNPs and genes identified in GWAS. The database used has 150550 SNPs including proxy SNPs, 772 traits and 1486 publications with p-value<$1x10^{-5}$ (as of 22nd January, 2013)

	cis		trans	
	eSNP (%)	probe (gene)	eSNP (%)	probe (gene)
Total	6011 (100)	1013 (943)	238 (100)	59 (59)
Shared	1647 (27.4)	226 (221)	37 (15.5)	2 (2)
Monocytes	2741 (45.6)	451 (414)	133 (55.9)	55 (55)
B-cells	1623 (27.0)	336 (318)	68 (28.6)	2 (2)

4.5.1 cis eQTL associated with GWAS loci

The search of the cis eQTL dataset in relation to the GWAS disease associated loci identified many eQTL involving genes whose products are involved in immune responses. This included autoimmune disorders and inflammation, which are described below.

Ankylosing spondylitis, ulcerative colitis and Crohn's disease

Ankylosing spondylitis (AS) and ulcerative colitis (UC) are both chronic inflammatory diseases that occur in the axial skeleton or bowel respectively. The susceptible GWAS SNP for the two diseases, rs10781507 and its proxy SNP rs10781500 on chr9 (r^2 0.967, $p_{GWAS}=1x10^{-6}$ with AS, $p_{GWAS}=3x10^{-19}$ with UC, only highest p-values shown), are associated with expression of *INPP5E* that encodes inositol polyphosphate-5-phosphatase in both cell types (Figure 4.19, $p_{mono}=1.12x10^{-34}$, $p_{B-cell}=1.31x^{-33}$, Evans *et al.*, 2011, McGovern *et al.*, 2010, Anderson *et al.*, 2011). These two disease susceptibility SNPs are also associated with monocyte specific expression of *CARD9* ($p_{mono}=1.04x10^{-24}$, Barrett *et al.*, 2009,McGovern *et al.*, 2010, Anderson *et al.*, 2011). *CARD9* encodes a protein that contains a caspase-associated recruitment domain (CARD), and controls innate anti-fungal immunity through non-TLR signalling such as *NOD2*-mediated activation (Hsu *et al.*, 2007, Gross *et al.*, 2006). This monocyte specific cis eQTL was identified as a susceptibility locus of Crohn's disease, a major inflammatory bowel disease (IBD) (Franke *et al.*, 2010, Julia *et al.*, 2013). Other loci that are associated with susceptibility to Crohn's disease were identified among shared eQTL including rs4360063 ($p_{GWAS}=1x10^{-10}$) which associates with *ERAP2* expression (Figure 4.19, $p_{mono}=1.05x10^{-110}$ and $p_{B-cell}=2.51x10^{-115}$, Franke *et al.*, 2010). *ERAP2* encodes endoplasmic reticulum amino peptidase 2 and is involved with trimming of precursor peptides in endoplasmic reticulum for HLA class I antigen presentation (Saveanu *et al.*, 2005). The reduction of *ERAP2* expression was shown to associate with a haplotype that has a non-synonymous variant for an alternative splicing leading to the truncated protein, leading to nonsense-mediated mRNA decay and reduction of HLA class I molecule expression (Andres *et al.*, 2010).

Coronary heart disease

Inflammation in cardiovascular disease is a common pathological mechanism that contributes to the development of coronary heart disease (CHD) and atherosclerosis. rs1412444 on chr10, a locus associated with CHD susceptibility ($p_{GWAS}=3x10^{-13}$), is associated in monocytes with expression of *LIPA* encoding lysosomal acid lipase A (Figure 4.19, $p_{mono}=8.41x10^{-40}$, Coronary Artery Disease (C4D) Genetics Consortium, 2011, Wild *et al.*, 2011). Deficiency of lysosomal acid lipase in the metabolism plasma is associated with cholesterol ester storage disease, hypercholesterolemia and increased atherosclerosis (Goldstein *et al.*, 1975). Unlike AS and UC, the cis eQTL identified in the risk loci of CHD was not identified in other disease GWAS.

Rheumatoid arthritis, systemic lupus erythematosus and multiple sclerosis

Rheumatoid arthritis (RA) and systemic lupus erythematosus (SLE) are both systemic autoimmune diseases where inflammation occurs in multiple tissues and organs. A risk locus for RA and SLE is tagged by rs13277113 which is a cis eQTL in B-cells for expression of *BLK* and *C8ORF13* (also known as *FAM167A*) on chr8 (p_{GWAS}=6x10^{-9} with RA, p_{GWAS}=1x10^{-10} with SLE, Figure 4.19, Gregersen *et al.*, 2009, Hom *et al.*, 2008). rs13277113 significantly associated with both genes, which are located adjacent to each other on chr8 (p_{B-cell}=7.43x10^{-78} for *C8ORF13* and p_{B-cell}=3.57x10^{-14} for *BLK*).

The monocyte specific cis eQTL for *CD40* on chr20 was identified in the loci associated with susceptibility to RA and also mlultiple sclerosis (MS) (rs4810485, p_{mono}=5.95x10^{-15}, p_{GWAS}=8x10^{-9} to RA, p_{GWAS}=1x10^{-7} to MS, Raychaudhuri *et al.*, 2008, Bahlo *et al.*, 2009). *CD40* encodes a member of TNF receptor superfamily and is expressed in APCs for signal transduction of immune responses.

Type 1 diabetes

Type 1 diabetes (T1D) is associated with autoimmune destruction of β cells. The T1D risk locus rs10743823 is associated with *CLECL1* expression in both cell types (p_{mono}=1.79x10^{-80}, p_{B-cell}=2.01x10^{-11}, p_{GWAS}=5x10^{-9}, Wellcome Trust Case Control Consortium, 2007). *CLECL1* encodes a transmembrane C-type lectin-like 1 protein and may enhance IL4 production as a T-cell co-stimulatory molecule (Ryan *et al.*, 2002). MS risk variant rs10466829 is in high LD with rs10743823 (r^2 0.967) which is associated with *CLECL1* expression in monocytes (p_{mono}=1.79x10^{-80}, p_{GWAS}=1x10^{-8}, Sawcer *et al.*, 2011).

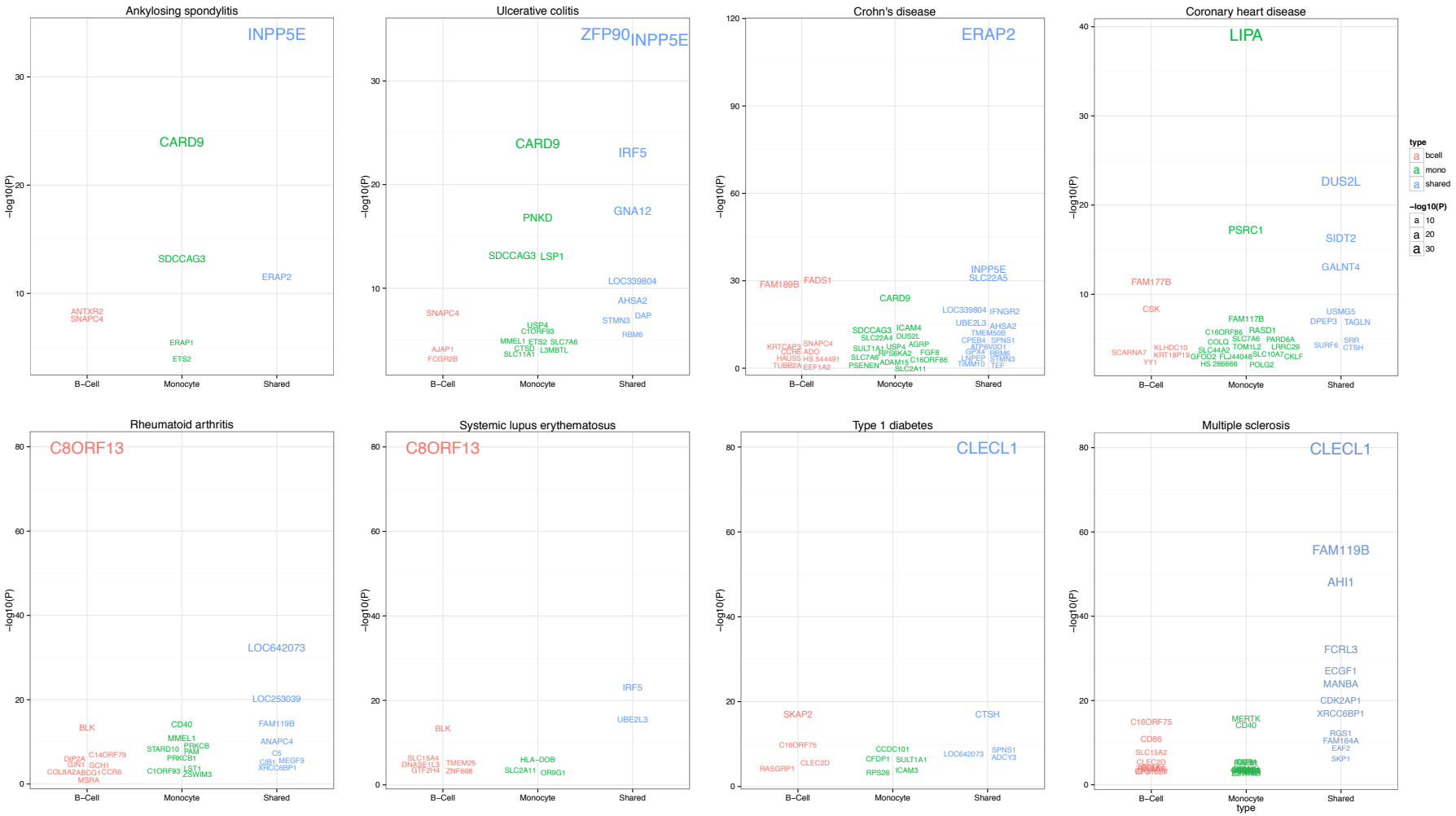


Figure 4.19: Examples of cis eQTL genes identified in disease susceptiblity GWAS loci. The X and Y axes show eQTL cell types (red; B-cells, green; monocytes, blue; shared) and -$\log_{10}$p-value of cis eQTL associated genes respectively.

4.5.2 trans eQTL associated with GWAS loci

The majority of these trans eQTL are found to involve 12q13.2 (shared and B-cell specific) and 6q21.32 (monocyte specific trans eQTL). It is striking that both these loci show evidence of disease association.

4.5.3 trans eQTL in 12q13.2 identified in GWAS

Shared and B-cell specific trans eQTL eSNPs in 12q13.2 overlap with previously identified susceptibility loci for several autoimmune diseases including T1D, asthma and vitiligo. Although the loci were associated in trans to different genes, all eSNPs were located within a 150kb region of chr12 between rs2069408 and rs1117147.

In order to determine the detail of this region, a haplotype block analysis was performed using Haploview (Barrett *et al.*, 2005). The Haploview program computes pairwise LD statistics of genetic markers and visualises population haplotype patterns using genotype data. Gabriel and colleagues defined a haplotype block as no evidence of historical recombination ('strong LD') with the following definition; fewer than 5% of all internal allele pairs show strong evidence of recombination (Gabriel *et al.*, 2002).

Based on the population stratification analysis for this volunteer population which showed high concordance with the HapMap CEU population, the LD data of this reference population was used to determine haplotype structure in the 12q13.2 region. The region from rs2069408 to rs1117147 was selected for the analysis and a haplotype block was found in the 78kb region from rs772921 to rs2292239 (Figure 4.20). Therefore, SNPs in this haplotype block are in strong LD, assuming that the entire region may be involved in the association of eSNPs and gene expression.

There are two genes associated with 10 disease susceptibility loci in trans: B-cell specific expression of *IP6K2* on chr3 and *LOC389386* on chr6. The same eSNPs in 12q13.2 are found to associate with both genes in trans. rs11171739 and rs10876864, 69.5kb away from each other, are in strong LD (r^2 =0.863) and are associated with both *IP6K2* and *LOC389386* expression (Figure 4.21). *IP6K2* encodes inositol hexakisphosphate kinase 2 which is involved in apoptosis as a positive regulator (Morrison *et al.*, 2002). *LOC389386* (also known as *LAP3P2*, encoding leucine aminopeptidase 3 pseudogene 2), is a pseudogene for which there is some evidence of transcription. The RNA fragment JF803844 (the Human mRNA GenBank) maps to 1.5kb upstream of the gene *CDKN1A* encoding cyclin-dependent kinase inhibitor 1A on chr3. *CDK2*, encoding cyclin-dependent kinase 2, is located in 12q13.2 where the trans eQTL are found. CDKN1A (also known as

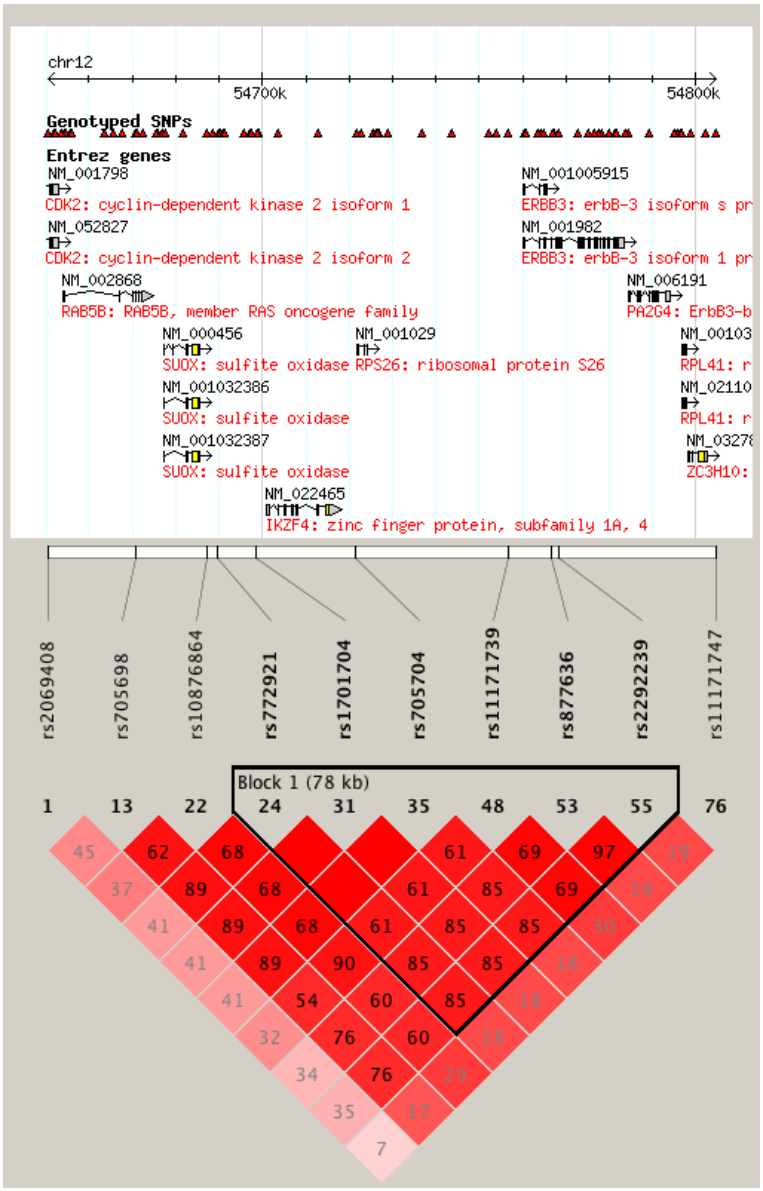


Figure 4.20: Haplotype analysis of disease susceptibility loci and B-cell specific trans eQTL in 12q13.2. The diagram shows the genomic map (top) and the solid spline of LD (bottom). (Top) The top line shows the genomic location in 12q13.2. The triangles shows genotyped SNPs in CEU population which is used as a reference. Entrez genes show genes encoded in the region with exons (squares) and introns (lines) in direction (arrows). (Bottom) The solid spline of LD displays reciprocal LD relationships of SNPs using D′ between rs2069408 to rs11171747 where D′ is a normalised measure of allele association. A haplotype block enclosed in black lines is defined when 95% of informative comparisons using D′ across SNPs are strong LD. The diamonds in red with values of D′ show LD relationships of each SNP pair. No value is shown when D′ is 1. The numbers below SNP rs numbers show distance in kb from the far left SNP.

p21) binds to and inhibits the activity of CDK2 during regulation of cell cycle from G1 to S phase transition (Gartel and Radhakrishnan, 2005). There was no cis association identified with *CDK2* expression in 12q13.2 and it remains unclear whether *CDKN1A* expression in an allele specific manner may be involved with *CDK2* expression and its regulation in cell cycle.

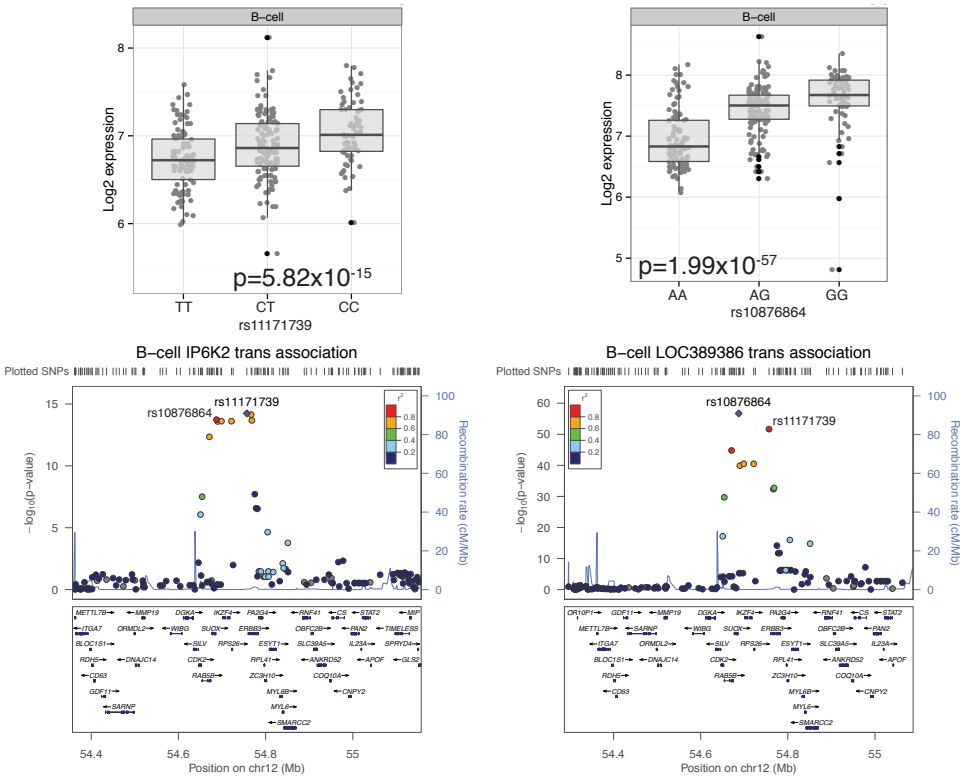


Figure 4.21: B-cell trans association in 12q13.2 with *IP6K2* (left) and *LOC389386* (right). The box plots of the two genes are shown with peak SNPs rs11171739 and rs10876864 respectively. P-values for rs11171739; *IP6K2* - p_{B-cell}=5.82x10^{-15}, *LOC389386* - p_{B-cell}=2.04x10^{-52}. P-values for rs10876864; *IP6K2* - p_{B-cell}=1.83x10^{-14}, *LOC389386* - p_{B-cell}=1.99x10^{-57}.

T1D susceptibility markers (peak SNP rs11171739, p_{GWAS}=1x10^{-11} with T1D) form trans eQTL in a 10kb region to rs2292239 in B-cells with *IP6K2* (p_{B-cell}=5.82x10^{-15}) and *LOC389386* (p_{B-cell}=2.04x10^{-52}, Figure 4.21, Wellcome Trust Case Control Consortium, 2007). rs2292239 showed a strong association with T1D susceptibility in several studies (p_{GWAS}=2x10^{-20}; Todd *et al.*, 2007, p_{GWAS}=3x10^{-16}; Cooper *et al.*, 2008, p_{GWAS}=3x10^{-27}; Barrett *et al.*, 2009, p_{GWAS}=3x10^{-27}; Plagnol *et al.*, 2011, eQTL association in this study; p_{B-cell}=1.78x10^{-33} to *LOC389386* and p_{B-cell}=2.12x10^{-14} to *IP6K2*). This SNP is in relatively strong LD with rs11171739 and is included in the same haplotype block (r^2 0.692). It is located in an intron of *ERBB3* which encodes a member of the epidermal growth factor receptor family of receptor tyrosine kinases. It is reported that rs2292239 is associated with *RPS26* expression, encoding ribosomal protein S26, in human liver tissue

($p=4.03x10^{-22}$, Schadt *et al.*, 2008). The other T1D susceptibility marker SNP rs1701704, 58kb away from rs11171739, is associated with both *LOC389386* and *IP6K2*, although the two SNPs are not in strong LD (r^2 0.562, Figure 4.20, $p_{B-cell}=3.17x10^{-41}$ and $2.44x10^{-14}$ respectively, $p_{GWAS}=5x10^{-18}$, Hakonarson *et al.*, 2008, Plagnol *et al.*, 2011).

rs1701704 is also associated with susceptibility to asthma which is a common chronic inflammatory disease of the airways and is thought to be affected by both genetic and environmental factors ($p_{GWAS}=2x10^{-13}$, Hirota *et al.*, 2011). The majority of asthma susceptibility loci were identified in multiple genomic locations by linkage mapping studies and GWAS including *ORMDL3* on chr17, IL4 on chr5, and TNF on chr6 (Ober and Hoffjan, 2006, Weiss *et al.*, 2009). This GWAS locus, rs1701704, is also shared with vitiligo. The cause of vitiligo is thought to be an autoimmune response to melanocytes resulting in cell death or dysfunction and leads to depigmentation of patches of skin. The disease has a complex aetiology but has some genetic component based on epidemiology studies (Alkhateeb *et al.*, 2003, Jin *et al.*, 2012). Vitiligo susceptibility marker rs10876864 forms B-cell specific trans eQTL with *IP6K2* and *LOC389386* ($p_{GWAS}=8x10^{-12}$, Jin *et al.*, 2012, Tang *et al.*, 2013). rs10876864 is again found as a T1D susceptibility GWAS locus as described above.

4.5.4 trans eQTL in the HLA region and relevance to disease

Based on the haplotypic analysis using Haploview, reported disease susceptibility SNP markers in the MHC class II region are in strong LD with monocyte specific trans eQTL to *AOAH*, *DEF8*, *ARHGAP24* and *RPL34* (Figure 4.22). The locus encodes proteins involved in antigen presentation in APCs which are central to adaptive immune responses. MHC class II molecules are expressed on the surface of professional immune cells including monocytes, macrophages, dendritic cells and B-cells and are encoded on chr6. They consist of α- and β- polypeptide chains which form a heterodimeric complex including peptide binding and transmembrane domains. Each peptide chain consists of two sub chains, where $\alpha 1$ and $\beta 1$ regions form membrane distal domains for peptide binding and presentation to T cell Receptors (TCR) on $CD4^+$ T cells, and $\alpha 2$ and $\beta 2$ regions form transmembrane domains. In contrast to the MHC class I molecules which present cytosolic antigens to $CD8^+$ T cells, the class II molecules present extracellularly derived antigens to $CD4^+$ T cells for downstream signal transduction. Exogenous proteins are processed in APCs through phagocytosis, endocytosis, lysis and are loaded onto the MHC class II molecules in endoplasmic reticulum and presented to the cellular membrane.

Possession of specific MHC class II alleles has been shown to predispose to or protect against autoimmune disease (*HLA-DRB1*: RA (Stastny, 1978, Wordsworth *et al.*, 1989), *HLA-DRB1*: UC (Satsangi *et al.*, 1996), *HLA-DRB1*1501*: MS (Olerup and Hillert, 1991) and *HLA-DQB1* and *-DRB1*: T1D (Todd *et al.*, 1987, Baisch *et al.*, 1990)). These alleles are extremely polymorphic compared to non-MHC genes, however structural studies on the MHC class II molecules provide insight into the molecular mechanism of these disease susceptibilities (Jones *et al.*, 2006). For example, *HLA-DRB1*0401*, **0404* and **0101* alleles are associated with RA and encode proteins HLA-DR4.1, -DR4.4, -DR1. These molecules share peptide-binding grooves (epitopes) in a β1 DRB1 chain and are selective to specific peptides (Gregersen *et al.*, 1987, Hammer *et al.*, 1995). T1D susceptibility alleles *HLA-DQB1* and *HLA-DRB1* are suggested to be involved in autoreactive response to insulin (Todd *et al.*, 1987, Baisch *et al.*, 1990). HLA-DR4 encoded in *HLA-DRB1*0401* and *DRB1*0405* alleles is involved with the destruction of β cell by autoreactive T cells which recognise an insulin epitope in pancreas (Kent *et al.*, 2005). Thus the observed monocyte trans eQTL in 6p21.32 is intriguing to investigate the involvement of trans regulated genes that may help to elucidate mechanisms of disease susceptibility and development.

Monocyte specific trans eQTL within a 75kb region of 6p21.32 were associated with four genes: *AOAH* on chr7, *DEF8* on chr16, *ARHGAP24* and *RPL34* on chr4 (Figure 4.23). The locus is near *HLA-DQA1* and associated with multiple autoimmune and inflammatory diseases including T1D, asthma, IBD and RA.

T1D susceptibility associated locus rs9272346 is in strong LD with rs1063355 with which a trans eQTL association was found with *AOAH* (r^2 0.870, Figure 4.22, $p_{GWAS}=5x10^{-134}$ and $6x10^{-129}$ to rs9272346, Wellcome Trust Case Control Consortium, 2007, Cooper *et al.*, 2008). rs1063355 is associated with monocyte specific expression of *AOAH* at $p_{mono}=1.58x10^{-15}$ (Figures 4.23). *AOAH* encoding acyloxyacyl hydrolase was identified to be highly associated with the several disease susceptibility loci. The gene product of *AOAH* modulates host inflammatory responses to Gram-negative bacteria because of its ability to breakdown bacterial LPS (Hagen *et al.*, 1991, Lu *et al.*, 2005). The most significant trans eQTL for *AOAH* is rs28366298 which is 43.5kb away from the T1D locus rs9272346 but not in LD (r^2 0.434, $p_{mono}=1.56x10^{-43}$, Figures 4.23 and 4.22). This trans eQTL was previously described in a study of whole blood gene expression (rs2395185, $p_{mono}=5.71x10^{-39}$ in this dataset, $p=7.0x10^{-38}$ in Fehrmann *et al.*, 2011). These two SNPs are about 127kb apart but the LD structure between them is not available from the 1000

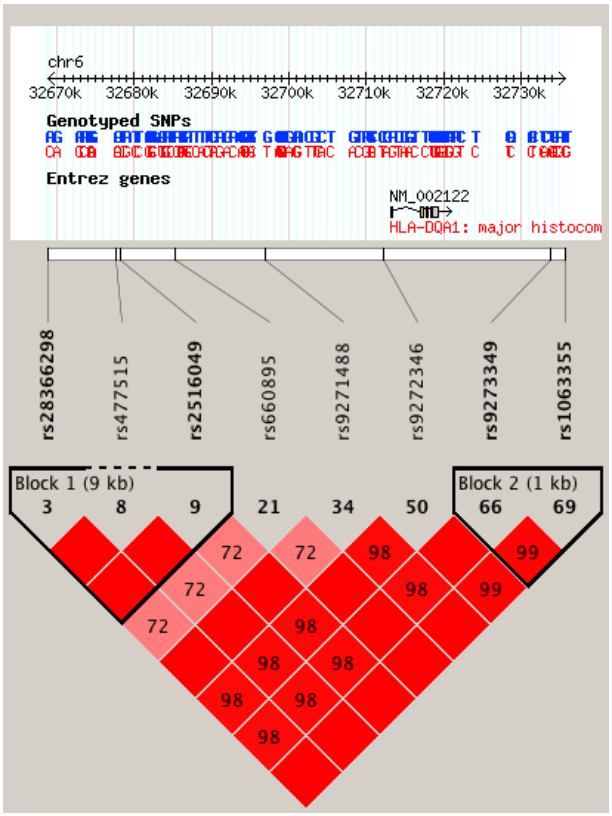


Figure 4.22: Haplotype analysis of disease susceptible loci and monocyte specific trans eQTL in 6p21.32. The diagram shows the genomic map (top) and the solid spline of LD (bottom). (Top) The top line shows the genomic location in 6p21.32. The triangles shows genotyped SNPs in CEU population which is used as a reference. Entrez genes show genes encoded in the region with exons (squares) and introns (lines) in direction (arrows). (Bottom) The solid spline of LD displays reciprocal LD relationships of SNPs using D′ between rs28366298 to rs1063355 where D′ is a normalised measure of allele association. A haplotype block enclosed in black lines is defined when 95% of informative comparisons using D′ across SNPs are strong LD. The diamonds in red with values of D′ show LD relationships of each SNP pair. No value is shown when D′ is 1. The numbers below SNP rs numbers show distance in kb from the far left SNP.

Genomes Pilot 1 dataset. *AOAH* expression is shown to associate in the innate immune response, however, involvement with autoimmunity is unknown.

rs477515 which is associated with susceptibility to early onset IBD is associated with monocyte expression of *AOAH* (p_{mono}= 5.71x10^{-38}), *RPL34*(p_{mono}=4.56x10^{-16}) and *DEF8* (p_{mono}=4.66x10^{-12}) in trans (Figure 4.23, p_{GWAS}=1x10^{-8} to IBD, Kugathasan *et al.*, 2008). *RPL34* encodes a ribosomal protein L34, a component of S60. *DEF8* encodes a protein differentially expressed in FDCP 8 homolog (FDCP - factor-dependent cell progenitors), however the function of the gene product is not known. The most significant trans eSNP rs35265698 for *DEF8* is in strong LD with a RA susceptibility locus rs660895 (r^2 0.879, p_{mono}=8.96x10^{-17} with rs35265698, p_{mono}=8.17x10^{-13} with rs660895, p_{GWAS}=1x10^{-108} with RA locus rs660895, Plenge *et al.*, 2007). rs477515 is in strong LD with rs9271488 which is associated with monocyte expression of *ARHGAP24*, encoding a Rho GTPase-

activating protein involved in angiogenesis and cell migration (r^2 0.901, p_{mono}=4.59x10^{-14} to rs9271488, Figure 4.23, Su *et al.*, 2004). The *ARHGAP24* peak trans eSNP rs28366298 is shared with association to *AOAH* expression (p_{mono}=4.59x10^{-14} to *ARHGAP24*). These SNPs are located in a non-coding region between *HLA-DRB1* and *-DQA1* in 6p21.32. The genes associated with this locus in trans are not directly involved in the mechanism and development of autoimmunity and inflammation.

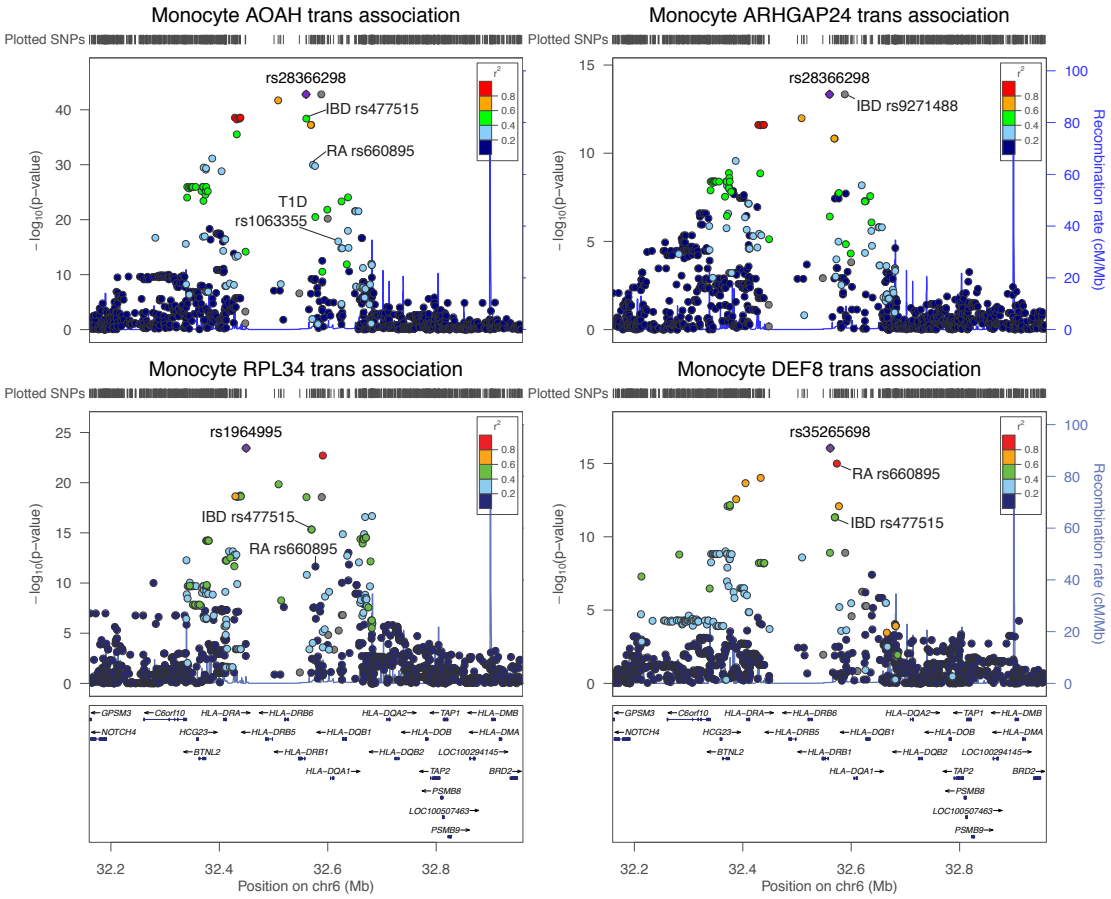


Figure 4.23: Monocyte trans eQTL in 6p21.32 with associated genes. *AOAH*; rs28366298 p_{mono}=1.56x10^{-43}, *ARHGAP24*; rs28366298 p_{mono}=4.59x10^{-14}, *RPL34*; rs1964995 p_{mono}=3.6x10^{-24}, *DEF8*; rs35265698 p_{mono}=8.96x10^{-17}. GWAS associated SNPs and the disease names are shown in the plot.

Trans eQTL for monocytes and B-cells have been identified in disease associated genetic loci. Several autoimmune diseases have susceptibility loci within the same small regions; monocyte specific trans eQTL in the MHC class II locus and B-cell specific on 12q13.2, which imply shared susceptibility and mechanisms to autoimmune diseases. Products of the eQTL associated genes do not have clear connections to disease susceptibility or development. However, these genes might help leading to other genes which may be critical for mechanisms of disease development.

4.6 HLA

Genetic variation in HLA alleles is the most variable in mammals and has been identified as being crucial in susceptibility to autoimmune and infectious diseases as well as transplant rejection. For example, more than 90% of celiac disease patients carry the *HLA-DQB1*02* isoforms (Sollid *et al.*, 1989, Bugawan *et al.*, 1989). Heterodimers of the alleles present toxic gluten peptides for activation of $CD4^+$ T cells followed by inflammatory responses in the intestine (Lundin *et al.*, 1993). HLA class I *B35* and *Cw04* confer rapid progression of AIDS (Acquired immunodeficiency syndrome) with patients infected with HIV-1 (Human immunodeficiency virus type-1) (Carrington *et al.*, 1999). It is important to determine genes underlying the mechanisms of disease susceptibility and development in association to HLA alleles. However, the genetic complexity of HLA alleles caused by LD in the extended MHC loci including both HLA and non-HLA genes makes identification of causal alleles difficult (Walsh *et al.*, 2003). In the previous section monocyte specific trans eQTL in the MHC class II locus have been identified. Here the association between HLA alleles and gene expression of the immune cell types was investigated in order to ascertain whether any specific HLA haplotype was associated with gene expression.

4.6.1 HLA type imputation

Defining HLA types has been established by several methods. The first and oldest method is serological testing based on complement dependent cytotoxicity. Historically this reliable and rapid method has been used for typing of the class I molecules, however it can only detect low level 2 digit resolution and requires a large volume of blood. As the number of HLA alleles typed has increased, the nomenclature of HLA alleles is now based on up to 6 digits (Marsh *et al.*, 2010). Methods based on molecular biological techniques have emerged based on DNA sequencing and PCR (sequence specific primer (SSP)-PCR and sequence specific oligonucleotide probes (SSOP)-PCR) for higher resolution. The more HLA typing in a particular population has been performed, the more accurate the HLA alleles are typed. The combination of classical HLA genes, the increased number of genotyped SNPs and the high resolution map of LD structure in the MHC region has facilitated haplotype analysis of the HLA alleles (Walsh *et al.*, 2003, Stewart *et al.*, 2004, Miretti *et al.*, 2005). Furthermore, dense SNP genotyping data from genome-wide arrays has become sufficient to impute HLA alleles (Leslie *et al.*, 2008, Dilthey *et al.*, 2013).

A computational method was used to impute HLA alleles for the volunteer population based on SNP genotyping data (HLA*IMP, *https://oxfordhla.well.ox.ac.uk/hla/*, Leslie *et al.*, 2008, Dilthey *et al.*, 2011). Imputation of the HLA alleles used 7587 SNPs on chr6:24921129-34535328 (no MAF cutoff applied, NCBI36/hg18 position) in 283 individuals to impute HLA alleles (class I alleles; *HLA-A*, *-B* and *-C*, class II alleles; *HLA-DQA*, *-DQB* and *-DRB*) to 2- and 4-digit resolution using the CEU population data (performed by Stephen Leslie and Alexander Dilthey, Wellcome Trust Centre for Human Genetics). The SNP data was phased into haplotypes for each individual using the HapMap CEU trios data as a reference based on the approximation algorithm modelled by Li and Stephens (Li and Stephens, 2003). The quality of the imputation was measured by positive predicted values (PPV) as accuracy which was calculated on a Bayesian approach and the cutoff was set to 0.7 (Appendix Table 4). The averages of the quality scores were 97.4%, 0.98 and 0.98 for chromosome call rate in the volunteer population, PPV and the call rate of the entire imputation including reference dataset. Validation of the imputation for the 2 digit alleles is shown in Figure 4.24. This was implemented in the HLA*IMP program, which used the training reference data to validate the HLA allele calling. The 206 imputed HLA alleles (74 and 132 for 2- and 4-digit alleles respectively) with PPV>0.7 were used for association analysis with the gene expression data for monocytes and B-cells.

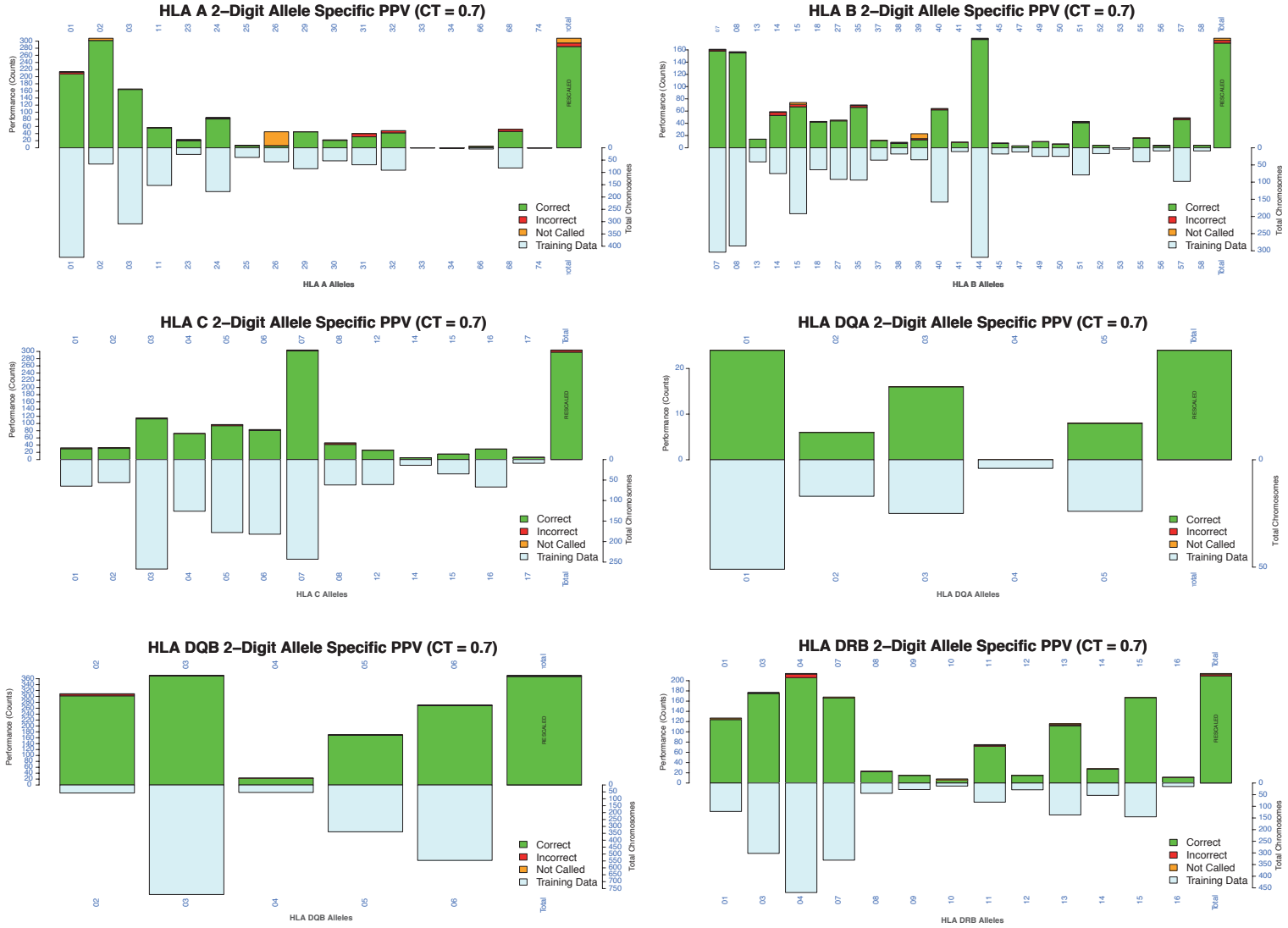


Figure 4.24: Validation of HLA 2 digit alleles imputation after filtering at cutoff threshold (CT) >0.7. PPV: positive predicted value. The number of chromosomes is shown as correct (green), incorrect (red) and not called (orange). The number of the reference data with HLA alleles is shown (light blue). Provided by Stephen Leslie and Alexander Dilthey.

4.6.2 Association of HLA alleles and gene expression

The association of the imputed HLA alleles for six classical MHC loci with gene expression from monocytes and B-cells in the volunteer population (29022 probes each) was determined in a linear model using an R package MatrixEQTL (*http://cran.r-project.org/web/packages/MatrixEQTL/index.html*, Shabalin, 2012). This program allows ultra fast eQTL association analysis using a large matrix operation. This analysis focused on trans association with genes outside of the MHC region (7.6Mbp long in chromosome 6p) because most of the MHC genes were already removed from the expression data due to a presence of SNPs in their sequence. Initially the 2-digit HLA alleles were determined for association in order to avoid false positive detection because some alleles were imputed in one or a few samples. The analysis identified a total of 208 associations in trans with FDR cutoff 0.05, which include 112, 80 and 16 associations respectively in monocytes only, B-cells only and shared between the two cell types (Table 4.8).

Table 4.8: HLA allele (2-digit) associations with gene expression. FDR cutoff 0.05 applied.

			Number of HLA alleles involved (Unique alleles)
	Association	Probe (Gene)	2-digit
monocytes	112	88 (81)	105 (44)
B-cells	80	57 (55)	63 (22)
shared	16	5 (5)	18 (15)
total	208	150 (141)	186 (81)

Cell type specificity

The cell type specificity in the association of HLA alleles with monocyte and B-cell genes was further determined (Table 4.9). The prevalence between monocytes and B-cells in the number of associations with either class I or II was subtle. However, there were more associations with class II alleles in monocytes than B-cells (57.1% in monocytes and 14.3% in B-cells).

4.6.3 HLA association and trans eQTL

Regulation of the HLA loci is essential for the adaptive immune system and involved in causality to autoimmune diseases. Some of the genes associated with HLA alleles were also associated with SNPs in the MHC region in trans. Overall, the most significant association was found in the HLA class II alleles with monocyte gene expression.

Table 4.9: Cell type specificity in association to 2-digit HLA class I and II alleles. The proportion of association with class II alleles with monocyte genes is larger than those with B-cell genes.

		Number of HLA association (%)			
HLA class		Total	Monocytes	B-cells	shared
class I	A	34	18	16	0
	B	94	55	35	4
	C	16	8	6	2
	Total class I (%)	144 (100)	81 (56.25)	57 (39.58)	6 (4.17)
class II	DQA	15	10	1	4
	DQB	6	4	0	2
	DRB	21	10	5	6
	Total class II (%)	42 (100)	24 (57.14)	6 (14.29)	12 (28.57)
	Total	186 (100)	105 (56.45)	63 (33.87)	18 (9.68)

Monocyte specific association

The most significant association with HLA alleles was *AOAH* expression in monocytes which was significantly associated with all three class II alleles (p_{mono}=4.05x10^{-39} with *HLA-DQA1*3*, p_{mono}=7.24x10^{-38} with *HLA-DRB1*4*, p_{mono}=4.74x10^{-23} with *HLA-DQB1*3*, Table 4.10). As described in the previous section for trans eQTL analysis in monocytes, *AOAH* is associated with SNPs in the MHC class II locus. Combining the genotyping information of trans peak eSNP rs28366298 and the *HLA-DQA* alleles to *AOAH* expression, the samples with major homozygous (CC) and *HLA-DQA1*3* have shown down-regulation of *AOAH* in monocytes but not in B-cells (Wilcoxon rank sum test, p=7.89x10^{-13}, Figure 4.25A). Further analysis using the 4-digit HLA allele data, this association was observed in *HLA-DQA1*301* (p_{mono}=7.81x10^{-38}). The same manner of analysis was applied to *HLA-DRB1*4* in 4-digit resolution which was significantly associated with *AOAH* expression in monocytes; the samples with *HLA-DRB1*401* and major homozygous (CC) for rs28366298 had significantly down-regulated *AOAH* expression in monocytes (p=7.02x10^{-11}, Figure 4.25A). In addition, *AOAH* expression in monocytes with *HLA-DRB1*4*, *HLA-DRB1*7*, *HLA-DRB1*9* alleles are significantly down-regulated (p=3.08x10^{-22}, Figure 4.25A). This suggests some HLA alleles share associations with gene expression.

The 4-digit resolution of *HLA-DQB1*3* showed that *AOAH* expression is significantly associated with *HLA-DQB1*302* together with the rs28366298 (CC) genotype (p=3.84x10^{-10}, appendix Figure 2). This analysis showed that *AOAH* expression was significantly associated with not only this eSNP but also HLA alleles in a cell type specific manner.

ARHGAP24 expression in monocytes is associated with trans eQTL in the MHC class II locus. *ARHGAP24* encodes Rho GTPase-activating protein and is involved in angiogenesis and cell migration (Su *et al.*, 2004). *HLA-DQA1*3* was significantly associated with *ARHGAP24* expression in monocytes (p_{mono}=3.03x10^{-7}, Table 4.10). Trans eSNP rs28366298 (CC) genotype with *HLA-DQA1*3* significantly up-regulated gene expression in both cell types (p_{mono}=7.11x10^{-11}, p_{B-cell}=7.62x10^{-8}, appendix Figure 1).

HLA association shared in the two cell types

DEF8 showed significant associations with *HLA-DQA1*3* and *HLA-DRB1*4* alleles in both cell types (p_{mono}=4.3x10^{-17}, p_{B-cell}=6.95x10^{-13} with *HLA-DQA1*3*, p_{mono}=2.63x10^{-17}, p_{B-cell}=6.23x10^{-13} with *HLA-DRB1*4*, Table 4.10). Further association with these alleles in 4-digit resolution revealed that the gene expression was associated with *HLA-DQA1*301* and *HLA-DRB1*401* (p_{mono}=9.92x10^{-18}, p_{B-cell}=2.65x10^{-12} with *HLA-DQA1*301*, p_{mono}=1.2x10^{-11}, p_{B-cell}=3.75x10^{-9} with *HLA-DRB1*401*, Figure 4.25B).

DEF8 was shown trans eQTL association in monocytes with rs35265698 (p_{mono}=8.96x10^{-17}). However, the association with HLA alleles was shown to be shared in the two cell types. *DEF8* expression with the peak eSNP rs35265698 minor homozygous genotype (GG) and *HLA-DQA1*301* was significantly down-regulated in both cell types (Wilcoxon rank sum test, p_{mono}=1.02x10^{-10}, p_{B-cell}=5.45x10^{-7}, Figure 4.25B). The *HLA-DRB1*401* allele with rs35265698 (GG) was also associated with down-regulation of *DEF8* (p_{mono}=1.4x10^{-6}, p_{B-cell}=2.99x10^{-5}).

Another example of a shared HLA association is *RPL34* which was shown to be a monocyte specific trans eQTL in the MHC class II locus (the most significant SNP rs1964995). Gene expression was associated with all three class II alleles, but was less significant with *HLA-DQB1* (p-values shown in the Table 4.10 and appendix Figure 1). The association with rs1964995 genotypes and the HLA alleles in *RPL34* expression was significant (Table 4.10).

Table 4.10: HLA allele association to cell type specific trans eQTL in the MHC region with a FDR (false discovery rate) cutoff of 0.05. trans eQTL are shown under the gene names. Some HLA alleles show only significant association to 2-digit alleles (NS: not significant). The significance of association between gene expression and HLA alleles is shown in each cell type. mono: monocytes.

Gene	HLA		Cell type	t-stat		p-value		FDR	
Name	2-digit	4-digit	specificity	2-digit	4-digit	2-digit	4-digit	2-digit	4-digit
Monocyte specific trans eQTL to chr6 SNPs									
AOAH	*DQA1.1*	*DQA1.101*	mono	8.71	5.85	$2.66\text{x}10^{-16}$	$1.40\text{x}10^{-8}$	$8.34\text{x}10^{-11}$	$6.94\text{x}10^{-4}$
Gene position: chr7	*DQA1.3*	*DQA1.301*	mono	-15.38	-15.03	$4.05\text{x}10^{-39}$	$7.81\text{x}10^{-38}$	$2.29\text{x}10^{-32}$	$1.47\text{x}10^{-31}$
SNP: rs28366298	*DQB1.3*	*DQB1.302*	mono	-10.83	-7.50	$4.74\text{x}10^{-23}$	$8.57\text{x}10^{-13}$	$2.98\text{x}10^{-17}$	$1.73\text{x}10^{-7}$
Position: chr6:32668837	*DQB1.5*	*DQB1.501*	mono	5.98	5.12	$6.74\text{x}10^{-9}$	$5.77\text{x}10^{-7}$	$3.66\text{x}10^{-4}$	$1.28\text{x}10^{-2}$
$p_{mono}=1.56\text{x}10^{-43}$	*DRB1.1*	NS	mono	5.36		$1.75\text{x}10^{-7}$		$5.19\text{x}10^{-3}$	
	DRB1.4	*DRB1.401*	mono	-15.04	-7.58	$7.24\text{x}10^{-38}$	$5.00\text{x}10^{-13}$	$1.47\text{x}10^{-31}$	$1.23\text{x}10^{-7}$
		DRB1.404	mono		-4.74		$3.48\text{x}10^{-6}$		$4.45\text{x}10^{-2}$
ARHGAP24									
Gene position: chr4	*DQA1.1*	NS	mono	-5.70		$3.08\text{x}10^{-8}$		$1.24\text{x}10^{-3}$	
SNP: rs28366298	*DQA1.3*	*DQA1.301*	mono	5.25	5.67	$3.03\text{x}10^{-7}$	$3.56\text{x}10^{-8}$	$8.01\text{x}10^{-3}$	$1.40\text{x}10^{-3}$
Position: chr6:32668837	*DRB1.4*	NS	mono	5.22		$3.43\text{x}10^{-7}$		$8.58\text{x}10^{-3}$	
$p_{mono}=4.59\text{x}10^{-14}$									
Shared trans eQTL to chr6 SNPs									
DEF8	*DQA1.1*	NS	mono	4.86		$1.95\text{x}10^{-6}$		$2.97\text{x}10^{-2}$	
Gene position: chr16	*DQA1.3*	*DQA1.301*	shared (mono)	-8.97	-9.18	$4.30\text{x}10^{-17}$	$9.92\text{x}10^{-18}$	$1.52\text{x}10^{-11}$	$4.31\text{x}10^{-12}$
SNP: rs35265698			shared (B-cell)	-7.53	-7.32	$6.95\text{x}10^{-13}$	$2.65\text{x}10^{-12}$	$1.87\text{x}10^{-7}$	$5.75\text{x}10^{-7}$
Position: chr6:32669312	*DQB1.3*	NS	mono	-5.88		$1.17\text{x}10^{-8}$		$6.06\text{x}10^{-4}$	
$p_{mono}=8.96\text{x}10^{-17}$	*DRB1.4*	*DRB1.401*	shared (mono)	-9.04	-6.09	$2.63\text{x}10^{-17}$	$3.75\text{x}10^{-9}$	$9.91\text{x}10^{-12}$	$2.12\text{x}10^{-4}$
$p_{B-cell}=6.24\text{x}10^{-13}$			shared (B-cell)	-7.55	-7.08	$6.23\text{x}10^{-13}$	$1.20\text{x}10^{-11}$	$1.76\text{x}10^{-7}$	$2.33\text{x}10^{-6}$
RPL34	*DQA1.5*	*DQA1.501*	B-cell	-5.79	-5.76	$1.86\text{x}10^{-8}$	$2.22\text{x}10^{-8}$	$1.19\text{x}10^{-3}$	$1.36\text{x}10^{-3}$
Gene position: chr4	*DQA1.1*	*DQA1.102*	mono	-5.40	-6.94	$1.41\text{x}10^{-7}$	$2.75\text{x}10^{-11}$	$4.34\text{x}10^{-3}$	$4.32\text{x}10^{-6}$
SNP: rs1964995	*DQA1.3*	*DQA1.301*	shared (mono)	6.90	6.83	$3.39\text{x}10^{-11}$	$5.22\text{x}10^{-11}$	$5.17\text{x}10^{-6}$	$7.02\text{x}10^{-6}$
Position: chr6:32557389			shared (B-cell)	5.18	5.26	$4.22\text{x}10^{-7}$	$2.84\text{x}10^{-7}$	$1.22\text{x}10^{-2}$	$9.20\text{x}10^{-3}$
$p_{mono}=3.60\text{x}10^{-24}$	*DQB1.6*	*DQB1.602*	shared (mono)	-8.74	-5.62	$2.16\text{x}10^{-16}$	$4.50\text{x}10^{-8}$	$7.18\text{x}10^{-11}$	$1.70\text{x}10^{-3}$
$p_{B-cells}=4.38\text{x}10^{-15}$	*DQB1.6*	NS	shared (B-cell)	-4.79		$2.67\text{x}10^{-6}$		$4.70\text{x}10^{-2}$	
	DRB1.13	NS	shared (mono)	-5.91		$9.74\text{x}10^{-9}$		$5.09\text{x}10^{-4}$	
			shared (B-cell)	-4.83		$2.29\text{x}10^{-6}$		$4.16\text{x}10^{-2}$	

	DRB1.4	NS	shared (mono)	6.73	9.55×10^{-11}	1.12×10^{-5}
			shared (B-cell)	4.96	1.24×10^{-6}	2.70×10^{-2}

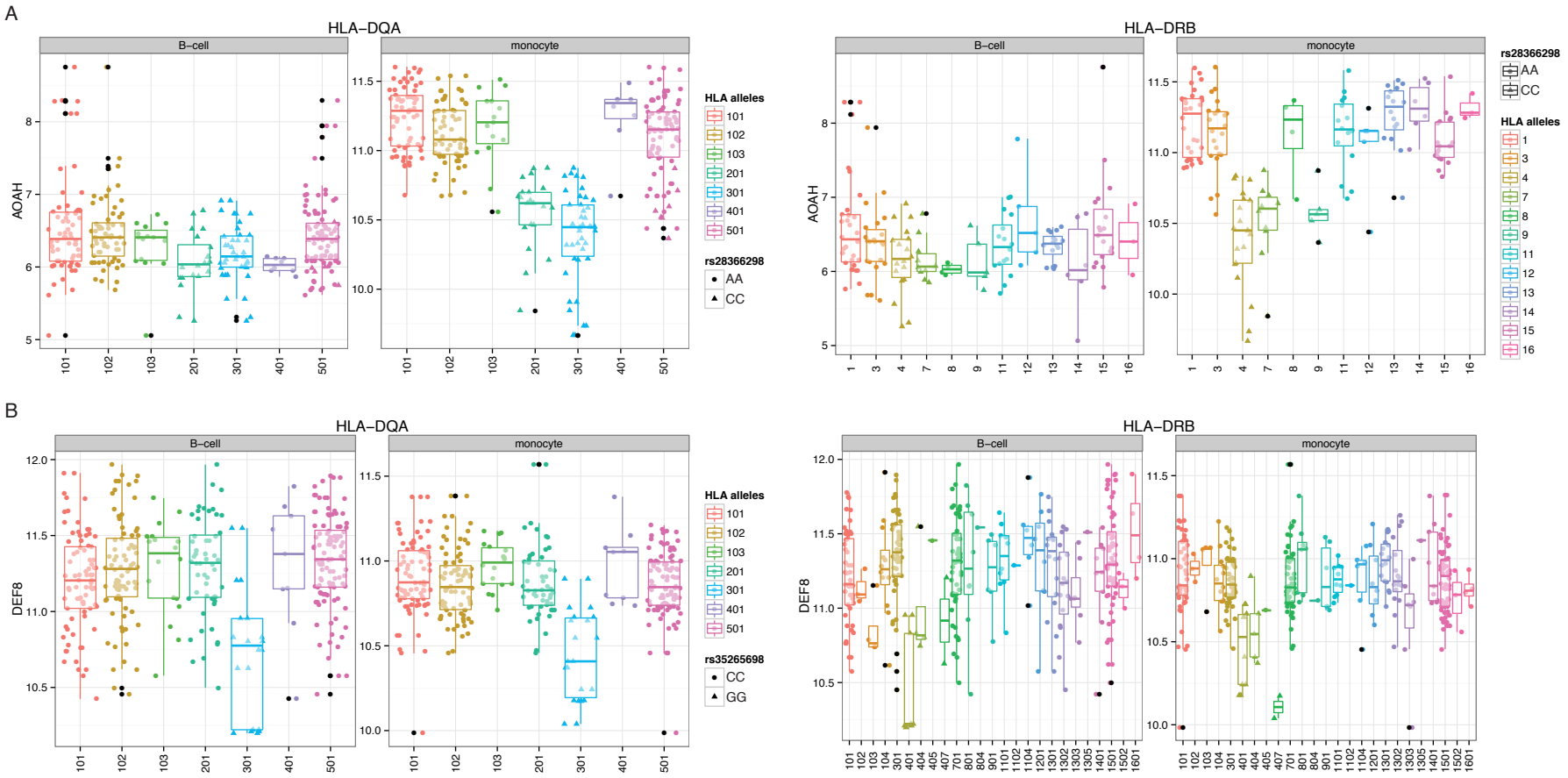


Figure 4.25: *AOAH* (A) and *DEF8* (B) associations with *HLA-DQA1* in 4-digit and *HLA-DRB1* in 2- (for *AOAH*) or 4-digit (for *DEF8*) resolution in a cell type specific manner. Homozygous genotypes of the eQTL SNPs (rs28366298 for *AOAH*, rs35265698 for *DEF8*) are shown. Black circles show outliers of box plots.

4.6.4 cis HLA association

The analysis using the genes in the MHC region identified significant cis associations between HLA alleles and 20 probes in monocytes and B-cells with a FDR cutoff of 0.05, although the validity of the gene expression data is uncertain due to extreme polymorphism. However, some associations were found interesting because of their associations with disease susceptibility. The table of the cis associations can be found in the appendix chapter (appendix Table 5). For example, expression of *PSORS1C3* which encodes psoriasis susceptibility locus 1 candidate 3 is significantly associated with *HLA-C*3* and *HLA-B*40* in B-cells (p_{B-cell}=7.91x10^{-15} and 9.42x10^{-9} respectively). The gene is located approximately 100kb away from *HLA-C* in 6p21 and is associated with susceptibility to psoriasis, a common chronic cutaneous inflammatory disorder (Trembath *et al.*, 1997, Nestle *et al.*, 2009). This locus has been identified by several GWAS for psoriasis susceptibility (Capon *et al.*, 2008,Strange *et al.*, 2010, Ellinghaus *et al.*, 2010, Huffmeier *et al.*, 2010). Although the expressed gene is predicted to be a non-protein coding gene, this association with *HLA-C* may suggest involvement at the transcriptional level for disease susceptibility. *HLA-A*1* was significantly associated with expression of the HLA class I gene *ZFP57* encoding a zinc finger protein in the two cell types (p_{B-cell}=3.29x10^{-6}, p_{mono}=1.70x10^{-6}). This result is consistent with a previous study showing that the common MHC haplotype affects genes in 6p21 including *ZFP57* (Vandiedonck *et al.*, 2011). These results suggest a cis association between HLA alleles and genes in 6p21, however the interpretation of the results may need caution and further validation may be required.

HLA alleles were successfully imputed from the genotyping data of the volunteer study individuals using the haplotype inference. The imputed HLA alleles have shown associations with gene expression in monocytes and B-cells where the loci have been linked to susceptibility to autoimmune diseases. The genes with cis and trans eQTL in 6p21.3 were identified to be associated with specific HLA alleles as well as the eQTL genotypes.

4.7 Discussion

The data presented here has shown that eQTL from highly purified primary human immune cells are cell type specific and related to immune function.

4.7.1 Population stratification

Genotyping data from healthy volunteers of European ancestry determined population stratification to verify the genetic background of the volunteer cohort. As a heritable component of gene expression has been shown by previous studies, this is one of the important steps for eQTL mapping analysis (Goring *et al.*, 2007, Spielman *et al.*, 2007, Emilsson *et al.*, 2008, Price *et al.*, 2011). Using a population with the same genetic background may help to reduce false positive associations.

4.7.2 cis eQTL mapping in two immune cell types

Purity of cells greatly contributes to identification of cell type specific eQTL as the majority, 78.1%, of eQTL were cell type specific which is consistent with a previous study comparing three cell types (Dimas *et al.*, 2009). As described in the previous chapter, the positive cell separation method was superior to the negative in terms of cell purity. The majority of cis associations involve genes expressed at similar levels across the two cell types (<0.5 difference in $\log_2$ average expression). This implies that cell type specific eQTL arise not only from differences in gene expression between cell types but also from functional genetic variants acting in a cell type specific manner.

Cell type specific eQTL were more distally spread from TSS than shared eQTL, consistent with the location of cell type specific enhancers and modulators (Dimas *et al.*, 2009). This supports the idea that functional variants with cell type specificity may lie distally from a core promoter element. This implies that cell type specific eQTL may utilise more regulatory elements collectively even though the effect sizes of the cell type specific eQTL were smaller than those of shared eQTL.

Genes associated with cis eQTL in a cell type specific manner included immunologically important genes and those associated with autoimmune diseases. For example, *BLK* encoding a B lymphoid tyrosine kinase is exclusively expressed in B-cells and thymocytes and the eQTL is associated with susceptibility to SLE. Monocyte specific eQTL was associated with *CD40* expression which encodes a TNF receptor family member that mediates signal transduction for immune responses. Pathway enrichment analysis did not identify a particular pathway. Approximately a third of the genes tested for the analysis associated with SNPs, which may be too high to find a contrast to the background genes for the enrichment analysis.

The directional effects of eQTL shared in both cell types have not been observed in previous eQTL studies. Some eQTL were observed to affect gene expression in an opposite

manner between monocytes and B-cells. Although the effect size of the directional eQTL was smaller than the non-directional shared eQTL, some genes involved encode immunologically important proteins such as Selectin-L encoded by *SELL*, a cell surface molecule required for migration of leukocytes during inflammation (McEver *et al.*, 1995, Xu *et al.*, 2008). This finding implies that directional eQTL may differentially serve in transcription for activation or repression in different cell types where a different set of transcriptional mechanisms is active.

The replication of cis eQTL with previously published studies has shown a relatively low concordance with eQTL data from LCLs (Dixon *et al.*, 2007, Stranger *et al.*, 2007a). The difference from primary B-cell eQTL may relate to transformation of LCLs and long term culturing modifying their expression profiles (Caliskan *et al.*, 2011). It has been reported that genetic control of whole blood and LCLs is predominantly independent, especially non-housekeeping genes, and it has been speculated that cell type specific eQTL account for this (Powell *et al.*, 2012). However a comparison of different eQTL studies is generally difficult due to the different platforms for expression and genotyping used (Majewski and Pastinen, 2011).

4.7.3 trans eQTL

Detection of trans eQTL associations has been shown to be difficult in previous studies because of small population size and tissue specificity (Price *et al.*, 2011, Grundberg *et al.*, 2012). In this dataset, trans eQTL have shown more cell type specificity than cis eQTL. Several cell type specific trans master regulators including MHC, 12q15 in monocytes and 9q31.2 in B-cells, were identified. Although functional characterisation of these trans eQTL remain to be explored, genes associated with the trans eQTL may help to imply the possible biological pathways involved. The monocyte specific *LYZ* locus that showed strong eQTL in both cis and trans regulates expression of many genes in trans. The trans regulated genes may regulate activation of downstream signalling pathways. For example, TF *CREB1* may play a key role in activating immune related genes and *RAB27A* encoding a GTPase is required to tether lysosome at a membrane (Elstak *et al.*, 2011). These genes showed expression correlation to *LYZ* in an allele specific manner. *YEATS4*, which showed a cis association with 12q13.2 encodes a putative TF but did not show expression correlation to the trans associated genes. The B-cell specific trans eQTL in 9q31.2 showed a moderate cis association to *KLF4* encoding a dual function transcription activator and suppressor. This eQTL association in trans includes genes encoding receptors and ligands

and showed correlation in gene expression (examples; *PLAUR*, *CLEC12A*, *TNFSF13B*). Modulation of expression levels for these receptors and ligands in an allele specific manner may lead to a differential response in signal transduction. However, it is difficult to interpret a set of genes involved in trans eQTL because not many genes have direct interaction with transcription or the immune system. In addition, associations between eQTL and genes associated in trans may be hidden due to a small effect size or unknown reasons.

4.7.4 GWAS and eQTL

While searching for eQTL in disease susceptibility associated loci identified from GWAS, several eQTL in cis and trans, including shared and cell type specific associations, were identified in susceptibility associated loci for autoimmune diseases. Biological interpretation of results from GWAS is often difficult because loci associated with susceptibility are frequently located in gene deserts or flank many possible candidate genes (Freedman *et al.*, 2011). eQTL from this dataset that overlap with GWAS loci may help to contribute functional characterisation and understanding of disease susceptibility.

Both cis and trans eQTL were found in GWAS susceptibility loci for autoimmune diseases. cis eQTL may be relatively straightforward to functionally characterise because they may function as transcriptional modulators or create or disrupt sites for epigenetic modification. Some examples were shown previously (e.g. *SORT1* and cholesterol levels; Musunuru *et al.*, 2010). The eQTL with *ERAP2* shared in the two cell types was found in a locus associated with susceptibility to Crohn's disease. Chronic inflammation in Crohn's disease may be involved in the expression level of *ERAP2* which is associated with expression of HLA class I molecules.

In contrast, trans eQTL suggest indirect interaction for disease susceptibility between the eQTL and genes involved. For example, the MHC class II region which is strongly involved in many autoimmune diseases showed trans associations to genes including *AOAH* in monocytes. The eSNPs in the MHC class II region are in intergenic as well as GWAS identified loci, therefore these variants may modulate enhancer or insulator elements. *AOAH* is involved in digestion of bacteria-derived LPS which is important in the innate immune system. Acyloxyacyl hydrolase is shown to limit LPS and control polyclonal antibody production by B-cells (Lu *et al.*, 2005). It is difficult to postulate functional interaction between the eQTL with GWAS associations and the genes involved for autoimmunity. The B-cell specific trans eQTL at 12q13.2 involving *CDK2* and a

potential trans gene *CDKN1A* was associated with several autoimmune diseases including T1D. *CDK2* and its inhibitor *CDKN1A* are involved in cell cycle and apoptosis, suggesting this association may modulate dysregulation of cell homeostasis as shown in tumourigenesis (Gartel and Radhakrishnan, 2005).

4.7.5 HLA

The increased number of genotypes in MHC loci and the advanced knowledge of the HLA types have made imputation of the HLA loci without serotyping possible. Analysis with the imputed HLA alleles of the volunteer study cohort has identified trans associations with gene expression in a cell type specific manner. The analysis identified genes involved in trans eQTL in monocytes in the MHC class II region including *AOAH*. Investigating specific HLA and eSNP alleles together have shown either up- or down-regulation of the associated genes. For example, *AOAH* expression in monocyte with *HLA-DRB1*4*, *-DRB1*7* and *-DRB1*9* was significantly down-regulated. These HLA alleles are linked to *HLA-DRB4* encoding a supertypic HLA-DR53 serotype which is positively associated with autoimmune diseases (Knowles *et al.*, 1986, Kobayashi *et al.*, 1990, Morling *et al.*, 1991, Stunz *et al.*, 1989). Since the *AOAH* product, acyloxyacyl hydrolase, functions in a part of LPS hydrolysis, the reduced expression of the gene may be involved with predisposition to inflammation.

Briefly cis association was explored with genes in the MHC region. Due to extreme polymorphisms in the MHC region, the gene expression data used needs further verification. However, the association of *HLA-C*3* and *PSORS1C3* in B-cells implies involvement with psoriasis susceptibility. No known non-synonymous mutation has been identified for this psoriasis susceptibility locus although many association studies identified association with this locus (Nestle *et al.*, 2009). Because some of the HLA alleles have been identified as susceptibility loci for autoimmune diseases and infectious diseases, the association with eQTL in immune cells may elucidate biological pathways for disease susceptibility and development.

4.7.6 Methodology and limitation of eQTL mapping

Methods to improve eQTL mapping analysis is described below. Imputation of SNPs at a genome-wide scale is readily available using reference genotyping data from The 1000 Genomes Project (Howie *et al.*, 2009). This would increase resolution of LD structures and improve detection of eQTL signals.

This study dataset used a MAF cutoff >1%. rs10816519 in 9q31.2 and its B-cell trans association with *KLF4* is intriguing to follow up. However, the current dataset does not have any minor homozygous individuals. Increased sample size may help to identify regulatory variants by including enough samples for SNPs with MAF between 1-5% to be analysed more informatively.

Use of the microarray platform for gene expression quantification is a powerful tool to quantify thousands of transcripts. However SNP confounding probe sequences can cause detection of false positive signals due to effects on oligonucleotide probe hybridisation to transcribed RNA. Thus the probes without SNPs (MAF>1% in the 1000 Genomes Project data, Interim Release 23/11/2010) were only included in this dataset (29022 out of 47321 probes). The Illumina BeadChip gene expression platform used in this dataset uses 50-mer oligonucleotide sequences to hybridise their target genes (Barbosa-Morais *et al.*, 2010). The affymetrix GeneArray uses 25-mer oligonucleotide sequences and multiple probes for each transcript, however SNPs in probe sequences can still cause false positive signals (Benovoy *et al.*, 2008). Highly polymorphic loci such as HLA become problematic to design SNP-free probes. This could be overcome by using whole transcriptome sequencing technology (Tang *et al.*, 2010, Majewski and Pastinen, 2011). Initial studies using transcriptome sequencing analysed eQTL mapping, however the technology is not cost effective quantifying hundreds of samples (Pickrell *et al.*, 2010, Montgomery *et al.*, 2010).

4.7.7 Summary

The dataset presented in this chapter consists of two highly purified immune cell types and the eQTL mapping analysis was performed in order to identify regulatory variation in cis and trans in the context of the immune response and inflammation. The identified eQTL were analysed in a cell type specific manner, with association with GWAS disease susceptibility loci and with imputed HLA alleles. These information may help to understand functional characterisation of genetic variants with emphasis on susceptibility and development of autoimmune diseases.

5 Paired eQTL analysis of monocytes and differentiated macrophages

5.1 Introduction

Macrophages have important roles in both innate and adaptive immune responses. They play a particular role in inflammation and tissue repair (Murray and Wynn, 2011). Dysregulation of these events is associated with chronic inflammation which is observed in many autoimmune and common diseases states (Ingersoll *et al.*, 2011). Chronic inflammatory and autoimmune diseases, including Crohn's disease, RA and MS, are involved with pro-inflammatory cytokines produced by inflammatory cytokines TNFα, IL12, and IL23 (Murray and Wynn, 2011). For example, Crohn's disease patients carry a population of inflammatory $CD14^+$ macrophages in the intestine distinct from normal tissue macrophages and they produce IL23 and TNFα to mediate the chronic inflammation state (Kamada *et al.*, 2008). Macrophages are involved in maintenance of tissue homeostasis and dysregulation may result in a pathological state. Obesity is associated with transition of adipose tissue macrophages (ATM) into inflammatory macrophages (Lumeng *et al.*, 2007). Homeostasis and insulin sensitivity are normally controlled by ATM with production of anti-inflammatory cytokines including IL10 (Olefsky and Glass, 2010). High fat diet changes adipocyte metabolism and gene expression resulting in production of pro-inflammatory cytokine TNFα and chemokines to recruit inflammatory macrophages as well as transition of ATM to inflammatory macrophages.

Circulating monocytes can differentiate into inflammatory macrophages by stimuli including exogenous PAMPs or endogenous cell death. Macrophages display heterogeneity reflected by their plasticity in order to accommodate immunological responses (Gordon and Taylor, 2005). Profiling the shift in gene expression level that associates with differentiation from monocytes to macrophages has been previously investigated at only

a small scale (Martinez *et al.*, 2006, Gratchev *et al.*, 2008, Szanto *et al.*, 2010). These studies have described distinctive gene expression changes that are caused by different types of activation in *ex vivo*. Gene enrichment analysis on differentially expressed genes highlighted biological pathways that are involved in lipid metabolism, chemokine production, and a different repertoire of G protein-coupled receptors (Martinez *et al.*, 2006). The expression of *PPARG*, encoding peroxisome proliferator activated receptor-γ, is required for differentiation of ATMs (Odegaard *et al.*, 2007). Gene profiling of macrophage differentiation showed STAT6 interaction with the promoter region of *PPARG*, suggesting that STAT6 regulates *PPARG* and a downstream mechanism for ATM differentiation (Szanto *et al.*, 2010). There has been limited investigation into genetic variation that influences these gene expression changes during macrophage activation. This chapter describes analysis of changes in gene expression during macrophage differentiation *ex vivo* from monocytes on a large scale and investigates how genetic variation may modulate this.

Objectives

This chapter describes analysis of paired gene expression data of human primary $CD14^+$ monocytes and *ex vivo* differentiated macrophages. The main aims of this analysis are to:

1. describe changes in gene expression during the monocytes to macrophage transition.
2. investigate the effect of genetic variation on macrophage gene expression.
3. investigate the effect that underlying genetic variation plays in this transition.

5.2 Sample collection of the paired monocytes and macrophages population

5.2.1 $CD14^+$ monocyte collection of healthy European individuals

A further cohort of healthy individuals of European descent were recruited in Oxfordshire between June and July, 2012. The median age of the population was 39 years (interquartile range (IQR) +/-18, Table 5.1).

Whole blood from 64 individuals was collected and processed by separating $CD14^+$ positive monocytes using a method described in section 3.2. $5x10^5$ $CD14^+$ monocyte cells were immediately harvested in the RLT reagent and were stored at −80 °C for later RNA

extraction. For macrophage differentiation, 4x10^5 $CD14^+$ monocyte cells were cultured with TNFα and granulocyte macrophage colony stimulating factor (GM-CSF) over 18 days and harvested in RLT reagent before RNA extraction and quantification (Grabstein *et al.*, 1986, see Materials and Methods, Chapter 2 for details). The median concentration of total RNA was 21.2 and 49.7 µg/µL (IQR +/-6.9 and +/-30.1 respectively) for $CD14^+$ monocytes and macrophages respectively.

		Total	Male	Female
	Sample number (%)	64	39 (60.9)	25 (39.1)
Age (years)	Range	19-66	19-66	20-59
	Median	39	39	39
	Interquartile range	18	14	22

Table 5.1: Demographics of the 64 healthy individuals of European descent.

5.2.2 *ex vivo* differentiation of monocytes into macrophages

In order to determine the extent of macrophage differentiation, expression of monocyte and macrophage CD markers were assessed using qPCR. The expression of CD14, a highly expressed myeloid cell marker, and CD68, a transmembrane glycoprotein highly expressed in macrophages and dendritic cells (DCs) were used. qPCR results were normalised to expression in monocytes and showed that the differentiated macrophages expressed CD68 significantly higher than monocytes (Figure 5.1). This result confirmed that the macrophages have been successfully differentiated. The expression of CD14 was equally expressed between the two cell types. The RNA samples were subsequently processed for gene expression profiling.

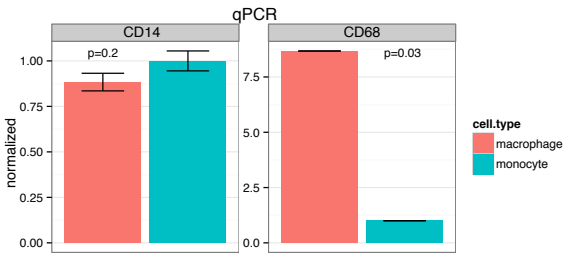


Figure 5.1: qPCR of CD14 and CD68 in differentiated macrophages in comparison with $CD14^+$ monocytes. The macrophages values are normalised with the monocytes. Mean and standard errors are shown. CD68 expression in macrophages is significantly higher than in monocytes (p=0.029, Wilcoxon rank sum test) and CD14 expression is at similar levels (p=0.2).

5.3 Population stratification analysis

Examining potential population stratification is important to perform when investigating the genetic determinants of gene expression at a population level. Gene expression has a heritable component, thus removing outliers based on the genetic background of a population can avoid skewed results in analyses. This section describes the analysis of the genetic structure of the cohort used to study monocyte and macrophage expression. The principals of genotyping, its quality control and population stratification used in this dataset was described in Section 4.3.

5.3.1 Genotyping

Genomic DNA was extracted from whole blood of 63 individuals and was hybridised onto the Illumina OminiExpress BeadChips. One of the 64 individuals was not genotyped due to unavailability of the DNA sample. The genotyping data was scanned and normalised on the GT module of Illumina Genome Studio. The 730525 SNPs which passed the threshold on GenCall scores >0.15 were used in further analyses.

5.3.2 Quality control of genotyping

Quality control of genotyping data on the 63 individuals was performed in order to remove any SNPs or samples which have bias caused by technical artefacts or underlying population structure.

SNP filtering

Due to the small sample size, a MAF>5% cutoff was used for genetic stratification analysis. A SNP call rate >0.98 with MAF>5% criteria left 590113 SNPs. The Hardy-Weinberg test is used to determine if biallelic SNPs deviated from the Hardy-Weinberg equilibrium (HWE) using a statistical significant test (Wigginton *et al.*, 2005). There were 92 SNPs identified at HWE $p<1x10^{-6}$ and removed from further analysis. Thus the genotyping dataset included 590021 SNPs for population stratification analysis.

Sample filtering

A sample call rate threshold of >0.98 was applied to the genotyping dataset and no individual failed. The average sample call rate was 99.95%. The reported sex was correct for all individuals by testing the heterozygosity rate of SNPs on the X chromosome.

The heterozygosity rate in this population was determined to identify any outliers. The mean heterozygosity rate in this population was 35.68% and the standard deviation was +/-0.74%. No individual failed this threshold. Relatedness of each individual in the population using an identity by descnet (IBD) test was calculated to determine any sample duplication or related individuals. No individual was duplicated or related to each other closer than second-degree relatives. Thus 63 individuals were retained for the next stage of analysis.

5.3.3 Population stratification

In order to assess population structure, multidimensional scaling (MDS) with 75691 SNPs which were not in LD (r^2<0.2) was calculated in comparison to the defined HapMap and 1000 Genomes populations. MDS visualises the similarity of samples in a dataset at a low dimensional space. The reference populations included 14 populations from Europe, Africa, Southeast Asia and Latin America (Appendix Table 6). A pairwise IBD matrix in this population was used to determine MDS for differences in ancestry. The first and second components (C1 and C2) of the MDS analysis with 14 populations describe that the current population significantly overlaps with the European populations (Figure 5.2). Further resolution of the MDS analysis with the European populations has shown that this population consists of individuals of Northern and Western European decent as they overlap with CEU and GBR (British from England and Scotland) populations (Figure 5.3).

Final dataset

The quality control and the population stratification analysis retained the 63 individuals without any outliers. For further analyses with the gene expression data, MAF 10% was chosen as a cutoff because the small population size of the current dataset may be susceptible to higher false positive results. The final genotyping dataset consists of 63 individuals with 515282 SNPs after filtering with the primary criteria of MAF>10%, SNP call rate>98%, HWE p>1x10^{-6} and sample call rate >98%.

5.4 Gene expression profiling

The gene expression of $CD14^{+}$ monocytes and macrophages was quantified using the Illumina Human HT-12v4 BeadChips at the Sanger Wellcome Trust Centre and the Wellcome Trust Centre for Human Genetics respectively. The 63 paired samples of

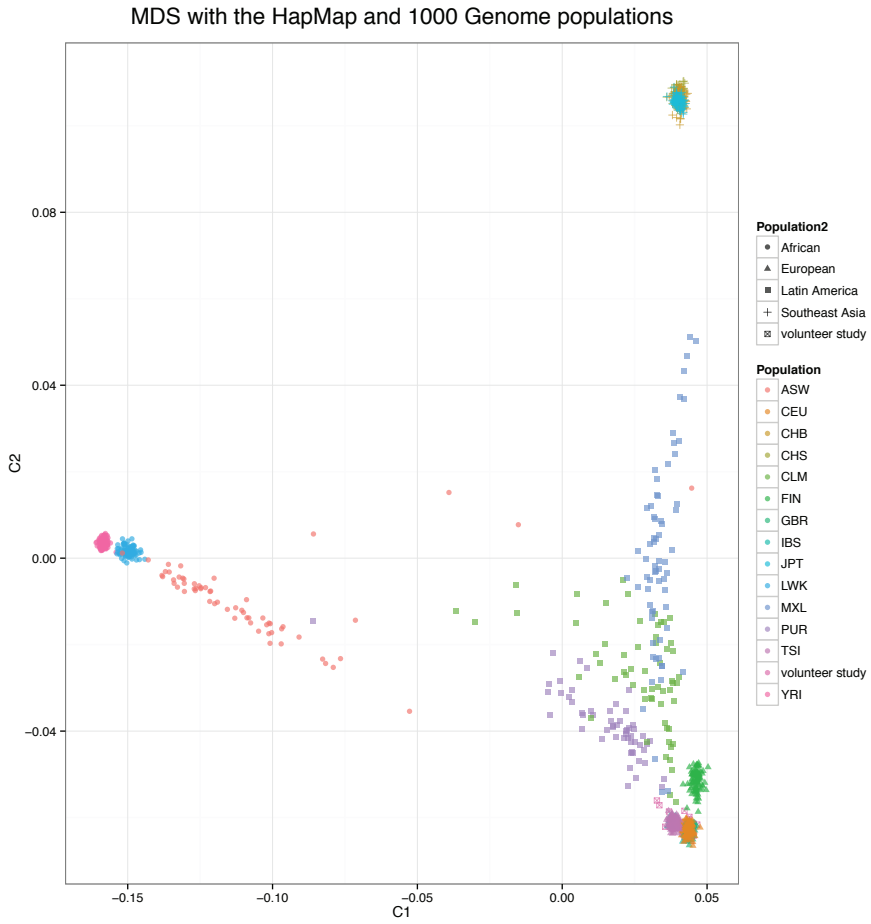


Figure 5.2: MDS plot of population stratification for the 63 individuals. Circles - Africans (ASW - African ancestry from southwest of the USA, LWK - Luhya from Kenya, YRI - Yoruban from Nigeria), triangles - European (CEU - Northern and western European ancestry from Utah, USA, FIN - Finnish from Finland, GBR -British from England and Scotland, IBS - Iberian in Spain, TSI - Toscani in Italy), squares - Latin America (CLM - Colombian in Medellin, Colombia, MXL - Mexican from Los Angeles, California, USA, PUR - Puerto Rican in Puerto Rico), crosses - Southeast Asia (CHB - Han Chinese in Beijing, China, CHS - Han Chinese in South China, JPT - Japanese from Tokyo, Japan), squares with x - volunteer study; the 63 healthy volunteers in this study.

monocytes and macrophages were analysed for gene expression. The details of the Illumina BeadChip technology for gene expression profiling are described in Section 3.4.

5.4.1 Normalisation of microarray data

The raw signal intensity of all 126 microarray data was obtained from the GX module of Illumina GenomeStudio and were subsequently analysed.

Filtering probes

Probe filtering was separately applied to the monocyte and macrophage datasets. Probes with detection scores >0.99 in at least 5% of the datasets (n=4 in the monocytes and macrophage datasets respectively) were used for subsequent analysis. There were 18422



Figure 5.3: MDS plot of population stratification within Europeans. Orange circles - CEU, dark yellow circles - FIN, green circles - GBR, light blue circles - IBS, blue circles - TSI, purple triangles - volunteer study (the 63 healthy volunteers in this study).

and 18717 probes that passed the filtering in the monocyte and macrophage datasets respectively. Out of these probes, 15910 probes were the same between the two datasets. Probes which met the criteria in either dataset but not in the other were included because these probes may be specifically expressed in only one cell type. There were 21229 probes remaining after filtering.

Normalisation

The monocyte and macrophage expression samples (n=126) were combined and normalised using Robust Spline Normalisation (RSN) after $\log_2$ transformation. For this dataset, two iterative procedures were required. One outlier was removed from further analysis. The density plots of the raw and normalised data show that the peaks of signal intensity were shifted to form one large peak after normalisation (Figure 5.4). Clustering of the samples shows that the cluster is tighter with the normalised data and forms two discrete clusters corresponding to the monocyte and the macrophage datasets (Figure 5.5). The

distribution of expression values of each sample before and after normalisation shows that the median is normalised to the same level throughout the samples (Figure 5.6). The variability outside of the interquartile range in the macrophages is greater than in the monocytes.

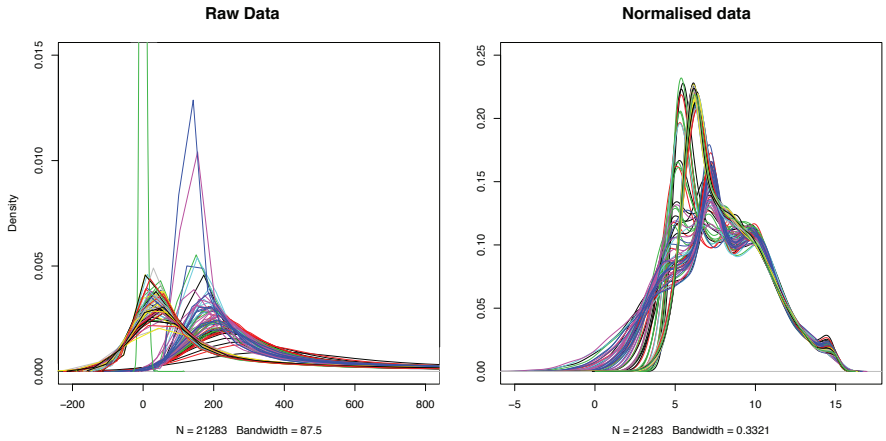


Figure 5.4: Density plot of signal intensity before (left panel) and after (right panel) normalisation. The raw data forms two large peaks corresponding to monocytes (left) and macrophages (right). The normalised data forms different peaks but in a same range of the signal intensity.

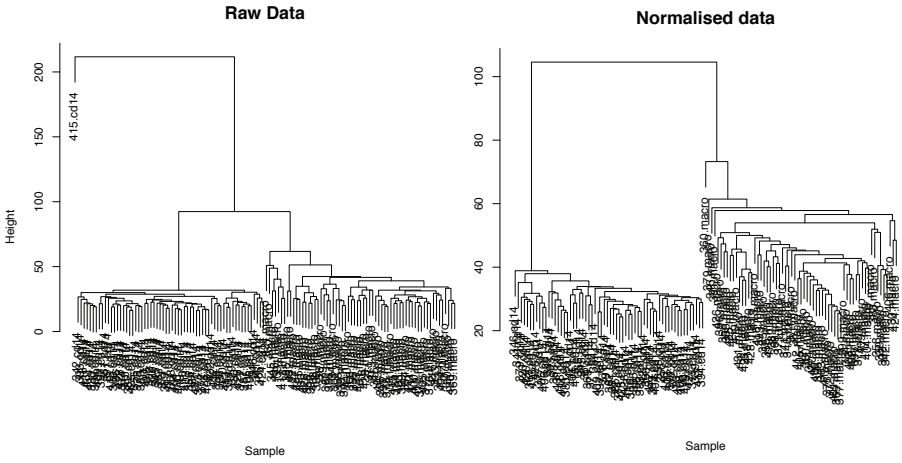


Figure 5.5: The clusters of the monocyte and macrophage expression data before (left panel) and after (right panel) normalisation. The outlier is located outside of either clusters in the raw data. The normalisation of the data after removal of the outlier forms two clusters corresponding to the monocytes (left) and the macrophages (right).

MA plots

Normalisation of the dataset was analysed using a MA plot. Four randomly selected samples (two monocytes and two macrophages) were plotted based on the ratio of $\log_2$ signal intensity in each probe (M) and the average (A) between two samples (Figure 5.7). The median values shown by the red lines in each plot approach zero (blue lines) in the

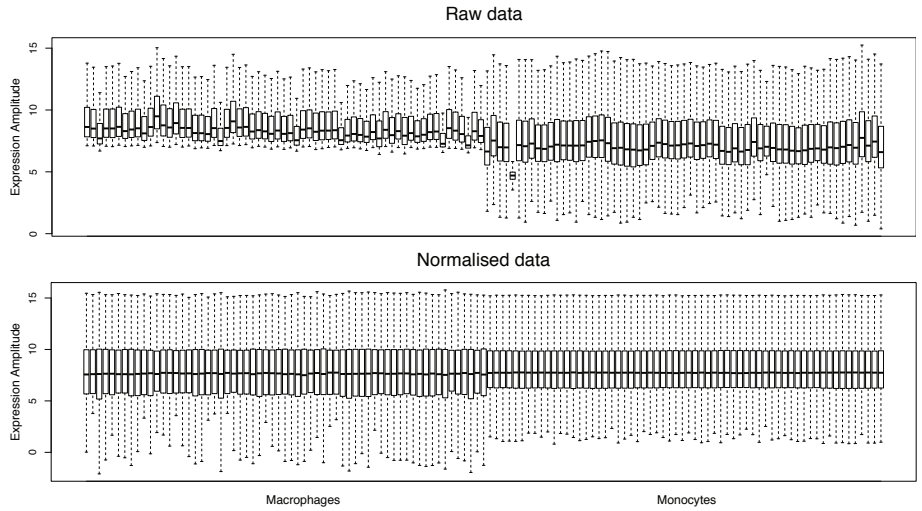


Figure 5.6: The distribution of expression values before (top panel) and after (bottom panel) normalisation. Variability outside of the interquartile ranges is greater in the macrophages (left half) than the monocytes (right half).

same cell types. However, in comparisons between different cell types the median values do not approach zero even after normalisation especially at low gene expression. They retained shift in distribution of gene expression as would be expected with differential gene expression between cell types.

Principal component analysis

Principal component analysis (PCA) explains the variation between samples by reduced dimensionality of the dataset. PCA was performed on the monocyte and macrophage datasets before and after normalisation in order to assess their variation. The cell types were clearly the main variation in the dataset both before and after normalisation (Figure 5.8). The largest variation is explained by the first component, PC1, which accounted for 55.13% and 70.02% of the variation in the raw and normalised data respectively. The outlier that was identified in the raw data was found far out from either cell type clusters. After normalisation, the second component (PC2) was reduced from 28.67% to 3.58%. Thus, the PCA demonstrated that the variation in the dataset was mainly due to the cell types. However, PC2 in the normalised data identifies two potential outliers in the macrophage data (IDs 360 and 379) which were not detected as outliers during the normalisation step (Figure 5.8).

Further sample filtering

In order to analyse the dataset in a pairwise manner, the macrophage sample arising from the same individual as the monocyte outlier was also removed from the dataset.

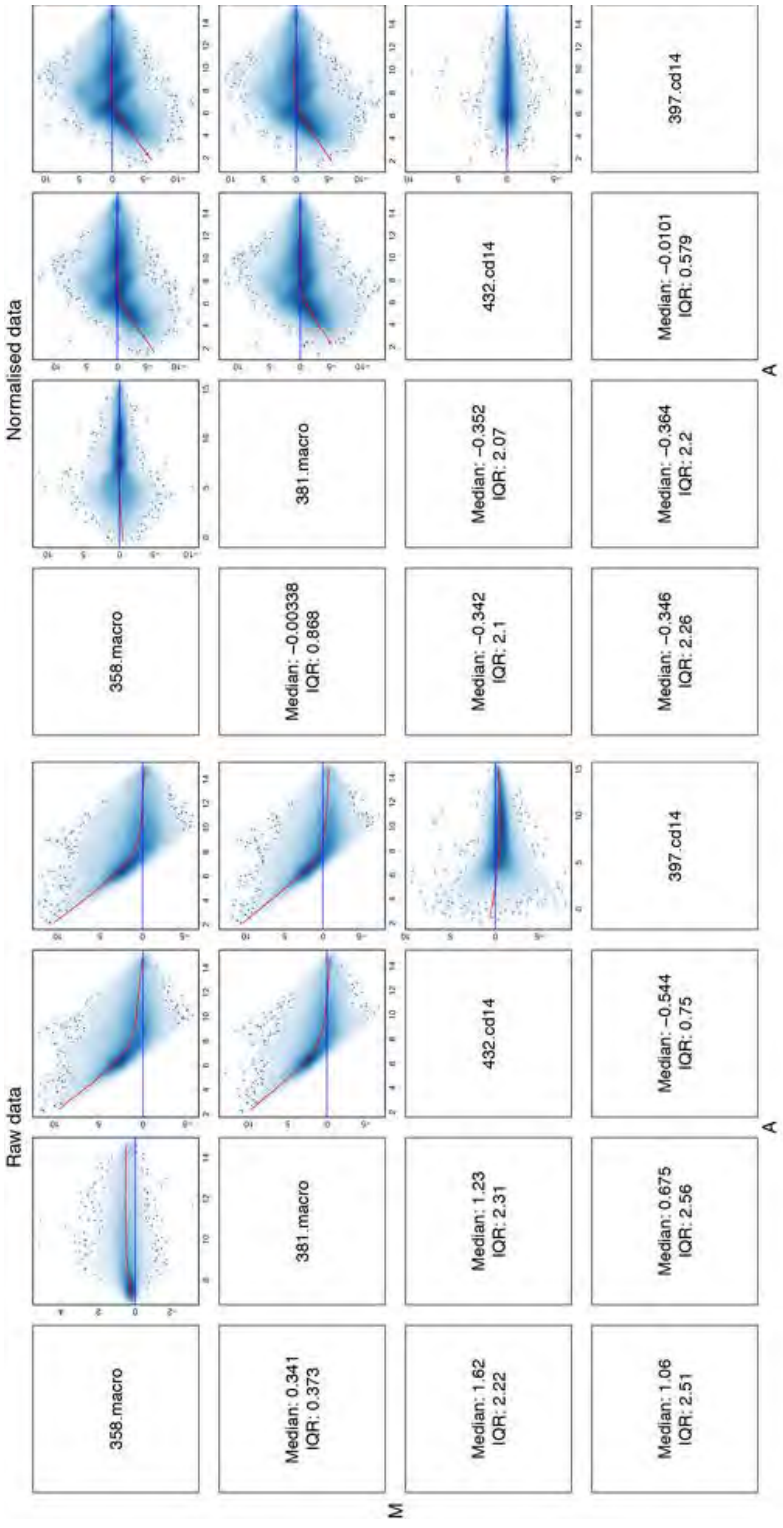


Figure 5.7: Pairwise MA plots of the monocyte and macrophage expression data from randomly selected samples before (left panel) and after (right panel) normalisation. The Y and X axes show the ratio of $\log_2$ signal intensity (M) and the average between the two samples (A). The median values (red lines) in each plot show the efficacy of the normalisation. macro; macrophages, cd14; $CD14^+$ monocytes. Medians and interquartile ranges (IQR) for each pair are shown.

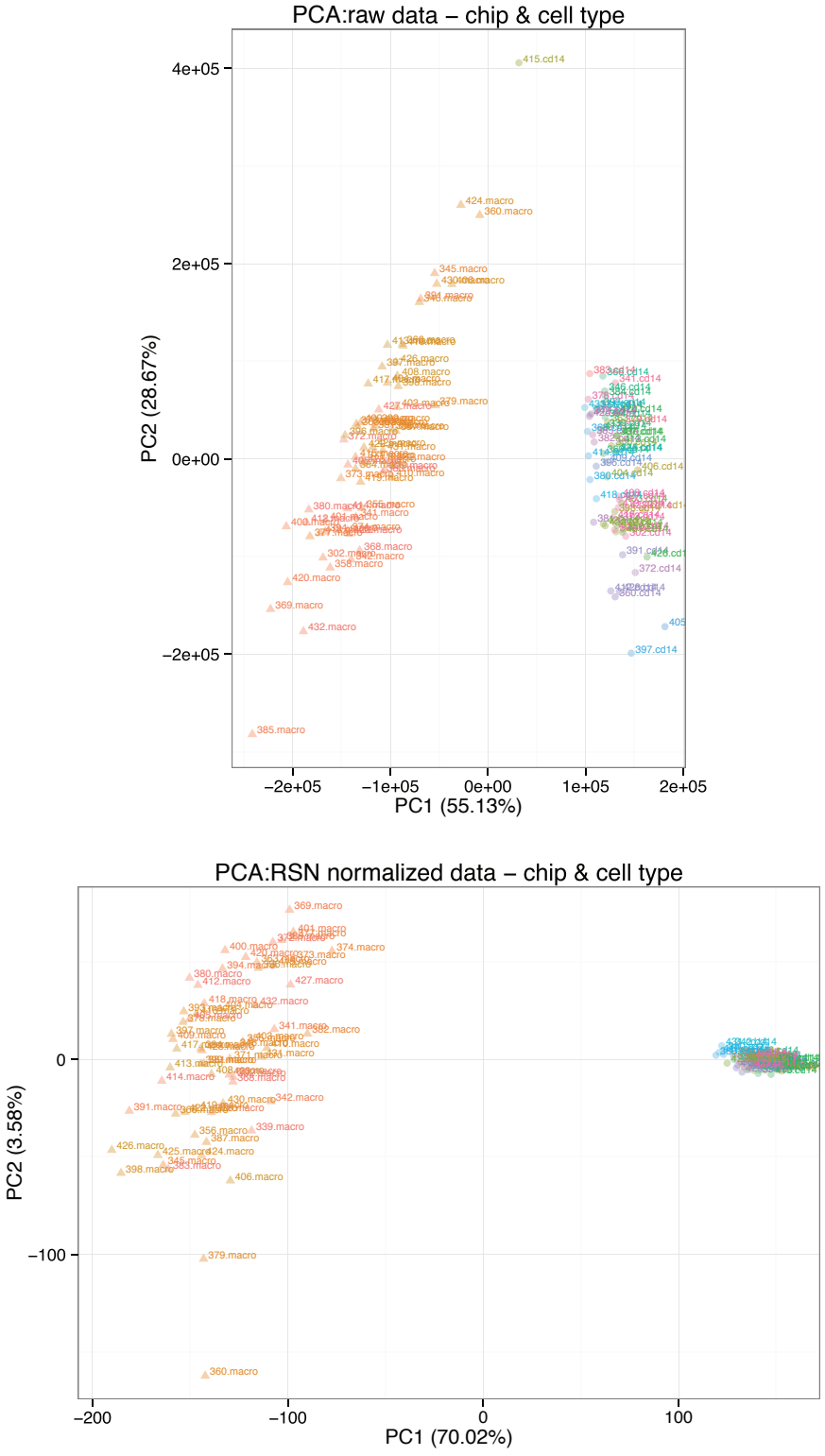


Figure 5.8: PCA plots of the expression data in the monocytes and macrophages before (top panel) and after (bottom panel) normalisation. The samples are coloured by microarray chips (n=42) and are shown by cell types (circles - monocytes, triangles - macrophages). Percentages of variance explained by PCs before and after normalisation are shown: PC1 - 55.13% and 70.02%, PC2 - 28.67% and 3.58% in the raw and normalised data respectively.

Normalisation of the remaining 124 microarray samples was performed with the 21229 probes. All the quality control plots were very similar to the data described above (Figures 5.4, 5.5, 5.6). There were no outliers detected and all quality control steps assured that the normalisation was efficiently performed.

Final dataset

The final dataset of the gene expression for further analysis consists of the 124 samples, the 62 paired dataset of the monocytes and the macrophages, with 21229 probes.

5.5 Differential gene expression analysis

Differential gene expression analysis of the monocytes and macrophages was performed using the R package limma (described in the Section 3.5.1). Limma is designed to compute a contrast of gene expression using an estimated coefficient of a linear model in a comparison of two or more groups. In this analysis, a moderated paired t-test was used for the paired monocytes and macrophages data.

5.5.1 Processing the data

The normalised gene expression data of the 62 paired samples was computed using limma. In summary, 71% of probes tested (15072 probes out of 21229) were significantly differentially expressed between two cell types at an adjusted p-value cutoff $1.1\text{x}10^{-10}$ (Bonferroni corrected for multiple testing). A contrast for the analysis was made for macrophages to monocytes, giving moderated t-statistics that are positive for macrophage up-regulated genes and negative for monocyte up-regulated genes. This analysis was focused on the change of gene expression in macrophage transition. In order to avoid confusion, the results are described in a manner of either up- or down-regulation in macrophages. A summary of the analysis statistics is shown in Table 5.2. The number of probes significant in monocytes is larger in total, but there are more probes in macrophages that are more significant (Figure 5.9). The B statistics (log-odds) of differential expression is larger in macrophages (Figure 5.10). The MA plot of the differentially expressed genes between the monocytes and macrophages shows that the less expressed genes on average between the two cell types are likely to be up-regulated in monocytes, which implies that these genes may be down-regulated in the macrophage due to differentiation (Figure 5.11). The biological interpretation of these genes is described

in the next sections, focused on the characteristics of macrophage differentiation and activation for immune responses.

Table 5.2: Summary of differentially expressed probes in macrophages contrasted to monocytes. Thresholds for cutoff are the adjusted p-value with FDR.

Threshold	Number of probes (%)			Collective (%)		
p (<)	Total	Up (%)	Down (%)	Total (%)	Up (%)	Down (%)
Total	21229	8115 (38.2)	13114 (61.8)			
$1.1x10^{-10}$	15072	5499(36.48)	9573(63.5)	100	100	100
$1x10^{-15}$	12499	4674(37.39)	7825(62.6)	82.9	85.0	81.7
$1x10^{-20}$	9664	3852(39.86)	5812(60.1)	64.1	70.0	60.7
$1x10^{-30}$	4786	2342(48.93)	2444(51.1)	31.8	42.6	25.5
$1x10^{-40}$	1743	1070(61.39)	673(38.6)	11.6	19.5	7.0
$1x10^{-50}$	444	376(84.68)	68(15.3)	2.9	6.8	0.7

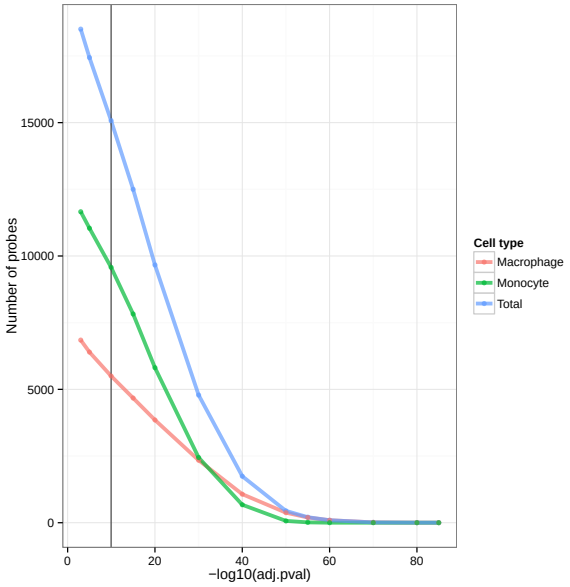


Figure 5.9: Numbers of significant probes in macrophages and monocytes differential gene expression analysis. The vertical grey line indicates $1.1x10^{-10}$ p-value cutoff.

5.5.2 Differentially expressed genes in macrophages compared to monocytes

Macrophages have multiple roles dependent upon development and anatomical location. For example, inflammatory macrophages can differentiate from circulating monocytes and play a major role in killing invading agents (Serbina *et al.*, 2008). They can alternatively differentiate into tissue macrophages through induction by cytokines IL4 and IL13 to maintain homeostasis and repair of tissues (Gordon and Martinez, 2010). Stimulating $CD14^{+}$ monocytes with TNFα and GM-CSF promotes differentiation into inflammatory

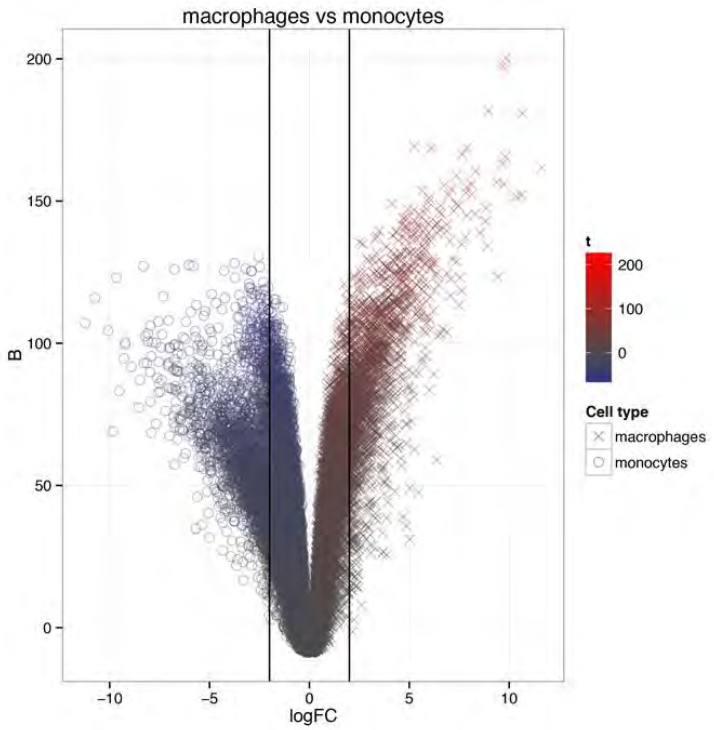


Figure 5.10: Volcano plot of differentially expressed genes in macrophages and monocytes. Probes up-regulated (red cross) and down-regulated (blue circle) in macrophages are plotted based on $\log_2$ fold change (x-axis) and B log-odds (y-axis). The moderated t-statistics are shown in two colours (red is up-regulated and blue is down-regulated). Vertical lines indicate a $\log_2$ fold change of +/-2.

macrophages and DCs (Grabstein *et al.*, 1986). This section describes the roles of genes differentially regulated in macrophage differentiation.

Pathway analysis of differentially expressed genes in monocytes and macrophages

In order to systematically interpret genes differentially expressed between monocytes and macrophages, a gene enrichment analysis was performed using the Reactome database (described in detail in Section 3.6). The significantly differentially expressed gene probes with available Entrez gene IDs (5428 and 8755 probes in up- and down-regulated macrophage genes respectively) were tested for enrichment in biological pathways with the background of a total of 20100 probes with FDR cutoff <0.01. 4552 and 7206 genes were significantly identified as enriched in 7 and 25 pathways in up- and down-regulated genes respectively (most significant pathways p_{up}=5.14x10^{-14} and p_{down}=5.81x10^{-7}). These pathways include subgroups. No pathway was significantly enriched by genes that are equally expressed between the two cell types. The results are summarised in Tables 5.3 and 5.4. The description of the pathways is given below, followed by individual genes.

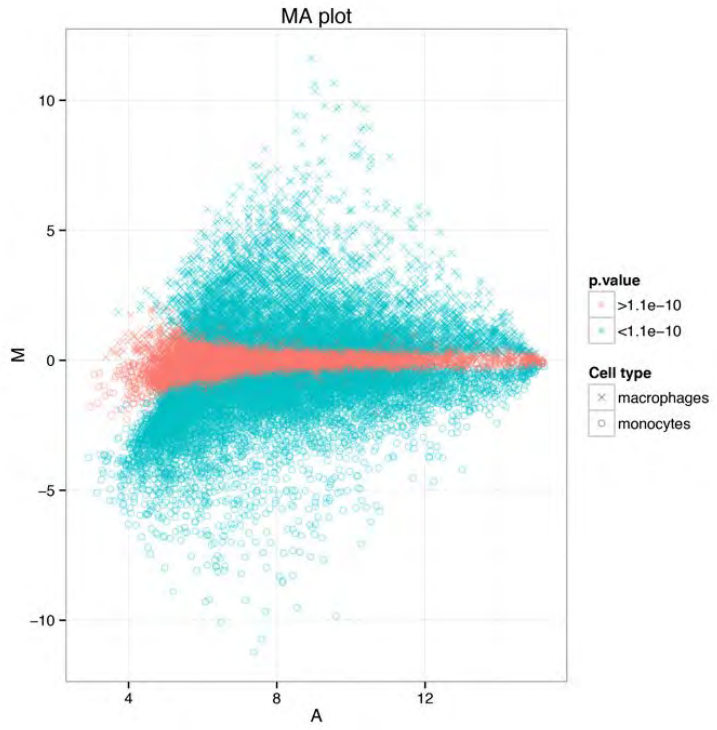


Figure 5.11: MA plot of differentially expressed genes in monocytes and macrophages. Probes that are up-regulated (cross) and down-regulated (circle) in macrophages are plotted based on average expression values across the two cell types (x-axis) and $\log_2$ fold changes (y-axis). Adjusted p-values (blue, $p<1.1x10^{-10}$, pink, $p>1.1x10^{-10}$) indicate distribution of the most significant probes.

Table 5.3: Reactome pathway enrichment analysis on the up-regulated genes in macrophages in comparison to monocytes. GeneRatio: over-represented genes in genes tested in each cell type. BgRatio: a ratio of genes in the pathway and background.

Pathway ID	Description	GeneRatio	BgRatio	p-value	FDR
1430728	Metabolism	410/4552	1004/15114	$5.14x10^{-14}$	$5.49x10^{-11}$
1428517	The citric acid (TCA) cycle and Respiratory electron transport	68/4552	119/15114	$7.53x10^{-10}$	$3.35x10^{-7}$
556833	Metabolism of lipids and lipoproteins	141/4552	344/15114	$9.54x10^{-6}$	$2.04x10^{-3}$
917937	Iron uptake and transport	20/4552	30/15114	$3.86x10^{-5}$	$5.88x10^{-3}$

Pathways enriched with the up- and down-regulated macrophage gene products

As described in Section 3.6, the Reactome pathway analysis identifies overrepresentation of genes compared to the background. Pathways consist of reactions in biological processes such as the cell cycle, mRNA processing and lipid metabolism, which contains more specific subgroups (Joshi-Tope *et al.*, 2005). The significantly enriched pathways are described below followed by the individual genes in respect to their biological roles.

Table 5.4: Reactome pathway enrichment analysis on the down-regulated macrophage genes in comparison to monocytes. GeneRatio: over-represented genes in genes tested in each cell type. BgRatio: a ratio of genes in the pathway and background.

Pathway ID	Description	GeneRatio	BgRatio	p-value	FDR
72172	mRNA Splicing	77/7206	109/15114	$9.26x10^{-7}$	$2.67x10^{-4}$
168249	Innate Immune System	184/7206	298/15114	$5.81x10^{-7}$	$2.67x10^{-4}$
1280215	Cytokine Signaling in Immune system	149/7206	239/15114	$3.07x10^{-6}$	$3.55x10^{-4}$
166054	Activated TLR4 signaling	60/7206	82/15114	$2.28x10^{-6}$	$3.14x10^{-4}$
449147	Signaling by Interluekins	69/7206	102/15114	$3.48x10^{-5}$	$1.82x10^{-4}$
168898	Toll Receptor Cascades	69/7206	99/15114	$7.2x10^{-6}$	$4.15x10^{-4}$

Pathways enriched with the macrophage differentially up-regulated genes were involved in metabolism (pathway ID 1430728, $p=5.14x10^{-14}$) and respiratory transport (pathway ID 1428517, $p=7.53x10^{-10}$, Table 5.3). The pathway of metabolism of lipids and lipoproteins was of interest as activated macrophages are highly associated with the pathologenesis of obesity in adipose tissues (Olefsky and Glass, 2010).

The macrophage differentially down-regulated expressed genes were enriched in pathways related to the innate immune system which includes TLR and cytokine signalling pathways (Table 5.4). Examples of the subgroups under Toll receptor cascades pathway (pathway ID 168898, $p=7.20x10^{-6}$) are the activated TLR4 signalling pathway (pathway ID 166054, $p=5.36x10^{-4}$) and the MyD88 dependent cascade on the plasma membrane (pathway ID 166058, $p=2.28x10^{-6}$). The down-regulated genes are, in other words, monocyte up-regulated genes. This result agrees with previously known roles of monocytes in innate immunity (Serbina *et al.*, 2008).

Genes relating to their biological roles

Genes with distinctive roles in inflammation in monocytes and macrophages may be detected by the differential gene expression analysis. Genes previously identified in association with disease susceptibility to diabetes and obesity are found to be significantly up-regulated in macrophages (Table 5.5). *PPARγ* is associated with T2D susceptibility and is involved in alternative macrophage differentiation (Altshuler *et al.*, 2000, Odegaard *et al.*, 2007). Genes that are involved in fatty acid metabolism are significantly up-regulated in macrophages; these genes are shown enriched in a lipid metabolism pathway by the Reactome analysis above ($p=5.14x10^{-14}$). *APOE*, encoding apolipoprotein E, a ligand for LDL receptors in plasma, is involved in lipid metabolism and mediates lipid

antigen presentation (Hatters *et al.*, 2006, Elzen *et al.*, 2005). Macrophages are also involved in cell apoptosis which is required for cell clearance and tissue repair (Nagata *et al.*, 2010). *CASP6* and *CASP3* encode apoptosis related protease and are essential for cell apoptosis (Thornberry and Lazebnik, 1998).

In addition to the genes involved in biological roles, genes encoding macrophage markers were determined to ensure cell maturation as macrophages. Up-regulation of *CD68* in macrophages supported maturation of cell differentiation ($p=1.66x10^{-46}$, Table 5.5). Other cell markers for macrophages were also up-regulated, which include genes encoding integrins, transmembrane receptors, and a galectin, a family of endogenous lectins. *ITGAM* and *ITGAX* encode integrinα M and X which form a heterodimeric transmembrane receptor with a β chain and are involved in signal transduction and cell adhesion in the innate immune response (Han *et al.*, 2010). *LGALS3* encodes a galactoside-binding soluble lectin also known as MAC-2 and is involved in the inflammation response (Rabinovich *et al.*, 2007). Up-regulation of these genes supported maturation of cell differentiation.

Macrophage down-regulated genes also equate to monocyte up-regulated genes. Therefore, expression of these genes are expected to highlight characteristic features of monocytes (Table 5.6). Circulating monocytes function as a part of the surveillance system for infection. PRRs including *TLRs* and *NOD2*, vital for the innate immune system to detect PAMPs, are down-regulated in macrophages.

A group of genes encoding G protein coupled receptors (GPCRs) are down-regulated in macrophages. G protein coupled receptors consist of seven transmembrane domains and are the largest protein family involved in many physiological roles in signal transmission (Marinissen and Gutkind, 2001). A GPCR subfamily of adenosine purinergic receptors was shown to increase a level of immunosuppressive cAMP in immune cells and decrease tissue damage by inflammation (Ohta and Sitkovsky, 2001). Martinez and colleagues reported that different repertoires of GPCR subfamilies were differentially expressed in different stages of macrophage differentiation (Martinez *et al.*, 2006). Agonists for GPCRs are remarkably varied and adenosine is just one example. It is difficult to classify the GPCR genes down-regulated in macrophages as they were not significantly enriched in pathways by the Reactome analysis. Further analysis of these genes may identify their associations to pathways specific to the diverse immune response.

Genes involved in the inflammatory response

Expression of cytokines and chemokines was previously reported in activated macrophages (Martinez *et al.*, 2006). The analysis presented here reveals multiple examples of differential ligand and receptor expression involved in cytokines and their receptors between macrophages and monocytes (Figure 5.12). These genes were selected based on the names of gene products and coloured onto a cytokine-cytokine receptor interaction pathway drawn by the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (ID hsa04060, Kanehisa and Goto, 2000). For example, pro-inflammatory IL1A is produced by activated macrophages as an acute cytokine and binds to IL1R1 which is down-regulated in macrophages. Another IL1 family member, IL1B, is down-regulated in macrophages, which binds to the same receptor IL1R1 (Dinarello, 2011). IL1A and TNF synergistically activate downstream pathways including transcriptional regulator NFκB for regulation of the inflammatory response (Gabay *et al.*, 2010).

Monocytes are attracted to a site of inflammation by chemokines which are secreted by infected or damaged tissues (Thelen and Stein, 2008). A chemokine ligand gene (*CCL22*) is up-regulated in macrophages while chemokine receptors such as *CCR2* and *CX3CR1* are down-regulated which agrees with a previous study of monocyte maturation (Phillips *et al.*, 2005). Various types of chemokine ligands are used to recruit different monocyte populations in response to different signals. These may explain how inflammatory macrophages secrete chemokine ligands to attract more monocytes which express receptors.

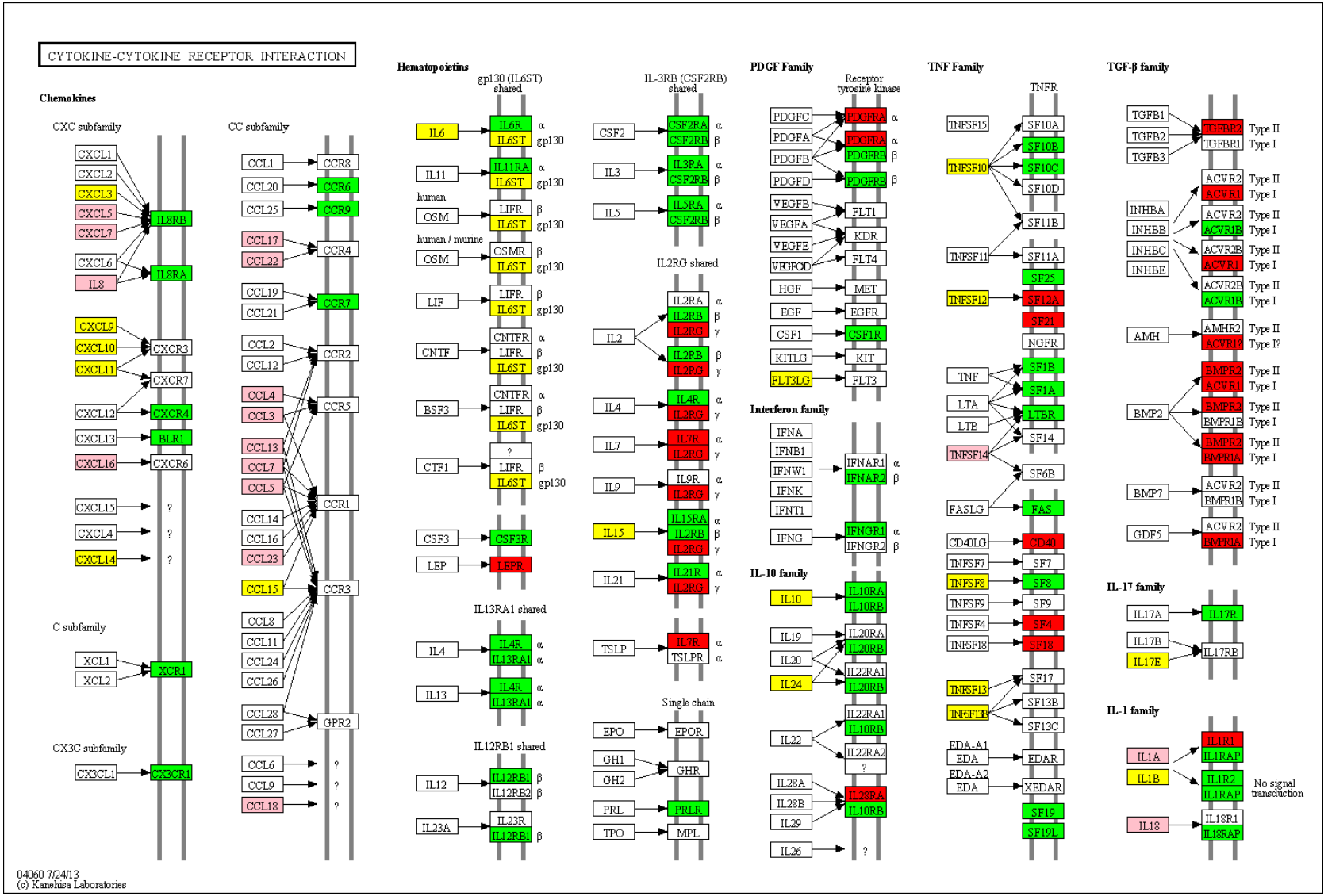


Figure 5.12: Differentially expressed genes in macrophages and monocytes in relation to a pathway for cytokines and their receptors. The differentially expressed genes involved in cytokines and their receptors are manually mapped onto a KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway for cytokine-cytokine receptor interaction (ID hsa04060, Kanehisa and Goto, 2000). Receptors up-(red) and down-(green) regulated, ligands up-(pink) and down-(yellow) regulated in macrophages are shown. KEGG pathway is a collection of manually curated pathway maps based on molecular interaction and biological reaction networks (*http://www.genome.jp/kegg/pathway.html*).

Table 5.5: Differentially up-regulated genes in macrophages. macro: macrophage, mono: monocyte, logFD: log fold changes, t: moderated t-statistics, B: B statistics (log-odds), adj.P.Val: adjusted p-values by FDR.

Macrophage up-regulated genes									Average expression	
TargetID	Gene product	RefSeq_ID	Location	logFC	t	B	P.Value	adj.P.Val	macro	mono
Inflammatory response										
CCL22	chemokine (C-C motif) ligand 22	NM_002990	16q13	10.35	97.24	151.65	$7.47\text{x}10^{-71}$	$6.34\text{x}10^{-68}$	14.21	3.86
SLC7A8	solute carrier family 7 (amino acid transporter light chain, L system), member 8	NM_182728	8	4.82	75.85	136.28	$4.68\text{x}10^{-64}$	$1.20\text{x}10^{-61}$	10.80	5.97
TREM2	triggering receptor expressed on myeloid cells 2	NM_018965	6p21.1	6.38	83.49	142.24	$1.12\text{x}10^{-66}$	$4.76\text{x}10^{-64}$	12.30	5.92
TNFRSF21	TNF receptor superfamily, member 21	NM_014452	6p21.1	5.18	66.90	128.45	$1.25\text{x}10^{-60}$	$1.99\text{x}10^{-58}$	12.22	7.04
LGALS3	lectin, galactoside-binding, soluble, 3	NM_002306	14q22.3	3.42	39.60	95.85	$1.66\text{x}10^{-46}$	$3.51\text{x}10^{-45}$	14.50	11.09
Pro-inflammatory cytokine										
IL1A	interleukin 1, α	NM_000575	2q14	2.47	14.18	37.32	$2.39\text{x}10^{-21}$	$5.22\text{x}10^{-21}$	6.93	4.46
Anti-inflammatory cytokine										
IL10	interleukin 10	NM_000572	1q31-q32	1.55	21.72	60.03	$4.17\text{x}10^{-31}$	$1.81\text{x}10^{-30}$	13.72	12.17
TLR pathway										
TLR1	toll-like receptor 1	NM_003263	4p14	1.72	14.37	38.00	$1.22\text{x}10^{-21}$	$2.76\text{x}10^{-21}$	11.37	9.65
Cell adhesion and attachment										
LIMS1	LIM and senescent cell antigen-like domains 1	NM_004987	2q12.3	5.79	11.17	26.38	$1.23\text{x}10^{-16}$	$2.16\text{x}10^{-16}$	10.84	5.06
ITGB5	integrin, β 5	NM_002213	3q21.2	5.99	96.16	150.97	$1.51\text{x}10^{-70}$	$1.24\text{x}10^{-67}$	13.41	7.42

ITGAM	integrin, α M (complement component 3 receptor 3 subunit)	NM_000632	16p11.2	1.09	26.98	72.62	$1.62x10^{-36}$	$1.14x10^{-35}$	13.56	12.47
ITGAX	integrin, α X (complement component 3 receptor 4 subunit)	NM_000887	16p11.2	0.87	10.50	23.77	$1.64x10^{-15}$	$2.73x10^{-15}$	8.19	7.31
Cell apoptosis										
CASP6	caspase 6, apoptosis-related cysteine peptidase	NM_001226	4q25	3.06	16.18	43.99	$3.24x10^{-24}$	$8.60x10^{-24}$	9.97	6.90
CASP3	caspase 3, apoptosis-related cysteine peptidase	NM_032991	4q34	0.96	17.52	48.19	$5.08x10^{-26}$	$1.52x10^{-25}$	10.82	9.86
Endocytosis										
MSR1	macrophage scavenger receptor 1	NM_138715	8p22	4.28	53.26	114.23	$1.90x10^{-54}$	$1.25x10^{-52}$	10.03	5.75
Lysosomes										
RAB7B	member RAS oncogene family	NM_177403	1q32	7.03	75.40	135.91	$6.82x10^{-64}$	$1.66x10^{-61}$	11.19	4.17
Myeloid cell differentiation										
APOC1	apolipoprotein C-I	NM_001645	19q13.2	8.97	160.73	181.75	$1.15x10^{-84}$	$8.14x10^{-81}$	14.82	5.85
CD68	CD68 molecule	NM_001251	17p13	1.13	25.80	69.99	$2.21x10^{-35}$	$1.38x10^{-34}$	14.45	13.32
ADAMDEC1	ADAM-like, decysin 1	NM_014479	8p12	6.97	109.50	158.92	$4.12x10^{-74}$	$6.74x10^{-71}$	12.63	5.65
Fatty acid metabolism										
APOE	apolipoprotein E	NM_000041	19q13.31	10.65	158.36	180.89	$2.96x10^{-84}$	$1.57x10^{-80}$	14.87	4.22
FABP3	fatty acid binding protein 3, muscle and heart (mammary-derived growth inhibitor)	NM_004102	1p33-p32	6.73	103.49	155.47	$1.47x10^{-72}$	$1.83x10^{-69}$	11.03	4.30

PPARG	peroxisome proliferator-activated receptor γ	NM_015869	3p25	4.02	47.24	106.76	$3.24x10^{-51}$	$1.32x10^{-49}$	10.68	6.65

Table 5.6: Differentially down-regulated genes in macrophages. macro: macrophage, mono: monocyte, t: moderated t-statistics, B: B statistics (log-odds), adj.P.Val: adjusted p-values by FDR.

Monocyte up-regulated genes									Average expression	
TargetID	Gene product	RefSeq_ID	Location	logFC	t	B	P.Value	adj.P.Val	macro	mono
Inflammatory response										
CCR2	chemokine (C-C motif) receptor 2	NM_000647	3p21	-6.03	-65.95	127.56	$3.06x10^{-60}$	$4.64x10^{-58}$	6.39	12.42
CX3CR1	chemokine (C-X3-C motif) receptor 1	NM_001337	3p21.3	-10.73	-54.66	115.85	$3.77x10^{-55}$	$2.79x10^{-53}$	2.22	12.95
ICAM3	intercellular adhesion molecule 3	NM_002162	19p13.3	-6.22	-43.39	101.49	$6.14x10^{-49}$	$1.79x10^{-47}$	7.42	13.65
IRF7	interferon regulatory factor 7	NM_004029	11p15.5	-3.47	-41.75	99.11	$6.50x10^{-48}$	$1.63x10^{-46}$	8.21	11.68
TNFRSF1A	TNF receptor superfamily, member 1A	NM_001065	-12p13.2	1.75	-35.32	88.72	$1.95x10^{-43}$	$2.88x10^{-42}$	10.46	12.21
PECAM1	platelet/endothelial cell adhesion molecule 1	NM_000442	17q23.3	-1.85	-35.32	88.84	$1.74x10^{-43}$	$2.58x10^{-42}$	10.73	12.58
IL6R	interleukin 6 receptor	NM_000565	1q21	4.07	-26.55	71.68	$4.15x10^{-36}$	$2.78x10^{-35}$	3.93	8.00
LTBR	lymphotoxin β receptor (TNFR superfamily, member 3)	NM_002342	12p13	-1.68	-29.81	78.57	$4.52x10^{-39}$	$4.11x10^{-38}$	10.60	12.28
NLRP12	NLR family, pyrin domain containing 12	NM_144687	19q13.42	-7.97	-17.62	48.47	$3.84x10^{-26}$	$1.16x10^{-25}$	3.73	11.70
Anti-inflammatory response										
IL10RA	interleukin 10 receptor, α	NM_001558	11q23	-2.19	-36.31	90.52	$3.27x10^{-44}$	$5.27x10^{-43}$	9.96	12.16
Lymphocyte activation										
SELL	selectin L	NM_000655	1q23	-11.23	-47.52	107.13	$2.34x10^{-51}$	$9.30x10^{-50}$	1.76	12.99
CD79B	CD79b molecule, immunoglobulin-associated β	NM_000626	17q23	-4.30	-34.05	86.60	$1.59x10^{-42}$	$2.12x10^{-41}$	4.19	8.49

MHC transactivator										
CIITA	class II, major histocompatibility complex, transactivator	NM_000246	16p13	-3.86	-24.74	67.52	$2.52x10^{-34}$	$1.42x10^{-33}$	4.47	8.33
Monocyte differentiation										
KLF4	Kruppel-like factor 4 (gut)	NM_004235	9q31	-5.01	-32.09	83.00	$5.65x10^{-41}$	$6.72x10^{-40}$	5.97	10.99
Pathogen recognition receptors										
NOD2	nucleotide-binding oligomerization domain containing 2	NM_022162	16q12	-5.62	-16.78	45.91	$4.84x10^{-25}$	$1.36x10^{-24}$	5.62	11.24
TLR pathway										
CD14	CD14 molecule	NM_000591	5q31.3	-4.66	-12.27	30.50	$2.05x10^{-18}$	$3.92x10^{-18}$	9.90	14.56
MYD88	myeloid differentiation primary response 88	NM_002468.3	3p22	-1.35	-33.88	86.29	$2.17x10^{-42}$	$2.82x10^{-41}$	10.66	12.01
TLR2	toll-like receptor 2	NM_003264	4q32	-5.52	-23.98	65.72	$1.51x10^{-33}$	$7.95x10^{-33}$	1.83	7.35
TLR4	toll-like receptor 4	NM_138554	9q33.1	-0.86	-11.18	26.39	$1.22x10^{-16}$	$2.13x10^{-16}$	8.71	9.57
TLR5	toll-like receptor 5	NM_003268	1q32.3-q42	-4.58	-34.95	88.20	$3.27x10^{-43}$	$4.74x10^{-42}$	6.62	11.19
TLR6	toll-like receptor 6	NM_006068	4p16.1	-1.53	-22.52	62.09	$5.44x10^{-32}$	$2.53x10^{-31}$	8.50	10.03
TLR7	toll-like receptor 7	NM_016562	Xp22.3	-4.97	-19.59	54.27	$1.25x10^{-28}$	$4.48x10^{-28}$	5.41	10.38
TLR8	toll-like receptor 8	NM_016610	Xp22	-3.51	-25.05	68.26	$1.22x10^{-34}$	$7.10x10^{-34}$	7.49	11.00
TLR10	toll-like receptor 10	NM_001017388	4p14	-3.69	-22.56	62.21	$4.83x10^{-32}$	$2.26x10^{-31}$	3.33	7.02

MHC genes

MHC genes on chromosome 6 (chr6) are crucial for antigen presentation and induce the T-cell mediated adaptive immune response. The genes encoded in the MHC region were analysed to see if they were differentially expressed in either cell type. All significantly differentially expressed HLA genes were down-regulated in macrophages (Table 5.7). The most significantly down-regulated gene was *HLA-F* which encodes a class I heavy chain to be a part of a heterodimer class I molecule and has few polymorphisms in its coding sequence (Geraghty *et al.*, 1990). MHC class II genes are expressed in APCs including monocytes, macrophages, DCs, B-cells and T-cells. These genes are also down-regulated in macrophages. For example, *HLA-DRB4* encodes a β4 subunit for a MHC class II HLA-DR cell surface receptor.

MHC genes are known for hyperpolymorphism critical for effective antigen recognition and immunity. Microarray probes are difficult to design without SNPs. Although this analysis shows differentially expressed genes between macrophages and monocytes involving MHC genes, these expression values may require further validation.

Table 5.7: Significantly differentially down-regulated *HLA* genes in macrophages in comparison to monocytes. macro; macrophages, mono; monocytes, logFD: log fold changes, t: moderated t-statistics, B: B statistics (log-odds), adj.P.Val: adjusted p-values by FDR.

							Average expression	
Name	RefSeq_ID	logFC	t	B	P.Value	adj.P.Val	macro	mono
class I								
HLA-F	NM_001098479	-1.78	-36.23	90.40	$3.70x10^{-44}$	$5.93x10^{-43}$	10.51	12.29
HLA-E	NM_005516	-1.09	-31.14	81.19	$3.38x10^{-40}$	$3.47x10^{-39}$	13.27	14.36
HLA-G	NM_002127	1.65	25.37	69.01	$5.82x10^{-35}$	$3.49x10^{-34}$	12.47	10.81
HLA-C	NM_002117	-1.00	-11.05	25.91	$1.95x10^{-16}$	$3.39x10^{-16}$	10.21	11.21
HLA-H	NR_001434	0.50	8.54	15.99	$3.78x10^{-12}$	$5.56x10^{-12}$	13.01	12.51
class II								
HLA-DOB	NM_002120	-3.45	-24.99	68.11	$1.42x10^{-34}$	$8.17x10^{-34}$	3.03	6.48
HLA-DRB4	NM_021983	-1.58	-21.82	60.30	$3.21x10^{-31}$	$1.41x10^{-30}$	12.73	14.31
HLA-DPB1	NM_002121	-1.84	-16.32	44.45	$2.04x10^{-24}$	$5.50x10^{-24}$	8.38	10.22
HLA-DRA	NM_019111	-0.90	-15.43	41.56	$3.60x10^{-23}$	$8.98x10^{-23}$	14.01	14.91
HLA-DOA	NM_002119	-3.04	-13.24	34.04	$6.17x10^{-20}$	$1.27x10^{-19}$	6.13	9.17
HLA-DPA1	NM_033554	-0.73	-12.61	31.75	$5.95x10^{-19}$	$1.17x10^{-18}$	13.82	14.55
HLA-DRB6	NR_001298	-0.73	-12.18	30.16	$2.88x10^{-18}$	$5.46x10^{-18}$	12.93	13.66
HLA-DQA1	XM_936128	-1.19	-11.51	27.65	$3.49x10^{-17}$	$6.27x10^{-17}$	11.47	12.66
HLA-DQB1	NM_002123	-1.97	-9.71	20.69	$3.52x10^{-14}$	$5.55x10^{-14}$	7.85	9.83

5.6 eQTL mapping analysis

Mapping eQTL using the expression data of two closely related immune cell types may identify specific eQTL that are highly associated with each cell type, especially eQTL that are important for macrophage activation. This section describes eQTL mapping of human primary monocytes and macrophages in related to cell type specificity.

Gene expression data often includes experimental noise such as batch effect even after data normalisation. Therefore mathematical methods are used to filter these artefact effects by decomposing microarray data into different dimensions as components. Such methods include integrating multivariate dimension reduction techniques, a singular value decomposition (SVD) and an independent component analysis (ICA). Components representing noise are used to normalise data. The step is considered to improve interpretation of microarray data and helps to identify patterns of gene expression in relation to biological function (Alter *et al.*, 2000). Briefly, SVD is a linear transformation of expression data from genes on an array and the sample number of arrays into unique reduced dimensional data which are sorted by the rank of variance explained (Alter *et al.*, 2000). ICA is a method to decompose expression data into statistically independent modes (Hyvarinen and Oja, 2000).

These methods can be applied to gene expression data for eQTL analysis to filter experimental noise (Biswas *et al.*, 2008, Fehrmann *et al.*, 2011, Leek and Storey, 2007). In order to enhance discovery of eQTL associations, Biswas and colleagues used both SVD and ICA on gene expression data for eQTL analysis because two methods identify different source of variation in gene expression data (Biswas *et al.*, 2008). Principal component analysis (PCA) is a mathematical procedure to decompose a set of observations into a set of values of linearly uncorrelated variables called principal components (PCs), which is almost equivalent or often identical to SVD. Each PC explains a unique variable.

The eQTL analysis in this section used the R package MatrixeQTL that incorporates a linear regression model with PCs of gene expression data as covariates in order to maximise the number of eQTL detected in an analysis (Shabalin, 2012). As described above, PCs explain variations that are derived from experimental noise, environmental effect, individual gene expression variation or genetic effect. Studies applying PCA to their gene expression data argue that top ranked PCs explaining a larger part of the total variation explain sample differences that may be caused by environmental variation as well as experimental variation (Biswas *et al.*, 2008, Fehrmann *et al.*, 2011). Therefore these PCs derived from noise can be filtered from gene expression data for eQTL analysis and

the remaining PCs explain genetic effects in the data. The advantages of incorporating PCs are to reduce noise that is irrelevant from true variation of gene expression derived from genetic effect and to detect more eQTL signals (Ringner, 2008).

5.6.1 Data processing

The final gene expression and genotype datasets that passed quality control contain both autosomal and sex chromosomes. For this eQTL analysis, only autosomal chromosomes were included. Further filtering was thus performed on the datasets.

Probe filtering

The expression probes were further filtered based on the presence of SNPs in the probe sequences. Using a method described in Section 4.2, 7116 out of 21229 probes were identified with SNPs in the probe sequences. A further 496 probes found on sex chromosomes were also excluded. Thus 13617 probes remained for this eQTL analysis.

Genotyping filtering

A small MAF in a small population may cause detection of false positive results when the variance of gene expression is particularly large. Thus, MAF>10% was applied to the genotyping data. After applying the MAF cutoff >10%, removal of outliers of the expression data and SNPs on sex chromosomes, the genotyping data had 503391 SNPs and 62 samples for eQTL analysis.

eQTL analysis with PCA

The eQTL analysis using the gene expression and genotyping data of the 62 samples was performed separately for monocytes and macrophages. For local cis eQTL analysis, shorter distances than the previous eQTL analysis, 250kbp, 500kbp and 1Mbp, were assessed for their effect on the rate of eQTL discovery. This is because this analysis intended to focus on detection of eQTL that may be changed during macrophage differentiation, and a smaller window encompasses promoter and enhancer elements. Trans eQTL for this analysis was not considered due to the small sample size.

The R package MatrixeQTL performs PCA on the gene expression data and incorporates a various number of PCs for eQTL analysis. The numbers of eQTL using different numbers of PCs vary because the incorporation of PCs will filter variance explained by the PCs out of gene expression. Therefore, when all PCs are used no eQTL is detected because PCs

that account for variability of gene expression caused by the genetic effect are included. I decided to use different numbers of PCs for each expression data. Each gene expression dataset was independently measured and contain different causes of variability. The macrophages underwent an incubation step which may be the largest source of variance in the sample collection step and was not underwent in the monocytes.

Thus, the eQTL analysis was performed using up to 50 PCs in a stepwise manner in order to chose an appropriate number of PCs to incorporate. FDR<0.05 was used to cutoff the eQTL associations. The analysis using PCs detected various numbers of eQTL in different cis distances at FDR<0.05 (Figure 5.13). The number of PCs that maximised the detection of eQTL varied depending on the cis distances (Table 5.8). This is partially explained by the differing number of tests performed in different cis distances and because FDR was calculated for each set of tests. Selecting the number of PCs used for eQTL analysis can be decided by the number of unique probes whereby the more genes that are detected with eQTL the more successful the PCs that are used as covariates. Compared to the eQTL analysis without PCs, the results with PCs markedly increased the number of eQTL detected (2.6 and 1.7 times as many in macrophages and monocytes respectively). For further analysis, the results with cis distance 250kbp were used with 14 PCs for macrophages and 10 PCs for monocytes, based on these leading to the maximised number of probes with associations observed.

cis	PC	Total	eQTL	Probes	SNPs
Macrophage					
250kb	14	15911	12687	2084	10784
500Kb	13	18451	10795	1792	9207
1Mb	13	21533	8798	1539	7567
Monocyte					
250kb	10	21014	19383	2732	15893
500Kb	9	24007	17267	2443	14268
1Mb	9	27394	14298	2101	11939

Table 5.8: The cis eQTL analysis results with PCs which maximised the number of eQTL detected. Total; total eQTL with p-value threshold<$1x10^{-3}$ for cis. eQTL; eQTL under FDR<0.05. Probes and SNPs; unique number of probes and SNPs under FDR<0.05.

5.6.2 Cell type specific eQTL

Although monocytes and macrophages share the same cell lineage, some eQTL were identified as cell type specific. The two eQTL mapping results were analysed to determine eQTL which were specific to either cell type. Unique eSNPs (SNP-probe associations) were

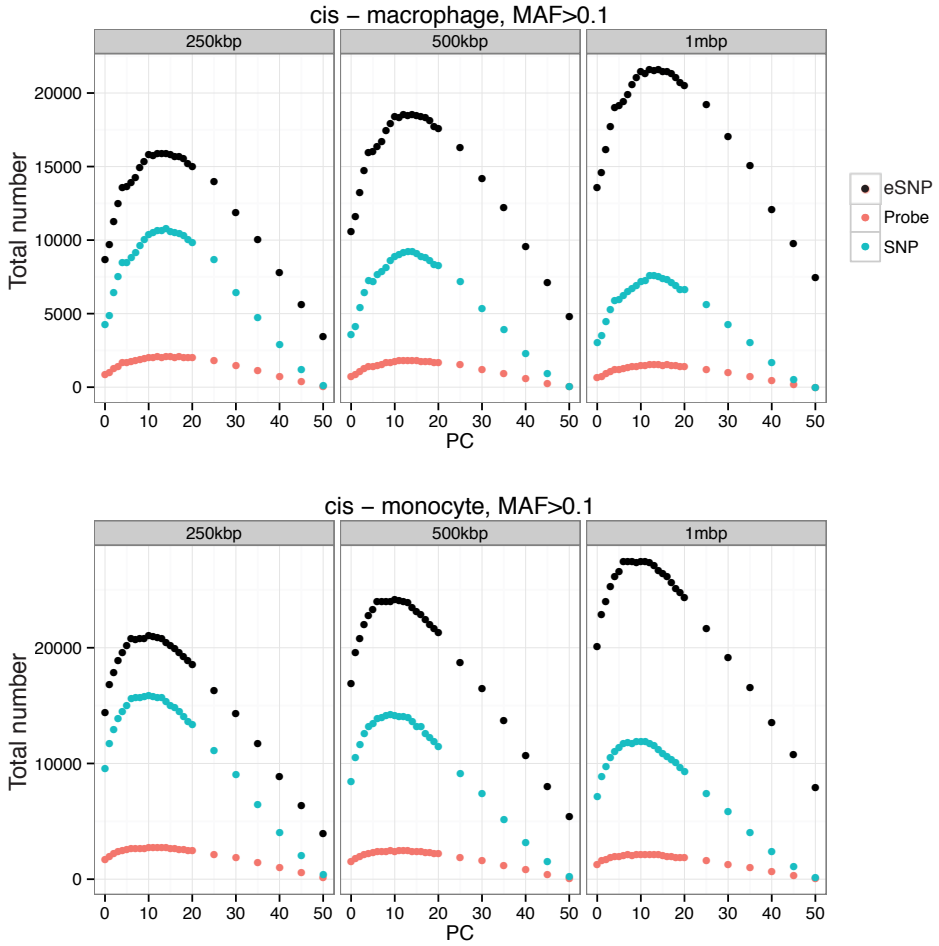


Figure 5.13: eQTL analysis using PCs of gene expression data as covariates in macrophages (top panels) and monocytes (bottom panels) in cis. All eSNPs (black), the number of unique probes (pink) and SNPs (blue) are selected with FDR<0.05.

compared between monocytes and macrophages for investigating novel eSNP associations especially in macrophages.

Overview of eSNP associations between monocytes and macrophages

In this cis association analysis, 7655 (23.87%) and 14351 (44.75%) eSNPs were unique to macrophages and monocytes respectively and 11064 (31.38%) eSNPs were identified as shared (Figure 5.14). 68.6% of the total cis eQTL were cell type specific. Genes that were found to be expressed at a similar level (normalised fold change difference within $\log_2$+/-0.5; 29.2% of the total probes tested for the eQTL analysis) were also analysed. 19.25% of the total eSNPs in cis were found to be cell type specific when the expression levels between the two cell types were similar (Table 5.9). This proportion was far smaller than in the B-cell and monocyte dataset presented in Chapter 4 which showed 85.2% of eSNPs within $\log_2$+/-0.5 gene expression difference between the two cell types.

Probes involved in eQTL in cis were classified by their differential gene expression states in macrophages using the results of the previous section (Table 5.10). All proportions of probes, classified as up- and down-regulated or not differentially expressed, showing associations with eSNPs were similar ranging between 26-34%.

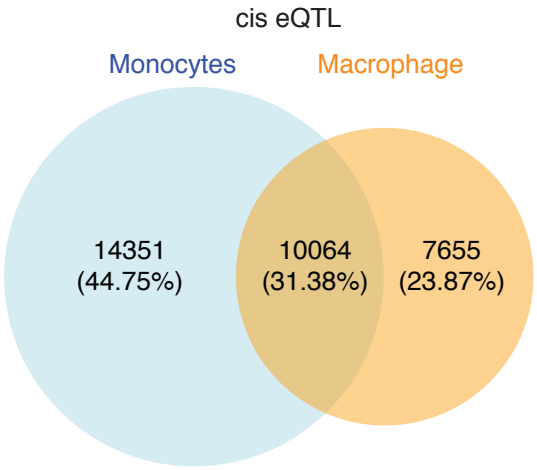


Figure 5.14: Venn diagram of the cell type specific cis eSNPs in macrophages and monocytes. cis distance is defined at 250kbp and FDR cutoff<0.05

Table 5.9: cis eSNPs identified with genes expressed in a similar level between macrophages and monocytes. Percentages show a proportion of the total eSNPs for each cell type or shared between two cell types. The number of eSNPs with genes whose average gene expression values are within +/-0.5 and 1 are shown.

	Macrophages (%)	Monocytes (%)	shared (%)	total (%)
<0.5	2295 (29.98)	3880 (27.04)	3188 (31.68)	9363 (29.20)
<1	3972 (51.89)	6732 (46.91)	5684 (56.48)	16388 (51.10)

Table 5.10: Differentially expressed probes involved in eQTL. The total number of eQTL associated probes were classified by the macrophage differential gene expression analysis.

	Total	Up.regulated	Down.regulated	Not.significant
Macrophage	1745 (100)	595 (34.10)	579 (33.18)	571 (32.72)
Monocyte	2543 (100)	668 (26.27)	1199 (47.15)	676 (26.58)
Shared	742 (100)	239 (32.21)	291 (39.22)	212 (28.57)

Distance of cis eQTL from transcription start sites

The eQTL analysis in this dataset defined a +/-250kbp margin from the probe start sites as cis association. The distance between SNPs and the transcription start sites (TSSs) within +/-250kbp was considered to determine the relationship of SNPs and TSS for the cell type specific eQTL. Information for the TSS was obtained from the Homo sapiens gene dataset (GRCh37.p11) in the Ensembl 72 Gene Database using the R package biomaRt (*http://www.biomart.org/*). The Wilcoxon rank sum tests on the distances from TSS to eQTL

showed that the distribution of the cell type specific eQTL was significantly different from the shared eQTL (p_{macro}=1.93x10^{-10} and p_{mono}=4.98x10^{-9} to shared respectively). By contrast, there was no significant difference between macrophages and monocytes specific eQTL (p=0.16). Plots based on the distance from TSS to eQTL showed that the cell type specific eQTL distally located away from the TSS (Figure 5.15).

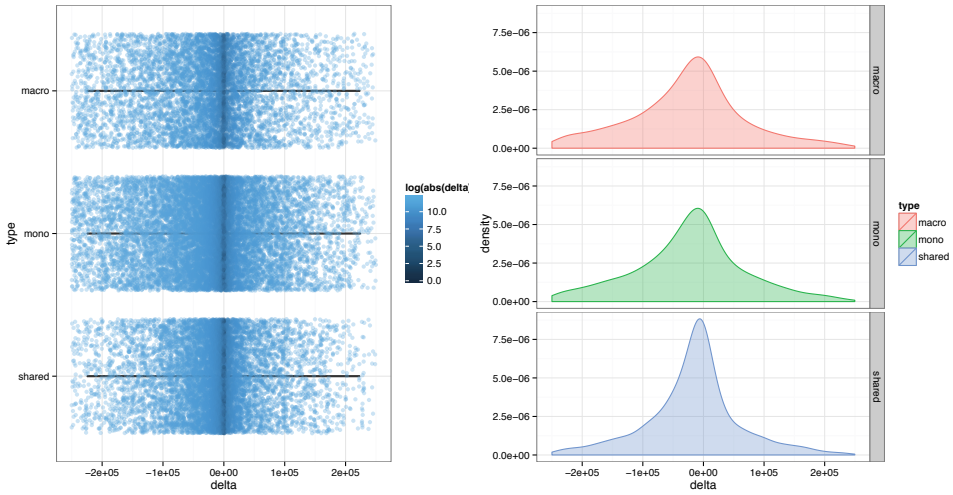


Figure 5.15: Distance from TSS to eQTL in monocytes, macrophages and shared datasets. The eQTL location is shown within +/-250kbp from TSS in box plots (A) and density plots (B). macro; macrophages, mono; monocytes, shared; shared between the two cell types.

5.6.3 Examples of cis eQTL in the two cell types

Representative examples of cell type specific eQTL are described in this section, which were identified throughout the genome (Figure 5.16). The proportion of the cell type specific eQTL is smaller in this dataset (68.6%) than the eQTL analysis on B-cells and monocytes in the previous chapter (78%, Chapter 4). The products of genes significantly associating with eSNPs include many genes that are involved in immune responses in both macrophages and monocytes.

Macrophage specific cis eQTL

The number of macrophage specific cis eQTL is less than monocytes (23.87% and 44.75%). The eSNPs specific to macrophages are shown in Table 5.11. The genes significantly associated with SNPs in cis are involved in the immune response and fatty acid metabolism. rs4555370 on chromosome 2 (chr2) is associated with *STAT4*, which encodes a TF essential for signal transduction of immune responses including IFN-γ production and Th1 activation in myeloid cells (p_{mac}=1.74x10^{-7}, Yamamoto *et al.*, 1994, Fukao

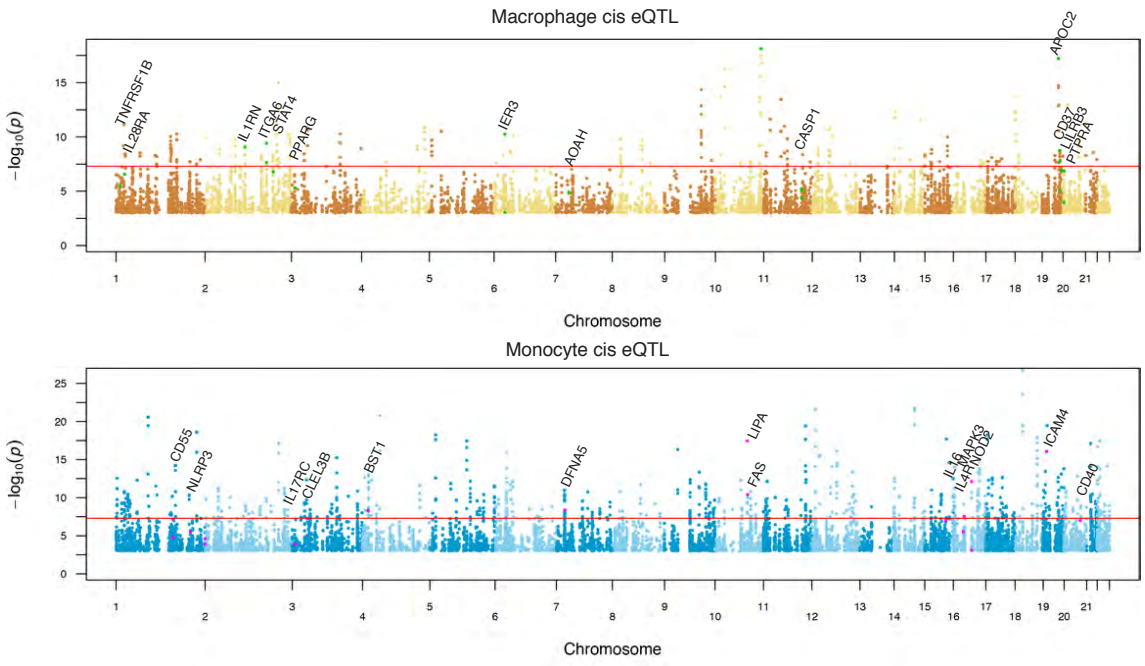


Figure 5.16: Manhattan plots of macrophage (top) and monocyte (bottom) specific cis eQTL. Examples of eQTL association are shown in green for macrophages (top) and magenta for monocytes (bottom). Red line; genome-wide p-value cutoff $5x10^{-8}$.

et al., 2001). Individuals with the minor homozygous allele expressed significantly less *STAT4* in macrophages (Figure 5.17A). *LILRB3* expression is significantly associated with rs11672654 on chr19 (p_{mac}=$1.19x10^{-7}$, Figure 5.17). The gene expression is further down-regulated in the minor allele homozygotes. *LILRB3* gene product, a member of the leukocyte immunoglobulin-like receptor family, binds to MHC class I molecules in APCs and is involved in inhibitory roles of the inflammatory response (Brown *et al.*, 2004). rs3807605 on chr7 is associated in cis with *AOAH* expression in macrophages (p_{mac}=$1.29x10^{-5}$, Table 5.11). In the previous chapter, the locus on chr6 showed a trans association to *AOAH* in monocytes (Section 4.5.4).

Interestingly, two genes involved in lipid metabolism showed cis eQTL uniquely in macrophages. *APOC2* encoding apolipoprotein C-II, a cofactor for lipoprotein lipase, in macrophages is greatly up-regulated with rs20495 alleles in cis (p_{mac}=$5.95x10^{-18}$, Jong *et al.*, 1999). The gene locus on chr19 is identified to associate with cholesterol levels and risk of CHD (Lusis *et al.*, 1986, Willer *et al.*, 2008). Macrophage specific cis eQTL on *PPARG* showed an association with rs17036160 on chr3 (p_{mac}=$5x10^{-6}$, Figure 5.17), although the MAF of the SNP is low and this dataset had only one minor homozygotes (CEU MAF 0.097). *PPARG* encodes a peroxisome proliferator-activated receptor γ which is known as a negative regulator of macrophage activation (Ricote *et al.*, 1998). Decreased expression of *PPARG* is involved in insulin resistance in muscle and adipose tissue

where tissue resident macrophages normally control homeostasis (Hevener *et al.*, 2003, Odegaard *et al.*, 2007). A recent study has shown that *KLF9* regulates activation of *PPARγ2* at the middle stage of adipogenesis (Pei *et al.*, 2011). Cis eQTL association with genes involved in lipid metabolism suggests that a genetic component is involved in disease susceptibility to obesity including diabetes. This is further investigated with loci in GWAS in the following section.

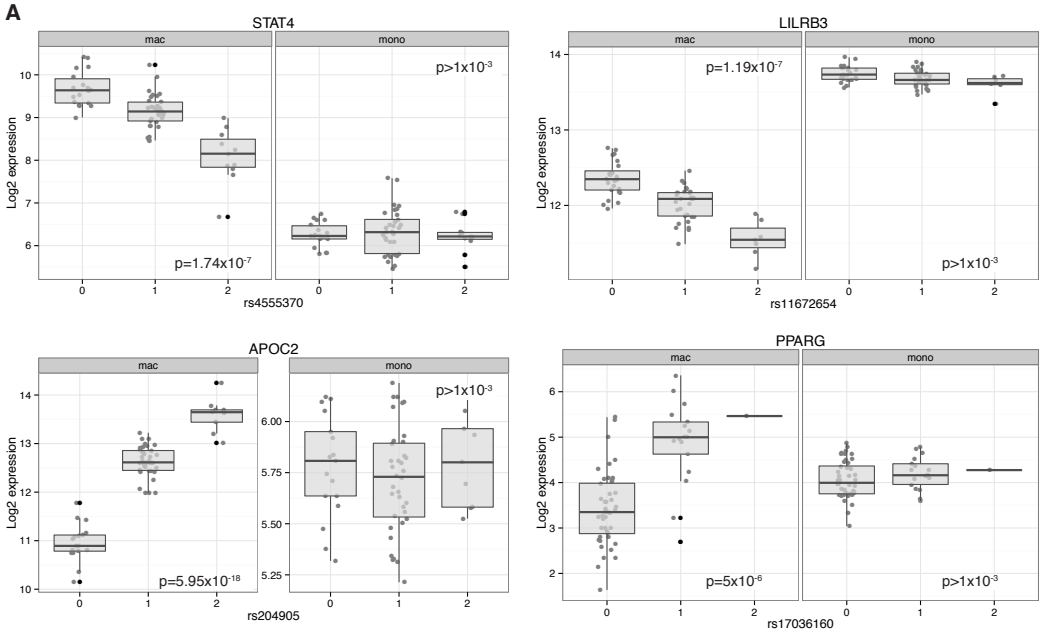


Figure 5.17

Monocyte specific cis eQTL

The eSNPs specific to monocytes are associated with genes that are involved with the immune response such as inflammatory cytokines (Table 5.12). The monocyte specific cis eQTL include eSNPs that were found in the previous chapter (Chapter 4). For example, *ICAM4* with rs281437 (p_{mono}=8.70x10^{-17}) and *CD40* in association with rs745307 (p_{mono}=2.43x10^{-13}) were replicated in this monocyte dataset (Table 5.12), although the degrees of significance are smaller than the previous dataset due to the smaller sample size. Genes involved in inflammatory disease and lipid metabolism were associated with monocyte specific eQTL in cis. *NOD2* (also known as *CARD15*) expression in monocytes was associated with rs1981760 on chr16 (p_{mono}=7.83x10^{-13}, Figure 5.18). The gene product serves as a pathogen recognition receptor (PRR) to recognise bacterial peptidoglycans and activates NFκB signal transduction (Ogura *et al.*, 2001b, Kobayashi *et al.*, 2005). The gene mutations are associated with Crohn's disease susceptibility since the failure of recognition of

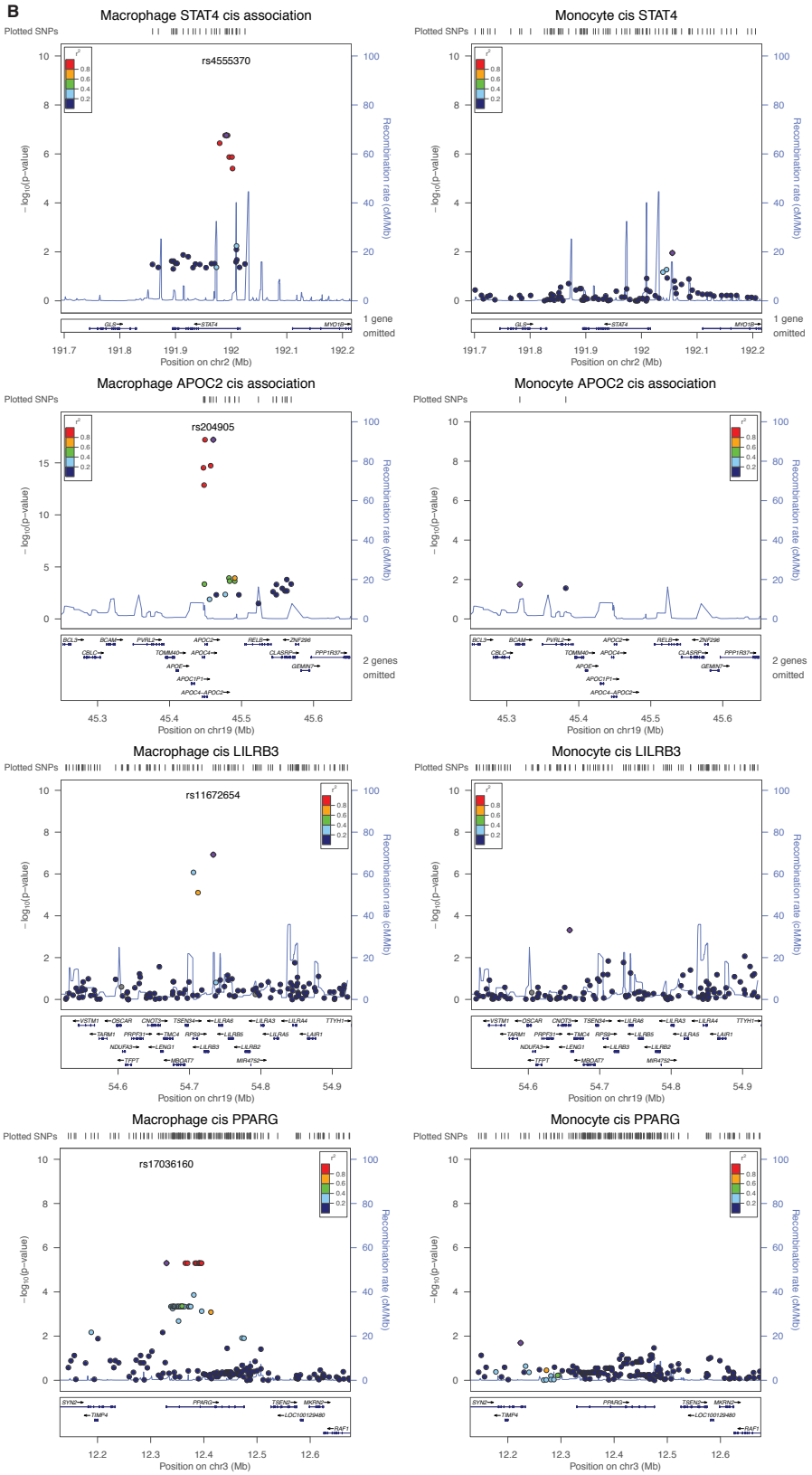


Figure 5.17: Examples of macrophage specific eQTL in comparison to monocytes. Box (A) and association plots (B) of example macrophage specific cis eQTL. (A) The number of minor alleles is shown in the box plots (0 - major homozygous, 1 - heterozygous, 2 - minor homozygous) for macrophages (mac) and monocytes (mono).

bacterial components is associated with disease development (Hugot *et al.*, 2001, Ogura *et al.*, 2001a). rs1412444 on chr10 is associated with *LIPA* expression (p_{mono}=4.23x10^{-11}, Figure 5.18). *LIPA* encodes a lysosomal lipase A to break down cholesteryl esters and triglycerides, gene mutations in *LIPA* cause cholesteryl ester storage disease (Goldstein *et al.*, 1975).

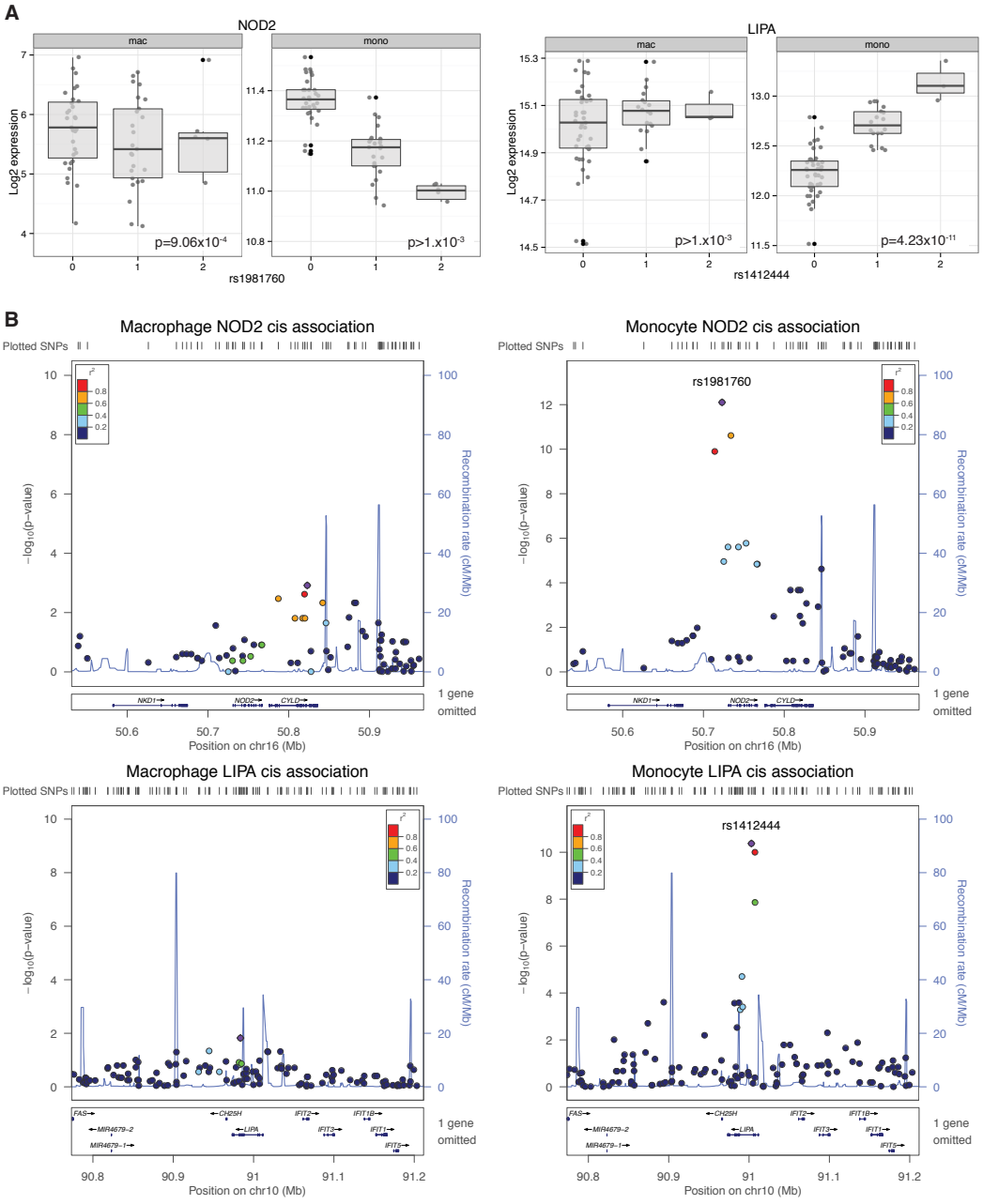


Figure 5.18: Examples of monocyte specific cis eQTL in a comparison to macrophages. (A) Box plots of example gene expression levels associating with SNPs in three genotypes. (B) Association plots. (A) The number of minor alleles is shown in the box plots (0 - major homozygous, 1 - heterozygous, 2 - minor homozygous) for macrophages (mac) and monocytes (mono).

Shared cis eQTL

Shared cis eQTL showed associations with genes that are important for the immune response (Table 5.13, Figure 5.19). rs10239340 on chr7 is associated with *IRF5* in cis in both monocytes and macrophages (p_{mono}=5.99x10^{-25}, p_{mac}=2.56x10^{-21}, Figure 5.20). The eQTL association is maintained in macrophages, even though the expression level of *IRF5* is markedly decreased. *IRF5* encodes an interferon regulatory factor that has versatile roles in the immune system including virus mediated activation of interferon and the higher expression levels promote polarisation of inflammatory macrophages (Krausgruber *et al.*, 2011). The cis association of *LYZ*, encoding lysosome enzyme, and rs1384, in almost perfect LD to rs10784774, was found in both monocytes and macrophages (r^2 0.967, p_{mono}=3.11x10^{-19}, p_{mac}=8.50x10^{-10}, Figure 5.20). The *LYZ* expression level was not significantly different between two cell types and the effect size of the eSNP in macrophages is smaller (macrophages t-statistics 7.69 compared to monocytes 14.25). Interestingly, the most significant eSNP for macrophage *LYZ* expression is rs589448 which is 4.4kb away from the gene and not in strong LD with rs1384 (r^2 0.45, p_{mono}=1.30x10^{-13}, p_{mac}=9.01x10^{-13}). *ERAP2*, encoding endoplasmic reticulum aminopeptidase 2, is involved in trimming HLA class I peptides (Saveanu *et al.*, 2005). *ERAP2* is associated with rs4360063 on chr5 (p_{mono}=2.27x10^{-19}, p_{mac}=5.34x10^{-16}, Figure 5.20). This association was identified in both B-cells and monocytes in the previous chapter and the SNP is associated with susceptibility to Crohn's disease (Figure 4.19, Franke *et al.*, 2010). rs4559 in chr12 is associated with *STAT6* encoding a TF and signal transducer in the STAT family in both cell types. *STAT6* is required for IL4 mediated signalling and development of Th2 cells (p_{mono}=5.21x10^{-21}, p_{mac}=1.43x10^{-13}, Takeda *et al.*, 1996). Except *LYZ* association, the three examples were also shared with B-cells as described in the previous chapter (Chapter 4).

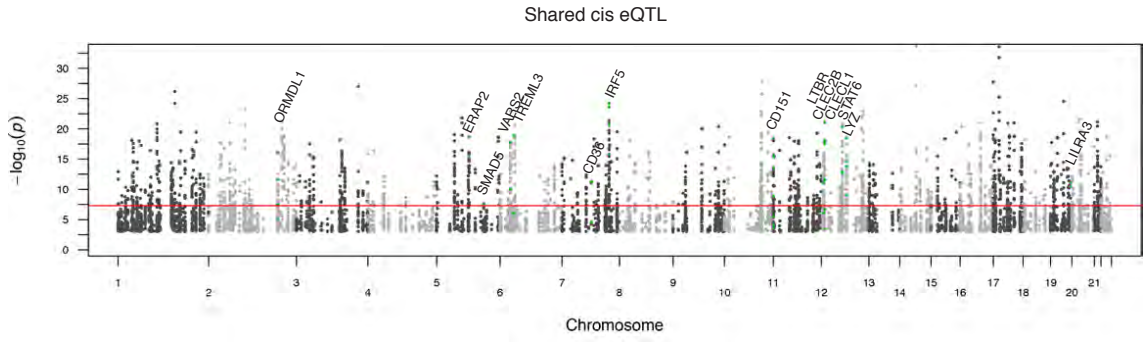


Figure 5.19: Manhattan plot of shared cis eQTL between monocytes and macrophages. The examples of eQTL association are shown in green. Red line; genome-wide p-value 5x10^{-8}.

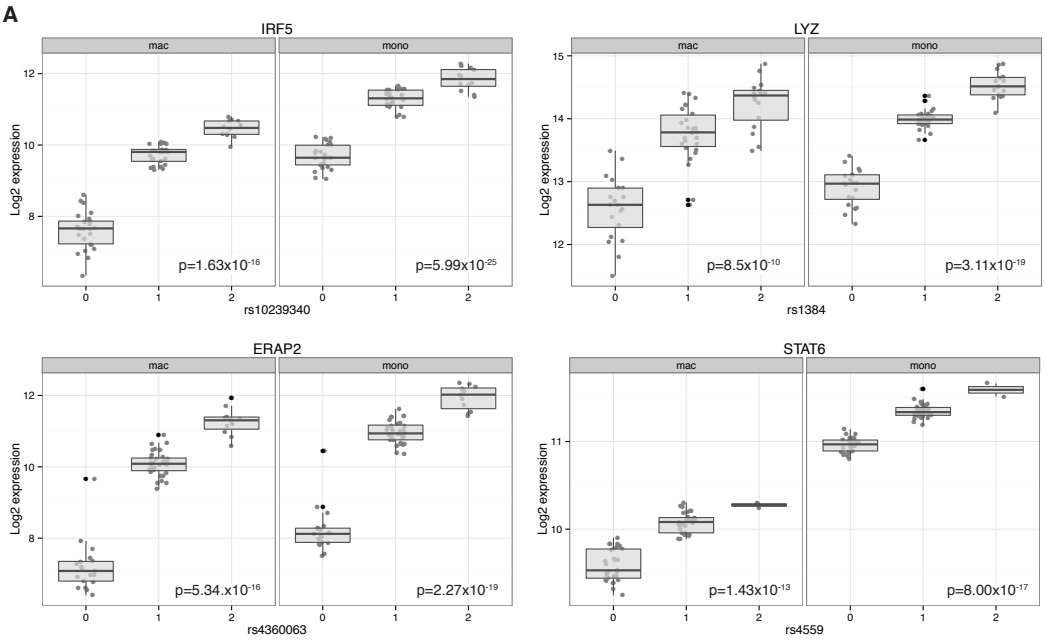

A
IRF5
mac
mono
Log2 expression
p=1.63x10-16
p=5.99x10-25
rs10239340
LYZ
p=8.5x10-10
p=3.11x10-19
rs1384
ERAP2
p=5.34.x10-16
p=2.27x10-19
rs4360063
STAT6
p=1.43x10-13
p=8.00x10-17
rs4559

Figure 5.20

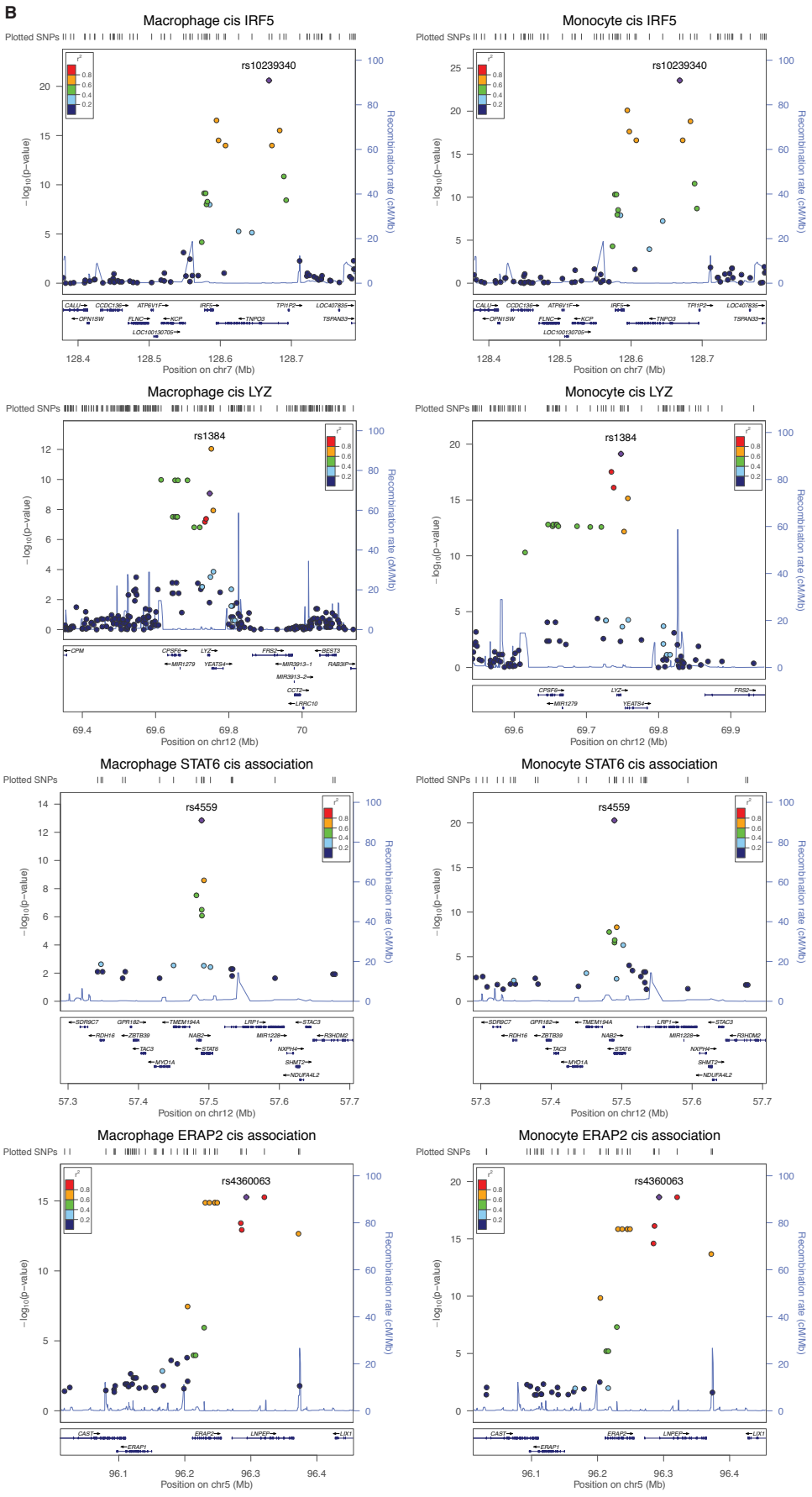


Figure 5.20: Shared cis eQTL between monocytes and macrophages. Box (A) and association plots (B) of example shared cis eQTL. (A) The number of minor alleles is shown in the box plots (0 - major homozygous, 1 - heterozygous, 2 - minor homozygous) for macrophages (mac) and monocytes (mono).

Table 5.11: Macrophage unique cis eQTL in comparison to monocytes. FDR cutoff 0.05. macro: macrophages, mono: monocytes.

Macrophage specific cis eQTL									Average expression	
SNP	SNP position	Gene	Gene product	REFSEQ ID	Location	t.stat	p.value	FDR	macro	mono
rs204905	chr19: 45460764	*APOC2*	apolipoprotein C-II	NM_000483	19q13.2	13.79	$5.95\text{x}10^{-18}$	$1.01\text{x}10^{-13}$	12.31	5.75
rs4711228	chr6: 30756274	*IER3*	immediate early response 3	NM_003897	6p21.3	-8.49	$5.57\text{x}10^{-11}$	$1.07\text{x}10^{-7}$	9.28	11.63
rs1056222	chr8: 24263345	*ADAMDEC1*	ADAM-like, decysin 1	NM_014479	8p12	-8.15	$1.75\text{x}10^{-10}$	$2.81\text{x}10^{-7}$	12.63	5.65
rs1076597	chr2: 173327273	*ITGA6*	integrin, α 6	NM_000210	2q31.1	7.93	$3.75\text{x}10^{-10}$	$5.40\text{x}10^{-7}$	10.00	4.47
rs2234677	chr2: 113875509	*IL1RN*	interleukin 1 receptor antagonist	NM_173842	2q14.2	7.69	$8.40\text{x}10^{-10}$	$1.13\text{x}10^{-6}$	9.92	9.42
rs354021	chr19: 49839015	*CD37*	CD37 molecule	NM_001774	19q13	-6.79	$1.85\text{x}10^{-8}$	$1.63\text{x}10^{-5}$	8.86	11.92
rs11672654	chr19: 54733333	*LILRB3*	leukocyte immunoglobulin-like receptor, subfamily B (with TM and ITIM domains), member 3	NM_006864	19q13.4	-6.26	$1.19\text{x}10^{-7}$	$7.91\text{x}10^{-5}$	12.10	13.70
rs677344	chr20: 2846289	*PTPRA*	protein tyrosine phosphatase, receptor type, A	NM_080841	20p13	6.22	$1.36\text{x}10^{-7}$	$8.83\text{x}10^{-5}$	10.59	10.34
rs4555370	chr2: 191991459	*STAT4*	signal transducer and activator of transcription 4	NM_003151	2q32	-6.15	$1.74\text{x}10^{-7}$	$1.08\text{x}10^{-4}$	9.12	6.29
rs7552167	chr1: 24518643	*IL28RA*	of IFNLR1 interferon, lambda receptor 1	NM_173064	1p36.11	-6.00	$2.85\text{x}10^{-7}$	$1.61\text{x}10^{-4}$	6.62	5.04
rs641941	chr1: 12206652	*TNFRSF1B*	tumor necrosis factor receptor superfamily, member 1B	NM_001066	1p36.22	-5.28	$3.38\text{x}10^{-6}$	$1.17\text{x}10^{-3}$	12.30	13.78
rs17036160	chr3: 12329783	*PPARG*	peroxisome proliferator-activated receptor γ	NM_015869	3p25	5.17	$5.00\text{x}10^{-6}$	$1.59\text{x}10^{-3}$	3.86	4.11
rs1792753	chr11: 104930676	*CASP1*	caspase 1, apoptosis-related cysteine peptidase	NM_033294	11q23	5.02	$8.35\text{x}10^{-6}$	$2.42\text{x}10^{-3}$	10.43	12.93

rs3807605	chr7: 36764123	*AOAH*	acyloxyacyl hydrolase (neutrophil)	NM_001637	7p14	-4.89	$1.29x10^{-5}$	$3.36x10^{-3}$	9.57	12.02

Table 5.12: Monocyte unique cis eQTL in comparison to macrophage. FDR cutoff 0.05. macro: macrophages, mono: monocytes.

Monocyte specific cis eQTL									Average expression	
SNP	SNP position	Gene	Gene product	REFSEQ ID	Location	t.stat	p.value	FDR	macro	mono
rs281437	chr19: 10397238	*ICAM4*	intercellular adhesion molecule 4 (Landsteiner-Wiener blood group)	NM_001544	19p13.2	12.34	$8.70x10^{-17}$	$2.93x10^{-13}$	5.47	7.19
rs745307	chr20: 44747086	*CD40*	CD40 molecule, TNF receptor superfamily member 5	NM_001250	20q12	9.88	$2.43x10^{-13}$	$3.52x10^{-10}$	9.15	7.98
rs1981760	chr16: 50723074	*NOD2*	nucleotide-binding oligomerization domain containing 2	NM_022162	16q12	-9.53	$7.84x10^{-13}$	$1.03x10^{-9}$	5.62	11.24
rs1412444	chr10: 91002927	*LIPA*	lipase A, lysosomal acid, cholesterol esterase	NM_000235	10q23.2	8.39	$4.23x10^{-11}$	$3.82x10^{-8}$	15.03	12.41
rs2255365	chr7: 24561661	*DFNA5*	deafness, autosomal dominant 5	NM_004403	7p15	7.09	$4.27x10^{-9}$	$2.33x10^{-6}$	12.39	7.39
rs3213710	chr4: 15717321	*BST1*	bone marrow stromal cell antigen 1	NM_004334	4p15	7.05	$4.97x10^{-9}$	$2.66x10^{-6}$	5.20	10.71
rs11150581	chr16: 30030699	*MAPK3*	mitogen-activated protein kinase 3	NM_001040056	16p11	-6.54	$3.10x10^{-8}$	$1.30x10^{-5}$	9.51	10.82
rs13963	chr3: 45077123	*CLEC3B*	C-type lectin domain family 3, member B	NM_003278	3p22	6.38	$5.61x10^{-8}$	$2.19x10^{-5}$	6.06	8.28
rs1800682	chr10: 90749963	*FAS*	Fas cell surface death receptor	NM_152877	10q24.1	-6.37	$5.84x10^{-8}$	$2.28x10^{-5}$	8.40	8.75
rs17875563	chr15: 81604293	*IL16*	interleukin 16	NM_172217	15q26.3	6.21	$1.03x10^{-7}$	$3.70x10^{-5}$	6.47	9.03
rs2057768	chr16: 27322095	*IL4R*	interleukin 4 receptor	NM_000418	16p12	5.24	$3.23x10^{-6}$	$6.98x10^{-4}$	9.50	10.37
rs1652333	chr1: 207470460	*CD55*	CD55 molecule, decay accelerating factor for complement (Cromer blood group)	NM_000574	1q32	-5.24	$3.23x10^{-6}$	$6.98x10^{-4}$	6.50	10.02
rs4925546	chr1: 247602968	*NLRP3*	NLR family, pyrin domain containing 3	NM_001079821	1q44	-4.65	$2.42x10^{-5}$	$3.55x10^{-3}$	7.10	12.21

rs395042	chr3: 9942234	*IL17RC*	interleukin 17 receptor C	NM_153461	3p25.3	-4.13	$1.39x10^{-4}$	$1.40x10^{-2}$	6.03	7.08

Table 5.13: Cis eQTL shared between macrophages and monocytes. FDR cutoff 0.05. macro: macrophages, mono: monocytes.

Shared cis eQTL						macrophages		monocytes	Average expression		
SNP	SNP position	Gene	Gene product	REFSEQ ID	Location	Cell	t.stat	p.value	FDR	mac	mono
rs10239340	chr7: 128668510	*IRF5*	interferon regulatory factor 5	NM_002200	7q32	mono	19.43	$5.99x10^{-25}$	$4.47x10^{-20}$	9.01	10.75
						macro	16.86	$2.56x10^{-21}$	$2.50x10^{-16}$		
rs4559	chr12: 57489648	*STAT6*	signal transducer and activator of transcription 6, interleukin-4 induced	NM_003153	12q13	mono	15.74	$5.21x10^{-21}$	$8.00x10^{-17}$	9.85	11.17
						macro	10.33	$1.43x10^{-13}$	$5.92x10^{-10}$		
rs1968871	chr6: 41185430	*TREML3*	TREML3P triggering receptor expressed on myeloid cells-like 3, pseudogene	XM_929970	6p21.1	mono	11.64	$7.69x10^{-16}$	$1.98x10^{-12}$	5.10	7.37
						macro	5.77	$6.43x10^{-7}$	$3.06x10^{-4}$		
rs4360063	chr5: 96293153	*ERAP2*	endoplasmic reticulum aminopeptidase 2	NM_022350	5q15	mono	14.36	$2.27x10^{-19}$	$1.97x10^{-15}$	9.41	10.30
						macro	12.18	$5.34x10^{-16}$	$5.07x10^{-12}$		
rs1384	chr12: 69747834	*LYZ*	lysozyme	NM_000239	12q15	mono	14.25	$3.11x10^{-19}$	$2.44x10^{-15}$	13.50	13.79
						macro	7.69	$8.50x10^{-10}$	$1.14x10^{-6}$		
rs4075289	chr11: 830670	*CD151*	CD151 molecule (Raph blood group)	NM_139029	11p15.5	mono	-14.18	$3.73x10^{-19}$	$2.80x10^{-15}$	9.85	8.45
						macro	-12.32	$3.63x10^{-16}$	$3.76x10^{-12}$		
rs10743823	chr12: 9881125	*CLECL1*	C-type lectin-like 1	NM_172004	12p13.31	mono	-13.88	$8.80x10^{-19}$	$5.55x10^{-15}$	11.07	8.84
						macro	-6.07	$2.27x10^{-7}$	$1.35x10^{-4}$		
rs7957621	chr12: 10007761	*CLEC2B*	C-type lectin domain family 2, member B	NM_005127	12p13	mono	-13.49	$2.75x10^{-18}$	$1.48x10^{-14}$	8.19	9.73
						macro	-9.41	$2.75x10^{-12}$	$7.53x10^{-9}$		
rs4667297	chr2: 190528996	*ORMDL1*	ORM1-like 1 (S. cerevisiae)	NM_016467	2q32	mono	12.37	$7.80x10^{-17}$	$2.71x10^{-13}$	7.08	6.59
						macro	9.45	$2.35x10^{-12}$	$6.67x10^{-9}$		
rs12706949	chr7: 80307224	*CD36*	CD36 molecule (thrombospondin receptor)	NM_001001548	7q11.2	mono	-8.95	$5.77x10^{-12}$	$6.34x10^{-9}$	11.97	12.97

						macro	-4.64	$2.91x10^{-5}$	$6.27x10^{-3}$		
rs2364480	chr12: 6495275	*LTBR*	lymphotoxin β receptor (TNFR superfamily, member 3)	NM_002342	12p13	mono	5.60	$9.06x10^{-7}$	$2.39x10^{-4}$	10.60	12.28
						macro	9.08	$8.08x10^{-12}$	$2.04x10^{-8}$		
rs103294	chr19: 54797848	*LILRA3*	leukocyte immunoglobulin-like receptor, subfamily A (without TM domain), member 3	NM_006865	19q13.4	macro	-9.07	$8.30x10^{-12}$	$2.09x10^{-8}$	9.23	11.01
						macro	-8.43	$3.62x10^{-11}$	$3.36x10^{-8}$		
rs2906829	chr5: 135456085	*SMAD5*	SMAD family member 5	NM_005903	5q31	mono	6.58	$2.77x10^{-8}$	$1.18x10^{-5}$	9.93	8.86
						macro	6.61	$3.50x10^{-8}$	$2.80x10^{-5}$		

5.6.4 Macrophage specific and shared eQTL associated with GWAS loci

The pathogenesis of autoimmune and inflammatory diseases that have been studied in GWAS are proposed to involve dysregulated macrophages including obesity and diabetes (Olefsky and Glass, 2010, Murray and Wynn, 2011). Macrophage specific and shared eQTL were investigated for overlap with GWAS disease- or trait associated loci. Using the NHGRI GWAS catalogue following the same method described in the previous chapter, in total 150550 SNPs with GWAS associations were interrogated with 7655 and 10064 eSNPs specific to macrophage and shared with monocytes respectively (proxy SNPs r^2>0.8, Hindorff *et al.*, 2009, Section 4.5). 427 and 391 unique eSNPs for macrophage specific and shared between two cell types were found in GWAS disease and trait associated loci or their proxy SNPs (Table 5.15).

Table 5.14: Macrophage specific and shared eQTL associated with GWAS loci.

	Total eQTL	eSNP found in GWAS loci (%)	unique eSNP	Probe (Gene)
Macrophage	7655	717 (9.37)	427	113 (110)
Shared	10064	1022 (10.16)	391	94 (88)

Macrophage specific eQTL associated with GWAS

The cis eQTL specific to macrophages were identified in the GWAS loci that have been associated with susceptibility to autoimmune and metabolic related diseases including T2D, T1D, cholesterol levels, CHD, Crohn's disease and early onset IBD (Table 5.15). *THADA* expression was associated with rs17030845 proxy SNP to rs7578597, T2D susceptibility associated locus (p_{mac}=1.39x10^{-9}, Figure 5.21, Zeggini *et al.*, 2008). The locus encoding *THADA*, a thyroid adenoma associated gene, on 2p21 is disrupted in thyroid adenoma by aberrant arrangements to make a fusion with *PPARG* locus on 3p25 (Rippe *et al.*, 2003, Drieschner *et al.*, 2006, Drieschner *et al.*, 2007). Another T2D susceptibility associated locus rs1801282 on chr3 is associated with *PPARG* expression in macrophages (p_{mac}=5x10^{-6}, Figure 5.17, Scott *et al.*, 2007, Zeggini *et al.*, 2007). rs1801282 is in strong LD with rs17036160. GWAS loci associated with risks to cardiovascular diseases were found in macrophage specific eQTL. The locus in chr2p23.3 is associated with traits involved with CHD including C-reactive protein (CRP) and cholesterol levels, and inflammatory Crohn's disease (Ridker *et al.*, 2008, Teslovich *et al.*, 2010, Franke *et al.*, 2010). *SNX17* expression is associated with these GWAS loci in a macrophage specific

manner (peak SNP rs1260326, p_{mac}=6.48x10^{-7}, Figure 5.21). A sorting nexin 17 encoded in *SNX17* belongs to a sorting nexin family and is involved in low density lipoprotein and amyloid precursor protein recycling in early endosome (Kerkhof *et al.*, 2005, Lee *et al.*, 2008). rs1050382 in strong LD with a CHD risk locus rs6504218 in chr17q23.3 is associated with *PECAM1* expression (p_{mac}=1.48x10^{-6}, Figure 5.21, Coronary Artery Disease (C4D) Genetics Consortium, 2011). *PECAM1* encodes a platelet endothelial cell adhesion glycoprotein and a member of immunoglobulin gene superfamily and is involved in leukocyte migration and adhesion, angiogenesis and neutrophil clearance (Newman *et al.*, 1990, DeLisser *et al.*, 1997, Luu *et al.*, 2003).

Shared eQTL associated with GWAS

Autoimmune disease associated loci were found in the shared eQTL. The shared associations with *IRF5* and *ERAP2* which are described in the B-cell and monocyte dataset in the previous section were found in autoimmune disease susceptibility loci. The shared eQTL association of rs4728142 and *IRF5* on chr7 was associated with the susceptibility to autoimmune diseases (p_{mac}=6.68x10^{-5}, p_{mono}=6.7x10^{-6}, SLE; Graham *et al.*, 2006, Han *et al.*, 2009, IBD; Jostins *et al.*, 2012, Anderson *et al.*, 2011, Figure 5.20, Table 5.16). Higher expression of *IRF5* is characteristic in inflammatory macrophages and is associated with promoter polymorphisms including eSNP rs4728142 (Krausgruber *et al.*, 2011). rs4360063 associated with *ERAP2* is in strong LD with rs2549794, a locus associated with susceptibility to Crohn's disease on chr5 (Figure 5.20, Franke *et al.*, 2010). A role of ERAP2 in trimming HLA class I peptides in endoplasmic reticulum is important for antigen presentation, suggesting disruption of this process may be involved in autoimmune response in the disease. Shared eQTL associations were found in neurodegenerative disorder associated loci. MS susceptibility associated locus rs703842 on chr12 is found with shared eQTL with *FAM119B* which is also called *METTL21B* and encodes methyltransferase like 21B (p_{mac}=1.25x10^{-15}, p_{mono}=1.72x10^{-21}, Figure 5.22, Bahlo *et al.*, 2009). This protein belongs to a lysine methyltransferase family which is involved in methylation to regulatory chaperones and their disruption is predicted to associate with degenerative neuromuscular diseases (Kernstock *et al.*, 2012, Cloutier *et al.*, 2013). Another MS susceptibility locus rs10466829 on chr12 is associated with *CLECL1* expression (p_{mac}=1.06x10^{-7}, p_{mono}=8.8x10^{-19}, Figure 5.22, Sawcer *et al.*, 2011). The *CLECL1* locus encodes a C-type lectin-like transmembrane protein which is involved in pathogen recognition as well as T cell activation (Cambi and Figdor, 2003). The

locus associated in HDL cholesterol levels and prostate cancer was found in association with *LILRA3* on chr19 (rs103294 , p_{mac}=8.30x10^{-12}, p_{mono}=3.62x10^{-11}, Figure 5.22, Teslovich *et al.*, 2010, Xu *et al.*, 2012). *LILRA3* encodes a leukocyte immunoglobulin-like receptor however, its function in these disease susceptibility is largely unknown. Autoimmune diseases including diabetes and obesity are highly involved in dysregulation of inflammatory macrophages. The locus encoding *SULT1A1* on chr16 is associated with susceptibility to autoimmune diseases, including T1D, Crohn's disease, IBD and body mass index (BMI) (Barrett *et al.*, 2009, Franke *et al.*, 2010, Imielinski *et al.*, 2009, Thorleifsson *et al.*, 2009, Willer *et al.*, 2009). The most significant SNP rs40831 associated with *SULT1A1* expression in macrophages is in strong LD with these GWAS disease associated loci (p_{mac}=1.86x10^{-6}, p_{mono}=7.04x10^{-6}, Figure 5.22). *SULT1A1* encodes a phenol sulfotransferase enzyme, however the role of *SULT1A1* in disease susceptibility is not well known and may be involved in hormone sulfation (Kester *et al.*, 1999, Harris *et al.*, 2000).

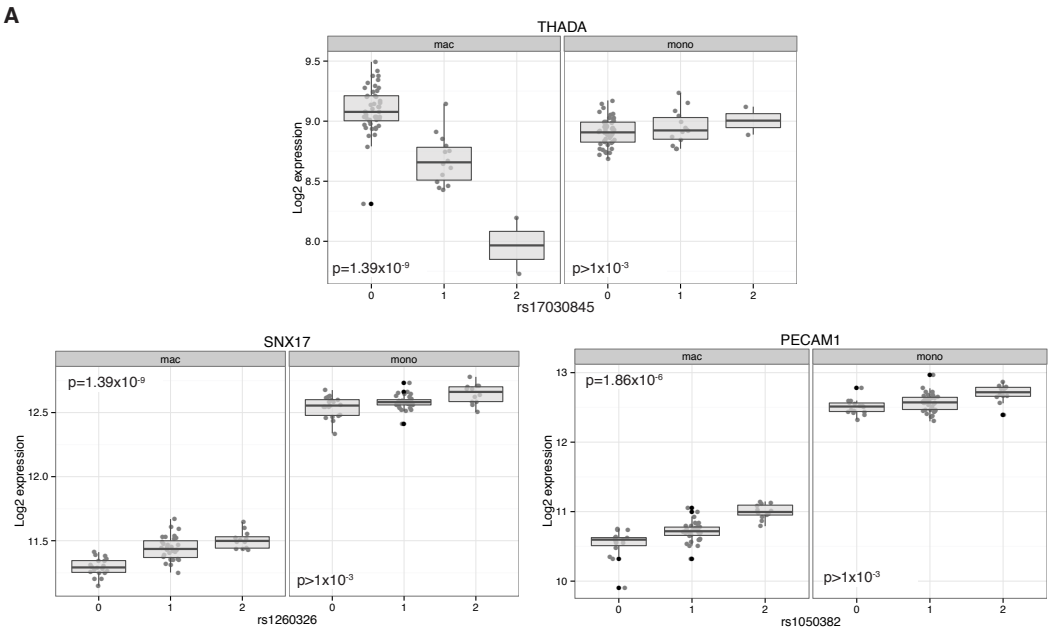


Figure 5.21

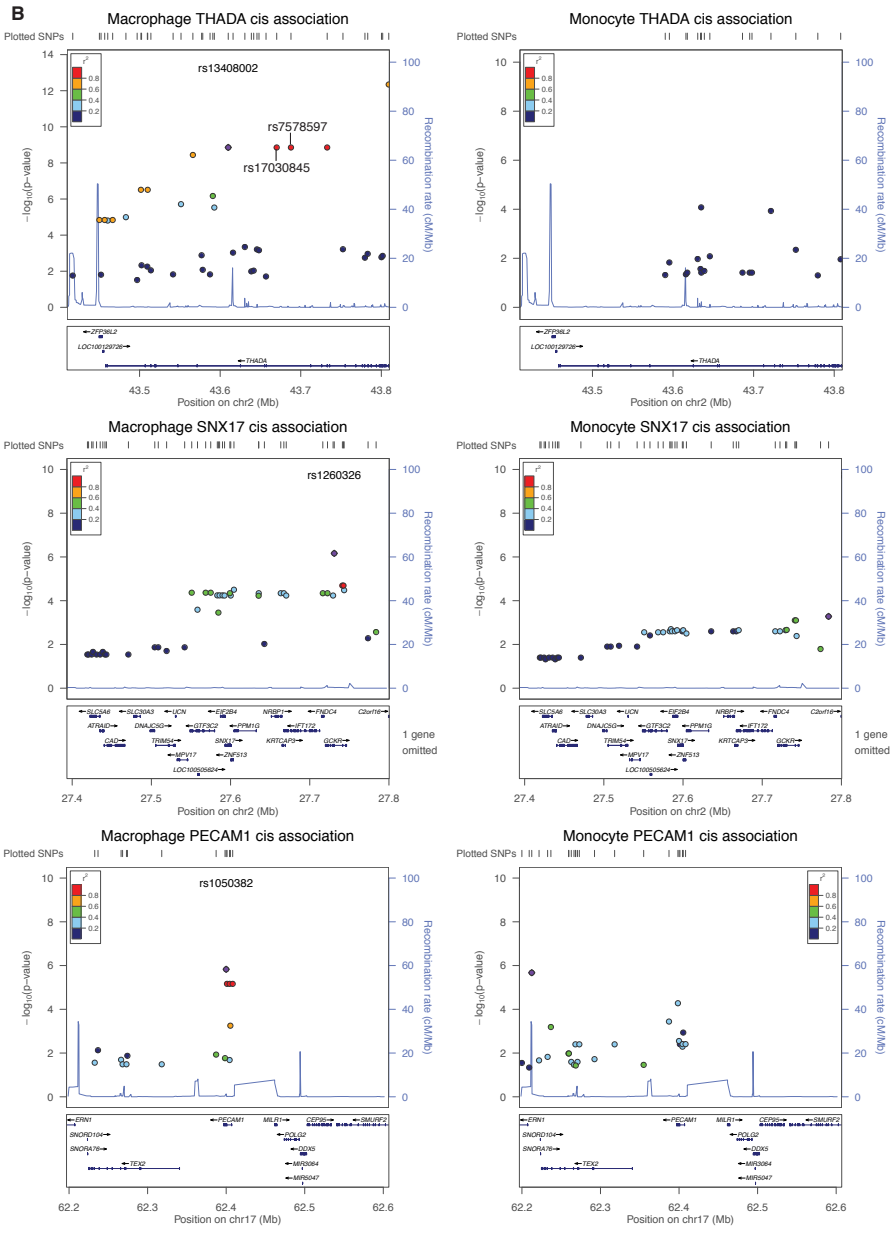


Figure 5.21: Box (A) and association plots (B) of example shared cis eQTL. (A) The number of minor alleles is shown in the box plots (0 - major homozygous, 1 - heterozygous, 2 - minor homozygous) for macrophages (mac) and monocytes (mono).

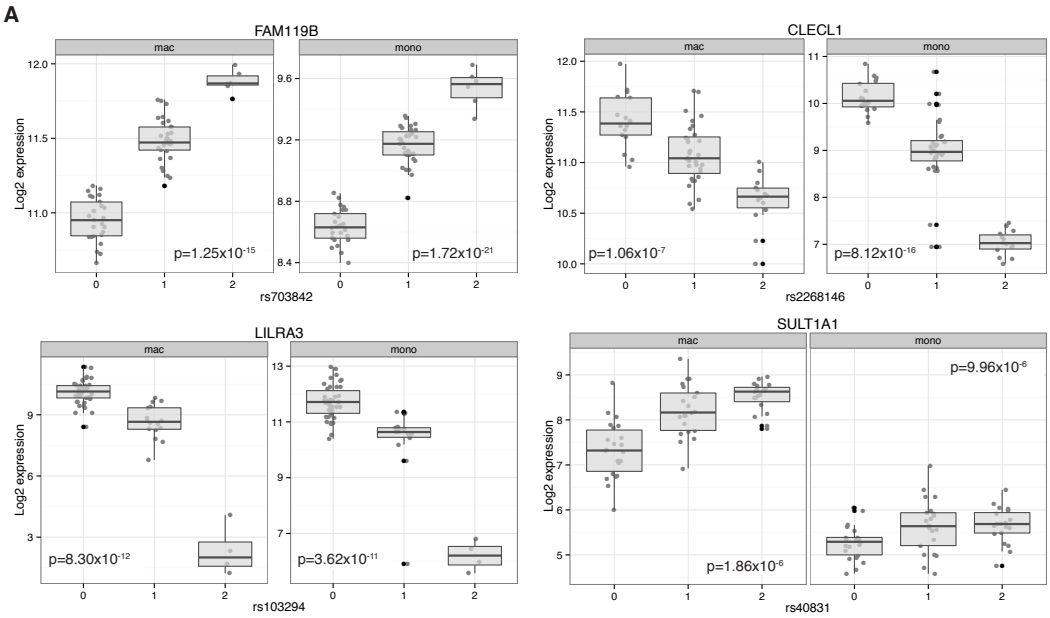

A
FAM119B
mac
mono
Log2 expression
p=1.25x10-15
p=1.72x10-21
rs703842
CLECL1
p=1.06x10-7
p=8.12x10-16
rs2268146
LILRA3
p=8.30x10-12
p=3.62x10-11
rs103294
SULT1A1
p=9.96x10-6
p=1.86x10-6
rs40831

Figure 5.22

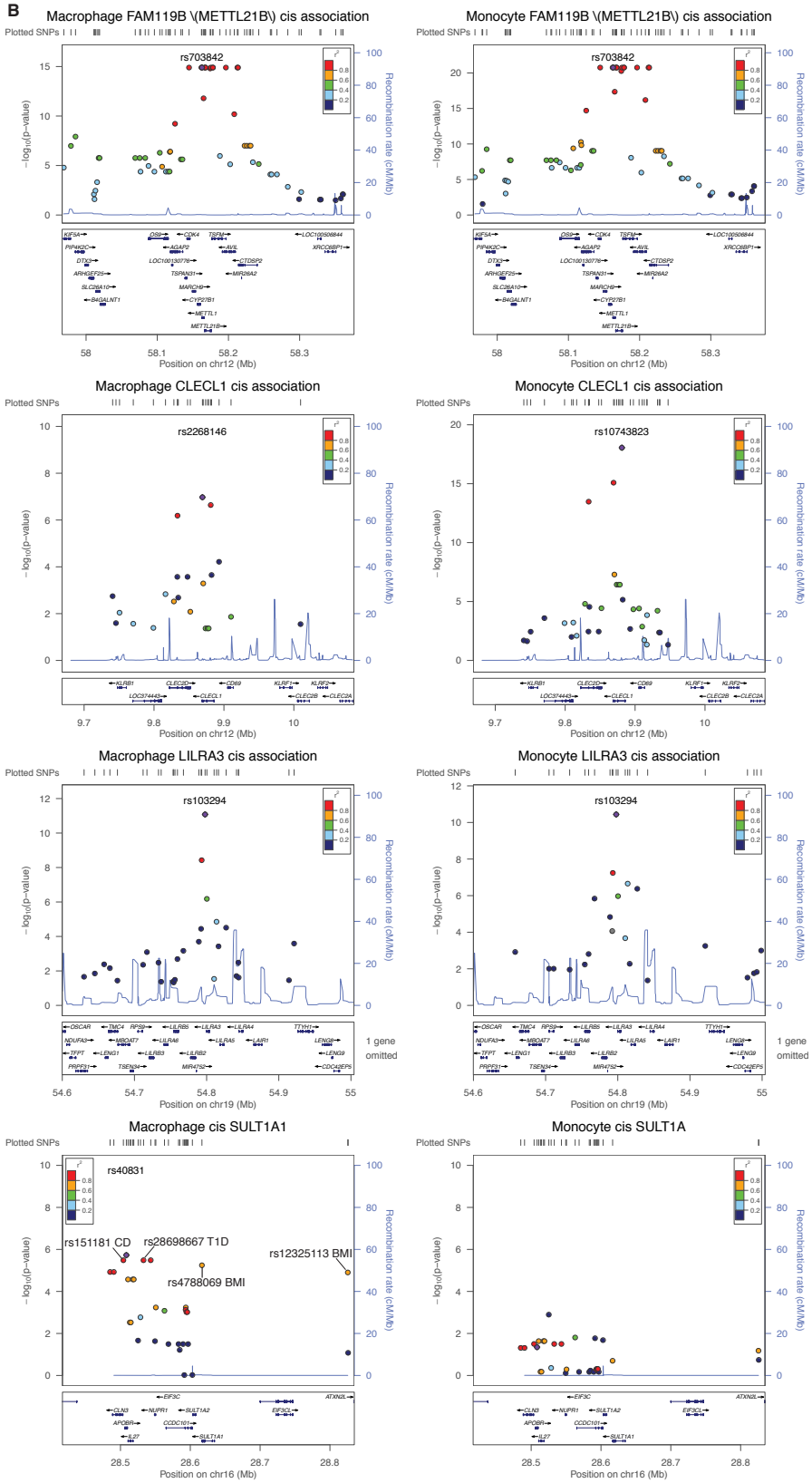


Figure 5.22: Box (A) and association plots (B) of example shared cis eQTL. (A) The number of minor alleles is shown in the box plots (0 - major homozygous, 1 - heterozygous, 2 - minor homozygous) or macrophages (mac) and monocytes (mono).

Table 5.15: Macrophage specific eQTL associated with GWAS loci. Proxy SNPs (r^2>0.8) are shown if an eSNP in macrophages is different from GWAS SNP. IBD: inflammatory bowel disease, SLE: systematic lupus erythematosus, BMI: Body Mass Index, *THADA*: thyroid adenoma associated, *PPARG*: peroxisome proliferator-activated receptor γ, *SNX17*: sorting nexin 17, *PECAM1*: platelet/endothelial cell adhesion molecule 1.

Disease or trait	Reference	Region	SNP	p_{GWAS}	Proxy SNP	Gene	t.stat	p.value	FDR
T2D	Zeggini *et al.*, 2008	2p21	rs7578597	$1x10^{-9}$	rs17030845	*THADA*	-7.55	$1.39x10^{-9}$	$1.76x10^{-6}$
	Scott *et al.*, 2007 Zeggini *et al.*, 2007	3p25.2	rs1801282	$2x10^{-6}$		*PPARG*	5.17	$5.00x10^{-6}$	$1.59x10^{-3}$
C-reactive protein	Ridker *et al.*, 2008	2p23.3	rs780094	$7x10^{-15}$	rs1260326	*SNX17*	5.75	$6.84x10^{-7}$	$3.22x10^{-4}$
Cholesterol, total	Teslovich *et al.*, 2010	2p23.3	rs1260326	$5x10^{-40}$		*SNX17*	5.75	$6.84x10^{-7}$	$3.22x10^{-4}$
Crohn's disease	Franke *et al.*, 2010	2p23.3	rs1260326	$7x10^{-6}$	rs780093	*SNX17*	5.75	$6.84x10^{-7}$	$3.22x10^{-4}$
CHD	Coronary Artery Disease (C4D) Genetics Consortium, 2011	17q23.3	rs6504218	$1x10^{-6}$	rs1050382	*PECAM1*	5.53	$1.48x10^{-6}$	$5.97x10^{-4}$

Table 5.16: Shared eQTL between macrophages and monocytes associated with GWAS loci. Proxy SNPs (r^2>0.8) are shown if an eSNP in macrophages is different from GWAS SNP. IBD: inflammatory bowel disease, UC: ulcerative colitis, SLE: systematic lupus erythematosus, *FAM119B*: *METTL21B*, methyltransferase like 21B, *CLECL1*: C-type lectin-like 1, *IRF5*: interferon regulatory factor 5, *ERAP2*: endoplasmic reticulum aminopeptidase 2, *LILRA3*: leukocyte immunoglobulin-like receptor, subfamily A (without TM domain), member 3, *PSRC1*: proline/serine-rich coiled-coil 1, *SULT1A1*: sulfotransferase family, cytosolic, 1A, member 1.

Disease or trait	Reference	Region	SNP	p_{GWAS}	Proxy SNP	Gene	Cell	t.stat	p.value	FDR
MS	Bahlo *et al.*, 2009	12q14.1	rs703842	$5x10^{-11}$		*FAM119B*	mono	16.16	$1.72x10^{-21}$	$3.04x10^{-17}$
							macro	11.89	$1.25x10^{-15}$	$1.05x10^{-11}$
	Sawcer *et al.*, 2011	12p13.31	rs10466829	$1x10^{-8}$	rs10743823	*CLECL1*	mono	-13.88	$8.80x10^{-19}$	$5.55x10^{-15}$
							macro	-6.29	$1.06x10^{-7}$	$7.18x10^{-5}$
RA	Raychaudhuri *et al.*, 2008	12q13.3	rs1678542	$9x10^{-8}$	rs11172254	*FAM119B*	mono	5.14	$4.64x10^{-6}$	$9.35x10^{-4}$
							macro	4.82	$1.61x10^{-5}$	$3.97x10^{-3}$
SLE and UC	Han *et al.*, 2009	7q32.1	rs4728142	$8x10^{-19}$		*IRF5*	mono	-5.03	$6.70x10^{-6}$	$1.27x10^{-3}$
	Anderson *et al.*, 2011						macro	-4.39	$6.68x10^{-5}$	$1.17x10^{-2}$
Crohn's disease	Franke *et al.*, 2010	5q15	rs2549794	$1x10^{-10}$	rs4360063	*ERAP2*	mono	14.36	$2.27x10^{-19}$	$1.97x10^{-15}$
							macro	12.18	$5.34x10^{-16}$	$5.07x10^{-12}$
T1D	Wellcome Trust Case Control Consortium, 2007	12p13.31	rs3764021	$5x10^{-8}$	rs10743823	*CLECL1*	mono	-13.88	$8.80x10^{-19}$	$5.55x10^{-15}$
							macro	-6.07	$2.27x10^{-7}$	$1.35x10^{-4}$
HDL cholesterol	Teslovich *et al.*, 2010	19q13.42	rs386000	$4x10^{-16}$	rs103294	*LILRA3*	macro	-9.07	$8.30x10^{-12}$	$2.09x10^{-8}$
							mono	-8.43	$3.62x10^{-11}$	$3.36x10^{-8}$
Prostate cancer	Xu *et al.*, 2012	19q13.42	rs103294	$5x10^{-16}$		*LILRA3*	macro	-9.07	$8.30x10^{-12}$	$2.09x10^{-8}$
							mono	-8.43	$3.62x10^{-11}$	$3.36x10^{-8}$
Total cholesterol	Aulchenko *et al.*, 2009	1p13.3	rs646776	$9x10^{-22}$		*PSRC1*	mono	4.43	$5.07x10^{-5}$	$6.40x10^{-3}$

	Teslovich *et al.*, 2010	1p13.3	rs629301	$6x10^{-131}$	rs646776	*PSRC1*	macro mono	4.21 4.43	$1.16x10^{-4}$ $5.07x10^{-5}$	$1.76x10^{-2}$ $6.40x10^{-3}$
T1D nephropathy	Barrett *et al.*, 2009	16p11.2	rs4788084	$3x10^{-13}$	rs28698667	*SULT1A1*	macro mono	4.21 5.02	$1.16x10^{-4}$ $7.04x10^{-6}$	$1.76x10^{-2}$ $1.32x10^{-3}$
Crohn's disease	Franke *et al.*, 2010	16p11.2	rs151181	$2x10^{-11}$		*SULT1A1*	macro mono	5.30 4.48	$3.24x10^{-6}$ $4.38x10^{-5}$	$1.13x10^{-3}$ $5.69x10^{-3}$
BMI	Thorleifsson *et al.*, 2009	16p11.2	rs7498665	$3x10^{-10}$	rs12325113	*SULT1A1*	macro mono	4.91 4.90	$1.17x10^{-5}$ $1.06x10^{-5}$	$3.13x10^{-3}$ $1.85x10^{-3}$
	Willer *et al.*, 2009			$5x10^{-11}$			macro	4.90	$1.24x10^{-5}$	$3.28x10^{-3}$
	Speliotes *et al.*, 2010		rs7359397	$2x10^{-20}$	rs4788099	*SULT1A1*	mono	4.90	$1.06x10^{-5}$	$1.85x10^{-3}$
IBD (early onset)	Imielinski *et al.*, 2009	16p11.2	rs8049439	$2x10^{-9}$		*SULT1A1*	macro mono	4.90 4.90	$1.24x10^{-5}$ $1.06x10^{-5}$	$3.28x10^{-3}$ $1.85x10^{-3}$
							macro	4.90	$1.24x10^{-5}$	$3.28x10^{-3}$

Gene enrichment analysis for macrophage specific eQTL

In order to investigate gene enrichment in biological pathways, the genes with eQTL associations were analysed using the ReactomePA program. The method is described in Section 3.6. In this analysis, macrophage and shared eQTL datasets were considered. In total 7470 and 9932 eQTL specific to macrophage and shared between the two cell types respectively were tested for the gene enrichment. There was no enrichment identified in any biological pathway in this analysis (p-value cutoff 0.05). The eQTL associated genes in this dataset as well as the previous B-cell and monocyte dataset were not significantly enriched in pathways using the Reactome database.

Summary

The comparison of eQTL association between macrophages and monocytes revealed novel loci that are specific to cell type as well as the loci that are replicated in monocytes from the previous analysis. Genes associated with macrophage specific eQTL were involved in lipid metabolism as well as the immune response and inflammation, which were also found in the GWAS loci associated with susceptibility to diabetes and obesity.

5.7 Discussion

5.7.1 Sample collection

The collection for two myeloid cell types was conducted at the same time in order to establish a paired sample dataset. $CD14^+$ monocytes were separated from freshly purified PBMCs and were subsequently incubated for macrophages differentiation. This process aimed to reduce noise that could have been caused during technical steps.

An established method for macrophage differentiation was used for this dataset. GM-CSF stimulation of $CD14^+$ monocytes is widely used to differentiate into classical inflammatory macrophages (Grabstein *et al.*, 1986). The development of macrophage differentiation from monocytes *in vivo* is controlled by cytokines that are produced during inflammatory response. Additional stimulation by pro-inflammatory cytokines, such as IFNγ, TNFα and microbial stimulus LPS, facilitates further polarisation of macrophages (Santiago-Schwarz *et al.*, 1992, Mosser and Edwards, 2008). TNFα was used in this dataset to achieve an activated status of macrophages.

The degree of macrophage differentiation from monocytes was determined by quantification of gene expression for markers of macrophages. *CD68* was used as a marker for macrophages and higher expression compared to monocytes by qPCR was observed. Macrophage lineages are remarkably heterogeneous and plastic in order to accommodate various responses in different tissues and timing of the immune response (Gordon and Taylor, 2005, Mosser and Edwards, 2008). This dataset was primarily focused on classical inflammatory macrophages. Further polarisation of inflammatory macrophages are induced by inflammatory cytokines, which leads to differentiation into effector cells in the Th1 cellular immune response (Hsieh *et al.*, 1993).

This dataset consists of 62 paired samples of monocyte and macrophage expression profiles. Many studies quantified genome-wide gene expression profiles in monocytes and macrophages with various stimuli. The sample sizes of these studies were far smaller for each treatment as they focused on characterising the dynamics of heterogeneous macrophage differentiation by various stimuli (Martinez *et al.*, 2006, Gratchev *et al.*, 2008, Szanto *et al.*, 2010). To the best of my knowledge, the current dataset is one of the largest macrophage dataset paired with circulating CD14+ monocytes. One of the unique features in this dataset is that the samples were collected from a similar genetic background. As heritability of gene expression has been widely observed, this dataset may have a more homogeneous profile, which has contributed to the detection of gene signals during differentiation.

5.7.2 Differential gene expression analysis

The differential gene expression analysis was performed using the 62 paired samples in order to investigate the dynamic change of differentiation at a gene level. 71% of probes tested were significantly differentially expressed between monocytes and macrophages. Monocytes had a larger proportion of genes differentially expressed (63%) than macrophages probably because the high purity of $CD14^{+}$ monocytes without the incubation step helped to reduce variation in gene expression.

Noise in the macrophage data

The largest source of noise between the monocyte and macrophage datasets is the process of incubation for differentiation. It would be ideal to have a monocyte dataset with incubation without any stimulation. The monocyte dataset is actually a subset of the large population study which is used as a baseline with other datasets that were stimulated with

various stimuli and timing. Although the incubation process was carefully performed to reduce the noise as much as possible using the same batch of the medium, the ligands and plates, it is almost impossible to remove all noise caused by the incubation. The effect of the incubation process was apparent in the PCA plot during the gene expression normalisation. It may be possible to remove the noise caused by the incubation process, which may require further analysis. A control sample set could be obtained by processing samples without stimulation using the same incubation process as the macrophage differentiation. Using PCA, the data obtained from this control could be used to identify PCs explaining variance of incubation.

Genes differentially expressed in macrophages and monocytes

The genes differentially expressed between the two cell types indicate their distinct roles in the inflammatory response. The pathway analysis was used to classify genes into their roles. The significantly up-regulated genes in macrophages are enriched with genes that encode chemokines, cytokines, proteins involved in phagocytosis and fatty acid metabolism. Involvement in fatty acid metabolism as well as pro-inflammatory cytokines clearly suggests classically inflammatory macrophages have a major role in adipose tissue and inflammation response. By contrast, the significantly down-regulated genes in macrophages include receptors for chemokine, cytokine, adhesion receptors and PRRs, suggesting macrophages switched off genes required in circulating monocytes in the blood stream as part of the surveillance system. These results including the pathway analysis suggest that differentiation from monocytes into macrophages by inflammatory stimuli globally changes a series of gene expression networks.

5.7.3 eQTL analysis

The eQTL analysis used two closely related myeloid cell types in order to investigate genetic variation that affects gene expression in a cell type specific manner in the context of the innate immune response and inflammation. The analysis was performed separately for the monocyte and macrophage expression data. Due to the small population size, the analysis used SNPs with MAF>10%. The cis window +/-250kbp was applied for this analysis because the number of probes was the largest and the majority of functional transcriptional elements were likely to be located within those limits.

eQTL analysis incorporating PCA

The eQTL analysis incorporated PCA that investigated the gene expression data for reduced data dimensionality. PCA transformed variance of the gene expression data and the incorporation of the appropriate number of PCs helps to detect more eQTL associations. In this dataset, 1.7 times more SNP-gene associations were detected as a result of the analysis with PCs. The numbers of PCs were chosen based on their maximum detection of eQTL in cis for further comparison of the cell type specific eQTL analysis. A different number of PCs between macrophages (14 PCs) and monocytes (10 PCs) was chosen. This difference is likely because of the additional incubation step for macrophage differentiation leading to additional variables.

Shared eQTL

Shared eQTL association between the two cell types (31.4% of the total eQTL association) have been maintained even after macrophage differentiation. This suggests these eQTL may influence the inflammatory response throughout different stages as they are maintained. Investigating the shared associations further in more diversely activated macrophages may be intriguing. For example, *IRF5* is highly expressed in inflammatory macrophages and regulates expression of pro-inflammatory cytokines. Polarisation of inflammatory macrophages is promoted by IRF5 and Th1-Th17 responses which are critical for host defence and autoimmunity (Krausgruber *et al.*, 2011). The change of the gene expression level in *IRF5* is affected by polymorphisms in the promoter region and are associated with susceptibility to autoimmune diseases including SLE and inflammatory bowel disease (Graham *et al.*, 2006, Han *et al.*, 2009, Jostins *et al.*, 2012).

LYZ association in cis was maintained in macrophages which was shown as a master regulatory eQTL in trans in the previous chapter. This chapter did not describe trans association of the *LYZ* locus in macrophages due to the small sample size, the cis association in the two cell types suggests trans association including a TF *CREB1* may be preserved in macrophages. This trans association may be interesting to investigate in not only macrophages but also other cell types in the monocyte lineage such as tissue resident macrophages and dendritic cells.

Cell type specific eQTL

Cell type specific eQTL have been detected by the comparison of the unique eSNPs from the individual results of the eQTL analysis. The proportion of the cell type specific eSNPs

was 68.6% which is lower than the B-cell and monocyte eQTL analysis (78%). However, this proportion can be considered as high in a comparison to the cell types in the same lineage.

Cell type specific eQTL have been identified even when the gene expression levels were similar between monocytes and macrophages (27-29%). Cell type specific eQTL in the previous chapter on B-cells and monocytes indicate that these associations may reflect primarily a difference in their cellular functions and transcriptional factors. This observation implies that these eQTL associations may use different transcriptional machinery for the gene expression. That distance from TSSs for the cell type specific eQTL was significantly different from that for the shared between the two cell types, which supports this idea. However, the distribution of the distance of eQTL from TSSs did not differ between macrophages and monocytes.

The findings shown in this section are simply based on the data analysis. Experimental validation may confirm the identified associations as well as the replication in a different population. For example, the association in *PPARG* with rs17036160, which has a low MAF, may be confirmed by specifically investigating larger numbers of individuals with minor homozygous alleles compared to major homozygous allele controls.

eQTL association hidden in naive immune cells

Macrophage specific eQTL association can be revealed by cell activation if circulating $CD14^{+}$ monocytes are considered in a naive state. Detection of trans eQTL associations is regarded as more difficult to detect because they may be hidden without stimulation (Gaffney, 2013). The comparison of differentially activated macrophages may provide association of genetic variants that are associated with disease susceptibility. For example, a comparative eQTL analysis on inflammatory, regulatory and tissue repair macrophages may determine specific eQTL association which corresponds to their cellular functions. The cis eQTL analysis has identified association to *PPARG* in macrophage and the SNP is associated with susceptibility to T2D (Scott *et al.*, 2007, Saxena *et al.*, 2007, Zeggini *et al.*, 2007, Manning *et al.*, 2012). This association requires further validation by other methods such as qPCR and in a larger cohort due to the small MAF of the SNP.

The eQTL analysis between normal tissue macrophages and dysregulatory activated macrophages (e.g. classical inflammatory cells in adipose tissues in obesity) from the same individuals may be intriguing to investigate the functional roles of genetic susceptibility to autoimmunity.

5.7.4 Summary

The analysis presented in this chapter was performed to investigate differential gene expression between macrophages and monocytes and the association with genetic variation. The comparative eQTL analysis revealed unique eQTL association in macrophages which may inform GWAS investigating autoimmune disease susceptibility. To my knowledge, this is one of the largest paired expression profiles of macrophages and monocytes and the first to map eQTL in paired samples of monocytes and macrophages differentiated from the same individuals. Although the sample size was limited, this dataset highlights the potential to investigate genetic variants modulating macrophage gene expression in relation to autoimmune and inflammatory diseases.

6 Functional characterisation of eQTL

6.1 Introduction

Recent GWAS have identified many new loci associated with complex diseases. However it has been difficult to demonstrate a causal relationship between genetic variation in these loci and disease susceptibility because the effect sizes of the risk variants are modest and fine mapping has been challenging (Freedman *et al.*, 2011). Also the identified loci in the majority of instances are in non-coding regions of the genome and it is difficult to predict their functional consequences. One of the goals of the ENCODE project is to characterise such non-coding regions of the human genome in terms of transcription factor binding and enhancer regions (The ENCODE Project Consortium; Dunham *et al.*, 2012). The project used different cell types to functionally map non-coding regions important for gene expression using several techniques including chromatin immunoprecipitation (ChIP) and DNaseI hypersensitivity assays (Gerstein *et al.*, 2012, Thurman *et al.*, 2012). Mapped regulatory non-coding sequences tend to have a conserved sequence, for example involving transcription factor binding motifs. Single nucleotide differences in the conserved regulatory sequence may modulate gene expression through less or more efficient binding of proteins. The information collected in the ENCODE project may contribute to understanding roles of single nucleotide variants that are associated with disease susceptibility.

eQTL mapping associates transcript abundance with DNA variants as explored in the previous chapters (Chapters 4 and 5). Screening mapped eQTL overlaying TFBSs may help to identify regulatory DNA variants with associating genes and may aid understanding of their functional consequences.

Functional characterisation of putative regulatory variants involving eQTL and GWAS loci may involve a variety of techniques. Reporter assays using a luciferase gene are widely used to detect enhancer activity of the region of interest. Electrophoretic mobility shift assay (EMSA) is used to detect effects of a single nucleotide difference on protein binding.

The details of these techniques are described in the results sections. However, few cases were functionally validated in terms of allelic effect of associated gene expression levels because the effect of individual variants tend to be modest as mentioned above. The GWAS locus which is associated with LDL cholesterol levels and susceptibility to myocardial infarction (MI) is one of the few successful cases of functional validation (Musunuru *et al.*, 2010). The study demonstrated that the susceptibility SNP creates a transcription factor binding site (TFBS) of *SORT1* and modifies hepatic expression of the gene using luciferase assays.

In Chapter 4, analysis of cell type specific eQTL identified cis and trans eQTL that are related to the immune systems. This included identification of novel master regulatory eQTL. In this chapter, 12q15, one of these master regulatory regions, was investigated functionally characterised. The specific aims of this chapter are:

1. *In silico* sequence analysis of the 12q15 region in search of functional elements using the ENCODE database

2. Investigation of transcriptional enhancer activity of sequences containing eSNPs using luciferase reporter assays

3. Determination of allelic differences in protein binding for eSNPs using EMSA

6.2 eQTL signal in *LYZ* region in 12q15

The 12q15 region in monocytes has shown strong eQTL signals with 73 probes (72 genes) throughout the genome in both cis and trans (Figures 4.14 and 4.4). The peak SNP rs10784774 has the strongest association with *LYZ* expression in cis ($p=8.9x10^{-78}$, Figure 4.15). The allele frequency of the peak SNP rs10784774 in the HapMap CEU population is A/G=0.491/0.509 (ancestral/derived alleles). Trans eQTL signals were detected for numerous loci of the genome (the most significant associations included *CREB1* chromosome 2 (chr2); $p=2.03x10^{-67}$, *DUSP19* chr2; $p=2.21x10^{-59}$, *RAB27A* chr15; $p=2.03x10^{-57}$, *NOD1* chr7; $p=2.33x10^{-20}$, Figure 4.15). The second most significant SNP rs1384 ($p=8.9x10^{-78}$) which is located in 3'UTR of *LYZ* is in almost perfect LD with rs10784774 (r^2 0.967, Figure 6.1).

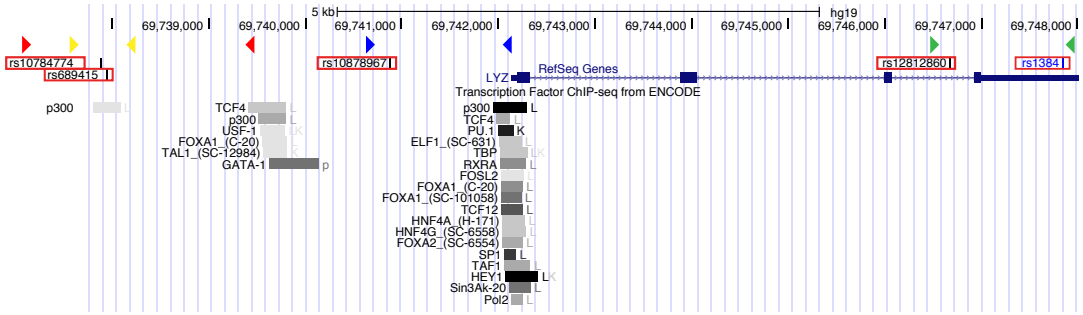


Figure 6.1: A diagram of the 12q15 LYZ promoter region. Taken and modified from the UCSC Genome browser (assembly GRCh37hg19). The most significant SNPs are labeled with red rectangles. rs10784774 and rs10878967 in upstream of the promoter, rs12812860 in the last intron and rs1384 in 3'UTR of *LYZ*. The transcription factor (TF) ChIP-seq data track from the ENCODE project displays several TFBSs in the *LYZ* promoter region including p300. The colour of TFBS rectangles corresponds to cluster scores. The colour filled triangles indicate primer annealing positions (red; rs10784774 pLYZ 2kb, yellow; rs10784774 300 TFBS, blue; rs10878967 pLYZ 1.5kb, green; rs1384 and rs12812860 3'UTR LYZ). See Table 2.2 for primer sequences.

6.2.1 Imputation of 12q15 locus

Fine mapping of any genetic association can be aided by higher density information on the genetic variation present in an associated locus and this was sought for the 12q15 region. Advances in genotype imputation methodology have contributed to increased resolution of the genotyping information from the typed genotyping data without further expensive physical genotyping (Marchini and Howie, 2010). The genotype imputation helps not only to increase power of GWAS but also permits meta-analysis using different populations which are genotyped on different platforms (Li *et al.*, 2009). The accuracy of imputation has been improved as the more common DNA variants have been identified by the efforts of the International HapMap and the 1000 Genomes Projects. Imputation in a population consisting of unrelated individuals is based on the linkage disequilibrium that creates shared haplotypes. The IMPUTE2 program uses the information of shared haplotypes as a reference haplotype panel to infer missing genotypes between typed genotypes with great precision (*http://mathgen.stats.ox.ac.uk/impute/impute_v2.html*, Howie *et al.*, 2009).

The imputation of the 12q15 locus was performed using the 1000 Genomes Project data as a reference panel (the 1000 Genomes Phase I integrated variant set, version 3, released in March 2012) in order to fine map the region of interest. Due to the limitation of computer power and memory space, chr12 only was separately imputed using the healthy volunteer cohort with 283 individuals data in NCBI37/hg19 assembly. The 31953 typed genotypes resulted in 813568 SNPs including the typed SNPs after imputation with the confident score cutoff >0.9.

The local eQTL association to *LYZ* expression in monocytes was performed with the imputed genotype data in chr12 using a linear model with PLINK. The analysis identified two SNPs with strong association which were in the same LD block as rs10784774 ($p=8.9x10^{-78}$, Figure 6.2). rs10878967 is about 1kb 5' upstream from the start codon of *LYZ*. rs12812860 is located in the last intron of *LYZ* (Figure 6.1). All four SNPs locate within 10kbp (chr12:69737879-69747834 in the UCSC genome browser NCBI37 assembly). The genotypes with imputation improved the resolution of the haplotype structures around rs10784774 but did not identify SNPs with higher significance.

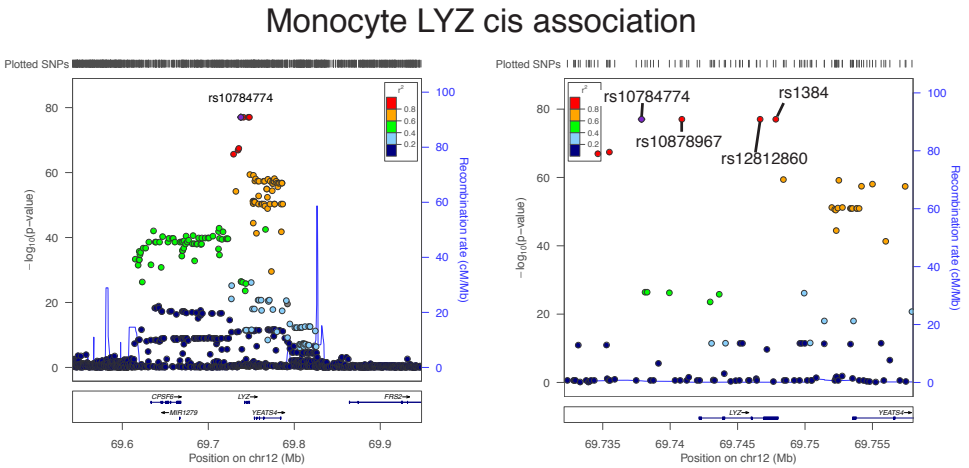


Figure 6.2: Local association plots of monocyte *LYZ* expression in cis with imputed SNPs. Both panels show same locus with different scales (left: 400kb and right: 30kb). rs10784774, rs10878967, rs12812860 and rs1384 are in almost perfect LD as coloured with red ($r^2>0.8$ in relation to the peak SNP rs10784774). Two additional haplotype blocks are observed from the plot: chr12:69615090-69722880 (green circles, $2.4x10^{-40}<p>4.4x10^{-34}$, $0.4<r^2>0.6$) and chr12:69748397-69786543 (orange circles, $4.0x10^{-60}<p>4.7x10^{-51}$, $0.6<r^2>0.8$).

6.2.2 Regulatory sequences in *LYZ* eQTL

The ENCODE project has generated a set of transcription factor ChIP-seq data in 72 cell lines for 119 DNA binding proteins which has mapped DNA-binding motifs in the human genome. In order to determine if the eSNPs may overlay DNA-binding motifs, the *LYZ* locus was investigated with the ENCODE transcription factor (TF) ChIP-seq data (Gerstein *et al.*, 2012). The TF ChIP-seq data uses a cluster score across the dataset of the different TF ChIP-seq experiments. The score is defined as a strength of the ChIP-seq signal in a cluster of short sequence reads which reflects a frequency of a TF occupying the region inside of the cluster.

rs10784774 overlaps with a p300 TFBS in the HepG2 (hepatocellular carcinoma cell line) ChIP-seq data. p300 is a main component of the p300-CBP co-activator family and

is involved in up-regulation of gene expression by relaxing chromatin structure and recruiting RNA polymerase II to the gene promoter (Eckner *et al.*, 1994, Eckner *et al.*, 1996, Yao *et al.*, 1998). Although the cluster score is not strong (156 out of 1000), this evidence promoted work to determine whether this SNP affects binding of transcription factor(s) as an enhancer.

rs1384 is positioned in the 3'UTR of *LYZ* and does not overlap with TFBSs (Figure 6.1). 3'UTR sequence in a gene can be a target of miRNA for mRNA degradation or prevention of protein translation. In order to compare miRNA accessibility to the *LYZ* 3'UTR between two alleles, miRNA target sites were searched using the miRNA database (Online microRNA prediction tool in the Segal laboratory, *http://genie.weizmann.ac.il/pubs/mir07/index.html*, Kertesz *et al.*, 2007). The database predicts miRNA bindings to target UTRs by calculating the free energy gained from the formation of the microRNA-target duplex. The results showed that there was no miRNA bound to the DNA sequence including rs1384 and no difference between two alleles.

rs10878967 is approximately 1.5kb upstream of the start codon of *LYZ*. Although it does not find any overlap with TFBSs, 1kb downstream of the SNP is a cluster of TFBSs confirmed by ChIP-seq, which includes a strong signal of p300 binding in the HepG2 cell line (Figure 6.1). The SNP and the DNA-binding motifs in the promoter region may be in a same LD.

In addition to TFBSs an epigenetic marker was identified in the *LYZ* eQTL. DNaseI hypersensitive sites indicate accessibility of transcriptional regulatory factors and active cis regulatory elements including enhancers, promoters, silencers, and insulators (Gross and Garrard, 1988, Gaszner and Felsenfeld, 2006). A DNaseI hypersensitive assay was conducted on 125 diverse cell types and tissues in the ENCODE project (Thurman *et al.*, 2012). Using this data, rs12812860 which is located in the last intron of three in *LYZ* was found to be located in a DNaseI hypersensitive site in the HL-60 cell line (promyelocytic leukaemia cell line).

6.2.3 Transcription factor binding motifs

Many SNPs identified from GWAS and eQTL studies postulate that the association of SNPs with disease susceptibility or gene expression regulation is caused by alteration of binding efficiency of proteins, notably TFs. TFs have specific favourable binding motifs that are normally conserved among or throughout species. The eSNPs in the *LYZ* eQTL were investigated to see if either alleles might modify TF binding motifs. JASPAR is a

bioinformatic tool to identify regulatory elements in eukaryotes (*http://jaspar.cgb.ki.se*, Sandelin *et al.*, 2004). The database is based on matrices of position-specific scoring of experimentally demonstrated TFBSs (Stormo, 2000, Wasserman and Sandelin, 2004).

The DNA sequences containing an eSNP with either allele were obtained from the UCSC Human Genome browser and analysed on the JASPAR CORE with human TFBS models under profile relative threshold score cutoff >80%. The TFBS identified in the positive strands were counted in this analysis. Unique TFBS were identified in only one allele, either ancestral or derived but no common function of identified TFs was identified (Table 6.1). The ancestral rs10784774 allele A creates a TFBS for RORA2 which is a zinc-coordinating nuclear hormone receptor and is involved in development and inflammation (Jetten, 2004). The ancestral rs10878967 allele T creates a Forkhead family of winged-helix TF FOXA1 which is implicated in breast and prostate cancers as well as epithelial cell differentiation and morphogenesis (Augello *et al.*, 2011), whereas the derived allele C creates TFBSs for ELK1 which is involved in various contexts including long-term memory and breast cancer and for AP1 which is involved in many cellular processes including proliferation, death and differentiation (Shaulian and Karin, 2002). The rs12812860 alleles create different TFBS of TFs involved in various cellular processes (Table 6.1). The rs1384 ancestral allele C creates a TFBS for ETS1 whose DNA binding domain is involved in chromosome translocation in tumours (Delattre *et al.*, 1992). Thus, the TFBS analysis of eQTL showed different accessibility of TFs in each allele.

Table 6.1: TFBS search on *LYZ* eSNPs using JASPAR. The ancestral (top) and derived (bottom) alleles were compared if different TFBSs were found. The scores and relative scores (out of 1) were calculated based on matrices of position-specificity and normalisation of scores on each nucleotide position respectively (Stormo, 2000, Wasserman and Sandelin, 2004).

SNP	Allele	TF	Score	Relative score	Reference
rs10784774	A	RORA2	8.48	0.80	Giguere *et al.*, 1994
	G	NA			
rs10878967	T	FOXA1	6.23	0.82	Carroll *et al.*, 2005
	C	ELK1	5.82	0.81	Treisman *et al.*, 1992
	C	AP1	6.70	0.86	Lee *et al.*, 1987
rs12812860	T	HOXA5	5.92	0.86	McGinnis *et al.*, 1984
	T	MZF1_1-4	4.99	0.81	Morris *et al.*, 1994
	T	SPIB	6.96	0.88	Ray *et al.*, 1992
	C	HIF1A::ARNT	7.61	0.89	Wenger *et al.*, 2005
	C	USF1	5.70	0.80	Bendall and Molloy, 1994
rs1384	C	ETS1	3.66	0.81	Woods *et al.*, 1992
	T	NA			

The detailed investigation of the *LYZ* eQTL using the publicly available functional data suggested that specific SNPs may be functionally linked to the allelic difference in

LYZ expression and modulate a master regulator acting in trans in the genome. Further experiments described in the following sections were therefore conducted to characterise the SNPs.

6.3 Assaying enhancer activity using the luciferase gene reporter system

Reporter gene systems have been used to investigate the functional significance of genetic sequences with respect to transcription regulation in eukaryotes. A gene or DNA fragment of interest is fused to a reporter gene in a single plasmid construct and is delivered into cells or tissues where an assay is held. Reporter genes have been discovered from organisms in the natural world. For example, *lacZ* in *Escherichia coli* encodes β-galactosidase that cleaves X-gal (5-bromo-4-chloro-indolyl-β-D-galactopyranoside) to produce intense blue colour. *gfp* gene isolated from jelly fish (*Aequrea victoria*) encodes green fluorescent protein (GFP) that exhibits bright green fluorescence on excitation by blue to ultraviolet range light (peak at 395 nm). Chemiluminescence reporter assays, which use an enzymatic reaction to produce photons in the presence of specific substrates, is widely adapted to quantify cis-acting transcriptional elements such as promoters and enhancers. The advantages of the chemiluminescence based assay over the fluorescence based assay are no requirement of excitation by high power laser and less background such as autofluorescence thus the assay is more easily carried out, rapid, safe and sensitive.

Luciferase is one of commonly used chemiluminescence enzymes. The well-studied luciferase genes were cloned from firefly (*Photinus pyralis*, Wet *et al.*, 1987) and sea pansy (*Renilla reniformis*, Lorenz *et al.*, 1991) and neither requires any post-translational modification. Firefly luciferase is a monomeric enzyme of 61kDa and catalyses two-step oxidation of substrate beetle luciferin to emit yellow to green light between 550-570 nm with ATP, Mg^{2+} and O_2. Renilla luciferase is a monomeric 36kDa enzyme and catalyses oxidation of coelenterate luciferin only with oxygen to emit blue light of 482 nm. Addition of coenzyme A stabilises the luminescence reaction by slowing down the decay of signal intensity. Two distinct luciferases are deployed in the Dual-Luciferase^{TM} reporter system (Promega) that uses the firefly gene for the study assay and the Renilla for an internal control of transfection.

Both firefly and Renilla luciferase reporter constructs available from Promega were used to study the enhancer activity of 12q15. The Firefly luciferase construct, pGL3-

Promoter, has firefly *luc+* gene with the SV40 promoter. A DNA fragment of interest for enhancer activity assay is inserted in 5' upstream of the SV40 promoter or 3' downstream of the *luc+* gene (Figure 6.3). The Renilla luciferase construct, pRL-TK, has Renilla *hluc* gene with the herpes simplex virus thymidine kinase (HSK-TK) promoter that provides low to moderate levels of hluc expression in transfected mammalian cells.

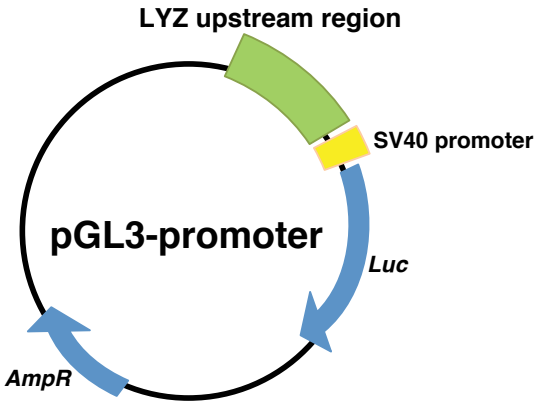


Figure 6.3: pGL3-promoter firefly luciferase reporter construct. *Luc*; lucifrase gene, *AmpR*; β-lactamase gene.

6.3.1 Cloning of eQTL locus flanking LYZ in 12q15

In order to investigate the enhancer activity of the *LYZ* eQTL (rs10784774, rs1384, rs12812860 and rs10878967), a cloning strategy was adopted as described below (Figure 6.1).The 2kb region spanning rs10784774 and a shorter region spanning only p300 TFBS (367bp long) were cloned in order to determine the activity of p300 TFBS with rs10784774. The eSNPs rs1384 and rs12812860 were cloned together in approximately a 1.5kb stretch including the LYZ 3'UTR and a part of the last intron. The region including eSNP rs10878967 covered approximately a 1.5kb region to include the start codon of *LYZ* and a number of TFBSs in the *LYZ* promoter.

The genomic DNA from individuals in the healthy volunteer population was used to amplify the enhancer elements based on their genotype using PCR (Chapter 2, Table 2.2). Both alleles of the eSNPs were amplified for comparison (Table 6.2). The PCR cycling conditions and enzymes used are described in the materials and methods (Chapter 2). The PCR products were subcloned into the pGL3-promoter firefly luciferase reporter construct (Promega) in the 5' upstream of the SV40 promoter (Figure 6.3). The plasmids were purified using midipreps with an endotoxin removal step following manufacturer's instructions (Qiagen, Chapter 2).

Table 6.2: *LYZ* eSNPs and alleles.

Allele	Derived	Ancestral	CEU frequency	
rs10784774	G	A	0.509	0.491
rs10878967	C	T	0.528	0.472
rs12812860	C	T	0.533	0.467
rs1384	T	C	0.509	0.491

6.3.2 *In vitro* luciferase reporter assay

Enhancer elements regulate gene expression by modulating transcriptional mechanisms in the cell. Transient transfection of the luciferase reporter constructs to cell lines allows measurement of enhancer activities.

Luciferase activity of rs10784774 region in different cell lines

The K562 human myelogenous leukaemia cell line (ATCC number CL-243, Lozzio and Lozzio, 1975) was first chosen to determine allelic difference of the enhancer activities quantitatively. The cell line resembles undifferentiated granulocytes including monocytes and erythrocytes and is highly efficient for transfection. At least three biological and two technical replications each were performed in the following experiments for transient co-transfection of firefly and Renilla luciferase constructs. The readings of firefly activity were normalised against that of Renilla, and then the values were normalised against the pGL3-promoter activity at each experiment. The significance of any observed difference between two alleles was determined by Wilcoxon rank sum test. The rs10784774 construct showed a significant difference between the G and A alleles suggesting this region with the G allele may have higher enhancer activity than the A allele, although the overall luciferase activity is lower than for the pGL3-promoter control vector ($p=8.46x10^{-3}$, top left of Figure 6.4). This result suggests it may be important to determine whether this enhancer activity is cell type specific as the *LYZ* eQTL was identified as cell type specific in the previous chapter (Chapter 4). Three additional cell lines, THP-1, Jurkat, and LCL GM18057, with different characteristics were transfected with the rs10784774 constructs to investigate if the enhancer activity was cell type specific. THP-1 is a human acute monocytic leukaemia cell line (ATCC number TIB-202, Tsuchiya *et al.*, 1980) and is characterised by lysozyme production, differentiation to macrophage and phagocytosis. Jurkat is a T cell type lymphoblastoid cell line isolated from a patient with acute lymphoblastic leukaemia and leukaemia transformed non-Hodgkin lymphoma (ATCC number TIB-152, Schneider *et al.*, 1977) and is characterised by expression of

complement receptors. GM18057 is one of the HapMap CEU LCLs and is immortalised by EBV infection as described in the previous chapters (Dausset *et al.*, 1990, International HapMap Consortium, 2003, Chapters 1 and 4). Variation in the enhancer activity was observed in the three cell lines. The allelic difference of the enhancer activity in the rs10784774 constructs was significant in THP-1 and Jurkat cell lines but not in the LCL GM18057 (THP-1; p=2.17x10^{-3}, Jurkat; p=1.56x10^{-4}, LCL GM18057; p=0.8, Figure 6.4).

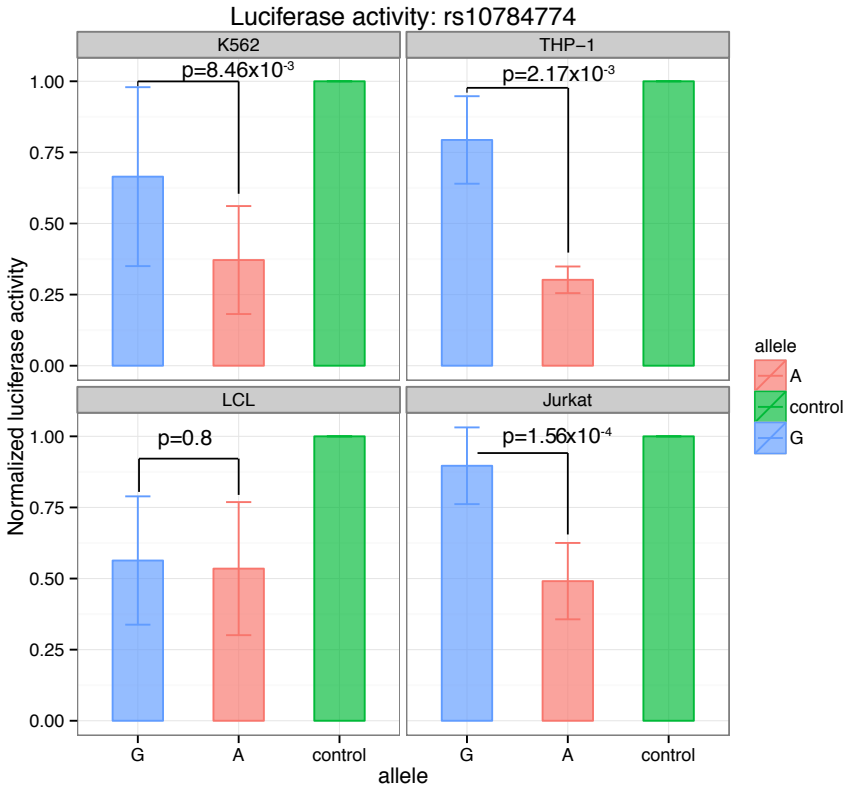


Figure 6.4: Luciferase activity of a region including rs10784774. Luciferase activity was determined during transient expression in K562 myelogenous leukaemia cell line (top left), THP-1 monocytic leukaemia cell line (top right), LCL GM18057 lymphoblastoid cell line (bottom left), Jurkat T cell type leukaemia cell line (bottom right). rs10784774 derived allele G (blue) and ancestral allele A (pink) compared after normalisation against control pGL3-promoter vector (green). Significance was determined by Wilcoxon rank sum test (p-values and standard deviation shown in the figure). K562, THP-1 and Jurkat show significant difference in luciferase activity between G and A alleles.

rs10784774 p300 TFBS

The constructs of the p300 TFBS with rs10784774 were determined to investigate if the region is critical for allelic enhancer activity. The constructs were tested using K562. No significant difference between the alleles was observed (p=1, Figure 6.5).

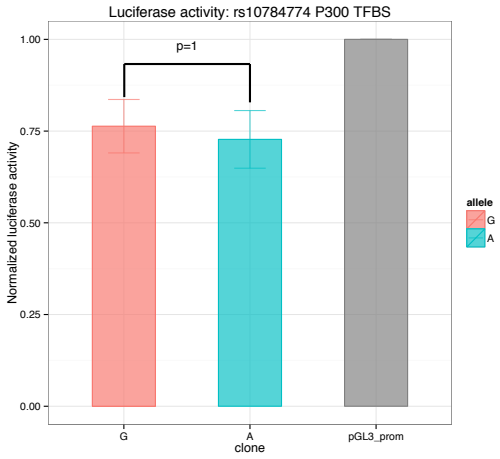


Figure 6.5: Luciferase activity of rs10784774 p300 TFBS (367bp long) in K562. rs10784774 derived allele G (pink) and ancestral allele A (turquoise) compared after normalisation against pGL3-promoter vector (grey). No significant difference was shown between the two alleles (Wilcoxon rank sum test p=1, standard deviation shown in the figure).

Additional eQTL in the *LYZ* region

The two regions including eSNPs rs10878967, rs1384 and rs12812860 were measured for their enhancer activities in comparison of the alleles (Figures 6.6 and 6.7). These two constructs were not significantly different (rs10878967; p=0.23, rs1384 and rs12812860; p=1, Figures 6.6 and 6.7). However, the rs10878967 constructs showed higher luciferase activity over the control vector, suggesting the region may have more capacity to up-regulate the luciferase reporter gene.

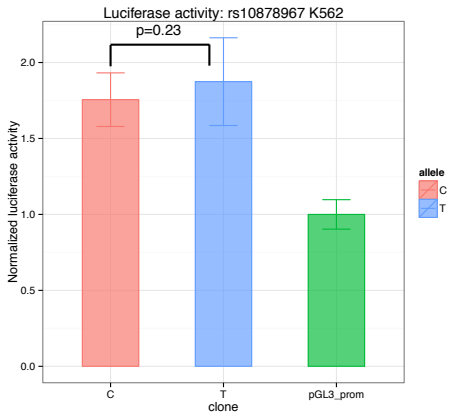


Figure 6.6: Luciferase activity of a region including rs10878967 in K562. rs10878967 derived allele C (pink) and ancestral allele T (blue) compared after normalisation against the control pGL3-promoter vector (green). No significant difference was shown between the two alleles (Wilcoxon rank sum test, p=0.23, standard deviation shown in the figure). The luciferase activity with the rs10878967 region is higher than with the control, which is only observed in this DNA fragment.

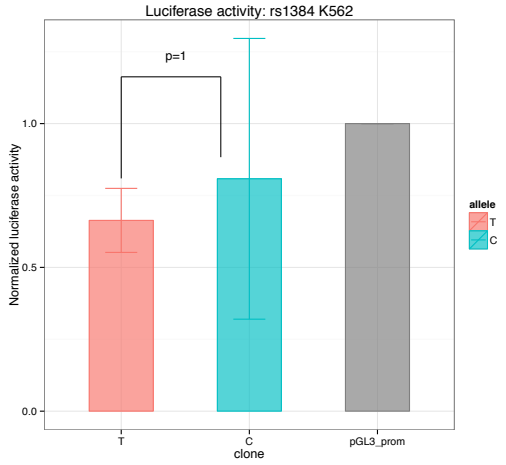


Figure 6.7: Luciferase activity of a region including rs1384 in K562. rs1384 derived allele T (pink) and ancestral allele C (turquoise) compared after normalisation against the control pGL3-promoter vector (grey). No significant difference was observed between the two alleles (Wilcoxon rank sum test, p=1, standard deviation shown in the figure).

Position effect of enhancers

Enhancers may function at a distance by formation of DNA loops and protein-protein interaction for example involving insulators (Gaszner and Felsenfeld, 2006). In order to determine the effect of position on enhancer activity in an allelic specific manner, the 2kb rs10784774 region was cloned in the 3' downstream of the firefly luciferase gene. In K562 and THP-1 cell lines, the allelic effect of the enhancer was observed only when cloned 5' upstream of the reporter gene, not in the 3' downstream (K562; p=4.11x10^{-5}, THP-1; p=2.17x10^{-3}, Figure 6.8).

In total 4 eSNPs of *LYZ* eQTL were investigated for their effects on enhancer activity using the luciferase reporter system. However, only the rs10784774 2kb construct revealed significant differences between the two alleles.

6.4 EMSA to identify SNP alleles bound differentially to cellular proteins

6.4.1 Background

Interaction of proteins and nucleic acids can be assessed experimentally by many techniques that have both advantages and limitations. EMSA is a sensitive technique to investigate DNA-protein interaction using oligonucleotide probes and purified proteins

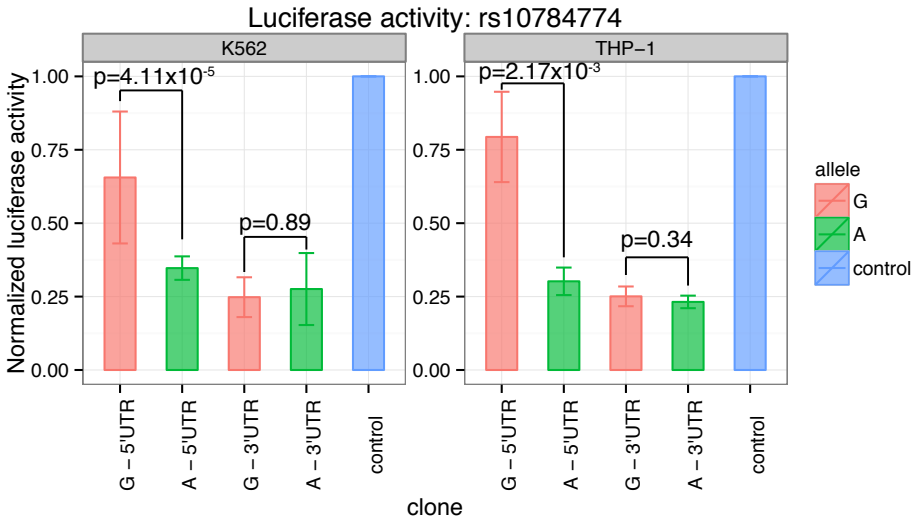


Figure 6.8: Position effect of luciferase activity with the rs10784774 region in K562 and THP-1. rs10784774 derived allele G (pink) and ancestral allele A (green) were tested in 5′ and 3′ UTR of the luciferase gene and were compared after normalisation against the control pGL3-promoter (blue) (Wilconxon rank sum test, p-values and standard deviation shown in the figure). A significant difference was observed in the 5′UTR constructs.

or crude nuclear extracts (Fried and Crothers, 1981, Garner and Revzin, 1981, Hellman and Fried, 2007). Electrophoretic mobility of labelled oligonucleotides (usually with radioisotopes or biotin) is monitored in the presence of proteins or nuclear extracts by running in a native gel. Probe-protein complexes can be detected by autoradiography or chemiluminescent imaging. EMSA is a qualitative and highly sensitive method of observing formation of non-equilibrium proteins to nucleic acid complexes.

Another method to detect DNA-protein interaction is a footprinting assay which exploits interference of enzymatic digestion or modification of a nucleic acid sequence bound by a protein (Brenowitz *et al.*, 1986). A classic footprinting protocol uses an end-labelled nucleic acid sequence with radioisotope binding to a protein followed by DNase I digestion and resolution of the fragments on a denaturing polyacrylamide gel. Interpretation of the image visualised by autoradiography is performed using ladders of bands and gaps which indicate sites of protection by protein(s) in the test. The advantage of this method over EMSA is that a protein of interest binds to nucleic acid sequences at the equilibrium condition for the footprint. Modified footprint methods are available for quantitative detection of binding (Brenowitz *et al.*, 1986). The limitations of the footprinting are that incomplete digestion and non-specific protein binding to the probe sequence make the image unclear and more difficult to interpret than EMSA.

Variants of DNA sequences may result in creation or loss of protein binding sites as described in the previous section of this chapter (Section 6.2.3). The SNPs in both alleles change their predicted TFBS. In order to determine if the alleles of the eSNPs in the LYZ promoter regions have different protein binding preferences, EMSA was conducted as described below.

6.4.2 Protein binding patterns of LYZ eSNPs

First the four LYZ eSNPs (rs10784774, rs1384, rs12812860, and rs10878967) were investigated to see if they have different protein binding patterns between the alleles. A SNP rs689415 was included as a control, which is located in 260bp closer to *LYZ* and is not in strong LD to rs10784774 (r^2 0.412, Figure 6.1). Probes were designed to be on average length 25-nt oligonucleotides with the eSNPs in the middle of the sequences as described in the materials and methods chapter (Table 2.4, Chapter 2). The short nucleotides were chosen to reduce protein binding around the SNP regions. In order to investigate protein binding patterns between two alleles, crude nuclear extract prepared from a K562 cell line was used to bind radiolabelled oligoduplexes. The results are shown in Figure 6.9. The binding patterns on rs1384 and rs10878967 showed stronger signals in the ancestral C and T alleles. This result prompted follow up of those SNPs using a competition EMSA and in different cell types.

6.4.3 Competition assay

Competition assay in EMSA is used to measure protein binding affinity for a specific probe by adding an excessive amount of a non-labelled probe. The competition assay on rs1384 and rs10878967 was performed by including unlabelled competitor probes corresponding to either the same allele or the other allele. Overall the binding complexes were relatively weak but the data suggested more allele specificity. The results of the rs1384 assay showed the allele C competed better than the T (Figure 6.10). The rs10878967 assay showed the allele T competed better than the allele C when they were competed against the opposite alleles (Figure 6.11). The comparison between non-specific competitors (6th lane from left to 13th lane in Figures 6.10 and 6.11) also demonstrated that the ancestral alleles C and T in rs1384 and rs10878967 had stronger protein binding that the derived alleles T and C.

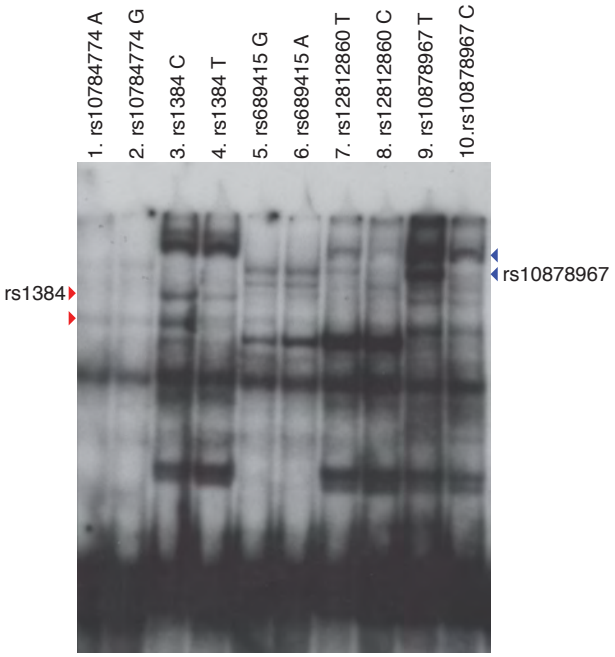


Figure 6.9: EMSA with LYZ eSNP probes with rs10784774, rs1384, rs689415, rs12812860 and rs10878967 using K562 nuclear extract. rs689415 is a SNP located in the LYZ promoter region and used for a control (lanes 5 and 6). The relative positions of the SNPs are shown in Figure 6.1. The ancestral alleles rs1384 C (lane 3, red arrow) and rs10878967 T (lane 9, blue arrow) showed stronger complexes.

rs10878967 competition assay in different cell lines

rs10878967 T allele showed a stronger binding pattern using proteins from nuclear extract prepared from K562. Because those eSNPs in *LYZ* were identified as monocyte specific, nuclear extracts from other types of cell lines were examined to see if rs10878967 still retains an allelic difference in protein binding. EMSA on nuclear extract prepared from THP-1 showed different protein band patterns between T and C alleles without competitors (1st and 8th lanes from left, Figure 6.12). The competitor assay using THP-1 showed the T allele was still stronger than the C allele (5th and 12th lanes in Figure 6.11). The EMSA using nuclear extract prepared from LCL GM18057 showed relatively vague band patterns (Figure 6.13). The T allele without competitor has a faint band (red arrow) that does not appear in the lanes with competitors, although it is difficult to distinguish the difference of protein binding patterns between the two alleles. The allelic difference in band patterns appeared in monocytic cell lines K562 and THP-1 but not in LCL which retains similar characteristics to B-cells.

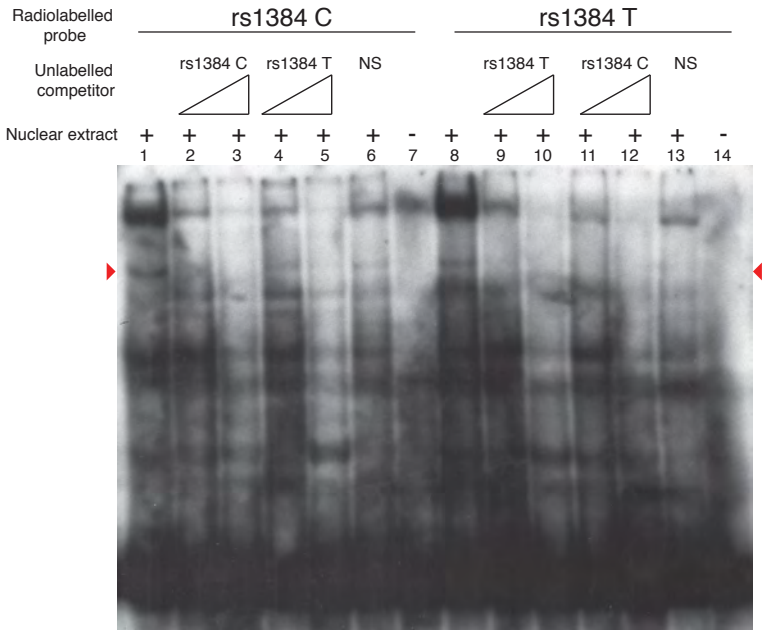


Figure 6.10: EMSA competition assay on rs1384 using K562 nuclear extract. Non-radiolabelled probes were used as competitors at 10X and 100X concentrations. NS; non-specific competitor. The C probe (lane 5) competed better for the complex labelled by red arrow than did T (lane 12) with 100X opposite allele competitors.

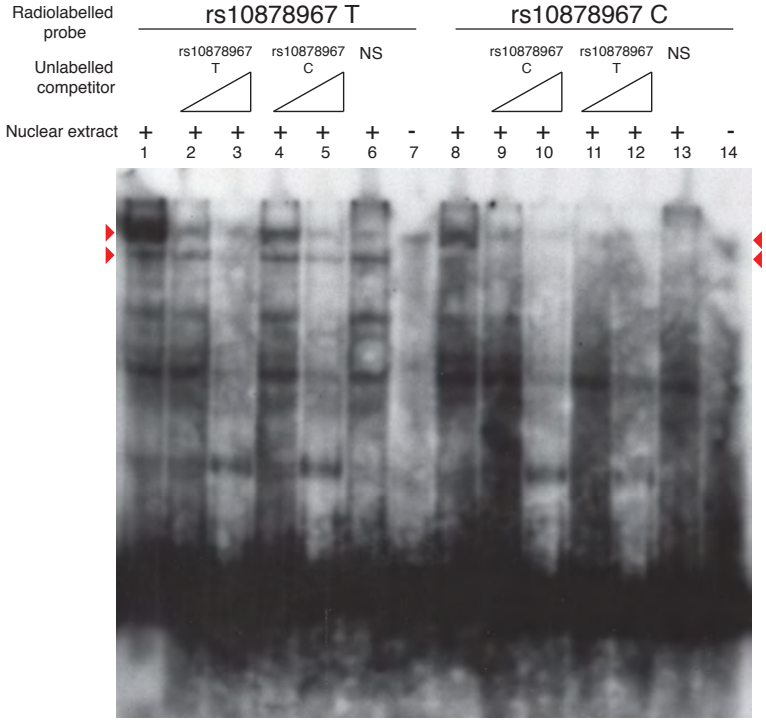


Figure 6.11: EMSA competition assay on rs10878967 using K562 nuclear extract. Non-radiolabelled probes were used as competitors at 10X and 100X concentrations. NS - non-specific competitor. The T probe (lane 5) competed better for the complex labelled by red arrows than C (lane 12) with 100X opposite allele competitors.

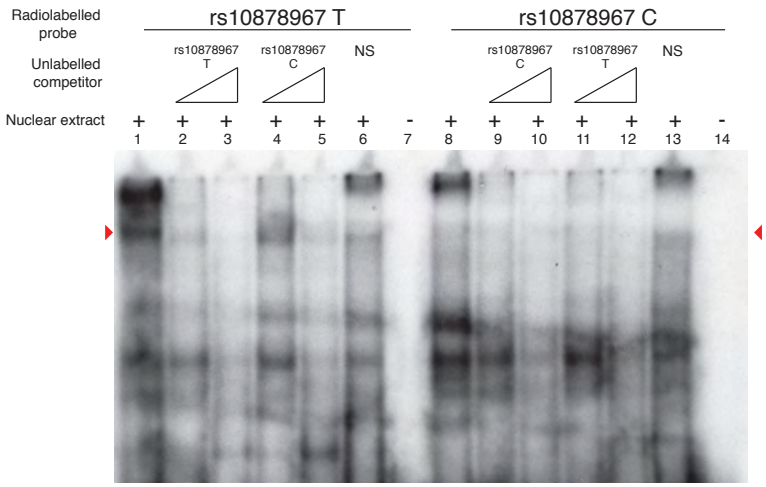


Figure 6.12: EMSA competition assay on rs10878967 with THP-1 nuclear extract. Non-radiolabelled probes were used as competitors at 10X and 100X concentrations. NS - non-specific competitor. The T probe (lane 5) competed better for the complex labelled by red arrow than did C (lane 12) with 100X opposite allele competitors.

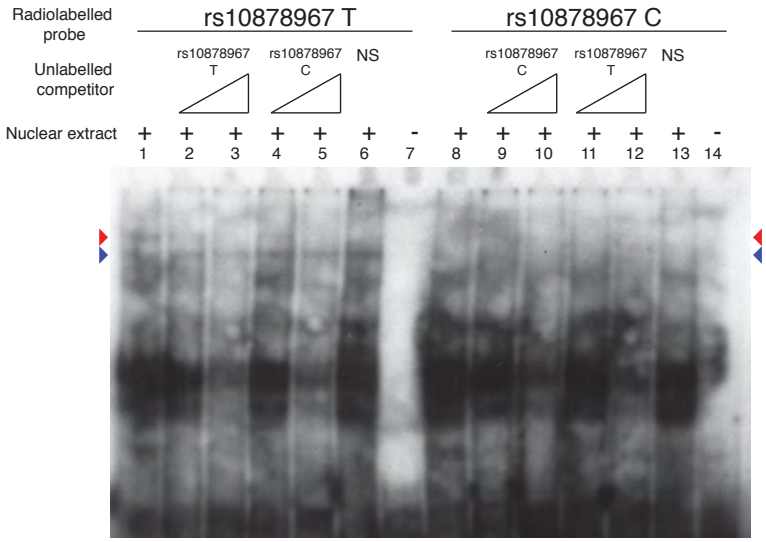


Figure 6.13: EMSA competition assay on rs10878967 with LCL GM18057 nuclear extract. Non-radiolabelled probes were used as competitors at 10X and 100X concentrations. NS; non-specific competitor. The T allele without competitor (lane 1) has a faint band indicated by red arrow. The all T allele lanes (lane 1-6) have another band indicated by blue arrow but may not be specific to the probe.

6.4.4 Antibody supershift assay

The identity of proteins that bind to the probes spanning eSNPs was determined by supershift assay. One limitation of EMSA using crude nuclear extracts is that proteins binding to DNA probes cannot be identified easily because estimation of protein sizes is normally difficult in a native gel with protein mobility affected by charge and hydrodynamic size at the pH of the running buffer. The supershift assay uses an antibody to the protein of interest to see if a DNA-protein complex shows shifted mobility due to the additional weight of the antibody or if binding between protein and DNA is abolished due to interference by the antibody.

rs1384 and rs10878967 were investigated to see if specific proteins of interest may bind to the probes. The antibodies used were anti-P300, anti-CREB1, anti-P65 and nonspecific control rabbit IgG. P300 and CREB1 are transcription cofactors and their binding sites are virtually ubiquitous among promoter regions including *LYZ* (Figure 6.1). P65 forms a heterodimeric NFκB transcription activator with P50 and is involved in many biological processes including inflammation and immunity in almost all cell types.

The supershift assay for rs10878967 with these antibodies using K562 nuclear extract showed supershift, a disappearance of the bands for the T allele in both anti-P300 and IgG lanes (Figure 6.14). The supershift on the IgG control lane raises questions about the specificity of the supershift by anti-P300.

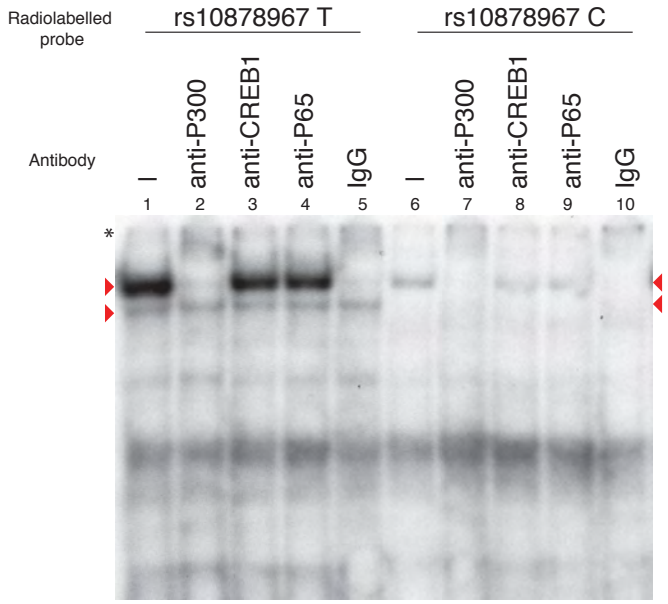


Figure 6.14: EMSA supershift assay on rs10878967 using K562 nuclear extract. IgG; negative control. Anti-P300 causes supershift (*) of the complexes (red arrow) in the T (lane 2) and the C (lane 7) with different signal strength (T>C), however disappearance of the complexes is shown in the IgG controls (lane 5 and 10).

The super shift assay for rs1384 with anti-P300 and anti-CREB1 did not show supershift of bands but a weak higher band appeared on the C allele with anti-P300 and IgG not in the T allele. The bands in the anti-P300 and IgG lanes with the C allele again raise a question about the specificity of the antibody binding for this interaction. Although the supershift assays did not identify proteins binding to the probes in both SNPs, there is potential to determine proteins that bind to one allele but not the other.

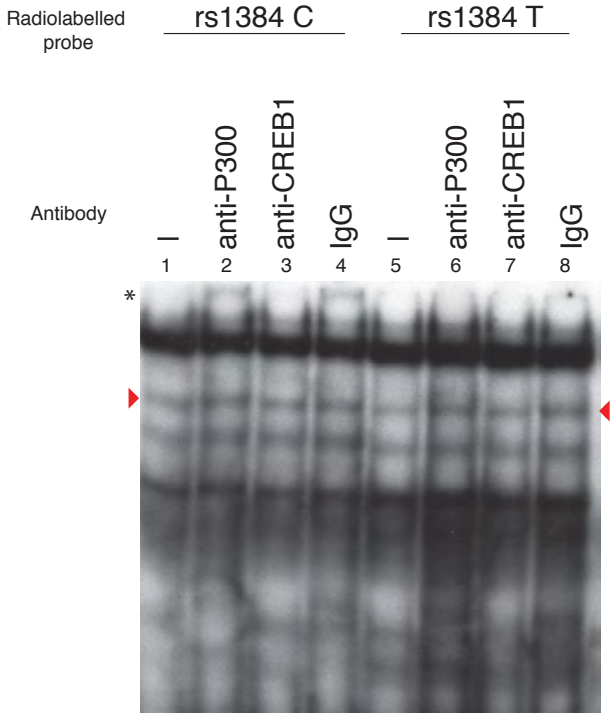


Figure 6.15: EMSA supershift assay on rs1384 using K562 nuclear extract. IgG; negative control. Supershift was observed in both anti-P300 and IgG lanes in the C alleles (*, lane 2,4). However, no clear band disappearance was observed (red arrow).

EMSA assays for the LYZ eSNPs have shown some evidence for different protein binding patterns between two alleles in two SNPs, rs10878967 and rs1384. The competition assay suggests that the ancestral alleles (rs10878967; T, rs1384; C) bind more strongly to proteins than do their derived alleles. The supershift assay was not able to identify proteins binding to the probes in an allele specific manner.

6.5 Discussion

6.5.1 Functional characterisation of GWAS and eQTL studies

The eQTL identified in a cell type specific manner in trans implied a function as a master regulator of genome-wide gene expression. Although the number of associated loci for

disease susceptibility or gene expression has been increasing, functional demonstration of their phenotypic effects has been difficult. In this chapter, one of the master regulators from the monocyte eQTL analysis, 12q15, was investigated in detail in order to try to demonstrate their functionality in an allele specific manner.

The most significant eSNP rs10784774 has shown association in both cis and trans with 73 probes. Protein products of the associated genes include TFs and immunity related proteins although they are not functionally associated in biological pathways as described in the previous chapter. Imputation of the 12q15 greatly improved the resolution of the association especially for *LYZ* and the nature of the LD structure (Figure 6.2). Two imputed SNPs, rs10878967 and rs12812860, were also included for further functional investigation.

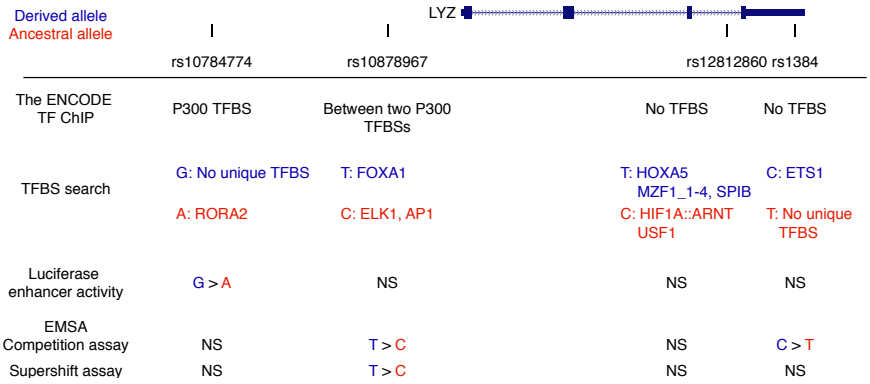


Figure 6.16: A summary of *LYZ* locus functional characterisation on the derived (blue) and ancestral (red) alleles of the eSNPs in the *LYZ* locus. The direction of the allele specificities for luciferase enhancer activity and EMSA are shown. NS - not significant.

6.5.2 Identification of regulatory sequence

Using publicly available data by the ENCODE project, TFBS were searched in the context of the LYZ eSNPs. The promoter SNPs rs10784774 and rs10878967 were identified to overlay or be flanking to p300 TFBS (The SNP positions are shown in Figure 6.16). rs1384 is located in 3'UTR of *LYZ*. 3'UTR of a gene carries regulatory sequences of mRNA, which defines stability of mRNA and provides protein binding sites for post-translational modification. Although the public database of miRNA did not find the binding sequence of known miRNA around rs1384, the SNP may affect stability of *LYZ* mRNA or protein binding for post-translational modification.

A search of TFBS in the *LYZ* eSNPs identified that single nucleotide difference added new TFBS in the locus which included that rs10784774 A allele creates RORA2, a zinc-coordinating nuclear hormone receptor (Jetten, 2004, Figure 6.16). Another example was

rs10878967 which created FOXA1 with ancestral T allele and AP1 with derived C allele, which are involved in epithelial cellular proliferation and cell differentiation (Augello *et al.*, 2011, Shaulian and Karin, 2002). .Experimental confirmation would be required to characterise this further for those TFBS. Each SNP difference probably contributes a small effect, however, coordinated binding of transcription factors to the locus may be required.

6.5.3 Luciferase reporter assay

The luciferase reporter assay is an easy method to determine the enhancer activity of a DNA sequence. The assay was used to test if the regions with the eSNPs had an enhancer activity by comparing the two alleles. The DNA fragments from 367bp to about 2kb containing rs10784774, rs10878967, rs1384 and rs12812860 were investigated for their enhancer activity (Figure 6.16). The 2kb DNA fragment around rs10784774 has shown allelic difference in the enhancer activity in a cell type specific manner. However, the 367bp DNA fragment containing p300 TFBS, the 1.5kb DNA fragment of rs1384 and rs10878967 did not show similar allelic enhancer activity. Only the DNA fragment with rs10878967 showed higher enhancer activity than the control. This may be explained by that region being closer to the *LYZ* start codon with a cluster of TFBS (Figure 6.1) and suggests the region has stronger enhancer activity than the other DNA fragments. Position of the fragment was found to affect enhancer activity in the rs10784774 2kb constructs, which implies the importance of a position of an enhancer from the promoter.

Technical improvement of the luciferase assay is possible to allow sensitive comparison of single nucleotide differences. The reporter plasmid used in this assay contains an SV40 promoter which is constitutively active and may make sensitive comparison between alleles difficult. Using a firefly luciferase reporter with a minimal promoter such as a 165 bp promoter of the p57 gene (Dauphinot *et al.*, 2001) and a late version of Promega pGL4 constructs (Dougherty and Sanders, 2005) may solve this problem. The pGL4 firefly luciferase gene (*luc2*) is modified to have fewer TFBS sites in the sequence and results in sensitive luciferase gene expression by an enhancer of interest and the promoter by reducing protein binding to the *luc2* gene.

6.5.4 EMSA

EMSA is a versatile and very sensitive method for identifying protein-DNA complexes *in vitro*. The radiolabelled probes including each SNP allele were compared by their affinity of binding to proteins by EMSA. Among SNPs associating in cis and trans eQTL in

monocytes, rs10878967 and rs1384 have demonstrated allelic differences in protein-probe complex patterns and the ancestral alleles showed stronger binding than the derived alleles using K562 nuclear extract (Figure 6.16). The two SNPs are located in the promoter and 3'UTR of LYZ implying that the regions may have more potential to modulate TFs and regulatory proteins than rs10784774. EMSA on rs10878967 in two additional nuclear extracts of cell lines showed that the monocytic cell line THP-1 also demonstrated higher affinity with the ancestral allele than with the derived, as well as in the myelogenous leukaemia cell line K562 compared to LCL. This suggests that the eSNPs identified in the monocyte specific eQTL may bind proteins specifically expressed in monocytic cell types.

The supershift assay was performed to identify proteins forming a complex to a probe. P300 and CREB1 which both act as transcription activators were investigated to see if they were bound to the probes. Although rs10878967 showed supershift by the P300 antibody, the negative control with supershift suggests that the binding of P300 to the probe may be in doubt. No supershift on rs1384 was observed. It is difficult to identify proteins that bind to the nucleotide sequence of interest by testing antibodies even though prediction could be made by sequence consensus to TFBS. Additionally a native gel used in this study separated proteins by their charges and hydrodynamic sizes and folding conformation not actual molecular mass, thus identification of actual protein sizes binding to probes is almost impossible. EMSA paired with SDS-PAGE with western detection or mass-spectrometry for protein identification is a recent approach to identify a protein binding to DNA probes (Woo *et al.*, 2002, Stead *et al.*, 2006, Stead and McDowall, 2007). A cost effective modified EMSA called multiplexed competition EMSA (Smith and Humphries, 2009) uses DNA competitors with consensus TFBS sequences, which may be able to screen proteins that bind to the probe sequences. Using those techniques it may be possible to identify specific proteins that bind to the probes in allele specific manner.

The results of the allelic comparison and competitor assays suggest that the ancestral alleles bind with stronger affinity to proteins than do the derived alleles in rs10878967 and rs1384. Protein binding in this experiment is non-equibrillium so quantification of binding complexes is not possible. Quantitative assay of protein binding affinity may be possible using modified EMSA. For example, fluorescent double labelling of probes binding to proteins can be quantify signals specific to each allele (Forwood and Jans, 2006).

EMSA detects *in vitro* condition of DNA-protein binding in an allele specific manner. In order to detect *in vivo* binding in an allele specific manner, ChIP and DNase

hypersensitivity assays combined with other techniques including a WGS would be suitable. For example, genetic variation at the promoter region of *LTA* encoding lymphotoxin-α modulates repression of the gene expression in an allele specific manner (Knight *et al.*, 2004). Knight and colleagues used a haplotype specific ChIP experiment, which uses a heterozygous genotype sample, to demonstrate an allele specific *in vivo* interaction of activated B-cell factor-1 (ABF-1) with the locus using qPCR or MALDI-TOF mass spectrometer. The WGS technology has been widely applied to ChIP and epigenetics studies as well as genome and transcriptome sequencing. Direct sequencing of products from ChIP or DNase hypersensitive assays can identify allelic differences in protein binding sites at a nucleotide scale (Furey, 2012). DNase-seq, a combination of a DNase I hypersensitive site assay and the WGS, can detect protein binding sequences at a nucleotide scale *in vivo*, allowing identification of an allelic difference of the binding site (Hesselberth *et al.*, 2009).

6.5.5 Summary

Functional characterisation of the monocyte eQTL around *LYZ* has been attempted in this chapter by *in silico* sequence analysis, determination of the enhancer activity and protein binding around the eSNPs in an allele specific manner. The assays conducted here investigated each eSNP individually and have not provided significant evidence of allelic differences in gene expression. More sensitive assays may be required to identify enhancer activity in an allele specific manner. A larger region taking account of haplotypic blocks and LD structure may be required to reveal allele specific differences in gene expression. Although the results obtained in this chapter suggest that the techniques used have great potential to characterise eQTL, further *in vivo* experiments may be essential.

7 General Discussion

Since the sequencing of the human genome in 2001, huge advances have been made in the understanding of human genetic variation and its impact on disease susceptibility. A key finding from genome-wide association studies (GWAS) of common diseases has been the importance of regulatory variants modulating gene expression.

This thesis has focused on investigation of genetic variation as a heritable determinant of gene expression with particular interest focus upon the immune system and susceptibility to disease. For this purpose, a large cohort of healthy volunteers of European ancestry were recruited and gene expression quantified for B-cells (a key cell type in the adaptive immune response), monocytes and macrophages (innate immune responses). The investigation of genetic variants associating with gene expression at a population scale was conducted in the context of the immune response and inflammation in relation to disease association.

In this chapter, the overall findings and limitations of this thesis will be assessed in the context of this field of research. Potential experiments and research areas for future work will be considered.

7.1 Findings and conclusions

7.1.1 Differential gene expression analysis

Many studies have used microarray platforms to study differential gene expression analysis. These have identified transcriptome signatures in different cell types and tissues, especially these involved in disease. For example, expression data analyses on tumours were able to distinguish different tumours and subtypes from transcriptome profiling (Ross *et al.*, 2000, Alizadeh *et al.*, 2000). The data described in this thesis has identified distinctive signatures of gene expression across immune cell types by conducting two differential gene expression analyses to compare B-cells and monocytes, and macrophages and monocytes. To the best of my knowledge, the data presented in this thesis may

be the first to describe differential gene expression in these primary cells from healthy individuals on a large scale. A high proportion of differentially expressed genes was expected in the B-cell and monocyte dataset because they are differentiated cell lineages and have distinctive roles in the immune system. However, the high proportion of genes differentially expressed between macrophages and monocytes observed was interesting as they share the same cell lineage as described below.

As the differential gene expression data produced large amounts of information, pathway analysis was a useful tool to annotate genes into groups of biological pathways. This analysis was also used for subsequent eQTL analysis. Gene enrichment analysis showed that monocyte genes were enriched in the pathways related to cellular functions in the immune system including the innate immune system, inflammatory cytokine responses and antigen presentation. By contrast, B-cell genes were enriched in pathways involved in gene expression and RNA metabolism. This is consistent with circulating B-cells being in an unactivated state compared to monocytes until they serve as APCs for downstream antibody production via $CD4^+$ helper T cell activation.

The pathway analysis on the dataset with paired monocytes and differentiated macrophages showed that the monocyte genes were enriched in the pathways of the immune systems. The lipid metabolism pathway was enriched in up-regulated macrophage genes. This result may reflect that inflammatory macrophages are involved in the pathological state of adipose tissues where normal adipose tissue macrophages (ATMs) maintain resolution of inflammation by producing anti-inflammatory cytokines (Lumeng *et al.*, 2007, Murray and Wynn, 2011).

Monocytes and macrophages are remarkably heterogenous even after differentiation into inflammatory or tissue resident types (Gordon and Taylor, 2005). CD14+ monocytes differentiate into dendritic cells as well as macrophages, however their cell fate based on the expression levels of CD markers are not well defined (Geissmann *et al.*, 2010). Commonly used stimuli for the differentiation include IFNγ alone or with TLR ligands (e.g. LPS) for an inflammatory phenotype and IL4 and IL13 for an alternatively activated phenotype (Mantovani *et al.*, 2002, Martinez *et al.*, 2009). This dataset used GM-CSF and pro-inflammatory TNFα as stimuli and treated the differentiated cells as a inflammatory state. The differential gene expression data was used to determine maturation of macrophage differentiation as *CD68*, *LGALS3* and *ITGAX* were significantly up-regulated. However, considering heterogeneity of the cell type, the definition of the macrophages used in this dataset may need care.

7.1.2 eQTL mapping analysis

eQTL mapping analysis is designed to investigate genetic determinants of gene expression using genetic variants as a marker. Cis eQTL were found more frequently than trans because the study was relatively underpowered to detect trans associations. Cis associations were located closer to DNA regulatory sequences as cis eQTL were shown to accumulate near transcription start sites (TSSs). These regions frequently contain DNA regulatory sequences including promoters and enhancers, and typically had a bigger effect size. More than 25% of unique genes tested for the analysis in monocytes and B-cells showed eQTL associations. A lower proportion of genes tested (approximately 10%) associated with SNP alleles in the macrophage and monocyte dataset, probably due to the smaller sample size.

Directional cis eQTL between monocytes and B-cells was an interesting finding. Here the direction of allelic association differed dependent on cell type. The genes involved are related in the immune response including *SELL* encoding a cell adhesion molecule and *DFNA5* encoding a tumour suppressor. This implies genetic variants may regulate gene expression in an opposite way in different cell types.

Various statistical methods to analyse eQTL associations are being developed. The analysis for the first dataset (Chapter 4) was applied to a linear model with permutation (n=1000) for significance tests. For the second dataset (Chapter 5), principal components (PCs) of gene expression were incorporated as covariates in a linear regression model to increase the rate of eQTL discovery. The advantage of incorporating PCs based on gene expression is to remove batch effects generated by microarray experiments and any other experimental artefacts (Liang *et al.*, 2013). The macrophage dataset used the greater number of PCs to maximise the detection of eQTL compared to the monocytes. This implies that other strong non-genetic components in the data were present, which include the effect of the incubation step for differentiation. Further investigation of the PCs used for the analysis may identify a specific PC that relates to the incubation effect as well as identify PCs with genetic determinants.

cis eQTL

Comparison of cis eQTL between different or similar cell types can determine cell type specificity. A high proportion of cell type specific eQTL were found when comparing B-cells and monocytes (78%) and also between macrophages and monocytes (68%). It is likely that many more cell type specific eQTL were identified in this dataset than other studies due to the high purity of the cell populations in the samples analysed. Moreover,

the cells were not confounded by EBV transformation, which emphasises an advantage of primary cells in this type of analysis compared to data from lymphoblastoid cell lines (LCLs).

eQTL mapping analysis in cis in the macrophage dataset has identified a large proportion of cell type specific associations in comparison to the deriving monocytes. This implies many eQTL may be absent in an unstimulated state. Activation of cells may result in uncovering hidden eQTL associations. It is logical to think that cell activation markedly changes gene expression levels, which lead to identification of macrophage specific eQTL. This may be secondary association to new genes, or altered chromatin usage and regulatory factor binding in an activated state.

eQTL associated genes were found to be related to roles in the immune responses and inflammation. This suggests that these immune responses are affected by genetic variation, which also implies connection to susceptibility associated loci in common and autoimmune diseases.

trans eQTL

In contrast to cis eQTL, analysis showed a higher proportion of cell type specificity in trans than cis (92.4% compared to 78% in the monocytes and B-cells eQTL). This difference suggests intermediate associations between trans eQTL and gene expression unique to each cell type.

Trans eQTL were found to have smaller effect sizes than cis in the B-cell and monocyte dataset. This may be reflecting secondary associations driven by cis associations. A larger study cohort may be needed in order to capture more trans association. Therefore, the macrophage and monocyte eQTL analysis (Chapter 5) considered only the cis eQTL analysis due to its small sample size.

Trans eQTL analysis identified master regulatory trans eQTL in a cell type specific manner. Other eQTL studies identified similar associations (Small *et al.*, 2011, Fehrmann *et al.*, 2011). These loci are also associated in cis to local genes, however they may not encode transcription factors (TFs). The monocyte specific master regulatory trans eQTL in 12q15 is associated with *LYZ* in cis. Lysozyme, encoded by *LYZ*, performs an important role as an initial defence for the innate immune system, where it functions to hydrolyse bacterial cell walls. By contrast, a B-cell specific master regulatory trans eQTL in 9q21 is associated with *KLF4* encoding a Kruppel-like factor 4 transcription activator and

repressor. The genes involved in these trans associations were not enriched in pathway analysis, therefore intermediate associations may be hidden.

The *LYZ* locus at 12q15 has retained cis eQTL association in macrophages. This implies the locus genetically influences cell activity or homeostasis after differentiation from monocytes. A preliminary trans analysis of the *LYZ* locus in macrophages suggests same genes are associated as monocytes including *CREB1* and *KILLIN*. However, the analysis requires further validation due to the small effect and sample size.

In order to identify the association between a cis gene in a master trans locus and trans associated genes, this study attempted to determine relationships between trans eQTL and both cis and trans associated genes involved by pathway analysis and correlation of allelic gene expression. Most significant trans genes were highly correlated in their expression with cis genes in an allele specific manner at both 12q15 in monocytes and 9p21 in B-cells.

GWAS and eQTL

One of the aims for eQTL analysis is to identify functional association between disease susceptibility associated variants and gene expression. Many new loci identified in GWAS explain relatively small susceptibility to diseases. This missing heritability in GWAS may be partially explained at transcriptional and epigenetic levels. SNP markers defining GWAS loci associated with susceptibility to autoimmune diseases were found in the eQTL datasets. Some of these autoimmune disease associated loci are shared between diseases as also shown by Fehrmann and colleagues (Fehrmann *et al.*, 2011).

A cis eQTL association with a GWAS locus provides evidence of a likely modulated gene and a functional mechanism of disease susceptibility. For example, shared cis eQTL with *ERAP2* was associated with inflammatory diseases including Crohn's disease and AS. Antigen presentation by HLA class I molecules requires peptide trimming by amino peptidases encoded by *ERAP2* and *ERAP1* in the endoplasmic reticulum (Saveanu *et al.*, 2005). The allelic specific expression of *ERAP2* may be involved in autoimmune dysregulation during antigen presentation in the MHC class I molecules (Andres *et al.*, 2010). Cell type specific eQTL identified in GWAS loci may specify tissue specific pathways involved in disease susceptibility. For example, cis eQTL with *CARD9* in monocytes was associated with AS, UC and Crohn's disease. *CARD9* is involved in the immune response against fungal infection and provides a connection between the innate immune response and the adaptive immune response mediated by $CD4^+$ helper T cells (Glocker *et al.*, 2009, LeibundGut-Landmann *et al.*, 2007).

Macrophages are integral to dysregulation of inflammation which leads to inflammatory autoimmune diseases (Olefsky and Glass, 2010, Murray and Wynn, 2011). Therefore, macrophage eQTL studies may help to find a connection between autoimmune disease susceptibility associated loci identified by GWAS and the genes involved. cis eQTL with *PPARG* is an interesting finding to follow up as the gene is involved in the body fat metabolism and T2D (Deeb *et al.*, 1998, Altshuler *et al.*, 2000). *PPARG* encodes peroxisome proliferator-activated receptor γ and a transcription factor important for adipogenesis (Spiegelman, 1998). Decreased expression of *PPARG* which is associated with a genetic variant is involved in insulin sensitivity and regulation of macrophage activation (Odegaard *et al.*, 2007).

Trans eQTL were also identified in GWAS disease susceptibility loci. The monocyte specific trans eQTL in the MHC class II region to *AOAH* was significantly associated with disease susceptibility associated loci to IBD, T1D and RA. The loci encode many genes involved in antigen presentation and are highly associated with many autoimmune diseases (Jones *et al.*, 2006). A precise mechanism between SNP alleles in the MHC class II and trans associated genes is limited to speculation because further sequencing validation may be required due to their extreme polymorphism. However, acyloxyacyl hydrolase encoded by *AOAH* was shown to control B-cell polyclonal antibody production by limiting LPS derived from Gram-negative bacteria (Lu *et al.*, 2005). The expression level of *AOAH* may be involved in controlling autoimmune responses.

HLA

The HLA alleles are the most variable in the human genome and have been associated with many autoimmune and infectious diseases. A SNP in an MHC class II molecule gene may change its antigen preference as shown in its association to RA susceptibility (Gregersen *et al.*, 1987, Hammer *et al.*, 1995, Raychaudhuri *et al.*, 2012). The association analysis between the imputed HLA alleles and gene expression was performed to investigate how HLA alleles affect gene expression and the mechanism of disease susceptibility in association to HLA alleles. The analysis identified that *HLA-DQA1* and *HLA-DRB1* were associated with monocyte *AOAH* expression in trans and further the trans eQTL SNP alleles to the gene were associated with particular HLA alleles, namely *HLA-DRB1*04*, **07* and **09*. As mentioned above, *AOAH* is a good candidate to investigate further for its involvement in autoimmune response.

By contrast, B-cell genes associated with HLA alleles were only found as shared with monocytes. This may be relate to activation state. *DEF8* associations to *HLA-DQA1* and *HLA-DRB1* were found in both monocytes and B-cells in a similar manner with the trans eQTL SNP alleles, although the function of the gene product is unclear.

The cis associations between the HLA alleles and the MHC genes were limited in this thesis because the extreme polymorphism in the MHC region needs careful validation of genotyping data as well as gene expression. Quantification using microarray may be confounded by underlying SNP variation. The B-cell expression of *PSORS1C3*, a susceptibility locus to psoriasis, was found to associate with the MHC class I *HLA-C* and *HLA-B* alleles where psoriasis susceptibility is associated (Trembath *et al.*, 1997, Nestle *et al.*, 2009, Strange *et al.*, 2010). This suggests specific HLA alleles are involved in *PSORS1C3* expression and psoriasis susceptibility, however more careful validation of the data is required.

7.1.3 Functional characterisation of eQTL

Functional characterisation of eQTL was attempted, focusing on the *LYZ* locus. Using a luciferase reporter assay and EMSA, an allelic difference was experimentally determined. Two SNPs in complete LD showed allelic difference in the separate experiments, however it was difficult to conclude that the sequences containing these SNPs function as regulatory elements. An *in silico* search for a functional sequence using the JASPAR database for transcription factor binding site (TFBS) search and the ENCODE ChIP-seq dataset for TF binding has shown these SNPs may be involved in transcription regulation. However, it is difficult to experimentally demonstrate the difference caused by a single nucleotide change because its effect size may be too small to detect *in vitro*.

7.2 Future work

This work has shown that genetic variants are associated with gene expression in a cell type specific manner. Genetic variants with association to common and autoimmune disease susceptibility are found to also show association with these regulatory variants and may help to understand mechanisms of disease susceptibility and development. In light of these findings, this section describes general and specific limitations and future directions generated from this work.

7.2.1 Analysis methods

Many statistical methods for eQTL mapping analysis are published, however it is difficult to find the best one. This may be because genome-wide eQTL analysis has been developed in just the past decade. Researchers may need to try different methods to analyse their datasets as each study population may have different characteristics, such as sample size, cell and tissue types, or the genetic background of the cohort studied. An initial eQTL study using three cell types showed 78% cell type specificity (Dimas *et al.*, 2009). However, reanalysis using a model which takes into account active and inactive eQTL in different cell types discovered that only 8% was cell type specific (Flutre *et al.*, 2013). This study dataset was smaller (n=75) and used LCLs as well as primary T cells and fibroblasts. It would be interesting to apply this method to analyse the volunteer dataset to see how many cell type specific eQTL are retained considering it has the larger sample size (n=283) and the high purity of the primary cell types.

Sample size was a limiting factor in the macrophage population analysis even though the data for the cis eQTL associations has resolved immunologically interesting genes. Expanding the size of population for macrophages as well as monocytes would be important to investigate eQTL analysis in the innate immune system including trans association.

Macrophages consist of a heterogenous cell population and two macrophage types, classical inflammatory and tissue specific regulatory macrophages, with separate roles in inflammation and autoimmunity (Geissmann *et al.*, 2010). Development of insulin resistance is involved with inflammatory macrophages where ATMs maintain tissue homeostasis (Olefsky and Glass, 2010). The deficiency of *PPARG* in mice impairs differentiation of alternative macrophage for tissue homeostasis and results in predisposition to obesity and insulin resistance (Odegaard *et al.*, 2007). PPARγ2, one of two isoforms of *PPARG*, with Pro12Ala substitution, causes reduced protein binding affinity in the promoter region and is associated with improved insulin sensitivity and lower BMI (Deeb *et al.*, 1998, Altshuler *et al.*, 2000). As a cis eQTL association with macrophage *PPARG* expression was shown in this study, it would be interesting to investigate eQTL mapping in ATMs in comparison to inflammatory macrophages. This association was not found in the previous eQTL studies which used adipose tissues (Emilsson *et al.*, 2008, Grundberg *et al.*, 2012). eQTL mapping analysis using purified cell types may have an advantage to identify cell type specific eQTL.

Fine mapping of eQTL can be performed using imputated SNPs. The *LYZ* locus, which was imputed using the IMPUTE2 program, showed a cleaner LD structure and associations were found on the promoter region. Genome-wide imputation has become a standard step for GWAS (Marchini and Howie, 2010). It is possible to identify more significant eQTL associations with imputed genotype data even though genotyping chips have carefully selected tagging SNPs to cover most common SNPs (Liang *et al.*, 2013). In addition, a haplotype based eQTL mapping analysis taking LD structure into account may help to identify functional regulatory variants. A combination of whole genome sequencing (WGS) and RNA-seq that sequence non-coding regions in depth and quantifies allelic expression simultaneously may be able to identify new polymorphisms and regulatory variants (Pastinen, 2010).

The ENCODE project archived a comprehensive dataset of epigenetics using TF ChIP-seq and DNaseI sensitivity assays for chromatin structure and modification. These experimental data were generated using WGS and can be used to extract information for allelic imbalance in TFBSs (Park, 2009, Pastinen, 2010). Combining these epigenetic markers onto eQTL associations may identify further loci that affect gene expression in an allele specific manner.

7.2.2 HLA association

HLA alleles are renowned for their involvement in autoimmune diseases. Association between the HLA alleles and gene expression was found in this dataset. Using this information as a reference, the expression level of *AOAH* could be investigated in a disease cohort such as T1D and RA. The *HLA-DQA1* and *HLA-DRB1* alleles are known to associate with susceptibility to these diseases (Lee *et al.*, 2001, Gregersen *et al.*, 1987, Hammer *et al.*, 1995). A role of *AOAH* in regulation of endogenous antibody response to B-cell polyclonal production may suggest its involvement in autoimmunity and the adaptive immune system (Lu *et al.*, 2005).

Cis association analysis with genes in the MHC region is interesting, however it is challenging to precisely quantify gene expression in the MHC region and data needs validation. A microarray platform uses DNA probes which often cannot avoid confounding in probe designing with SNPs and other polymorphisms in some regions of the MHC. A haplotype based exon array which has combination of tiling and splice junction probes was designed to overcome this problem (Vandiedonck *et al.*, 2011). RNA-

seq may also overcome this probe design problem, however still needs caution in mapping short sequencing reads to a hyperpolymorphic region (McKenna *et al.*, 2010).

7.2.3 Mechanism behind trans eQTL association

Trans eQTL associations are difficult to identify due to their small effect size compared to cis and are poorly replicated (Petretto *et al.*, 2006). A larger study population size may help to identify trans association. A meta-analysis of eQTL datasets with primary cell types (n=5311 and replication n=2775) identified trans association with *IKZF1* to a locus associated with SLE susceptibility (Westra *et al.*, 2013). The macrophage dataset was not large enough to investigate trans eQTL association even though stimulated samples which undergo a global gene expression change may uncover hidden trans association (Gaffney, 2013). A large cohort (n=414) consisting of $CD14^+$ monocytes stimulated with a TLR4 agonist LPS and a pro-inflammatory cytokine IFNγ has been collected and analysed for eQTL mapping by Benjamin Fairfax, myself and colleagues (manuscript in preparation). The data from this analysis is enriched with newly identified eQTL specific to stimulated samples including multiple master regulatory trans eQTL. Interestingly, trans eQTL in the *LYZ* locus was found in all conditions. The effect size of this eQTL in macrophages was smaller than that in monocytes. This new large study could assess their effect sizes compared to unstimulated samples to see if a stimulated state is involving for examining a different set of transcriptional machinery.

The genes involved in trans eQTL may interact directly with each other or indirectly. To find intermediate associations, a pathway type analysis may predict upstream or downstream genes that were not identified by eQTL mapping. A pathway analysis may help to identify candidate genes that may interact with genes associated in trans.

7.2.4 Disease specific eQTL analysis

Organ specific symptoms are shown in some autoimmune diseases. Therefore some eQTL analyses of disease associated variants have focused on using the primary cells or tissues involved. For example, psoriasis study used skin, the LDL cholesterol levels and CHD study used liver tissue, and obesity study used adipose tissue (Ding *et al.*, 2010, Musunuru *et al.*, 2010, Emilsson *et al.*, 2008). As this work has shown a high frequency of cell type specific eQTL association, studying disease with a disease associated cell type may be logical. However, accessibility to appropriate cell or tissue samples for specific eQTL analysis is sometimes limited.

Many GWAS for disease have focused on European populations. Some disease associated loci have been identified across different populations. For example, psoriasis with the *HLA-C* locus in the Chinese population, Crohn's disease with MHC class II loci in the Japanese and Ashkenazi Jewish populations and SLE with the *BLK* locus in the Chinese population (Zhang *et al.*, 2009, Yamazaki *et al.*, 2013, Kenny *et al.*, 2012, Han *et al.*, 2009). eQTL analysis on diverse populations reported that 34% of cis eQTL were shared in at least two populations (Stranger *et al.*, 2012). Different prevalence of some diseases are found across populations, where environmental factors including diet are thought to be involved in differing disease susceptibility. However, cell type specific eQTL mapping across diverse populations may find population specific loci which may associate with disease susceptibility loci of GWAS.

7.2.5 Identification of regulatory variation

It is known that single nucleotide substitutions in regulatory elements may be enough to change protein binding or susceptibility to chromatin modification and methylation. Experimental characterisation for eQTL is often difficult due to the small effect size. The methods used experimentally in this thesis were possibly not sensitive enough to detect the allelic effect.

Enhancer activity can be measured *in vivo* using reporter genes including a commonly used green fluorescent protein (GFP) and *lucZ* in zebrafish and mouse embryos. Enhancer elements are relatively conserved throughout mammalian species and some are ultraconserved up to teleost fish (Visel *et al.*, 2008). The effect of single nucleotide changes to these highly conserved regulatory elements may be visible using these reporter assay systems. Synthetic sequences are systematically tested to predict regulatory elements as well as assess the impact of copy number, position and functional combination with other TFBSs (Smith *et al.*, 2013b, Smith *et al.*, 2013a). They reported heterotypic enhancers are more effective for reporter gene expression than homotypic ones. A composite enhancer element consisting of short eQTL fragments containing rs10784774, rs10878967 and rs1384 can be constructed for each haplotype in the *LYZ* locus. Combined TFBSs in a specific haplotype may show stronger enhancer activity than testing each small eQTL.

7.3 Concluding remarks

GWAS for common and autoimmune diseases have identified hundreds of susceptibility associated loci in non-coding genetic variation in the past decade. However, the specific modulated genes and functional mechanism often remain under speculation. Missing heritability in GWAS can be partially explained at transcriptional and epigenetic levels. The work presented in this thesis aimed to investigate genetic variants in the context of gene expression regulation in immunologically important cell types. For this aim, highly purified primary human immune cell types from healthy individuals of European ancestry were investigated at a population scale. The gene expression profile obtained from these samples was analysed to identify expression signatures which were related to biological roles to each cell type. The eQTL mapping analysis was conducted with the genotyping data to identify genetic regulatory variants in a cell type specific manner. The analysis identified trans master regulatory loci as well as a high proportion of cell type specific eQTL, which were not identified in studies using immortalised cell lines. In addition, these eQTL were identified in common and autoimmune disease susceptibility loci. Functional characterisation for eQTL suggested allele specific effects *in vivo*. These findings have clearly shown genetic variants as a determinant of gene expression and their association to disease susceptibility and mechanisms. Further investigation may help to elucidate disease susceptibility and functional understanding of the impact of genetic variation on gene expression.

Appendix

Table 1: Reactome pathway enrichment analysis on genes significantly differentially expressed in B-cells. GeneRatio: over-represented genes in genes tested in each cell type. BgRatio: a ratio of genes in a pathway and background. qvalue: adjusted p-value by FDR.

ID	Description	GeneRatio	BgRatio	pvalue	qvalue
74160	Gene Expression	423/5654	785/13397	$7.63x10^{-12}$	$7.44x10^{-9}$
156842	Eukaryotic Translation Elongation	95/5654	134/13397	$1.41x10^{-11}$	$7.44x10^{-9}$
72764	Eukaryotic Translation Termination	91/5654	130/13397	$1.15x10^{-10}$	$1.52x10^{-8}$
156902	Peptide chain elongation	91/5654	130/13397	$1.15x10^{-10}$	$1.52x10^{-8}$
975956	Nonsense Mediated Decay Independent of the Exon Junction Complex	92/5654	135/13397	$8.73x10^{-10}$	$1.02x10^{-7}$
72689	Formation of a pool of free 40S subunits	94/5654	139/13397	$1.08x10^{-9}$	$1.14x10^{-7}$
927802	Nonsense Mediated Decay	100/5654	152/13397	$3.23x10^{-9}$	$2.84x10^{-7}$
975957	Nonsense Mediated Decay Enhanced by the Exon Junction Complex	100/5654	152/13397	$3.23x10^{-9}$	$2.84x10^{-7}$
212436	Generic Transcription Pathway	144/5654	238/13397	$7.63x10^{-9}$	$6.19x10^{-7}$
72706	GTP hydrolysis and joining of the 60S ribosomal subunit	98/5654	151/13397	$1.34x10^{-8}$	$1.01x10^{-6}$
157279	3' -UTR-mediated translational regulation	97/5654	150/13397	$2.09x10^{-8}$	$1.37x10^{-6}$
156827	L13a-mediated translational silencing of Ceruloplasmin expression	97/5654	150/13397	$2.09x10^{-8}$	$1.37x10^{-6}$
1640170	Cell Cycle	191/5654	336/13397	$3.22x10^{-8}$	$2.00x10^{-6}$
73886	Chromosome Maintenance	52/5654	70/13397	$4.84x10^{-8}$	$2.84x10^{-6}$
1799339	SRP-dependent cotranslational protein targeting to membrane	97/5654	154/13397	$1.42x10^{-7}$	$7.70x10^{-6}$
72737	Cap-dependent Translation Initiation	99/5654	158/13397	$1.53x10^{-7}$	$7.70x10^{-6}$
72613	Eukaryotic Translation Initiation	99/5654	158/13397	$1.53x10^{-7}$	$7.70x10^{-6}$
69278	Cell Cycle, Mitotic	160/5654	287/13397	$2.09x10^{-6}$	$1.00x10^{-4}$
72766	Translation	112/5654	192/13397	$4.29x10^{-6}$	$1.97x10^{-4}$
1538133	G0 and Early G1	17/5654	18/13397	$4.57x10^{-6}$	$2.01x10^{-4}$
69190	DNA strand elongation	22/5654	26/13397	$1.07x10^{-5}$	$4.51x10^{-4}$
157579	Telomere Maintenance	31/5654	43/13397	$6.86x10^{-5}$	$2.78x10^{-3}$
69481	G2/M Checkpoints	26/5654	35/13397	$1.18x10^{-4}$	$4.53x10^{-3}$
532600	Metabolism of RNA	147/5654	276/13397	$1.20x10^{-4}$	$4.53x10^{-3}$
68962	Activation of the pre-replicative complex	20/5654	25/13397	$1.30x10^{-4}$	$4.74x10^{-3}$
447010	Metabolism of mRNA	138/5654	258/13397	$1.49x10^{-4}$	$5.23x10^{-3}$

Table 2: Reactome pathway enrichment analysis on genes significantly differentially expressed in monocytes. GeneRatio: over-represented genes in genes tested in each cell type. BgRatio: a ratio of genes in a pathway and background. qvalue: adjusted p-value by FDR.

ID	Description	GeneRatio	BgRatio	pvalue	qvalue
1430728	Metabolism	508/4725	932/13397	$1.29x10^{-35}$	$1.07x10^{-32}$
168256	Immune System	413/4725	781/13397	$2.57x10^{-25}$	$1.06x10^{-22}$
168249	Innate Immune System	167/4725	289/13397	$2.66x10^{-15}$	$7.31x10^{-13}$
162582	Signal Transduction	362/4725	741/13397	$4.04x10^{-15}$	$8.33x10^{-13}$
109582	Hemostasis	176/4725	317/13397	$7.78x10^{-14}$	$1.28x10^{-11}$
1236975	Antigen processing-Cross presentation	55/4725	70/13397	$1.36x10^{-13}$	$1.62x10^{-11}$
1643685	Disease	320/4725	654/13397	$1.37x10^{-13}$	$1.62x10^{-11}$
199991	Membrane Trafficking	97/4725	153/13397	$1.03x10^{-12}$	$1.06x10^{-10}$
1428517	The citric acid (TCA) cycle and respiratory electron transport	75/4725	112/13397	$7.41x10^{-12}$	$6.79x10^{-10}$
556833	Metabolism of lipids and lipoproteins	167/4725	310/13397	$1.00x10^{-11}$	$8.25x10^{-10}$
1280218	Adaptive Immune System	238/4725	477/13397	$1.84x10^{-11}$	$1.38x10^{-9}$
1236978	Cross-presentation of soluble exogenous antigens (endosomes)	37/4725	45/13397	$1.25x10^{-10}$	$8.59x10^{-9}$
163200	Respiratory electron transport, ATP synthesis by chemiosmotic coupling, and heat production by uncoupling proteins.	55/4725	78/13397	$2.26x10^{-10}$	$1.44x10^{-8}$
109581	Apoptosis	77/4725	123/13397	$5.33x10^{-10}$	$3.14x10^{-8}$
1483257	Phospholipid metabolism	86/4725	142/13397	$6.09x10^{-10}$	$3.34x10^{-8}$
76002	Platelet activation, signaling and aggregation	89/4725	149/13397	$8.41x10^{-10}$	$4.33x10^{-8}$
168898	Toll Receptor Cascades	65/4725	100/13397	$1.23x10^{-9}$	$5.98x10^{-8}$
1236974	ER-Phagosome pathway	43/4725	58/13397	$1.70x10^{-9}$	$7.78x10^{-8}$
211733	Regulation of activated PAK-2p34 by proteasome mediated degradation	36/4725	46/13397	$2.94x10^{-9}$	$1.27x10^{-7}$
422475	Axon guidance	92/4725	159/13397	$4.33x10^{-9}$	$1.79x10^{-7}$
180534	Vpu mediated degradation of CD4	36/4725	47/13397	$8.28x10^{-9}$	$3.10x10^{-7}$
350562	Regulation of ornithine decarboxylase (ODC)	36/4725	47/13397	$8.28x10^{-9}$	$3.10x10^{-7}$
450408	Destabilization of mRNA by AUF1 (hnRNP D0)	38/4725	51/13397	$1.20x10^{-8}$	$4.31x10^{-7}$
1266738	Developmental Biology	121/4725	227/13397	$1.61x10^{-8}$	$5.52x10^{-7}$
69017	CDK-mediated phosphorylation and removal of Cdc6	35/4725	46/13397	$1.82x10^{-8}$	$5.98x10^{-7}$
1169091	Activation of NF-kappaB in B Cells	44/4725	63/13397	$2.30x10^{-8}$	$7.29x10^{-7}$
71387	Metabolism of carbohydrates	95/4725	171/13397	$4.02x10^{-8}$	$1.23x10^{-6}$
983169	Class I MHC mediated antigen processing & presentation	114/4725	214/13397	$4.34x10^{-8}$	$1.28x10^{-6}$
75815	Ubiquitin-dependent degradation of Cyclin D	35/4725	47/13397	$4.69x10^{-8}$	$1.29x10^{-6}$
69229	Ubiquitin-dependent degradation of Cyclin D1	35/4725	47/13397	$4.69x10^{-8}$	$1.29x10^{-6}$
611105	Respiratory electron transport	44/4725	64/13397	$4.85x10^{-8}$	$1.29x10^{-6}$
349425	Autodegradation of the E3 ubiquitin ligase COP1	36/4725	49/13397	$5.44x10^{-8}$	$1.33x10^{-6}$
166054	Activated TLR4 signalling	53/4725	82/13397	$5.49x10^{-8}$	$1.33x10^{-6}$
166016	Toll Like Receptor 4 (TLR4) Cascade	53/4725	82/13397	$5.49x10^{-8}$	$1.33x10^{-6}$
169911	Regulation of Apoptosis	38/4725	53/13397	$6.87x10^{-8}$	$1.57x10^{-6}$
195721	Signaling by Wnt	42/4725	61/13397	$9.15x10^{-8}$	$1.99x10^{-6}$
195253	Degradation of beta-catenin by the destruction complex	42/4725	61/13397	$9.15x10^{-8}$	$1.99x10^{-6}$

69613	p53-Independent G1/S DNA damage checkpoint	35/4725	48/13397	$1.14x10^{-7}$	$2.25x10^{-6}$
69610	p53-Independent DNA Damage Response	35/4725	48/13397	$1.14x10^{-7}$	$2.25x10^{-6}$
69601	Ubiquitin Mediated Degradation of Phosphorylated Cdc25A	35/4725	48/13397	$1.14x10^{-7}$	$2.25x10^{-6}$
1168372	Downstream Signaling Events Of B Cell Receptor (BCR)	71/4725	121/13397	$1.15x10^{-7}$	$2.25x10^{-6}$
69541	Stabilization of p53	36/4725	50/13397	$1.28x10^{-7}$	$2.46x10^{-6}$
76005	Response to elevated platelet cytosolic Ca2+	41/4725	60/13397	$1.81x10^{-7}$	$3.39x10^{-6}$
166520	Signalling by NGF	104/4725	196/13397	$2.17x10^{-7}$	$3.98x10^{-6}$
174113	SCF-beta-TrCP mediated degradation of Emi1	35/4725	49/13397	$2.65x10^{-7}$	$4.74x10^{-6}$
187037	NGF signalling via TRKA from the plasma membrane	74/4725	130/13397	$3.41x10^{-7}$	$5.98x10^{-6}$
1280215	Cytokine Signaling in Immune system	116/4725	226/13397	$4.48x10^{-7}$	$7.69x10^{-6}$
428157	Sphingolipid metabolism	33/4725	46/13397	$4.96x10^{-7}$	$8.34x10^{-6}$
450531	Regulation of mRNA Stability by Proteins that Bind AU-rich Elements	52/4725	84/13397	$5.58x10^{-7}$	$9.20x10^{-6}$
69563	p53-Dependent G1 DNA Damage Response	37/4725	54/13397	$6.48x10^{-7}$	$1.03x10^{-5}$
69580	p53-Dependent G1/S DNA damage checkpoint	37/4725	54/13397	$6.48x10^{-7}$	$1.03x10^{-5}$
114608	Platelet degranulation	38/4725	56/13397	$6.71x10^{-7}$	$1.04x10^{-5}$
162909	Host Interactions of HIV factors	67/4725	117/13397	$8.99x10^{-7}$	$1.37x10^{-5}$
373755	Semaphorin interactions	34/4725	49/13397	$1.17x10^{-6}$	$1.75x10^{-5}$
69615	G1/S DNA Damage Checkpoints	37/4725	55/13397	$1.31x10^{-6}$	$1.92x10^{-5}$
1222352	Latent infection of Homo sapiens with Mycobacterium tuberculosis	20/4725	24/13397	$1.80x10^{-6}$	$2.56x10^{-5}$
1222556	Phagosomal maturation (early endosomal stage)	20/4725	24/13397	$1.80x10^{-6}$	$2.56x10^{-5}$
166058	MyD88:Mal cascade initiated on plasma membrane	45/4725	72/13397	$2.19x10^{-6}$	$2.91x10^{-5}$
181438	Toll Like Receptor 2 (TLR2) Cascade	45/4725	72/13397	$2.19x10^{-6}$	$2.91x10^{-5}$
168179	Toll Like Receptor TLR1:TLR2 Cascade	45/4725	72/13397	$2.19x10^{-6}$	$2.91x10^{-5}$
168188	Toll Like Receptor TLR6:TLR2 Cascade	45/4725	72/13397	$2.19x10^{-6}$	$2.91x10^{-5}$
166166	MyD88-independent cascade	43/4725	68/13397	$2.34x10^{-6}$	$3.01x10^{-5}$
168164	Toll Like Receptor 3 (TLR3) Cascade	43/4725	68/13397	$2.34x10^{-6}$	$3.01x10^{-5}$
68827	CDT1 association with the CDC6:ORC:origin complex	35/4725	52/13397	$2.48x10^{-6}$	$3.13x10^{-5}$
187577	SCF(Skp2)-mediated degradation of p27/p21	36/4725	54/13397	$2.52x10^{-6}$	$3.13x10^{-5}$
174084	Autodegradation of Cdh1 by Cdh1:APC/C	37/4725	56/13397	$2.55x10^{-6}$	$3.13x10^{-5}$
381150	Diabetes pathways	52/4725	87/13397	$2.58x10^{-6}$	$3.13x10^{-5}$
382551	Transmembrane transport of small molecules	108/4725	215/13397	$4.08x10^{-6}$	$4.87x10^{-5}$
1660662	Glycosphingolipid metabolism	19/4725	23/13397	$4.29x10^{-6}$	$5.05x10^{-5}$
937061	NFkB and MAP kinases activation mediated by TLR4 signaling repertoire	40/4725	63/13397	$4.52x10^{-6}$	$5.25x10^{-5}$
983705	Signaling by the B Cell Receptor (BCR)	78/4725	146/13397	$4.90x10^{-6}$	$5.60x10^{-5}$

975155	MyD88 dependent cascade initiated on endosome	43/4725	70/13397	$7.01x10^{-6}$	$7.80x10^{-5}$
168181	Toll Like Receptor 7/8 (TLR7/8) Cascade	43/4725	70/13397	$7.01x10^{-6}$	$7.80x10^{-5}$
69202	Cyclin E associated events during G1/S transition	37/4725	58/13397	$8.75x10^{-6}$	$9.62x10^{-5}$
177929	Signaling by EGFR	62/4725	112/13397	$1.02x10^{-5}$	$1.11x10^{-4}$
975138	TRAF6 mediated induction of NFkB and MAP kinases upon TLR7/8 or 9 activation	42/4725	69/13397	$1.24x10^{-5}$	$1.33x10^{-4}$
437239	Recycling pathway of L1	24/4725	33/13397	$1.27x10^{-5}$	$1.34x10^{-4}$
174178	APC/C:Cdh1 mediated degradation of Cdc20 and other APC/C:Cdh1 targeted proteins in late mitosis/early G1	38/4725	61/13397	$1.49x10^{-5}$	$1.56x10^{-4}$
174154	APC/C:Cdc20 mediated degradation of Securin	37/4725	59/13397	$1.55x10^{-5}$	$1.58x10^{-4}$
69656	Cyclin A:Cdk2-associated events at S phase entry	37/4725	59/13397	$1.55x10^{-5}$	$1.58x10^{-4}$
168138	Toll Like Receptor 9 (TLR9) Cascade	43/4725	72/13397	$1.92x10^{-5}$	$1.93x10^{-4}$
1433557	Signaling by SCF-KIT	49/4725	85/13397	$2.00x10^{-5}$	$1.98x10^{-4}$
917937	Iron uptake and transport	21/4725	28/13397	$2.08x10^{-5}$	$2.04x10^{-4}$
1643713	Signaling by EGFR in Cancer	62/4725	114/13397	$2.16x10^{-5}$	$2.09x10^{-4}$
71291	Metabolism of amino acids and derivatives	72/4725	138/13397	$3.22x10^{-5}$	$3.05x10^{-4}$
449147	Signaling by Interleukins	53/4725	95/13397	$3.29x10^{-5}$	$3.07x10^{-4}$
381070	Activation of Chaperones by IRE1alpha	30/4725	46/13397	$3.32x10^{-5}$	$3.07x10^{-4}$
975871	MyD88 cascade initiated on plasma membrane	40/4725	67/13397	$3.77x10^{-5}$	$3.37x10^{-4}$
168142	Toll Like Receptor 10 (TLR10) Cascade	40/4725	67/13397	$3.77x10^{-5}$	$3.37x10^{-4}$
390918	Peroxisomal lipid metabolism	15/4725	18/13397	$3.94x10^{-5}$	$3.49x10^{-4}$
202733	Cell surface interactions at the vascular wall	39/4725	65/13397	$4.01x10^{-5}$	$3.51x10^{-4}$
199992	trans-Golgi Network Vesicle Budding	32/4725	51/13397	$5.71x10^{-5}$	$4.90x10^{-4}$
421837	Clathrin derived vesicle budding	32/4725	51/13397	$5.71x10^{-5}$	$4.90x10^{-4}$
176814	Activation of APC/C and APC/C:Cdc20 mediated degradation of mitotic proteins	37/4725	62/13397	$7.39x10^{-5}$	$6.15x10^{-4}$
176409	APC/C:Cdc20 mediated degradation of mitotic proteins	37/4725	62/13397	$7.39x10^{-5}$	$6.15x10^{-4}$
174184	Cdc20:Phospho-APC/C mediated degradation of Cyclin A	37/4725	62/13397	$7.39x10^{-5}$	$6.15x10^{-4}$
381119	Unfolded Protein Response	36/4725	60/13397	$7.87x10^{-5}$	$6.49x10^{-4}$
68867	Assembly of the pre-replicative complex	35/4725	58/13397	$8.36x10^{-5}$	$6.82x10^{-4}$
1445148	Translocation of GLUT4 to the Plasma Membrane	33/4725	54/13397	$9.35x10^{-5}$	$7.53x10^{-4}$
445144	Signal transduction by L1	19/4725	26/13397	$9.50x10^{-5}$	$7.53x10^{-4}$
400685	Sema4D in semaphorin signaling	19/4725	26/13397	$9.50x10^{-5}$	$7.53x10^{-4}$
70326	Glucose metabolism	32/4725	52/13397	$9.84x10^{-5}$	$7.72x10^{-4}$
372790	Signaling by GPCR	132/4725	288/13397	$1.21x10^{-4}$	$9.39x10^{-4}$
198753	ERK/MAPK targets	15/4725	19/13397	$1.26x10^{-4}$	$9.70x10^{-4}$
1483206	Glycerophospholipid biosynthesis	34/4725	57/13397	$1.45x10^{-4}$	$1.08x10^{-3}$
416572	Sema4D induced cell migration and growth-cone collapse	16/4725	21/13397	$1.55x10^{-4}$	$1.08x10^{-3}$
2206302	MPS I - Hurler syndrome	39/4725	68/13397	$1.59x10^{-4}$	$1.08x10^{-3}$

2206296	MPS II - Hunter syndrome	39/4725	68/13397	$1.59x10^{-4}$	$1.08x10^{-3}$
2206307	MPS IIIA - Sanfilippo syndrome A	39/4725	68/13397	$1.59x10^{-4}$	$1.08x10^{-3}$
2206282	MPS IIIB - Sanfilippo syndrome B	39/4725	68/13397	$1.59x10^{-4}$	$1.08x10^{-3}$
2206291	MPS IIIC - Sanfilippo syndrome C	39/4725	68/13397	$1.59x10^{-4}$	$1.08x10^{-3}$
2206305	MPS IIID - Sanfilippo syndrome D	39/4725	68/13397	$1.59x10^{-4}$	$1.08x10^{-3}$
2206290	MPS IV - Morquio syndrome A	39/4725	68/13397	$1.59x10^{-4}$	$1.08x10^{-3}$
2206308	MPS IV - Morquio syndrome B	39/4725	68/13397	$1.59x10^{-4}$	$1.08x10^{-3}$
2206280	MPS IX - Natowicz syndrome	39/4725	68/13397	$1.59x10^{-4}$	$1.08x10^{-3}$
2206285	MPS VI - Maroteaux-Lamy syndrome	39/4725	68/13397	$1.59x10^{-4}$	$1.08x10^{-3}$
2206292	MPS VII - Sly syndrome	39/4725	68/13397	$1.59x10^{-4}$	$1.08x10^{-3}$
2206281	Mucopolysaccharidoses	39/4725	68/13397	$1.59x10^{-4}$	$1.08x10^{-3}$
1630316	Glycosaminoglycan metabolism	39/4725	68/13397	$1.59x10^{-4}$	$1.08x10^{-3}$
164952	The role of Nef in HIV-1 replication and disease pathogenesis	18/4725	25/13397	$1.99x10^{-4}$	$1.34x10^{-3}$
381038	Activation of Chaperone Genes by XBP1(S)	27/4725	43/13397	$2.12x10^{-4}$	$1.41x10^{-3}$
77387	Insulin receptor recycling	13/4725	16/13397	$2.22x10^{-4}$	$1.46x10^{-3}$
196025	Formation of annular gap junctions	8/4725	8/13397	$2.38x10^{-4}$	$1.55x10^{-3}$
190873	Gap junction degradation	8/4725	8/13397	$2.38x10^{-4}$	$1.55x10^{-3}$
168180	TRAF6 Mediated Induction of proinflammatory cytokines	33/4725	56/13397	$2.50x10^{-4}$	$1.61x10^{-3}$
2132295	MHC class II antigen presentation	47/4725	87/13397	$2.53x10^{-4}$	$1.62x10^{-3}$
622312	Inflammasomes	14/4725	18/13397	$2.85x10^{-4}$	$1.81x10^{-3}$
983168	Antigen processing: Ubiquitination & Proteasome degradation	85/4725	177/13397	$3.00x10^{-4}$	$1.89x10^{-3}$
69304	Regulation of DNA replication	36/4725	63/13397	$3.11x10^{-4}$	$1.90x10^{-3}$
69300	Removal of licensing factors from origins	36/4725	63/13397	$3.11x10^{-4}$	$1.90x10^{-3}$
69052	Switching of origins to a post-replicative state	36/4725	63/13397	$3.11x10^{-4}$	$1.90x10^{-3}$
68949	Orc1 removal from chromatin	36/4725	63/13397	$3.11x10^{-4}$	$1.90x10^{-3}$
917977	Transferrin endocytosis and recycling	15/4725	20/13397	$3.39x10^{-4}$	$2.05x10^{-3}$
190236	Signaling by FGFR	51/4725	97/13397	$3.41x10^{-4}$	$2.05x10^{-3}$
1226099	Signaling by FGFR in disease	57/4725	111/13397	$3.53x10^{-4}$	$2.11x10^{-3}$
877300	Interferon gamma signaling	34/4725	59/13397	$3.65x10^{-4}$	$2.16x10^{-3}$
844456	The NLRP3 inflammasome	11/4725	13/13397	$3.73x10^{-4}$	$2.18x10^{-3}$
163210	Formation of ATP by chemiosmotic coupling	11/4725	13/13397	$3.73x10^{-4}$	$2.18x10^{-3}$
176408	Regulation of APC/C activators between G1/S and early anaphase	37/4725	66/13397	$4.32x10^{-4}$	$2.51x10^{-3}$
74752	Signaling by Insulin receptor	41/4725	75/13397	$4.43x10^{-4}$	$2.55x10^{-3}$
535734	Fatty acid, triacylglycerol, and ketone body metabolism	51/4725	98/13397	$4.70x10^{-4}$	$2.68x10^{-3}$
373760	L1CAM interactions	36/4725	64/13397	$4.71x10^{-4}$	$2.68x10^{-3}$
216083	Integrin cell surface interactions	28/4725	47/13397	$5.68x10^{-4}$	$3.20x10^{-3}$
109607	Extrinsic Pathway for Apoptosis	9/4725	10/13397	$5.74x10^{-4}$	$3.20x10^{-3}$
73887	Death Receptor Signalling	9/4725	10/13397	$5.74x10^{-4}$	$3.20x10^{-3}$
445355	Smooth Muscle Contraction	13/4725	17/13397	$6.36x10^{-4}$	$3.52x10^{-3}$
913531	Interferon Signaling	67/4725	137/13397	$6.73x10^{-4}$	$3.67x10^{-3}$
1236973	Cross-presentation of particulate exogenous antigens (phagosomes)	7/4725	7/13397	$6.77x10^{-4}$	$3.67x10^{-3}$
168643	Nucleotide-binding domain, leucine rich repeat containing receptor (NLR) signaling pathways	29/4725	50/13397	$8.40x10^{-4}$	$4.47x10^{-3}$
909733	Interferon alpha/beta signaling	29/4725	50/13397	$8.40x10^{-4}$	$4.47x10^{-3}$
174143	APC/C-mediated degradation of cell cycle proteins	38/4725	70/13397	$8.47x10^{-4}$	$4.47x10^{-3}$

453276	Regulation of mitotic cell cycle	38/4725	70/13397	$8.47x10^{-4}$	$4.47x10^{-3}$
425407	SLC-mediated transmembrane transport	60/4725	122/13397	$1.05x10^{-3}$	$5.49x10^{-3}$
500792	GPCR ligand binding	65/4725	134/13397	$1.05x10^{-3}$	$5.49x10^{-3}$
1679131	Trafficking and processing of endosomal TLR	11/4725	14/13397	$1.18x10^{-3}$	$6.14x10^{-3}$
71406	Pyruvate metabolism and Citric Acid (TCA) cycle	23/4725	38/13397	$1.29x10^{-3}$	$6.64x10^{-3}$
166658	Complement cascade	12/4725	16/13397	$1.40x10^{-3}$	$7.16x10^{-3}$
977606	Regulation of Complement cascade	8/4725	9/13397	$1.47x10^{-3}$	$7.50x10^{-3}$
211859	Biological oxidations	32/4725	58/13397	$1.48x10^{-3}$	$7.50x10^{-3}$
190333	Downstream signaling of activated FGFR	45/4725	88/13397	$1.58x10^{-3}$	$7.93x10^{-3}$
432720	Lysosome Vesicle Biogenesis	14/4725	20/13397	$1.64x10^{-3}$	$8.11x10^{-3}$
71403	Citric acid cycle (TCA cycle)	14/4725	20/13397	$1.64x10^{-3}$	$8.11x10^{-3}$
1474244	Extracellular matrix organization	18/4725	28/13397	$1.64x10^{-3}$	$8.11x10^{-3}$
198725	Nuclear Events (kinase and transcription factor activation)	15/4725	22/13397	$1.69x10^{-3}$	$8.30x10^{-3}$
444473	Formyl peptide receptors bind formyl peptides and many other ligands	6/4725	6/13397	$1.92x10^{-3}$	$9.26x10^{-3}$
196807	Nicotinate metabolism	6/4725	6/13397	$1.92x10^{-3}$	$9.26x10^{-3}$
390247	Beta-oxidation of very long chain fatty acids	6/4725	6/13397	$1.92x10^{-3}$	$9.26x10^{-3}$
2172127	DAP12 interactions	51/4725	103/13397	$2.00x10^{-3}$	$9.57x10^{-3}$

Table 3: Reactome pathway enrichment analysis on the correlated genes in B-cells and monocytes. GeneRatio: over-represented genes in genes tested in each cell type. BgRatio: a ratio of genes in a pathway and background. qvalue: adjusted p-value by FDR.

Pathway ID	Description	GeneRatio	BgRatio	pvalue	qvalue
B-cells					
156842	Eukaryotic Translation Elongation	61/1867	95/5654	$1.38x10^{-7}$	$1.38x10^{-7}$
72689	Formation of a pool of free 40S subunits	60/1867	94/5654	$1.38x10^{-7}$	$1.38x10^{-7}$
157279	3' -UTR-mediated translational regulation	61/1867	97/5654	$1.38x10^{-7}$	$1.38x10^{-7}$
156827	L13a-mediated translational silencing of Ceruloplasmin expression	61/1867	97/5654	$1.38x10^{-7}$	$1.38x10^{-7}$
156902	Peptide chain elongation	58/1867	91/5654	$1.38x10^{-7}$	$1.38x10^{-7}$
72764	Eukaryotic Translation Termination	58/1867	91/5654	$1.38x10^{-7}$	$1.38x10^{-7}$
72706	GTP hydrolysis and joining of the 60S ribosomal subunit	61/1867	98/5654	$1.85x10^{-7}$	$1.85x10^{-7}$
975956	Nonsense Mediated Decay Independent of the Exon Junction Complex	58/1867	92/5654	$2.09x10^{-7}$	$2.09x10^{-7}$
72737	Cap-dependent Translation Initiation	61/1867	99/5654	$2.53x10^{-7}$	$2.53x10^{-7}$
72613	Eukaryotic Translation Initiation	61/1867	99/5654	$2.53x10^{-7}$	$2.53x10^{-7}$
532600	Metabolism of RNA	82/1867	147/5654	$4.2x10^{-7}$	$4.2x10^{-7}$
447010	Metabolism of mRNA	78/1867	138/5654	$4.2x10^{-7}$	$4.2x10^{-7}$
1799339	SRP-dependent cotranslational protein targeting to membrane	59/1867	97/5654	$7x10^{-7}$	$7x10^{-7}$
927802	Nonsense-Mediated Decay	60/1867	100/5654	$9.76x10^{-7}$	$9.76x10^{-7}$
975957	Nonsense Mediated Decay Enhanced by the Exon Junction Complex	60/1867	100/5654	$9.76x10^{-7}$	$9.76x10^{-7}$
72766	Translation	65/1867	112/5654	$1.51x10^{-6}$	$1.51x10^{-6}$
392499	Metabolism of proteins	96/1867	193/5654	$3.02x10^{-5}$	$3.02x10^{-5}$

1643685	Disease	113/1867	257/5654	$4.5x10^{-3}$	$4.5x10^{-3}$
Monocytes					
392499	Metabolism of proteins	86/1641	167/4725	$3.61x10^{-3}$	$3.61x10^{-3}$
163200	Respiratory electron transport, ATP synthesis by chemiosmotic coupling, and heat production by uncoupling proteins.	35/1641	55/4725	$3.61x10^{-3}$	$3.61x10^{-3}$
74160	Gene Expression	111/1641	231/4725	$3.61x10^{-3}$	$3.61x10^{-3}$
1430728	Metabolism	219/1641	508/4725	$4.35x10^{-3}$	$4.35x10^{-3}$

Table 4: HLA imputation summary. Alleles: The number of alleles available to impute with HLA*IMP. Alleles imputed: alleles predicted in the volunteer population. Chromosome call rate: the number of chromosome in the volunteer population imputed after >0.7 cutoff. PPV: positive predicted value (accuracy), Call rate: call rate including the reference and the inference data (the volunteer population).

digit	HLA	Allele	Alleles imputed	Chromosome call rate	PPV	Call rate
2	*A*	17	15	98.93	0.97	0.96
	B	25	24	98.94	0.97	0.98
	C	13	13	99.29	0.98	1.00
	DQA1	5	5	99.47	1.00	1.00
	DQB1	5	5	100.00	0.99	1.00
	DRB1	13	12	99.29	0.99	0.99
4	*A*	27	22	95.94	0.97	0.95
	B	47	38	98.23	0.96	0.98
	C	21	19	98.94	0.98	0.99
	DQA1	7	7	97.70	0.98	1.00
	DQB1	18	16	99.82	0.97	0.99
	DRB1	34	30	88.69	0.94	0.92

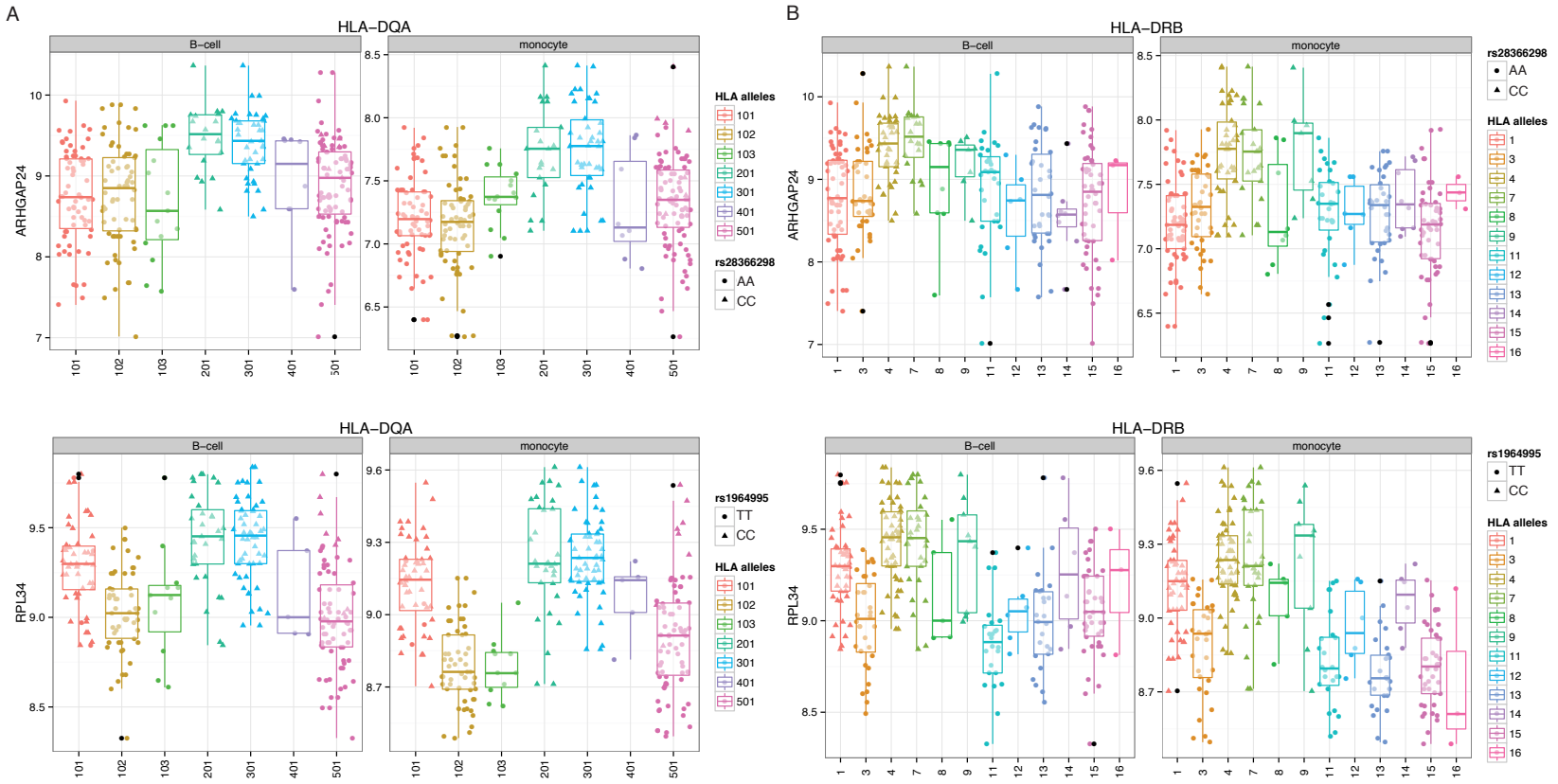


Figure 1: *HLA-DQA1* (A) and *HLA-DRB1* (B) association to *ARHGAP24* and *RPL34* trans eQTL gene expression. Homozygous genotypes of the eQTL SNPs and 4-digit resolution of *HLA-DQA1* and *HLA-DRB1* are plotted based on the gene expression in B-cells and monocytes.

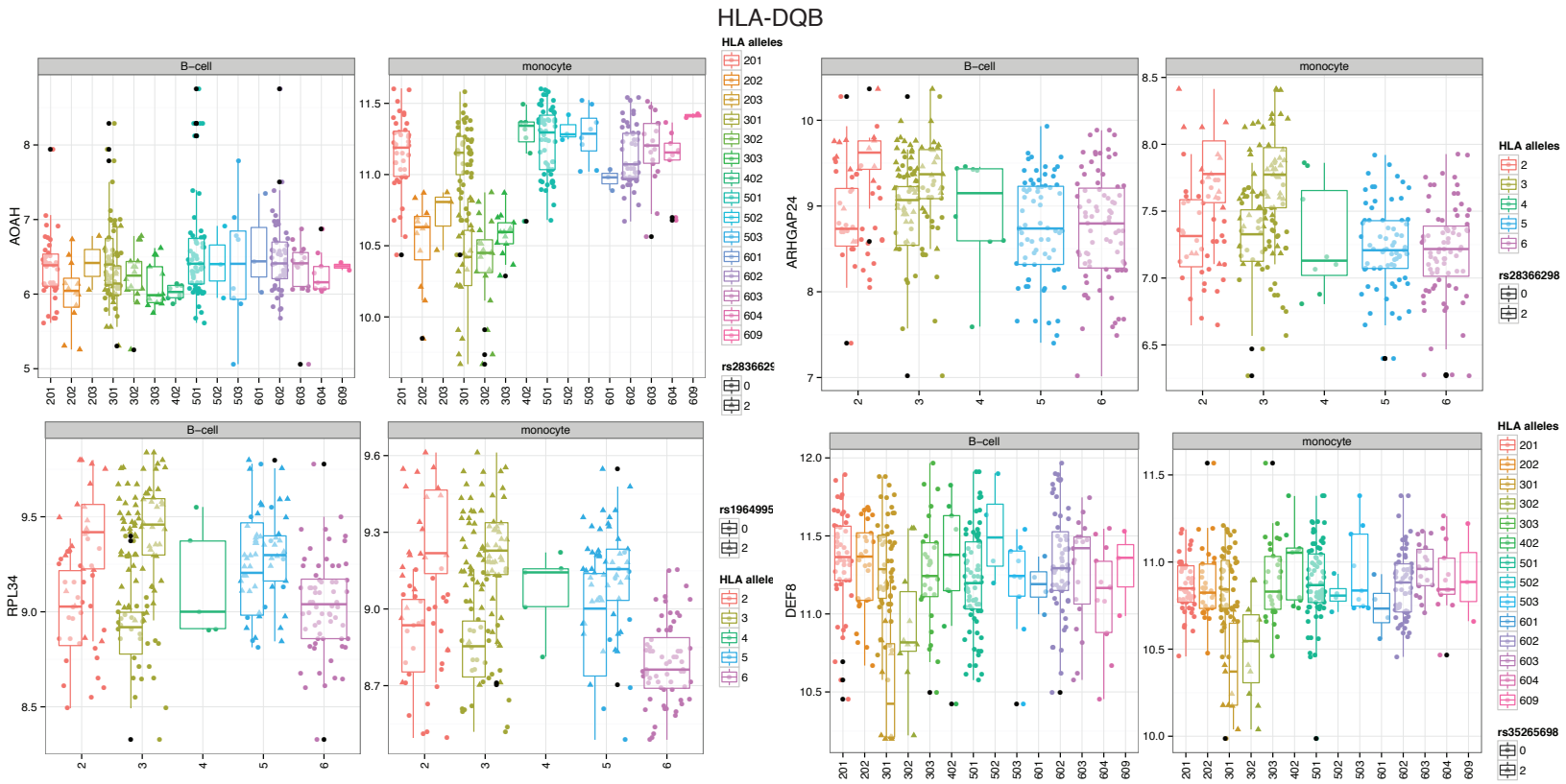


Figure 2: *HLA-DQB1* association to *AOAH, ARHGAP24, RPL34* and *DEF8* trans eQTL gene expression. Homozygous genotypes of the eQTL SNPs and HLA-DQB1 in 4-digit for *AOAH* and *DEF8* and 2-digit for *ARHGAP24* and *RPL34* are plotted based on the gene expression in B-cells and monocytes.

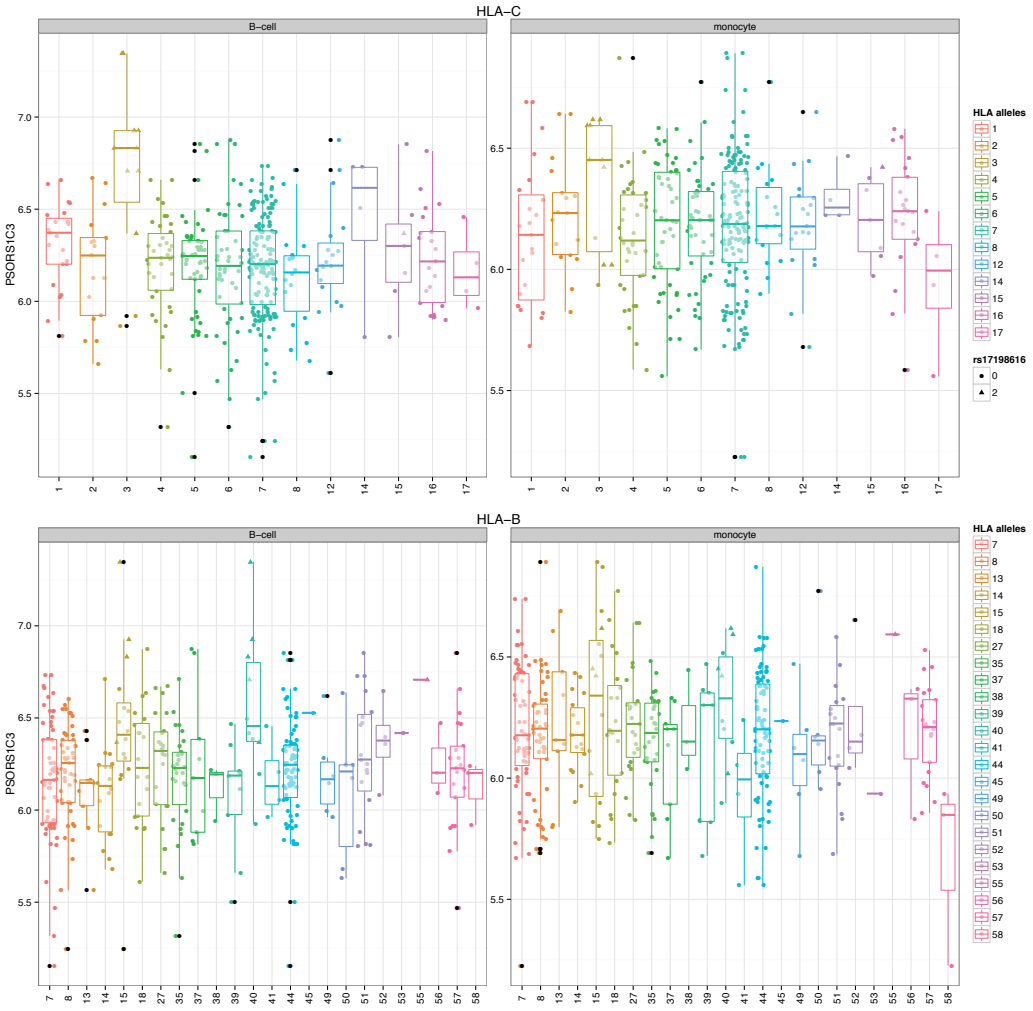


Figure 3: *HLA-C* (top) and *-B* (bottom) alleles in 2-digit resolution associated with *PSORS1C3* expression with eQTL SNP rs17198616 homozygous genotypes. Expression of *PSRRS1C3* in B-cells with *HLA-C*3* ($p=3.59x10^{-15}$) and *HLA-B*40* ($3.79x10^{-9}$) are significant.

Table 5: List of HLA allele association in cis in monocytes (mono) and B-cells (bcell).

Gene	REFSEQ_ID	*HLA*	Cell.type	t.stat	p.value	FDR
PSORS1C3	NR_026816	*HLA-C*3*	B-cell	8.22	7.91×10^{-15}	7.02×10^{-11}
		*HLA-B*40*	B-cell	5.92	9.42×10^{-9}	3.14×10^{-5}
		*HLA-B*15*	B-cell	4.35	1.93×10^{-5}	1.55×10^{-2}
VARS2	NM_020442	*HLA-C*7*	shared (mono)	-6.93	2.90×10^{-11}	1.29×10^{-7}
			shared (B-cell)	-5.61	4.85×10^{-8}	1.43×10^{-4}
		*HLA-DRB1*3*	shared (mono)	-5.40	1.45×10^{-7}	2.86×10^{-4}
			shared (B-cell)	-4.52	9.02×10^{-6}	9.42×10^{-3}
		*HLA-B*8*	shared (mono)	-5.32	2.15×10^{-7}	3.69×10^{-4}
			shared (B-cell)	-4.34	1.96×10^{-5}	1.55×10^{-2}
		*HLA-C*5*	mono	4.44	1.32×10^{-5}	1.12×10^{-2}
		*HLA-DQB1*2*	mono	-4.04	7.02×10^{-5}	4.35×10^{-2}
		*HLA-DQA1*5*	mono	-4.01	7.89×10^{-5}	4.62×10^{-2}
LST1	NM_007161	*HLA-C*5*	mono	-6.52	3.24×10^{-10}	1.15×10^{-6}
LST1	NM_205839	*HLA-A*33*	B-cell	4.78	2.89×10^{-6}	3.91×10^{-3}
ZFP57	XM_294093	*HLA-B*8*	B-cell	5.58	5.60×10^{-8}	1.57×10^{-4}
		*HLA-A*1*	shared (mono)	4.89	1.70×10^{-6}	1.93×10^{-3}
			shared (B-cell)	4.75	3.29×10^{-6}	3.98×10^{-3}
		*HLA-A*23*	mono	4.01	7.84×10^{-5}	4.62×10^{-2}
MDC1	NM_014641	*HLA-A*30*	B-cell	5.23	3.41×10^{-7}	7.90×10^{-4}
HLA-DOB	NM_002120	*HLA-DRB1*1*	mono	-5.17	4.51×10^{-7}	6.86×10^{-4}
		*HLA-DQB1*5*	mono	-4.47	1.13×10^{-5}	9.84×10^{-3}
NCR3	NM_147130	*HLA-C*8*	B-cell	-5.05	7.83×10^{-7}	1.60×10^{-3}
		*HLA-B*14*	B-cell	-4.98	1.10×10^{-6}	1.95×10^{-3}
HLA-F	NM_001098479	*HLA-A*11*	B-cell	4.92	1.49×10^{-6}	2.47×10^{-3}
		*HLA-A*26*	mono	-4.89	1.67×10^{-6}	1.93×10^{-3}
		*HLA-A*2*	mono	-4.25	2.96×10^{-5}	2.19×10^{-2}
HIST1H3I	NM_003533	*HLA-C*7*	B-cell	4.71	3.87×10^{-6}	4.58×10^{-3}
		*HLA-B*51*	B-cell	-4.63	5.52×10^{-6}	6.39×10^{-3}
PSMB9	NM_002800	*HLA-C*16*	mono	-4.55	7.89×10^{-6}	7.37×10^{-3}
		*HLA-DRB1*7*	B-cell	-4.55	8.04×10^{-6}	8.85×10^{-3}
		*HLA-DQA1*2*	B-cell	-4.53	8.89×10^{-6}	9.42×10^{-3}
ATP6V1G2	NM_130463	*HLA-DQA1*5*	mono	4.44	1.31×10^{-5}	1.12×10^{-2}
HLA-DPB1	NM_002121	*HLA-DRB1*3*	shared (mono)	4.35	1.92×10^{-5}	1.47×10^{-2}
			shared (B-cell)	4.27	2.67×10^{-5}	2.06×10^{-2}
TAP2	NM_018833	*HLA-DRB1*1*	mono	4.13	4.72×10^{-5}	3.35×10^{-2}
IER3	NM_003897	*HLA-B*8*	mono	4.11	5.16×10^{-5}	3.57×10^{-2}
OR10C1	NM_013941	*HLA-B*53*	mono	-4.09	5.62×10^{-5}	3.70×10^{-2}
SKIV2L	NM_006929	*HLA-C*16*	B-cell	4.07	6.20×10^{-5}	4.44×10^{-2}
HLA-DPB2	NR_001435	*HLA-DRB1*1*	B-cell	4.06	6.26×10^{-5}	4.44×10^{-2}
ZNRD1	NM_170783	*HLA-A*25*	mono	-4.04	6.85×10^{-5}	4.35×10^{-2}
PSORS1C2	NM_014069	*HLA-B*57*	mono	-4.04	6.90×10^{-5}	4.35×10^{-2}
LY6G5C	NM_025262	*HLA-A*68*	mono	-3.99	8.45×10^{-5}	4.78×10^{-2}

Table 6: Description of the HapMap and the 1000 Genome reference populations. N: numbers of individuals.

Population	Description	N
African		
ASW	HapMap African ancestry from southwest in the USA	61
LWK	Luhya from Kenya	97
YRI	Yoruban from Nigeria	88
European		
CEU	Northern and western European ancestry from Utah, USA	85
FIN	HapMap Finnish from Finland	93
GBR	British from England and Scotland	89
IBS	Iberian in Spain	14
TSI	Toscan in Italy	98
Southeast Asia		
CHB	Han Chinese in Beijing, China	97
CHS	Han Chinese in South China	100
JPT	Japanese from Tokyo, Japan	89
Latin America		
CLM	Colombian in Medellin, Colombia	60
MXL	HapMap Mexican from Los Angeles, California, USA	66
PUR	Puerto Rican in Puerto Rico	55

Bibliography

Abecasis, G. R., A. Auton, L. D. Brooks, M. A. DePristo, R. M. Durbin, R. E. Handsaker, H. M. Kang, G. T. Marth, and G. A. McVean (2012). "An integrated map of genetic variation from 1,092 human genomes". *Nature* 491, pp. 56–65.

Achiron, A., M. Gurevich, N. Friedman, N. Kaminski, and M. Mandel (2004). "Blood transcriptional signatures of multiple sclerosis: unique gene expression of disease activity". *Ann Neurol* 55, pp. 410–7.

Aird, I., H. H. Bentall, and J. A. Roberts (1953). "A relationship between cancer of stomach and the ABO blood groups". *Br Med J* 1, pp. 799–801.

Akino, K., M. Toyota, H. Suzuki, T. Imai, R. Maruyama, M. Kusano, N. Nishikawa, Y. Watanabe, Y. Sasaki, T. Abe, E. Yamamoto, I. Tarasawa, T. Sonoda, M. Mori, K. Imai, Y. Shinomura, and T. Tokino (2007). "Identification of DFNA5 as a target of epigenetic inactivation in gastric cancer". *Cancer Sci* 98, pp. 88–95.

Alfonso, C. and L. Karlsson (2000). "Nonclassical MHC class II molecules". *Annu Rev Immunol* 18, pp. 113–42.

Alizadeh, A. A., M. B. Eisen, R. E. Davis, C. Ma, I. S. Lossos, A. Rosenwald, J. C. Boldrick, H. Sabet, T. Tran, X. Yu, J. I. Powell, L. Yang, G. E. Marti, T. Moore, Jr. Hudson J., L. Lu, D. B. Lewis, R. Tibshirani, G. Sherlock, W. C. Chan, T. C. Greiner, D. D. Weisenburger, J. O. Armitage, R. Warnke, R. Levy, W. Wilson, M. R. Grever, J. C. Byrd, D. Botstein, P. O. Brown, and L. M. Staudt (2000). "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling". *Nature* 403, pp. 503–11.

Alkhateeb, A., P. R. Fain, A. Thody, D. C. Bennett, and R. A. Spritz (2003). "Epidemiology of vitiligo and associated autoimmune diseases in Caucasian probands and their families". *Pigment Cell Res* 16, pp. 208–14.

Alter, O., P. O. Brown, and D. Botstein (2000). "Singular value decomposition for genome-wide expression data processing and modeling". *Proc Natl Acad Sci U S A* 97, pp. 10101–6.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman (1990). "Basic local alignment search tool". *J Mol Biol* 215, pp. 403–10.

Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman (1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs". *Nucleic Acids Res* 25, pp. 3389–402.

Altshuler, D., J. N. Hirschhorn, M. Klannemark, C. M. Lindgren, M. C. Vohl, J. Nemesh, C. R. Lane, S. F. Schaffner, S. Bolk, C. Brewer, T. Tuomi, D. Gaudet, T. J. Hudson, M. Daly, L. Groop, and E. S. Lander (2000). "The common PPARgamma Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes". *Nat Genet* 26, pp. 76–80.

Altshuler, D. M., R. A. Gibbs, L. Peltonen, E. Dermitzakis, S. F. Schaffner, F. Yu, P. E. Bonnen, P. I. de Bakker, P. Deloukas, S. B. Gabriel, R. Gwilliam, S. Hunt, M. Inouye, X. Jia, A. Palotie, M. Parkin, P. Whittaker, K. Chang, A. Hawes, L. R. Lewis, Y. Ren, D. Wheeler, D. M. Muzny, C. Barnes, K. Darvishi, M. Hurles, J. M. Korn, K. Kristiansson, C. Lee, S. A. McCarrol, J. Nemesh, A. Keinan, S. B. Montgomery, S. Pollack, A. L. Price, N. Soranzo, C. Gonzaga-Jauregui, V. Anttila, W. Brodeur, M. J. Daly, S. Leslie, G. McVean, L. Moutsianas, H. Nguyen, Q. Zhang, M. J. Ghori, R. McGinnis, W. McLaren, F. Takeuchi, S. R. Grossman, I. Shlyakhter, E. B. Hostetter, P. C. Sabeti, C. A. Adebamowo, M. W. Foster, D. R. Gordon, J. Licinio, M. C. Manca, P. A. Marshall, I. Matsuda, D. Ngare, V. O. Wang, D. Reddy, C. N. Rotimi, C. D. Royal,

R. R. Sharp, C. Zeng, L. D. Brooks, and J. E. McEwen (2010). "Integrating common and rare genetic variation in diverse human populations". *Nature* 467, pp. 52–8.

Amberger, J., C. A. Bocchini, A. F. Scott, and A. Hamosh (2009). "McKusick's Online Mendelian Inheritance in Man (OMIM)". *Nucleic Acids Res* 37, pp. D793–6.

Anderson, C. A., F. H. Pettersson, G. M. Clarke, L. R. Cardon, A. P. Morris, and K. T. Zondervan (2010). "Data quality control in genetic case-control association studies". *Nat Protoc* 5, pp. 1564–73.

Anderson, C. A., G. Boucher, C. W. Lees, A. Franke, M. D'Amato, K. D. Taylor, J. C. Lee, P. Goyette, M. Imielinski, A. Latiano, C. Lagace, R. Scott, L. Amininejad, S. Bumpstead, L. Baidoo, R. N. Baldassano, M. Barclay, T. M. Bayless, S. Brand, C. Buning, J. F. Colombel, L. A. Denson, M. De Vos, M. Dubinsky, C. Edwards, D. Ellinghaus, R. S. Fehrmann, J. A. Floyd, T. Florin, D. Franchimont, L. Franke, M. Georges, J. Glas, N. L. Glazer, S. L. Guthery, T. Haritunians, N. K. Hayward, J. P. Hugot, G. Jobin, D. Laukens, I. Lawrance, M. Lemann, A. Levine, C. Libioulle, E. Louis, D. P. McGovern, M. Milla, G. W. Montgomery, K. I. Morley, C. Mowat, A. Ng, W. Newman, R. A. Ophoff, L. Papi, O. Palmieri, L. Peyrin-Biroulet, J. Panes, A. Phillips, N. J. Prescott, D. D. Proctor, R. Roberts, R. Russell, P. Rutgeerts, J. Sanderson, M. Sans, P. Schumm, F. Seibold, Y. Sharma, L. A. Simms, M. Seielstad, A. H. Steinhart, S. R. Targan, L. H. van den Berg, M. Vatn, H. Verspaget, T. Walters, C. Wijmenga, D. C. Wilson, H. J. Westra, R. J. Xavier, Z. Z. Zhao, C. Y. Ponsioen, V. Andersen, L. Torkvist, M. Gazouli, N. P. Anagnou, T. H. Karlsen, L. Kupcinskas, J. Sventoraityte, J. C. Mansfield, S. Kugathasan, M. S. Silverberg, J. Halfvarson, J. I. Rotter, C. G. Mathew, A. M. Griffiths, R. Gearry, T. Ahmad, S. R. Brant, *et al.* (2011). "Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47". *Nat Genet* 43, pp. 246–52.

Andres, A. M., M. Y. Dennis, W. W. Kretzschmar, J. L. Cannons, S. Q. Lee-Lin, B. Hurle, P. L. Schwartzberg, S. H. Williamson, C. D. Bustamante, R. Nielsen, A. G. Clark, and E. D. Green (2010). "Balancing selection maintains a form of ERAP2 that undergoes nonsense-mediated decay and affects antigen presentation". *PLoS Genet* 6, e1001157.

Antoniou, A., P. D. Pharoah, S. Narod, H. A. Risch, J. E. Eyfjord, J. L. Hopper, N. Loman, H. Olsson, O. Johannsson, A. Borg, B. Pasini, P. Radice, S. Manoukian, D. M. Eccles, N. Tang, E. Olah, H. Anton-Culver, E. Warner, J. Lubinski, J. Gronwald, B. Gorski, H. Tulinius, S. Thorlacius, H. Eerola, H. Nevanlinna, K. Syrjakoski, O. P. Kallioniemi, D. Thompson, C. Evans, J. Peto, F. Lalloo, D. G. Evans, and D. F. Easton (2003). "Average risks of breast and ovarian cancer associated with BRCA1 or BRCA2 mutations detected in case Series unselected for family history: a combined analysis of 22 studies". *Am J Hum Genet* 72, pp. 1117–30.

Arbour, N. C., E. Lorenz, B. C. Schutte, J. Zabner, J. N. Kline, M. Jones, K. Frees, J. L. Watt, and D. A. Schwartz (2000). "TLR4 mutations are associated with endotoxin hyporesponsiveness in humans". *Nat Genet* 25, pp. 187–91.

Arnold, L., A. Henry, F. Poron, Y. Baba-Amer, N. van Rooijen, A. Plonquet, R. K. Gherardi, and B. Chazaud (2007). "Inflammatory monocytes recruited after skeletal muscle injury switch into antiinflammatory macrophages to support myogenesis". *J Exp Med* 204, pp. 1057–69.

Augello, M. A., T. E. Hickey, and K. E. Knudsen (2011). "FOXA1: master of steroid receptor function in cancer". *EMBO J* 30, pp. 3885–94.

Aulchenko, Y. S., S. Ripatti, I. Lindqvist, D. Boomsma, I. M. Heid, P. P. Pramstaller, B. W. Penninx, A. C. Janssens, J. F. Wilson, T. Spector, N. G. Martin, N. L. Pedersen, K. O. Kyvik, J. Kaprio, A. Hofman, N. B. Freimer, M. R. Jarvelin, U. Gyllensten, H. Campbell, I. Rudan, A. Johansson, F. Marroni, C. Hayward, V. Vitart, I. Jonasson, C. Pattaro, A. Wright, N. Hastie, I. Pichler, A. A. Hicks, M. Falchi, G. Willemsen, J. J. Hottenga, E. J. de Geus, G. W. Montgomery, J. Whitfield, P. Magnusson, J. Saharinen, M. Perola, K. Silander, A. Isaacs, E. J. Sijbrands, A. G. Uitterlinden, J. C. Witteman, B. A. Oostra, P. Elliott, A. Ruokonen, C. Sabatti, C. Gieger, T. Meitinger, F. Kronenberg, A. Doring, H. E. Wichmann, J. H. Smit, M. I. McCarthy, C. M. van Duijn, and L. Peltonen (2009). "Loci influencing lipid levels and coronary heart disease risk in 16 European population cohorts". *Nat Genet* 41, pp. 47–55.

Baechler, E. C., F. M. Batliwalla, G. Karypis, P. M. Gaffney, W. A. Ortmann, K. J. Espe, K. B. Shark, W. J. Grande, K. M. Hughes, V. Kapur, P. K. Gregersen, and T. W. Behrens (2003). "Interferon-inducible gene expression signature in peripheral blood cells of patients with severe lupus". *Proc Natl Acad Sci U S A* 100, pp. 2610–5.

Bahlo, M., Australia (ANZgene), and New Zealand Multiple Sclerosis Genetics Consortium (2009). "Genome-wide association study identifies new multiple sclerosis susceptibility loci on chromosomes 12 and 20". *Nat Genet* 41, pp. 824–8.

Baisch, J. M., T. Weeks, R. Giles, M. Hoover, P. Stastny, and J. D. Capra (1990). "Analysis of HLA-DQ genotypes and susceptibility in insulin-dependent diabetes mellitus". *N Engl J Med* 322, pp. 1836–41.

Bakker, P. I. de, G. McVean, P. C. Sabeti, M. M. Miretti, T. Green, J. Marchini, X. Ke, A. J. Monsuur, P. Whittaker, M. Delgado, J. Morrison, A. Richardson, E. C. Walsh, X. Gao, L. Galver, J. Hart, D. A. Hafler, M. Pericak-Vance, J. A. Todd, M. J. Daly, J. Trowsdale, C. Wijmenga, T. J. Vyse, S. Beck, S. S. Murray, M. Carrington, S. Gregory, P. Deloukas, and J. D. Rioux (2006). "A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC". *Nat Genet* 38, pp. 1166–72.

Baltimore, D. (1970). "RNA-dependent DNA polymerase in virions of RNA tumour viruses". *Nature* 226, pp. 1209–11.

Barbosa-Morais, N. L., M. J. Dunning, S. A. Samarajiwa, J. F. Darot, M. E. Ritchie, A. G. Lynch, and S. Tavare (2010). "A re-annotation pipeline for Illumina BeadArrays: improving the interpretation of gene expression data". *Nucleic Acids Res* 38, e17.

Barnes, B. J., M. J. Kellum, K. E. Pinder, J. A. Frisancho, and P. M. Pitha (2003). "Interferon regulatory factor 5, a novel mediator of cell cycle arrest and cell death". *Cancer Res* 63, pp. 6424–31.

Barrett, J. C., B. Fry, J. Maller, and M. J. Daly (2005). "Haploview: analysis and visualization of LD and haplotype maps". *Bioinformatics* 21, pp. 263–5.

Barrett, J. C., S. Hansoul, D. L. Nicolae, J. H. Cho, R. H. Duerr, J. D. Rioux, S. R. Brant, M. S. Silverberg, K. D. Taylor, M. M. Barmada, A. Bitton, T. Dassopoulos, L. W. Datta, T. Green, A. M. Griffiths, E. O. Kistner, M. T. Murtha, M. D. Regueiro, J. I. Rotter, L. P. Schumm, A. H. Steinhart, S. R. Targan, R. J. Xavier, C. Libioulle, C. Sandor, M. Lathrop, J. Belaiche, O. Dewit, I. Gut, S. Heath, D. Laukens, M. Mni, P. Rutgeerts, A. Van Gossum, D. Zelenika, D. Franchimont, J. P. Hugot, M. de Vos, S. Vermeire, E. Louis, L. R. Cardon, C. A. Anderson, H. Drummond, E. Nimmo, T. Ahmad, N. J. Prescott, C. M. Onnie, S. A. Fisher, J. Marchini, J. Ghori, S. Bumpstead, R. Gwilliam, M. Tremelling, P. Deloukas, J. Mansfield, D. Jewell, J. Satsangi, C. G. Mathew, M. Parkes, M. Georges, and M. J. Daly (2008). "Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease". *Nat Genet* 40, pp. 955–62.

Barrett, J. C., D. G. Clayton, P. Concannon, B. Akolkar, J. D. Cooper, H. A. Erlich, C. Julier, G. Morahan, J. Nerup, C. Nierras, V. Plagnol, F. Pociot, H. Schuilenburg, D. J. Smyth, H. Stevens, J. A. Todd, N. M. Walker, and S. S. Rich (2009). "Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes". *Nat Genet* 41, pp. 703–7.

Bateson, W. (1905). "Experimental studies in the physiology of heredity". *Reports to the Evolution Committee of the Royal Society* 2, pp. 1–55, 80–99.

Batliwalla, F. M., W. Li, C. T. Ritchlin, X. Xiao, M. Brenner, T. Laragione, T. Shao, R. Durham, S. Kemshetti, E. Schwarz, R. Coe, M. Kern, E. C. Baechler, T. W. Behrens, P. K. Gregersen, and P. S. Gulko (2005a). "Microarray analyses of peripheral blood cells identifies unique gene expression signature in psoriatic arthritis". *Mol Med* 11, pp. 21–9.

Batliwalla, F. M., E. C. Baechler, X. Xiao, W. Li, S. Balasubramanian, H. Khalili, A. Damle, W. A. Ortmann, A. Perrone, A. B. Kantor, P. S. Gulko, M. Kern, R. Furie, T. W. Behrens, and P. K. Gregersen (2005b). "Peripheral blood gene expression profiling in rheumatoid arthritis". *Genes Immun* 6, pp. 388–97.

Bauernfeind, F. G., G. Horvath, A. Stutz, E. S. Alnemri, K. MacDonald, D. Speert, T. Fernandes-Alnemri, J. Wu, B. G. Monks, K. A. Fitzgerald, V. Hornung, and E. Latz (2009). "Cutting edge: NF-kappaB activating pattern recognition and cytokine

receptors license NLRP3 inflammasome activation by regulating NLRP3 expression". *J Immunol* 183, pp. 787–91.
Bell, G. I. and K. S. Polonsky (2001). "Diabetes mellitus and genetically programmed defects in beta-cell function". *Nature* 414, pp. 788–91.
Belle, T. L. van, K. T. Coppieters, and M. G. von Herrath (2011). "Type 1 diabetes: etiology, immunology, and therapeutic strategies". *Physiol Rev* 91, pp. 79–118.
Bendall, A. J. and P. L. Molloy (1994). "Base preferences for DNA binding by the bHLH-Zip protein USF: effects of MgCl2 on specificity and comparison with binding of Myc family members". *Nucleic Acids Res* 22, pp. 2801–10.
Bender, W., M. Akam, F. Karch, P. A. Beachy, M. Peifer, P. Spierer, E. B. Lewis, and D. S. Hogness (1983). "Molecular Genetics of the Bithorax Complex in Drosophila melanogaster". *Science* 221, pp. 23–9.
Benjamini, Yoav and Yosef Hochberg (1995). "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing". *Journal of the Royal Statistical Society. Series B (Methodological)* 57, pp. 289–300.
Benkirane, M., D. Y. Jin, R. F. Chun, R. A. Koup, and K. T. Jeang (1997). "Mechanism of transdominant inhibition of CCR5-mediated HIV-1 infection by ccr5delta32". *J Biol Chem* 272, pp. 30603–6.
Bennett, L., A. K. Palucka, E. Arce, V. Cantrell, J. Borvak, J. Banchereau, and V. Pascual (2003). "Interferon and granulopoiesis signatures in systemic lupus erythematosus blood". *J Exp Med* 197, pp. 711–23.
Benovoy, D., T. Kwan, and J. Majewski (2008). "Effect of polymorphisms within probe-target sequences on olignonucleotide microarray experiments". *Nucleic Acids Res* 36, pp. 4417–23.
Bernasconi, N. L., N. Onai, and A. Lanzavecchia (2003). "A role for Toll-like receptors in acquired immunity: up-regulation of TLR9 by BCR triggering in naive B cells and constitutive expression in memory B cells". *Blood* 101, pp. 4500–4.
Bird, A. (2002). "DNA methylation patterns and epigenetic memory". *Genes Dev* 16, pp. 6–21.
Birney, E., J. A. Stamatoyannopoulos, A. Dutta, R. Guigo, T. R. Gingeras, E. H. Margulies, Z. Weng, M. Snyder, E. T. Dermitzakis, R. E. Thurman, M. S. Kuehn, C. M. Taylor, S. Neph, C. M. Koch, S. Asthana, A. Malhotra, I. Adzhubei, J. A. Greenbaum, R. M. Andrews, P. Flicek, P. J. Boyle, H. Cao, N. P. Carter, G. K. Clelland, S. Davis, N. Day, P. Dhami, S. C. Dillon, M. O. Dorschner, H. Fiegler, P. G. Giresi, J. Goldy, M. Hawrylycz, A. Haydock, R. Humbert, K. D. James, B. E. Johnson, E. M. Johnson, T. T. Frum, E. R. Rosenzweig, N. Karnani, K. Lee, G. C. Lefebvre, P. A. Navas, F. Neri, S. C. Parker, P. J. Sabo, R. Sandstrom, A. Shafer, D. Vetrie, M. Weaver, S. Wilcox, M. Yu, F. S. Collins, J. Dekker, J. D. Lieb, T. D. Tullius, G. E. Crawford, S. Sunyaev, W. S. Noble, I. Dunham, F. Denoeud, A. Reymond, P. Kapranov, J. Rozowsky, D. Zheng, R. Castelo, A. Frankish, J. Harrow, S. Ghosh, A. Sandelin, I. L. Hofacker, R. Baertsch, D. Keefe, S. Dike, J. Cheng, H. A. Hirsch, E. A. Sekinger, J. Lagarde, J. F. Abril, A. Shahab, C. Flamm, C. Fried, J. Hackermuller, J. Hertel, M. Lindemeyer, K. Missal, A. Tanzer, S. Washietl, J. Korbel, O. Emanuelsson, J. S. Pedersen, N. Holroyd, R. Taylor, D. Swarbreck, N. Matthews, M. C. Dickson, D. J. Thomas, M. T. Weirauch, *et al.* (2007). "Identification and analysis of functional elements in 1human genome by the ENCODE pilot project". *Nature* 447, pp. 799–816.
Biswas, S., J. D. Storey, and J. M. Akey (2008). "Mapping gene expression quantitative trait loci by singular value decomposition and independent component analysis". *BMC Bioinformatics* 9, p. 244.
Bjorkman, P. J., M. A. Saper, B. Samraoui, W. S. Bennett, J. L. Strominger, and D. C. Wiley (1987). "The foreign antigen binding site and T cell recognition regions of class I histocompatibility antigens". *Nature* 329, pp. 512–8.
Bodmer, W. and C. Bonilla (2008). "Common and rare variants in multifactorial susceptibility to common diseases". *Nat Genet* 40, pp. 695–701.
Bodmer, W. F. (1972). "Evolutionary significance of the HL-A system". *Nature* 237, 139–45 passim.

Borel, C., S. Deutsch, A. Letourneau, E. Migliavacca, S. B. Montgomery, A. S. Dimas, C. E. Vejnar, H. Attar, M. Gagnebin, C. Gehrig, E. Falconnet, Y. Dupre, E. T. Dermitzakis, and S. E. Antonarakis (2011). "Identification of cis- and trans-regulatory variation modulating microRNA expression levels in human fibroblasts". *Genome Res* 21, pp. 68–73.

Borglum, A. D., A. Byskov, P. Ragno, A. L. Roldan, P. Tripputi, G. Cassani, K. Dano, F. Blasi, L. Bolund, and T. A. Kruse (1992). "Assignment of the urokinase-type plasminogen activator receptor gene (PLAUR) to chromosome 19q13.1-q13.2". *Am J Hum Genet* 50, pp. 492–7.

Botstein, D., R. L. White, M. Skolnick, and R. W. Davis (1980). "Construction of a genetic linkage map in man using restriction fragment length polymorphisms". *Am J Hum Genet* 32, pp. 314–31.

Bottazzo, G. F., B. M. Dean, J. M. McNally, E. H. MacKay, P. G. Swift, and D. R. Gamble (1985). "In situ characterization of autoimmune phenomena and expression of HLA molecules in the pancreas in diabetic insulitis". *N Engl J Med* 313, pp. 353–60.

Brem, R. B., G. Yvert, R. Clinton, and L. Kruglyak (2002). "Genetic dissection of transcriptional regulation in budding yeast". *Science* 296, pp. 752–5.

Brenowitz, M., D. F. Senear, M. A. Shea, and G. K. Ackers (1986). "Quantitative DNase footprint titration: a method for studying protein-DNA interactions". *Methods Enzymol* 130, pp. 132–81.

Brewerton, D. A., F. D. Hart, A. Nicholls, M. Caffrey, D. C. James, and R. D. Sturrock (1973). "Ankylosing spondylitis and HL-A 27". *Lancet* 1, pp. 904–7.

Brookes, A. J. (1999). "The essence of SNPs". *Gene* 234, pp. 177–86.

Brown, D., J. Trowsdale, and R. Allen (2004). "The LILR family: modulators of innate and adaptive immune pathways in health and disease". *Tissue Antigens* 64, pp. 215–25.

Bugawan, T. L., G. Angelini, J. Larrick, S. Auricchio, G. B. Ferrara, and H. A. Erlich (1989). "A combination of a particular HLA-DP beta allele and an HLA-DQ heterodimer confers susceptibility to coeliac disease". *Nature* 339, pp. 470–3.

Bullaughey, K., C. I. Chavarria, G. Coop, and Y. Gilad (2009). "Expression quantitative trait loci detected in cell lines are often present in primary tissues". *Hum Mol Genet* 18, pp. 4296–303.

Burton, P. R., A. L. Hansell, I. Fortier, T. A. Manolio, M. J. Khoury, J. Little, and P. Elliott (2009). "Size matters: just how big is BIG?: Quantifying realistic sample size requirements for human genome epidemiology". *Int J Epidemiol* 38, pp. 263–73.

Calin, G. A. and C. M. Croce (2006). "MicroRNA-cancer connection: the beginning of a new tale". *Cancer Res* 66, pp. 7390–4.

Caliskan, M., D. A. Cusanovich, C. Ober, and Y. Gilad (2011). "The effects of EBV transformation on gene expression levels and methylation profiles". *Hum Mol Genet* 20, pp. 1643–52.

Camacho, J. A. and N. Rioseco-Camacho (2009). "The human and mouse SLC25A29 mitochondrial transporters rescue the deficient ornithine metabolism in fibroblasts of patients with the hyperornithinemia-hyperammonemia-homocitrullinuria (HHH) syndrome". *Pediatr Res* 66, pp. 35–41.

Cambi, A. and C. G. Figdor (2003). "Dual function of C-type lectin-like receptors in the immune system". *Curr Opin Cell Biol* 15, pp. 539–46.

Campbell, R. D. and J. Trowsdale (1993). "Map of the human MHC". *Immunol Today* 14, pp. 349–52.

Cantero-Recasens, G., C. Fandos, F. Rubio-Moscardo, M. A. Valverde, and R. Vicente (2010). "The asthma-associated ORMDL3 gene product regulates endoplasmic reticulum-mediated calcium signaling and cellular stress". *Hum Mol Genet* 19, pp. 111–21.

Capon, F., M. J. Bijlmakers, N. Wolf, M. Quaranta, U. Huffmeier, M. Allen, K. Timms, V. Abkevich, A. Gutin, R. Smith, R. B. Warren, H. S. Young, J. Worthington, A. D. Burden, C. E. Griffiths, A. Hayday, F. O. Nestle, A. Reis, J. Lanchbury, J. N. Barker, and R. C. Trembath (2008). "Identification of ZNF313/RNF114 as a novel psoriasis susceptibility gene". *Hum Mol Genet* 17, pp. 1938–45.

Cardon, L. R. and J. I. Bell (2001). "Association study designs for complex diseases". *Nat Rev Genet* 2, pp. 91–9.

Cardon, L. R. and L. J. Palmer (2003). "Population stratification and spurious allelic association". *Lancet* 361, pp. 598–604.
Carlson, C. S., M. A. Eberle, M. J. Rieder, Q. Yi, L. Kruglyak, and D. A. Nickerson (2004). "Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium". *Am J Hum Genet* 74, pp. 106–20.
Carrington, M., G. W. Nelson, M. P. Martin, T. Kissner, D. Vlahov, J. J. Goedert, R. Kaslow, S. Buchbinder, K. Hoots, and S. J. O'Brien (1999). "HLA and HIV-1: heterozygote advantage and B*35-Cw*04 disadvantage". *Science* 283, pp. 1748–52.
Carroll, J. S., X. S. Liu, A. S. Brodsky, W. Li, C. A. Meyer, A. J. Szary, J. Eeckhoute, W. Shao, E. V. Hestermann, T. R. Geistlinger, E. A. Fox, P. A. Silver, and M. Brown (2005). "Chromosome-wide mapping of estrogen receptor binding reveals long-range regulation requiring the forkhead protein FoxA1". *Cell* 122, pp. 33–43.
Carroll, M. C., R. D. Campbell, D. R. Bentley, and R. R. Porter (1984). "A molecular map of the human major histocompatibility complex class III region linking complement genes C4, C2 and factor B". *Nature* 307, pp. 237–41.
Chapman, S. J. and A. V. Hill (2012). "Human genetic susceptibility to infectious disease". *Nat Rev Genet* 13, pp. 175–88.
Chen, G., M. H. Shaw, Y. G. Kim, and G. Nunez (2009). "NOD-like receptors: role in innate immunity and inflammatory disease". *Annu Rev Pathol* 4, pp. 365–98.
Chen, K. and N. Rajewsky (2007). "The evolution of gene regulation by transcription factors and microRNAs". *Nat Rev Genet* 8, pp. 93–103.
Cheung, V. G., L. K. Conlin, T. M. Weber, M. Arcaro, K. Y. Jen, M. Morley, and R. S. Spielman (2003). "Natural variation in human gene expression assessed in lymphoblastoid cells". *Nat Genet* 33, pp. 422–5.
Cheung, V. G., R. S. Spielman, K. G. Ewens, T. M. Weber, M. Morley, and J. T. Burdick (2005). "Mapping determinants of human gene expression by regional and genome-wide association". *Nature* 437, pp. 1365–9.
Cho, J. H. and P. K. Gregersen (2011). "Genomics and the multifactorial nature of human autoimmune disease". *N Engl J Med* 365, pp. 1612–23.
Cho, Y. J. and P. Liang (2008). "Killin is a p53-regulated nuclear inhibitor of DNA synthesis". *Proc Natl Acad Sci U S A* 105, pp. 5396–401.
Choy, E. (2012). "Understanding the dynamics: pathways involved in the pathogenesis of rheumatoid arthritis". *Rheumatology (Oxford)* 51 Suppl 5, pp. v3–11.
Choy, E., R. Yelensky, S. Bonakdar, R. M. Plenge, R. Saxena, P. L. De Jager, S. Y. Shaw, C. S. Wolfish, J. M. Slavik, C. Cotsapas, M. Rivas, E. T. Dermitzakis, E. Cahir-McFarland, E. Kieff, D. Hafler, M. J. Daly, and D. Altshuler (2008). "Genetic analysis of human traits in vitro: drug response and gene expression in lymphoblastoid cell lines". *PLoS Genet* 4, e1000287.
Clarke, L. and J. Carbon (1980). "Isolation of the centromere-linked CDC10 gene by complementation in yeast". *Proc Natl Acad Sci U S A* 77, pp. 2173–7.
Cloutier, P., M. Lavallee-Adam, D. Faubert, M. Blanchette, and B. Coulombe (2013). "A newly uncovered group of distantly related lysine methyltransferases preferentially interact with molecular chaperones to regulate their activity". *PLoS Genet* 9, e1003210.
Consortium, the MHC sequencing (1999). "Complete sequence and gene map of a human major histocompatibility complex. The MHC sequencing consortium". *Nature* 401, pp. 921–3.
Cookson, W., L. Liang, G. Abecasis, M. Moffatt, and M. Lathrop (2009). "Mapping complex disease traits with global gene expression". *Nat Rev Genet* 10, pp. 184–94.
Cooper, J. D., D. J. Smyth, A. M. Smiles, V. Plagnol, N. M. Walker, J. E. Allen, K. Downes, J. C. Barrett, B. C. Healy, J. C. Mychaleckyj, J. H. Warram, and J. A. Todd (2008). "Meta-analysis of genome-wide association study data identifies additional type 1 diabetes risk loci". *Nat Genet* 40, pp. 1399–401.
Corder, E. H., A. M. Saunders, N. J. Risch, W. J. Strittmatter, D. E. Schmechel, Jr. Gaskell P. C., J. B. Rimmler, P. A. Locke, P. M. Conneally, K. E. Schmader, and et al. (1994). "Protective effect of apolipoprotein E type 2 allele for late onset Alzheimer disease". *Nat Genet* 7, pp. 180–4.

Coronary Artery Disease (C4D) Genetics Consortium, the (2011). "A genome-wide association study in Europeans and South Asians identifies five new loci for coronary artery disease". *Nat Genet* 43, pp. 339–44.

Crick, F. (1970). "Central dogma of molecular biology". *Nature* 227, pp. 561–3.

Croft, D., G. O'Kelly, G. Wu, R. Haw, M. Gillespie, L. Matthews, M. Caudy, P. Garapati, G. Gopinath, B. Jassal, S. Jupe, I. Kalatskaya, S. Mahajan, B. May, N. Ndegwa, E. Schmidt, V. Shamovsky, C. Yung, E. Birney, H. Hermjakob, P. D'Eustachio, and L. Stein (2011). "Reactome: a database of reactions, pathways and biological processes". *Nucleic Acids Res* 39, pp. D691–7.

Crouser, E. D., D. A. Culver, K. S. Knox, M. W. Julian, G. Shao, S. Abraham, S. Liyanarachchi, J. E. Macre, M. D. Wewers, M. A. Gavrilin, P. Ross, A. Abbas, and C. Eng (2009). "Gene expression profiling identifies MMP-12 and ADAMDEC1 as potential pathogenic mediators of pulmonary sarcoidosis". *Am J Respir Crit Care Med* 179, pp. 929–38.

Cudworth, A. G. and J. C. Woodrow (1975). "Evidence for HL-A-linked genes in "juvenile" diabetes mellitus". *Br Med J* 3, pp. 133–5.

Dauphinot, L., C. De Oliveira, T. Melot, N. Sevenet, V. Thomas, B. E. Weissman, and O. Delattre (2001). "Analysis of the expression of cell cycle regulators in Ewing cell lines: EWS-FLI-1 modulates p57KIP2and c-Myc expression". *Oncogene* 20, pp. 3258–65.

Dausset, J., H. Cann, D. Cohen, M. Lathrop, J. M. Lalouel, and R. White (1990). "Centre d'etude du polymorphisme humain (CEPH): collaborative genetic mapping of the human genome". *Genomics* 6, pp. 575–7.

Deeb, S. S., L. Fajas, M. Nemoto, J. Pihlajamaki, L. Mykkanen, J. Kuusisto, M. Laakso, W. Fujimoto, and J. Auwerx (1998). "A Pro12Ala substitution in PPARgamma2 associated with decreased receptor activity, lower body mass index and improved insulin sensitivity". *Nat Genet* 20, pp. 284–7.

Delattre, O., J. Zucman, B. Plougastel, C. Desmaze, T. Melot, M. Peter, H. Kovar, I. Joubert, P. de Jong, G. Rouleau, and et al. (1992). "Gene fusion with an ETS DNA-binding domain caused by chromosome translocation in human tumours". *Nature* 359, pp. 162–5.

DeLisser, H. M., M. Christofidou-Solomidou, R. M. Strieter, M. D. Burdick, C. S. Robinson, R. S. Wexler, J. S. Kerr, C. Garlanda, J. R. Merwin, J. A. Madri, and S. M. Albelda (1997). "Involvement of endothelial PECAM-1/CD31 in angiogenesis". *Am J Pathol* 151, pp. 671–7.

Devlin, B. and N. Risch (1995). "A comparison of linkage disequilibrium measures for fine-scale mapping". *Genomics* 29, pp. 311–22.

Dib, C., S. Faure, C. Fizames, D. Samson, N. Drouot, A. Vignal, P. Millasseau, S. Marc, J. Hazan, E. Seboun, M. Lathrop, G. Gyapay, J. Morissette, and J. Weissenbach (1996). "A comprehensive genetic map of the human genome based on 5,264 microsatellites". *Nature* 380, pp. 152–4.

Dilthey, A., S. Leslie, L. Moutsianas, J. Shen, C. Cox, M. R. Nelson, and G. McVean (2013). "Multi-population classical HLA type imputation". *PLoS Comput Biol* 9, e1002877.

Dilthey, A. T., L. Moutsianas, S. Leslie, and G. McVean (2011). "HLA*IMP–an integrated framework for imputing classical HLA alleles from SNP genotypes". *Bioinformatics* 27, pp. 968–72.

Dimas, A. S., S. Deutsch, B. E. Stranger, S. B. Montgomery, C. Borel, H. Attar-Cohen, C. Ingle, C. Beazley, M. Gutierrez Arcelus, M. Sekowska, M. Gagnebin, J. Nisbett, P. Deloukas, E. T. Dermitzakis, and S. E. Antonarakis (2009). "Common regulatory variation impacts gene expression in a cell type-dependent manner". *Science* 325, pp. 1246–50.

Dinarello, C. A. (2011). "Interleukin-1 in the pathogenesis and treatment of inflammatory diseases". *Blood* 117, pp. 3720–32.

Ding, J., J. E. Gudjonsson, L. Liang, P. E. Stuart, Y. Li, W. Chen, M. Weichenthal, E. Ellinghaus, A. Franke, W. Cookson, R. P. Nair, J. T. Elder, and G. R. Abecasis (2010). "Gene expression in skin and lymphoblastoid cells: Refined statistical method reveals extensive overlap in cis-eQTL signals". *Am J Hum Genet* 87, pp. 779–89.

Dixon, A. L., L. Liang, M. F. Moffatt, W. Chen, S. Heath, K. C. Wong, J. Taylor, E. Burnett, I. Gut, M. Farrall, G. M. Lathrop, G. R. Abecasis, and W. O. Cookson (2007). "A genome-wide association study of global gene expression". *Nat Genet* 39, pp. 1202–7.

Djebali, S., C. A. Davis, A. Merkel, A. Dobin, T. Lassmann, A. Mortazavi, A. Tanzer, J. Lagarde, W. Lin, F. Schlesinger, C. Xue, G. K. Marinov, J. Khatun, B. A. Williams, C. Zaleski, J. Rozowsky, M. Roder, F. Kokocinski, R. F. Abdelhamid, T. Alioto, I. Antoshechkin, M. T. Baer, N. S. Bar, P. Batut, K. Bell, I. Bell, S. Chakrabortty, X. Chen, J. Chrast, J. Curado, T. Derrien, J. Drenkow, E. Dumais, J. Dumais, R. Duttagupta, E. Falconnet, M. Fastuca, K. Fejes-Toth, P. Ferreira, S. Foissac, M. J. Fullwood, H. Gao, D. Gonzalez, A. Gordon, H. Gunawardena, C. Howald, S. Jha, R. Johnson, P. Kapranov, B. King, C. Kingswood, O. J. Luo, E. Park, K. Persaud, J. B. Preall, P. Ribeca, B. Risk, D. Robyr, M. Sammeth, L. Schaffer, L. H. See, A. Shahab, J. Skancke, A. M. Suzuki, H. Takahashi, H. Tilgner, D. Trout, N. Walters, H. Wang, J. Wrobel, Y. Yu, X. Ruan, Y. Hayashizaki, J. Harrow, M. Gerstein, T. Hubbard, A. Reymond, S. E. Antonarakis, G. Hannon, M. C. Giddings, Y. Ruan, B. Wold, P. Carninci, R. Guigo, and T. R. Gingeras (2012). "Landscape of transcription in human cells". *Nature* 489, pp. 101–8.

Donath, M. Y. and S. E. Shoelson (2011). "Type 2 diabetes as an inflammatory disease". *Nat Rev Immunol* 11, pp. 98–107.

Donis-Keller, H., P. Green, C. Helms, S. Cartinhour, B. Weiffenbach, K. Stephens, T. P. Keith, D. W. Bowden, D. R. Smith, E. S. Lander, and et al. (1987). "A genetic linkage map of the human genome". *Cell* 51, pp. 319–37.

Dougherty, D. C. and M. M. Sanders (2005). "Comparison of the responsiveness of the pGL3 and pGL4 luciferase reporter vectors to steroid hormones". *Biotechniques* 39, pp. 203–7.

Draghici, S., P. Khatri, A. C. Eklund, and Z. Szallasi (2006). "Reliability and reproducibility issues in DNA microarray measurements". *Trends Genet* 22, pp. 101–9.

Drieschner, N., G. Belge, V. Rippe, M. Meiboom, S. Loeschke, and J. Bullerdiek (2006). "Evidence for a 3p25 breakpoint hot spot region in thyroid tumors of follicular origin". *Thyroid* 16, pp. 1091–6.

Drieschner, N., S. Kerschling, J. T. Soller, V. Rippe, G. Belge, J. Bullerdiek, and R. Nimzyk (2007). "A domain of the thyroid adenoma associated gene (THADA) conserved in vertebrates becomes destroyed by chromosomal rearrangements observed in thyroid adenomas". *Gene* 403, pp. 110–7.

Du, P., W. A. Kibbe, and S. M. Lin (2008). "lumi: a pipeline for processing Illumina microarray". *Bioinformatics* 24, pp. 1547–8.

Dubois, P. C., G. Trynka, L. Franke, K. A. Hunt, J. Romanos, A. Curtotti, A. Zhernakova, G. A. Heap, R. Adany, A. Aromaa, M. T. Bardella, L. H. van den Berg, N. A. Bockett, E. G. de la Concha, B. Dema, R. S. Fehrmann, M. Fernandez-Arquero, S. Fiatal, E. Grandone, P. M. Green, H. J. Groen, R. Gwilliam, R. H. Houwen, S. E. Hunt, K. Kaukinen, D. Kelleher, I. Korponay-Szabo, K. Kurppa, P. MacMathuna, M. Maki, M. C. Mazzilli, O. T. McCann, M. L. Mearin, C. A. Mein, M. M. Mirza, V. Mistry, B. Mora, K. I. Morley, C. J. Mulder, J. A. Murray, C. Nunez, E. Oosterom, R. A. Ophoff, I. Polanco, L. Peltonen, M. Platteel, A. Rybak, V. Salomaa, J. J. Schweizer, M. P. Sperandeo, G. J. Tack, G. Turner, J. H. Veldink, W. H. Verbeek, R. K. Weersma, V. M. Wolters, E. Urcelay, B. Cukrowska, L. Greco, S. L. Neuhausen, R. McManus, D. Barisani, P. Deloukas, J. C. Barrett, P. Saavalainen, C. Wijmenga, and D. A. van Heel (2010). "Multiple common variants for celiac disease influencing immune gene expression". *Nat Genet* 42, pp. 295–302.

Dudoit, S., Y. H. Yang, M. J. Callow, and T. P. Speed (2002). "Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments". *Statistica Sinica* 12, pp. 111–39.

Durbin, R. M., G. R. Abecasis, D. L. Altshuler, A. Auton, L. D. Brooks, R. A. Gibbs, M. E. Hurles, and G. A. McVean (2010). "A map of human genome variation from population-scale sequencing". *Nature* 467, pp. 1061–73.

Dymecki, S. M., J. E. Niederhuber, and S. V. Desiderio (1990). "Specific expression of a tyrosine kinase gene, blk, in B lymphoid cells". *Science* 247, pp. 332–6.

Eckner, R., M. E. Ewen, D. Newsome, M. Gerdes, J. A. DeCaprio, J. B. Lawrence, and D. M. Livingston (1994). "Molecular cloning and functional analysis of the adenovirus E1A-associated 300-kD protein (p300) reveals a protein with properties of a transcriptional adaptor". *Genes Dev* 8, pp. 869–84.

Eckner, R., T. P. Yao, E. Oldread, and D. M. Livingston (1996). "Interaction and functional collaboration of p300/CBP and bHLH proteins in muscle and B-cell differentiation". *Genes Dev* 10, pp. 2478–90.

Edwards, J. C., G. Cambridge, and V. M. Abrahams (1999). "Do self-perpetuating B lymphocytes drive human autoimmune disease?" *Immunology* 97, pp. 188–96.

Ellinghaus, E., D. Ellinghaus, P. E. Stuart, R. P. Nair, S. Debrus, J. V. Raelson, M. Belouchi, H. Fournier, C. Reinhard, J. Ding, Y. Li, T. Tejasvi, J. Gudjonsson, S. W. Stoll, J. J. Voorhees, S. Lambert, S. Weidinger, B. Eberlein, M. Kunz, P. Rahman, D. D. Gladman, C. Gieger, H. E. Wichmann, T. H. Karlsen, G. Mayr, M. Albrecht, D. Kabelitz, U. Mrowietz, G. R. Abecasis, J. T. Elder, S. Schreiber, M. Weichenthal, and A. Franke (2010). "Genome-wide association study identifies a psoriasis susceptibility locus at TRAF3IP2". *Nat Genet* 42, pp. 991–5.

Elstak, E. D., M. Neeft, N. T. Nehme, J. Voortman, M. Cheung, M. Goodarzifard, H. C. Gerritsen, P. M. van Bergen En Henegouwen, I. Callebaut, G. de Saint Basile, and P. van der Sluijs (2011). "The munc13-4-rab27 complex is specifically required for tethering secretory lysosomes at the plasma membrane". *Blood* 118, pp. 1570–8.

Elzen, P. van den, S. Garg, L. Leon, M. Brigl, E. A. Leadbetter, J. E. Gumperz, C. C. Dascher, T. Y. Cheng, F. M. Sacks, P. A. Illarionov, G. S. Besra, S. C. Kent, D. B. Moody, and M. B. Brenner (2005). "Apolipoprotein-mediated pathways of lipid antigen presentation". *Nature* 437, pp. 906–10.

Emilsson, V., G. Thorleifsson, B. Zhang, A. S. Leonardson, F. Zink, J. Zhu, S. Carlson, A. Helgason, G. B. Walters, S. Gunnarsdottir, M. Mouy, V. Steinthorsdottir, G. H. Eiriksdottir, G. Bjornsdottir, I. Reynisdottir, D. Gudbjartsson, A. Helgadottir, A. Jonasdottir, A. Jonasdottir, U. Styrkarsdottir, S. Gretarsdottir, K. P. Magnusson, H. Stefansson, R. Fossdal, K. Kristjansson, H. G. Gislason, T. Stefansson, B. G. Leifsson, U. Thorsteinsdottir, J. R. Lamb, J. R. Gulcher, M. L. Reitman, A. Kong, E. E. Schadt, and K. Stefansson (2008). "Genetics of gene expression and its effect on disease". *Nature* 452, pp. 423–8.

ENCODE Project Consortium, the (2004). "The ENCODE (ENCyclopedia Of DNA Elements) Project". *Science* 306, pp. 636–40.

Esteller, M. (2007). "Cancer epigenomics: DNA methylomes and histone-modification maps". *Nat Rev Genet* 8, pp. 286–98.

Esteller, M., J. M. Silva, G. Dominguez, F. Bonilla, X. Matias-Guiu, E. Lerma, E. Bussaglia, J. Prat, I. C. Harkes, E. A. Repasky, E. Gabrielson, M. Schutte, S. B. Baylin, and J. G. Herman (2000). "Promoter hypermethylation and BRCA1 inactivation in sporadic breast and ovarian tumors". *J Natl Cancer Inst* 92, pp. 564–9.

Evans, D. M., C. C. Spencer, J. J. Pointon, Z. Su, D. Harvey, G. Kochan, U. Oppermann, A. Dilthey, M. Pirinen, M. A. Stone, L. Appleton, L. Moutsianas, S. Leslie, T. Wordsworth, T. J. Kenna, T. Karaderi, G. P. Thomas, M. M. Ward, M. H. Weisman, C. Farrar, L. A. Bradbury, P. Danoy, R. D. Inman, W. Maksymowych, D. Gladman, P. Rahman, A. Morgan, H. Marzo-Ortega, P. Bowness, K. Gaffney, J. S. Gaston, M. Smith, J. Bruges-Armas, A. R. Couto, R. Sorrentino, F. Paladini, M. A. Ferreira, H. Xu, Y. Liu, L. Jiang, C. Lopez-Larrea, R. Diaz-Pena, A. Lopez-Vazquez, T. Zayats, G. Band, C. Bellenguez, H. Blackburn, J. M. Blackwell, E. Bramon, S. J. Bumpstead, J. P. Casas, A. Corvin, N. Craddock, P. Deloukas, S. Dronov, A. Duncanson, S. Edkins, C. Freeman, M. Gillman, E. Gray, R. Gwilliam, N. Hammond, S. E. Hunt, J. Jankowski, A. Jayakumar, C. Langford, J. Liddle, H. S. Markus, C. G. Mathew, O. T. McCann, M. I. McCarthy, C. N. Palmer, L. Peltonen, R. Plomin, S. C. Potter, A. Rautanen, R. Ravindrarajah, M. Ricketts, N. Samani, S. J. Sawcer, A. Strange, R. C. Trembath, A. C. Viswanathan, M. Waller, P. Weston, P. Whittaker, S. Widaa, N. W. Wood, G. McVean, J. D. Reveille, B. P. Wordsworth, M. A. Brown, and P. Donnelly (2011). "Interaction between ERAP1 and HLA-B27 in ankylosing spondylitis implicates peptide handling in the mechanism for HLA-B27 in disease susceptibility". *Nat Genet* 43, pp. 761–7.

Fairfax, B. P., S. Makino, J. Radhakrishnan, K. Plant, S. Leslie, A. Dilthey, P. Ellis, C. Langford, F. O. Vannberg, and J. C. Knight (2012). "Genetics of gene expression in primary immune cells identifies cell type-specific master regulators and roles of HLA alleles". *Nat Genet* 44, pp. 502–10.

Fearon, D. T. and R. M. Locksley (1996). "The instructive role of innate immunity in the acquired immune response". *Science* 272, pp. 50–3.

Fehrmann, R. S., R. C. Jansen, J. H. Veldink, H. J. Westra, D. Arends, M. J. Bonder, J. Fu, P. Deelen, H. J. Groen, A. Smolonska, R. K. Weersma, R. M. Hofstra, W. A. Buurman, S. Rensen, M. G. Wolfs, M. Platteel, A. Zhernakova, C. C. Elbers, E. M. Festen, G. Trynka, M. H. Hofker, C. G. Saris, R. A. Ophoff, L. H. van den Berg, D. A. van Heel, C. Wijmenga, G. J. Te Meerman, and L. Franke (2011). "Trans-eQTLs Reveal That Independent Genetic Variants Associated with a Complex Phenotype Converge on Intermediate Genes, with a Major Role for the HLA". *PLoS Genet* 7, e1002197.

Feinberg, A. P. (2007). "Phenotypic plasticity and the epigenetics of human disease". *Nature* 447, pp. 433–40.

Feinberg, A. P., R. Ohlsson, and S. Henikoff (2006). "The epigenetic progenitor origin of human cancer". *Nat Rev Genet* 7, pp. 21–33.

Ferreira, M. A., A. F. McRae, S. E. Medland, D. R. Nyholt, S. D. Gordon, M. J. Wright, A. K. Henders, P. A. Madden, P. M. Visscher, N. R. Wray, A. C. Heath, G. W. Montgomery, D. L. Duffy, and N. G. Martin (2011). "Association between ORMDL3, IL1RL1 and a deletion on chromosome 17q21 with asthma risk in Australia". *Eur J Hum Genet* 19, pp. 458–64.

Feuk, L., A. R. Carson, and S. W. Scherer (2006). "Structural variation in the human genome". *Nat Rev Genet* 7, pp. 85–97.

Fleming, A. and V.D. Allison (1922). "Observations on a Bacteriolytic Substance (Lysozyme) Found in Secretions and Tissues". *Br J Exp Pathol* 3, pp. 252–60.

Flutre, T., X. Wen, J. K. Pritchard, and M. Stephens (2013). "A statistical framework for joint eQTL analysis in multiple tissues."

Forwood, J. K. and D. A. Jans (2006). "Quantitative analysis of DNA-protein interactions using double-labeled native gel electrophoresis and fluorescence-based imaging". *Electrophoresis* 27, pp. 3166–70.

Foster, S. L., D. C. Hargreaves, and R. Medzhitov (2007). "Gene-specific control of inflammation by TLR-induced chromatin modifications". *Nature* 447, pp. 972–8.

Franke, A., D. P. McGovern, J. C. Barrett, K. Wang, G. L. Radford-Smith, T. Ahmad, C. W. Lees, T. Balschun, J. Lee, R. Roberts, C. A. Anderson, J. C. Bis, S. Bumpstead, D. Ellinghaus, E. M. Festen, M. Georges, T. Green, T. Haritunians, L. Jostins, A. Latiano, C. G. Mathew, G. W. Montgomery, N. J. Prescott, S. Raychaudhuri, J. I. Rotter, P. Schumm, Y. Sharma, L. A. Simms, K. D. Taylor, D. Whiteman, C. Wijmenga, R. N. Baldassano, M. Barclay, T. M. Bayless, S. Brand, C. Buning, A. Cohen, J. F. Colombel, M. Cottone, L. Stronati, T. Denson, M. De Vos, R. D'Inca, M. Dubinsky, C. Edwards, T. Florin, D. Franchimont, R. Gearry, J. Glas, A. Van Gossum, S. L. Guthery, J. Halfvarson, H. W. Verspaget, J. P. Hugot, A. Karban, D. Laukens, I. Lawrance, M. Lemann, A. Levine, C. Libioulle, E. Louis, C. Mowat, W. Newman, J. Panes, A. Phillips, D. D. Proctor, M. Regueiro, R. Russell, P. Rutgeerts, J. Sanderson, M. Sans, F. Seibold, A. H. Steinhart, P. C. Stokkers, L. Torkvist, G. Kullak-Ublick, D. Wilson, T. Walters, S. R. Targan, S. R. Brant, J. D. Rioux, M. D'Amato, R. K. Weersma, S. Kugathasan, A. M. Griffiths, J. C. Mansfield, S. Vermeire, R. H. Duerr, M. S. Silverberg, J. Satsangi, S. Schreiber, J. H. Cho, V. Annese, H. Hakonarson, M. J. Daly, and M. Parkes (2010). "Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci". *Nat Genet* 42, pp. 1118–25.

Franklin, R. E. and R. G. Gosling (1953). "Molecular configuration in sodium thymonucleate". *Nature* 171, pp. 740–1.

Frazer, K. A., D. G. Ballinger, D. R. Cox, D. A. Hinds, L. L. Stuve, R. A. Gibbs, J. W. Belmont, A. Boudreau, P. Hardenbol, S. M. Leal, S. Pasternak, D. A. Wheeler, T. D. Willis, F. Yu, H. Yang, C. Zeng, Y. Gao, H. Hu, W. Hu, C. Li, W. Lin, S. Liu, H. Pan, X. Tang, J. Wang, W. Wang, J. Yu, B. Zhang, Q. Zhang, H. Zhao, J. Zhou, S. B. Gabriel, R. Barry, B. Blumenstiel, A. Camargo, M. Defelice, M. Faggart, M. Goyette, S. Gupta,

J. Moore, H. Nguyen, R. C. Onofrio, M. Parkin, J. Roy, E. Stahl, E. Winchester, L. Ziaugra, D. Altshuler, Y. Shen, Z. Yao, W. Huang, X. Chu, Y. He, L. Jin, Y. Liu, W. Sun, H. Wang, Y. Wang, X. Xiong, L. Xu, M. M. Waye, S. K. Tsui, H. Xue, J. T. Wong, L. M. Galver, J. B. Fan, K. Gunderson, S. S. Murray, A. R. Oliphant, M. S. Chee, A. Montpetit, F. Chagnon, V. Ferretti, M. Leboeuf, J. F. Olivier, M. S. Phillips, S. Roumy, C. Sallee, A. Verner, T. J. Hudson, P. Y. Kwok, D. Cai, D. C. Koboldt, R. D. Miller, L. Pawlikowska, P. Taillon-Miller, M. Xiao, L. C. Tsui, W. Mak, Y. Q. Song, P. K. Tam, Y. Nakamura, T. Kawaguchi, T. Kitamoto, T. Morizono, A. Nagashima, Y. Ohnishi, A. Sekine, T. Tanaka, *et al.* (2007). "A second generation human haplotype map of over 3.1 million SNPs". *Nature* 449, pp. 851–61.

Freedman, M. L., A. N. Monteiro, S. A. Gayther, G. A. Coetzee, A. Risch, C. Plass, G. Casey, M. De Biasi, C. Carlson, D. Duggan, M. James, P. Liu, J. W. Tichelaar, H. G. Vikis, M. You, and I. G. Mills (2011). "Principles for the post-GWAS functional characterization of cancer risk loci". *Nat Genet* 43, pp. 513–8.

Freudenberg, J., H. S. Lee, B. G. Han, H. D. Shin, Y. M. Kang, Y. K. Sung, S. C. Shim, C. B. Choi, A. T. Lee, P. K. Gregersen, and S. C. Bae (2011). "Genome-wide association study of rheumatoid arthritis in Koreans: population-specific loci as well as overlap with European susceptibility loci". *Arthritis Rheum* 63, pp. 884–93.

Fried, M. and D. M. Crothers (1981). "Equilibria and kinetics of lac repressor-operator interactions by polyacrylamide gel electrophoresis". *Nucleic Acids Res* 9, pp. 6505–25.

Fu, J., M. G. Wolfs, P. Deelen, H. J. Westra, R. S. Fehrmann, G. J. Te Meerman, W. A. Buurman, S. S. Rensen, H. J. Groen, R. K. Weersma, L. H. van den Berg, J. Veldink, R. A. Ophoff, H. Snieder, D. van Heel, R. C. Jansen, M. H. Hofker, C. Wijmenga, and L. Franke (2012). "Unraveling the regulatory mechanisms underlying tissue-dependent genetic variation of gene expression". *PLoS Genet* 8, e1002431.

Fuchs, A., M. Cella, E. Giurisato, A. S. Shaw, and M. Colonna (2004). "Cutting edge: CD96 (tactile) promotes NK cell-target cell adhesion by interacting with the poliovirus receptor (CD155)". *J Immunol* 172, pp. 3994–8.

Fuchs, C. S., E. L. Giovannucci, G. A. Colditz, D. J. Hunter, F. E. Speizer, and W. C. Willett (1994). "A prospective study of family history and the risk of colorectal cancer". *N Engl J Med* 331, pp. 1669–74.

Fukao, T., D. M. Frucht, G. Yap, M. Gadina, J. J. O'Shea, and S. Koyasu (2001). "Inducible expression of Stat4 in dendritic cells and macrophages and its critical role in innate and adaptive immune responses". *J Immunol* 166, pp. 4446–55.

Furey, T. S. (2012). "ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions". *Nat Rev Genet* 13, pp. 840–52.

Gabay, C., C. Lamacchia, and G. Palmer (2010). "IL-1 pathways in inflammation and human diseases". *Nat Rev Rheumatol* 6, pp. 232–41.

Gabriel, S. B., S. F. Schaffner, H. Nguyen, J. M. Moore, J. Roy, B. Blumenstiel, J. Higgins, M. DeFelice, A. Lochner, M. Faggart, S. N. Liu-Cordero, C. Rotimi, A. Adeyemo, R. Cooper, R. Ward, E. S. Lander, M. J. Daly, and D. Altshuler (2002). "The structure of haplotype blocks in the human genome". *Science* 296, pp. 2225–9.

Gaffney, D. J. (2013). "Global properties and functional complexity of human gene regulatory variation". *PLoS Genet* 9, e1003501.

Gaffney, D. J., J. B. Veyrieras, J. F. Degner, R. Pique-Regi, A. A. Pai, G. E. Crawford, M. Stephens, Y. Gilad, and J. K. Pritchard (2012). "Dissecting the regulatory architecture of gene expression QTLs". *Genome Biol* 13, R7.

Garner, M. M. and A. Revzin (1981). "A gel electrophoresis method for quantifying the binding of proteins to specific DNA regions: application to components of the Escherichia coli lactose operon regulatory system". *Nucleic Acids Res* 9, pp. 3047–60.

Garrod, A.E. (1902). "The incidence of alkaptonuria: a study in chemical individuality". *Lancet* ii, pp. 1616–20.

Gartel, A. L. and S. K. Radhakrishnan (2005). "Lost in transcription: p21 repression, mechanisms, and consequences". *Cancer Res* 65, pp. 3980–5.

Gaszner, M. and G. Felsenfeld (2006). "Insulators: exploiting transcriptional and epigenetic mechanisms". *Nat Rev Genet* 7, pp. 703–13.

Geissmann, F., M. G. Manz, S. Jung, M. H. Sieweke, M. Merad, and K. Ley (2010). "Development of monocytes, macrophages, and dendritic cells". *Science* 327, pp. 656–61.

Geraghty, D. E., X. H. Wei, H. T. Orr, and B. H. Koller (1990). "Human leukocyte antigen F (HLA-F). An expressed HLA gene composed of a class I coding sequence linked to a novel transcribed repetitive element". *J Exp Med* 171, pp. 1–18.

Geremia, A., P. Biancheri, P. Allan, G. R. Corazza, and A. Di Sabatino (2013). "Innate and adaptive immunity in inflammatory bowel disease". *Autoimmun Rev*.

Gerrits, A., Y. Li, B. M. Tesson, L. V. Bystrykh, E. Weersing, A. Ausema, B. Dontje, X. Wang, R. Breitling, R. C. Jansen, and G. de Haan (2009). "Expression quantitative trait loci are highly sensitive to cellular differentiation state". *PLoS Genet* 5, e1000692.

Gerstein, M. B., A. Kundaje, M. Hariharan, S. G. Landt, K. K. Yan, C. Cheng, X. J. Mu, E. Khurana, J. Rozowsky, R. Alexander, R. Min, P. Alves, A. Abyzov, N. Addleman, N. Bhardwaj, A. P. Boyle, P. Cayting, A. Charos, D. Z. Chen, Y. Cheng, D. Clarke, C. Eastman, G. Euskirchen, S. Frietze, Y. Fu, J. Gertz, F. Grubert, A. Harmanci, P. Jain, M. Kasowski, P. Lacroute, J. Leng, J. Lian, H. Monahan, H. O'Geen, Z. Ouyang, E. C. Partridge, D. Patacsil, F. Pauli, D. Raha, L. Ramirez, T. E. Reddy, B. Reed, M. Shi, T. Slifer, J. Wang, L. Wu, X. Yang, K. Y. Yip, G. Zilberman-Schapira, S. Batzoglou, A. Sidow, P. J. Farnham, R. M. Myers, S. M. Weissman, and M. Snyder (2012). "Architecture of the human regulatory network derived from ENCODE data". *Nature* 489, pp. 91–100.

Giguere, V., M. Tini, G. Flock, E. Ong, R. M. Evans, and G. Otulakowski (1994). "Isoform-specific amino-terminal domains dictate DNA-binding properties of ROR alpha, a novel family of orphan hormone nuclear receptors". *Genes Dev* 8, pp. 538–53.

Glocker, E. O., A. Hennigs, M. Nabavi, A. A. Schaffer, C. Woellner, U. Salzer, D. Pfeifer, H. Veelken, K. Warnatz, F. Tahami, S. Jamal, A. Manguiat, N. Rezaei, A. A. Amirzargar, A. Plebani, N. Hannesschlager, O. Gross, J. Ruland, and B. Grimbacher (2009). "A homozygous CARD9 mutation in a family with susceptibility to fungal infections". *N Engl J Med* 361, pp. 1727–35.

Goldstein, J. L., S. E. Dana, J. R. Faust, A. L. Beaudet, and M. S. Brown (1975). "Role of lysosomal acid lipase in the metabolism of plasma low density lipoprotein. Observations in cultured fibroblasts from a patient with cholesteryl ester storage disease". *J Biol Chem* 250, pp. 8487–95.

Golub, T. R., D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander (1999). "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring". *Science* 286, pp. 531–7.

Gonzalez, E., H. Kulkarni, H. Bolivar, A. Mangano, R. Sanchez, G. Catano, R. J. Nibbs, B. I. Freedman, M. P. Quinones, M. J. Bamshad, K. K. Murthy, B. H. Rovin, W. Bradley, R. A. Clark, S. A. Anderson, J. O'Connell R, B. K. Agan, S. S. Ahuja, R. Bologna, L. Sen, M. J. Dolan, and S. K. Ahuja (2005). "The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility". *Science* 307, pp. 1434–40.

Gordon, S. and F. O. Martinez (2010). "Alternative activation of macrophages: mechanism and functions". *Immunity* 32, pp. 593–604.

Gordon, S. and P. R. Taylor (2005). "Monocyte and macrophage heterogeneity". *Nat Rev Immunol* 5, pp. 953–64.

Goring, H. H., J. E. Curran, M. P. Johnson, T. D. Dyer, J. Charlesworth, S. A. Cole, J. B. Jowett, L. J. Abraham, D. L. Rainwater, A. G. Comuzzie, M. C. Mahaney, L. Almasy, J. W. MacCluer, A. H. Kissebah, G. R. Collier, E. K. Moses, and J. Blangero (2007). "Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes". *Nat Genet* 39, pp. 1208–16.

Grabstein, K. H., D. L. Urdal, R. J. Tushinski, D. Y. Mochizuki, V. L. Price, M. A. Cantrell, S. Gillis, and P. J. Conlon (1986). "Induction of macrophage tumoricidal activity by granulocyte-macrophage colony-stimulating factor". *Science* 232, pp. 506–8.

Graham, R. R., S. V. Kozyrev, E. C. Baechler, M. V. Reddy, R. M. Plenge, J. W. Bauer, W. A. Ortmann, T. Koeuth, M. F. Gonzalez Escribano, B. Pons-Estel, M. Petri, M. Daly, P. K. Gregersen, J. Martin, D. Altshuler, T. W. Behrens, and M. E. Alarcon-Riquelme (2006). "A common haplotype of interferon regulatory factor 5 (IRF5) regulates splicing and

expression and is associated with increased risk of systemic lupus erythematosus". *Nat Genet* 38, pp. 550–5.

Gratchev, A., J. Kzhyshkowska, S. Kannookadan, M. Ochsenreiter, A. Popova, X. Yu, S. Mamidi, E. Stonehouse-Usselmann, I. Muller-Molinet, L. Gooi, and S. Goerdt (2008). "Activation of a TGF-beta-specific multistep gene expression program in mature macrophages requires glucocorticoid-mediated surface expression of TGF-beta receptor II". *J Immunol* 180, pp. 6553–65.

Gregersen, P. K., J. Silver, and R. J. Winchester (1987). "The shared epitope hypothesis. An approach to understanding the molecular genetics of susceptibility to rheumatoid arthritis". *Arthritis Rheum* 30, pp. 1205–13.

Gregersen, P. K., C. I. Amos, A. T. Lee, Y. Lu, E. F. Remmers, D. L. Kastner, M. F. Seldin, L. A. Criswell, R. M. Plenge, V. M. Holers, T. R. Mikuls, T. Sokka, L. W. Moreland, Jr. Bridges S. L., G. Xie, A. B. Begovich, and K. A. Siminovitch (2009). "REL, encoding a member of the NF-kappaB family of transcription factors, is a newly defined risk locus for rheumatoid arthritis". *Nat Genet* 41, pp. 820–3.

Gregory, C. D., C. Dive, S. Henderson, C. A. Smith, G. T. Williams, J. Gordon, and A. B. Rickinson (1991). "Activation of Epstein-Barr virus latent genes protects human B cells from death by apoptosis". *Nature* 349, pp. 612–4.

Grivennikov, S. I., F. R. Greten, and M. Karin (2010). "Immunity, inflammation, and cancer". *Cell* 140, pp. 883–99.

Gross, D. S. and W. T. Garrard (1988). "Nuclease hypersensitive sites in chromatin". *Annu Rev Biochem* 57, pp. 159–97.

Gross, O., A. Gewies, K. Finger, M. Schafer, T. Sparwasser, C. Peschel, I. Forster, and J. Ruland (2006). "Card9 controls a non-TLR signalling pathway for innate anti-fungal immunity". *Nature* 442, pp. 651–6.

Grundberg, E., T. Kwan, B. Ge, K. C. Lam, V. Koka, A. Kindmark, H. Mallmin, J. Dias, D. J. Verlaan, M. Ouimet, D. Sinnett, F. Rivadeneira, K. Estrada, A. Hofman, J. M. van Meurs, A. Uitterlinden, P. Beaulieu, A. Graziani, E. Harmsen, O. Ljunggren, C. Ohlsson, D. Mellstrom, M. K. Karlsson, O. Nilsson, and T. Pastinen (2009). "Population genomics in a disease targeted primary cell model". *Genome Res* 19, pp. 1942–52.

Grundberg, E., K. S. Small, A. K. Hedman, A. C. Nica, A. Buil, S. Keildson, J. T. Bell, T. P. Yang, E. Meduri, A. Barrett, J. Nisbett, M. Sekowska, A. Wilk, S. Y. Shin, D. Glass, M. Travers, J. L. Min, S. Ring, K. Ho, G. Thorleifsson, A. Kong, U. Thorsteindottir, C. Ainali, A. S. Dimas, N. Hassanali, C. Ingle, D. Knowles, M. Krestyaninova, C. E. Lowe, P. Di Meglio, S. B. Montgomery, L. Parts, S. Potter, G. Surdulescu, L. Tsaprouni, S. Tsoka, V. Bataille, R. Durbin, F. O. Nestle, S. O'Rahilly, N. Soranzo, C. M. Lindgren, K. T. Zondervan, K. R. Ahmadi, E. E. Schadt, K. Stefansson, G. D. Smith, M. I. McCarthy, P. Deloukas, E. T. Dermitzakis, and T. D. Spector (2012). "Mapping cis- and trans-regulatory effects across multiple tissues in twins". *Nat Genet* 44, pp. 1084–9.

Gunderson, K. L., S. Kruglyak, M. S. Graige, F. Garcia, B. G. Kermani, C. Zhao, D. Che, T. Dickinson, E. Wickham, J. Bierle, D. Doucet, M. Milewski, R. Yang, C. Siegmund, J. Haas, L. Zhou, A. Oliphant, J. B. Fan, S. Barnard, and M. S. Chee (2004). "Decoding randomly ordered DNA arrays". *Genome Res* 14, pp. 870–7.

Gusella, J. F., N. S. Wexler, P. M. Conneally, S. L. Naylor, M. A. Anderson, R. E. Tanzi, P. C. Watkins, K. Ottina, M. R. Wallace, A. Y. Sakaguchi, and et al. (1983). "A polymorphic DNA marker genetically linked to Huntington's disease". *Nature* 306, pp. 234–8.

Hagen, F. S., F. J. Grant, J. L. Kuijper, C. A. Slaughter, C. R. Moomaw, K. Orth, P. J. O'Hara, and R. S. Munford (1991). "Expression and characterization of recombinant human acyloxyacyl hydrolase, a leukocyte enzyme that deacylates bacterial lipopolysaccharides". *Biochemistry* 30, pp. 8415–23.

Hakonarson, H., H. Q. Qu, J. P. Bradfield, L. Marchand, C. E. Kim, J. T. Glessner, R. Grabs, T. Casalunovo, S. P. Taback, E. C. Frackelton, A. W. Eckert, K. Annaiah, M. L. Lawson, F. G. Otieno, E. Santa, J. L. Shaner, R. M. Smith, C. C. Onyiah, R. Skraban, R. M. Chiavacci, L. J. Robinson, C. A. Stanley, S. E. Kirsch, M. Devoto, D. S. Monos, S. F. Grant, and C. Polychronakos (2008). "A novel susceptibility locus for

type 1 diabetes on Chr12q13 identified by a genome-wide association study". *Diabetes* 57, pp. 1143–6.

Hammer, G. E., F. Gonzalez, M. Champsaur, D. Cado, and N. Shastri (2006). "The aminopeptidase ERAAP shapes the peptide repertoire displayed by major histocompatibility complex class I molecules". *Nat Immunol* 7, pp. 103–12.

Hammer, J., F. Gallazzi, E. Bono, R. W. Karr, J. Guenot, P. Valsasnini, Z. A. Nagy, and F. Sinigaglia (1995). "Peptide binding specificity of HLA-DR4 molecules: correlation with rheumatoid arthritis association". *J Exp Med* 181, pp. 1847–55.

Han, C., J. Jin, S. Xu, H. Liu, N. Li, and X. Cao (2010). "Integrin CD11b negatively regulates TLR-triggered inflammatory responses by activating Syk and promoting degradation of MyD88 and TRIF via Cbl-b". *Nat Immunol* 11, pp. 734–42.

Han, J. W., H. F. Zheng, Y. Cui, L. D. Sun, D. Q. Ye, Z. Hu, J. H. Xu, Z. M. Cai, W. Huang, G. P. Zhao, H. F. Xie, H. Fang, Q. J. Lu, J. H. Xu, X. P. Li, Y. F. Pan, D. Q. Deng, F. Q. Zeng, Z. Z. Ye, X. Y. Zhang, Q. W. Wang, F. Hao, L. Ma, X. B. Zuo, F. S. Zhou, W. H. Du, Y. L. Cheng, J. Q. Yang, S. K. Shen, J. Li, Y. J. Sheng, X. X. Zuo, W. F. Zhu, F. Gao, P. L. Zhang, Q. Guo, B. Li, M. Gao, F. L. Xiao, C. Quan, C. Zhang, Z. Zhang, K. J. Zhu, Y. Li, D. Y. Hu, W. S. Lu, J. L. Huang, S. X. Liu, H. Li, Y. Q. Ren, Z. X. Wang, C. J. Yang, P. G. Wang, W. M. Zhou, Y. M. Lv, A. P. Zhang, S. Q. Zhang, D. Lin, Y. Li, H. Q. Low, M. Shen, Z. F. Zhai, Y. Wang, F. Y. Zhang, S. Yang, J. J. Liu, and X. J. Zhang (2009). "Genome-wide association study in a Chinese Han population identifies nine new susceptibility loci for systemic lupus erythematosus". *Nat Genet* 41, pp. 1234–7.

Harris, R. M., R. H. Waring, C. J. Kirk, and P. J. Hughes (2000). "Sulfation of "estrogenic" alkylphenols and 17beta-estradiol by human platelet phenol sulfotransferases". *J Biol Chem* 275, pp. 159–66.

Hatters, D. M., C. A. Peters-Libeu, and K. H. Weisgraber (2006). "Apolipoprotein E structure: insights into function". *Trends Biochem Sci* 31, pp. 445–54.

Heinzen, E. L., D. Ge, K. D. Cronin, J. M. Maia, K. V. Shianna, W. N. Gabriel, K. A. Welsh-Bohmer, C. M. Hulette, T. N. Denny, and D. B. Goldstein (2008). "Tissue-specific genetic control of splicing: implications for the study of complex traits". *PLoS Biol* 6, e1.

Hellman, L. M. and M. G. Fried (2007). "Electrophoretic mobility shift assay (EMSA) for detecting protein-nucleic acid interactions". *Nat Protoc* 2, pp. 1849–61.

Hershey, A. D. and M. Chase (1952). "Independent functions of viral protein and nucleic acid in growth of bacteriophage". *J Gen Physiol* 36, pp. 39–56.

Hesselberth, J. R., X. Chen, Z. Zhang, P. J. Sabo, R. Sandstrom, A. P. Reynolds, R. E. Thurman, S. Neph, M. S. Kuehn, W. S. Noble, S. Fields, and J. A. Stamatoyannopoulos (2009). "Global mapping of protein-DNA interactions in vivo by digital genomic footprinting". *Nat Methods* 6, pp. 283–9.

Hevener, A. L., W. He, Y. Barak, J. Le, G. Bandyopadhyay, P. Olson, J. Wilkes, R. M. Evans, and J. Olefsky (2003). "Muscle-specific Pparg deletion causes insulin resistance". *Nat Med* 9, pp. 1491–7.

Hindorff, L. A., P. Sethupathy, H. A. Junkins, E. M. Ramos, J. P. Mehta, F. S. Collins, and T. A. Manolio (2009). "Potential etiologic and functional implications of genome-wide association loci for human diseases and traits". *Proc Natl Acad Sci U S A* 106, pp. 9362–7.

Hirota, T., A. Takahashi, M. Kubo, T. Tsunoda, K. Tomita, S. Doi, K. Fujita, A. Miyatake, T. Enomoto, T. Miyagawa, M. Adachi, H. Tanaka, A. Niimi, H. Matsumoto, I. Ito, H. Masuko, T. Sakamoto, N. Hizawa, M. Taniguchi, J. J. Lima, C. G. Irvin, S. P. Peters, B. E. Himes, A. A. Litonjua, K. G. Tantisira, S. T. Weiss, N. Kamatani, Y. Nakamura, and M. Tamari (2011). "Genome-wide association study identifies three new susceptibility loci for adult asthma in the Japanese population". *Nat Genet* 43, pp. 893–6.

Hirschhorn, J. N. (2003). "Genetic epidemiology of type 1 diabetes". *Pediatr Diabetes* 4, pp. 87–100.

Holzelova, E., C. Vonarbourg, M. C. Stolzenberg, P. D. Arkwright, F. Selz, A. M. Prieur, S. Blanche, J. Bartunkova, E. Vilmer, A. Fischer, F. Le Deist, and F. Rieux-Laucat (2004). "Autoimmune lymphoproliferative syndrome with somatic Fas mutations". *N Engl J Med* 351, pp. 1409–18.

Hom, G., R. R. Graham, B. Modrek, K. E. Taylor, W. Ortmann, S. Garnier, A. T. Lee, S. A. Chung, R. C. Ferreira, P. V. Pant, D. G. Ballinger, R. Kosoy, F. Y. Demirci, M. I. Kamboh, A. H. Kao, C. Tian, I. Gunnarsson, A. A. Bengtsson, S. Rantapaa-Dahlqvist, M. Petri, S. Manzi, M. F. Seldin, L. Ronnblom, A. C. Syvanen, L. A. Criswell, P. K. Gregersen, and T. W. Behrens (2008). "Association of systemic lupus erythematosus with C8orf13-BLK and ITGAM-ITGAX". *N Engl J Med* 358, pp. 900–9.

Horton, R., L. Wilming, V. Rand, R. C. Lovering, E. A. Bruford, V. K. Khodiyar, M. J. Lush, S. Povey, Jr. Talbot C. C., M. W. Wright, H. M. Wain, J. Trowsdale, A. Ziegler, and S. Beck (2004). "Gene map of the extended human MHC". *Nat Rev Genet* 5, pp. 889–99.

Horton, R., R. Gibson, P. Coggill, M. Miretti, R. J. Allcock, J. Almeida, S. Forbes, J. G. Gilbert, K. Halls, J. L. Harrow, E. Hart, K. Howe, D. K. Jackson, S. Palmer, A. N. Roberts, S. Sims, C. A. Stewart, J. A. Traherne, S. Trevanion, L. Wilming, J. Rogers, P. J. de Jong, J. F. Elliott, S. Sawcer, J. A. Todd, J. Trowsdale, and S. Beck (2008). "Variation analysis and gene annotation of eight MHC haplotypes: the MHC Haplotype Project". *Immunogenetics* 60, pp. 1–18.

Howie, B. N., P. Donnelly, and J. Marchini (2009). "A flexible and accurate genotype imputation method for the next generation of genome-wide association studies". *PLoS Genet* 5, e1000529.

Hsieh, C. S., S. E. Macatonia, C. S. Tripp, S. F. Wolf, A. O'Garra, and K. M. Murphy (1993). "Development of TH1 CD4+ T cells through IL-12 produced by Listeria-induced macrophages". *Science* 260, pp. 547–9.

Hsu, Y. M., Y. Zhang, Y. You, D. Wang, H. Li, O. Duramad, X. F. Qin, C. Dong, and X. Lin (2007). "The adaptor protein CARD9 is required for innate immune responses to intracellular pathogens". *Nat Immunol* 8, pp. 198–205.

Huang, J., D. Ellinghaus, A. Franke, B. Howie, and Y. Li (2012). "1000 Genomes-based imputation identifies novel and refined associations for the Wellcome Trust Case Control Consortium phase 1 Data". *Eur J Hum Genet* 20, pp. 801–5.

Hubner, N., C. A. Wallace, H. Zimdahl, E. Petretto, H. Schulz, F. Maciver, M. Mueller, O. Hummel, J. Monti, V. Zidek, A. Musilova, V. Kren, H. Causton, L. Game, G. Born, S. Schmidt, A. Muller, S. A. Cook, T. W. Kurtz, J. Whittaker, M. Pravenec, and T. J. Aitman (2005). "Integrated transcriptional profiling and linkage analysis for identification of genes underlying disease". *Nat Genet* 37, pp. 243–53.

Huffmeier, U., S. Uebe, A. B. Ekici, J. Bowes, E. Giardina, E. Korendowych, K. Juneblad, M. Apel, R. McManus, P. Ho, I. N. Bruce, A. W. Ryan, F. Behrens, J. Lascorz, B. Bohm, H. Traupe, J. Lohmann, C. Gieger, H. E. Wichmann, C. Herold, M. Steffens, L. Klareskog, T. F. Wienker, O. Fitzgerald, G. M. Alenius, N. J. McHugh, G. Novelli, H. Burkhardt, A. Barton, and A. Reis (2010). "Common variants at TRAF3IP2 are associated with susceptibility to psoriatic arthritis and psoriasis". *Nat Genet* 42, pp. 996–9.

Hugot, J. P., M. Chamaillard, H. Zouali, S. Lesage, J. P. Cezard, J. Belaiche, S. Almer, C. Tysk, C. A. O'Morain, M. Gassull, V. Binder, Y. Finkel, A. Cortot, R. Modigliani, P. Laurent-Puig, C. Gower-Rousseau, J. Macry, J. F. Colombel, M. Sahbatou, and G. Thomas (2001). "Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease". *Nature* 411, pp. 599–603.

Hunter, D. J. (2005). "Gene-environment interactions in human diseases". *Nat Rev Genet* 6, pp. 287–98.

Hyvarinen, A. and E. Oja (2000). "Independent component analysis: algorithms and applications". *Neural Netw* 13, pp. 411–30.

Idaghdour, Y., W. Czika, K. V. Shianna, S. H. Lee, P. M. Visscher, H. C. Martin, K. Miclaus, S. J. Jadallah, D. B. Goldstein, R. D. Wolfinger, and G. Gibson (2010). "Geographical genomics of human leukocyte gene expression variation in southern Morocco". *Nat Genet* 42, pp. 62–7.

Illumina, Inc. (2006). *Illuminas Genotyping Data Normalization Methods*.

— (2010). *Infinium Genotyping Data Analysis*.

Imbeaud, S., E. Graudens, V. Boulanger, X. Barlet, P. Zaborski, E. Eveno, O. Mueller, A. Schroeder, and C. Auffray (2005). "Towards standardization of RNA quality assessment using user-independent classifiers of microcapillary electrophoresis traces". *Nucleic Acids Res* 33, e56.

Imielinski, M., R. N. Baldassano, A. Griffiths, R. K. Russell, V. Annese, M. Dubinsky, S. Kugathasan, J. P. Bradfield, T. D. Walters, P. Sleiman, C. E. Kim, A. Muise, K. Wang, J. T. Glessner, S. Saeed, H. Zhang, E. C. Frackelton, C. Hou, J. H. Flory, G. Otieno, R. M. Chiavacci, R. Grundmeier, M. Castro, A. Latiano, B. Dallapiccola, J. Stempak, D. J. Abrams, K. Taylor, D. McGovern, G. Silber, I. Wrobel, A. Quiros, J. C. Barrett, S. Hansoul, D. L. Nicolae, J. H. Cho, R. H. Duerr, J. D. Rioux, S. R. Brant, M. S. Silverberg, K. D. Taylor, M. M. Barmuda, A. Bitton, T. Dassopoulos, L. W. Datta, T. Green, A. M. Griffiths, E. O. Kistner, M. T. Murtha, M. D. Regueiro, J. I. Rotter, L. P. Schumm, A. H. Steinhart, S. R. Targan, R. J. Xavier, C. Libioulle, C. Sandor, M. Lathrop, J. Belaiche, O. Dewit, I. Gut, S. Heath, D. Laukens, M. Mni, P. Rutgeerts, A. Van Gossum, D. Zelenika, D. Franchimont, J. P. Hugot, M. de Vos, S. Vermeire, E. Louis, L. R. Cardon, C. A. Anderson, H. Drummond, E. Nimmo, T. Ahmad, N. J. Prescott, C. M. Onnie, S. A. Fisher, J. Marchini, J. Ghori, S. Bumpstead, R. Gwillam, M. Tremelling, P. Delukas, J. Mansfield, D. Jewell, J. Satsangi, C. G. Mathew, M. Parkes, M. Georges, M. J. Daly, M. B. Heyman, G. D. Ferry, B. Kirschner, J. Lee, J. Essers, R. Grand, *et al.* (2009). "Common variants at five new loci associated with early-onset inflammatory bowel disease". *Nat Genet* 41, pp. 1335–40.

Ingersoll, M. A., A. M. Platt, S. Potteaux, and G. J. Randolph (2011). "Monocyte trafficking in acute and chronic inflammation". *Trends Immunol* 32, pp. 470–7.

Ingram, V. M. (1956). "A specific chemical difference between the globins of normal human and sickle-cell anaemia haemoglobin". *Nature* 178, pp. 792–4.

Innocenti, F., G. M. Cooper, I. B. Stanaway, E. R. Gamazon, J. D. Smith, S. Mirkov, J. Ramirez, W. Liu, Y. S. Lin, C. Moloney, S. F. Aldred, N. D. Trinklein, E. Schuetz, D. A. Nickerson, K. E. Thummel, M. J. Rieder, A. E. Rettie, M. J. Ratain, N. J. Cox, and C. D. Brown (2011). "Identification, replication, and functional fine-mapping of expression quantitative trait loci in primary human liver tissue". *PLoS Genet* 7, e1002078.

International HapMap Consortium, the (2003). "The International HapMap Project". *Nature* 426, pp. 789–96.

— (2005). "A haplotype map of the human genome". *Nature* 437, pp. 1299–320.

Iorio, M. V., M. Ferracin, C. G. Liu, A. Veronese, R. Spizzo, S. Sabbioni, E. Magri, M. Pedriali, M. Fabbri, M. Campiglio, S. Menard, J. P. Palazzo, A. Rosenberg, P. Musiani, S. Volinia, I. Nenci, G. A. Calin, P. Querzoli, M. Negrini, and C. M. Croce (2005). "MicroRNA gene expression deregulation in human breast cancer". *Cancer Res* 65, pp. 7065–70.

Jawaheer, D., M. F. Seldin, C. I. Amos, W. V. Chen, R. Shigeta, J. Monteiro, M. Kern, L. A. Criswell, S. Albani, J. L. Nelson, D. O. Clegg, R. Pope, Jr. Schroeder H. W., Jr. Bridges S. L., D. S. Pisetsky, R. Ward, D. L. Kastner, R. L. Wilder, T. Pincus, L. F. Callahan, D. Flemming, M. H. Wener, and P. K. Gregersen (2001). "A genomewide screen in multiplex rheumatoid arthritis families suggests genetic overlap with other autoimmune diseases". *Am J Hum Genet* 68, pp. 927–36.

Jersild, C., T. Fog, G. S. Hansen, M. Thomsen, A. Svejgaard, and B. Dupont (1973). "Histocompatibility determinants in multiple sclerosis, with special reference to clinical course". *Lancet* 2, pp. 1221–5.

Jetten, A. M. (2004). "Recent advances in the mechanisms of action and physiological functions of the retinoid-related orphan receptors (RORs)". *Curr Drug Targets Inflamm Allergy* 3, pp. 395–412.

Jin, W., R. M. Riley, R. D. Wolfinger, K. P. White, G. Passador-Gurgel, and G. Gibson (2001). "The contributions of sex, genotype and age to transcriptional variance in Drosophila melanogaster". *Nat Genet* 29, pp. 389–95.

Jin, Y., S. A. Birlea, P. R. Fain, T. M. Ferrara, S. Ben, S. L. Riccardi, J. B. Cole, K. Gowan, P. J. Holland, D. C. Bennett, R. M. Luiten, A. Wolkerstorfer, J. P. van der Veen, A. Hartmann, S. Eichner, G. Schuler, N. van Geel, J. Lambert, E. H. Kemp, D. J. Gawkrodger, A. P. Weetman, A. Taieb, T. Jouary, K. Ezzedine, M. R. Wallace, W. T. McCormack, M. Picardo, G. Leone, A. Overbeck, N. B. Silverberg, and R. A. Spritz (2012). "Genome-wide association analyses identify 13 new susceptibility loci for generalized vitiligo". *Nat Genet* 44, pp. 676–80.

Johnson, A. D., R. E. Handsaker, S. L. Pulit, M. M. Nizzari, C. J. O'Donnell, and P. I. de Bakker (2008). "SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap". *Bioinformatics* 24, pp. 2938–9.

Johnson, J. M., J. Castle, P. Garrett-Engele, Z. Kan, P. M. Loerch, C. D. Armour, R. Santos, E. E. Schadt, R. Stoughton, and D. D. Shoemaker (2003). "Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays". *Science* 302, pp. 2141–4.

Jones, E. Y., L. Fugger, J. L. Strominger, and C. Siebold (2006). "MHC class II proteins and disease: a structural perspective". *Nat Rev Immunol* 6, pp. 271–82.

Jong, M. C., M. H. Hofker, and L. M. Havekes (1999). "Role of ApoCs in lipoprotein metabolism: functional differences between ApoC1, ApoC2, and ApoC3". *Arterioscler Thromb Vasc Biol* 19, pp. 472–84.

Jong, S. de, K. R. van Eijk, D. W. Zeegers, E. Strengman, E. Janson, J. H. Veldink, L. H. van den Berg, W. Cahn, R. S. Kahn, M. P. Boks, and R. A. Ophoff (2012). "Expression QTL analysis of top loci from GWAS meta-analysis highlights additional schizophrenia candidate genes". *Eur J Hum Genet* 20, pp. 1004–8.

Joshi-Tope, G., M. Gillespie, I. Vastrik, P. D'Eustachio, E. Schmidt, B. de Bono, B. Jassal, G. R. Gopinath, G. R. Wu, L. Matthews, S. Lewis, E. Birney, and L. Stein (2005). "Reactome: a knowledgebase of biological pathways". *Nucleic Acids Res* 33, pp. D428–32.

Jostins, L., S. Ripke, R. K. Weersma, R. H. Duerr, D. P. McGovern, K. Y. Hui, J. C. Lee, L. P. Schumm, Y. Sharma, C. A. Anderson, J. Essers, M. Mitrovic, K. Ning, I. Cleynen, E. Theatre, S. L. Spain, S. Raychaudhuri, P. Goyette, Z. Wei, C. Abraham, J. P. Achkar, T. Ahmad, L. Amininejad, A. N. Ananthakrishnan, V. Andersen, J. M. Andrews, L. Baidoo, T. Balschun, P. A. Bampton, A. Bitton, G. Boucher, S. Brand, C. Buning, A. Cohain, S. Cichon, M. D'Amato, D. De Jong, K. L. Devaney, M. Dubinsky, C. Edwards, D. Ellinghaus, L. R. Ferguson, D. Franchimont, K. Fransen, R. Gearry, M. Georges, C. Gieger, J. Glas, T. Haritunians, A. Hart, C. Hawkey, M. Hedl, X. Hu, T. H. Karlsen, L. Kupcinskas, S. Kugathasan, A. Latiano, D. Laukens, I. C. Lawrance, C. W. Lees, E. Louis, G. Mahy, J. Mansfield, A. R. Morgan, C. Mowat, W. Newman, O. Palmieri, C. Y. Ponsioen, U. Potocnik, N. J. Prescott, M. Regueiro, J. I. Rotter, R. K. Russell, J. D. Sanderson, M. Sans, J. Satsangi, S. Schreiber, L. A. Simms, J. Sventoraityte, S. R. Targan, K. D. Taylor, M. Tremelling, H. W. Verspaget, M. De Vos, C. Wijmenga, D. C. Wilson, J. Winkelmann, R. J. Xavier, S. Zeissig, B. Zhang, C. K. Zhang, H. Zhao, M. S. Silverberg, V. Annese, H. Hakonarson, S. R. Brant, G. Radford-Smith, C. G. Mathew, J. D. Rioux, *et al.* (2012). "Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease". *Nature* 491, pp. 119–24.

Julia, A., E. Domenech, E. Ricart, R. Tortosa, V. Garcia-Sanchez, J. P. Gisbert, P. Nos Mateu, A. Gutierrez, F. Gomollon, J. L. Mendoza, E. Garcia-Planella, M. Barreiro-de Acosta, F. Munoz, M. Vera, C. Saro, M. Esteve, M. Andreu, A. Alonso, M. Lopez-Lasanta, L. Codo, J. L. Gelpi, A. C. Garcia-Montero, J. Bertranpetit, D. Absher, J. Panes, and S. Marsal (2013). "A genome-wide association study on a southern European population identifies a new Crohn's disease susceptibility locus at RBX1-EP300". *Gut*.

Kamada, N., T. Hisamatsu, S. Okamoto, H. Chinen, T. Kobayashi, T. Sato, A. Sakuraba, M. T. Kitazume, A. Sugita, K. Koganei, K. S. Akagawa, and T. Hibi (2008). "Unique CD14 intestinal macrophages contribute to the pathogenesis of Crohn disease via IL-23/IFN-gamma axis". *J Clin Invest* 118, pp. 2269–80.

Kanehisa, M. and S. Goto (2000). "KEGG: kyoto encyclopedia of genes and genomes". *Nucleic Acids Res* 28, pp. 27–30.

Kasowski, M., F. Grubert, C. Heffelfinger, M. Hariharan, A. Asabere, S. M. Waszak, L. Habegger, J. Rozowsky, M. Shi, A. E. Urban, M. Y. Hong, K. J. Karczewski, W. Huber, S. M. Weissman, M. B. Gerstein, J. O. Korbel, and M. Snyder (2010). "Variation in transcription factor binding among humans". *Science* 328, pp. 232–5.

Kataoka, T., N. Holler, O. Micheau, F. Martinon, A. Tinel, K. Hofmann, and J. Tschopp (2001). "Bcl-rambo, a novel Bcl-2 homologue that induces apoptosis via its unique C-terminal extension". *J Biol Chem* 276, pp. 19548–54.

Kato, M., S. Khan, E. d'Aniello, K. J. McDonald, and D. N. Hart (2007). "The novel endocytic and phagocytic C-Type lectin receptor DCL-1/CD302 on macrophages is colocalized with F-actin, suggesting a role in cell adhesion and migration". *J Immunol* 179, pp. 6052–63.

Keene, M. A., V. Corces, K. Lowenhaupt, and S. C. Elgin (1981). "DNase I hypersensitive sites in Drosophila chromatin occur at the 5' ends of regions of transcription". *Proc Natl Acad Sci U S A* 78, pp. 143–6.

Keidel, S., L. Chen, J. Pointon, and P. Wordsworth (2013). "ERAP1 and ankylosing spondylitis". *Curr Opin Immunol* 25, pp. 97–102.

Kenny, E. E., I. Pe'er, A. Karban, L. Ozelius, A. A. Mitchell, S. M. Ng, M. Erazo, H. Ostrer, C. Abraham, M. T. Abreu, G. Atzmon, N. Barzilai, S. R. Brant, S. Bressman, E. R. Burns, Y. Chowers, L. N. Clark, A. Darvasi, D. Doheny, R. H. Duerr, R. Eliakim, N. Giladi, P. K. Gregersen, H. Hakonarson, M. R. Jones, K. Marder, D. P. McGovern, J. Mulle, A. Orr-Urtreger, D. D. Proctor, A. Pulver, J. I. Rotter, M. S. Silverberg, T. Ullman, S. T. Warren, M. Waterman, W. Zhang, A. Bergman, L. Mayer, S. Katz, R. J. Desnick, J. H. Cho, and I. Peter (2012). "A genome-wide scan of Ashkenazi Jewish Crohn's disease suggests novel susceptibility loci". *PLoS Genet* 8, e1002559.

Kent, S. C., Y. Chen, L. Bregoli, S. M. Clemmings, N. S. Kenyon, C. Ricordi, B. J. Hering, and D. A. Hafler (2005). "Expanded T cells from pancreatic lymph nodes of type 1 diabetic subjects recognize an insulin epitope". *Nature* 435, pp. 224–8.

Kerkel, K., A. Spadola, E. Yuan, J. Kosek, L. Jiang, E. Hod, K. Li, V. V. Murty, N. Schupf, E. Vilain, M. Morris, F. Haghighi, and B. Tycko (2008). "Genomic surveys by methylation-sensitive SNP analysis identify sequence-dependent allele-specific DNA methylation". *Nat Genet* 40, pp. 904–8.

Kerkhof, P. van, J. Lee, L. McCormick, E. Tetrault, W. Lu, M. Schoenfish, V. Oorschot, G. J. Strous, J. Klumperman, and G. Bu (2005). "Sorting nexin 17 facilitates LRP recycling in the early endosome". *EMBO J* 24, pp. 2851–61.

Kernstock, S., E. Davydova, M. Jakobsson, A. Moen, S. Pettersen, G. M. Maelandsmo, W. Egge-Jacobsen, and P. O. Falnes (2012). "Lysine methylation of VCP by a member of a novel human protein methyltransferase family". *Nat Commun* 3, p. 1038.

Kertesz, M., N. Iovino, U. Unnerstall, U. Gaul, and E. Segal (2007). "The role of site accessibility in microRNA target recognition". *Nat Genet* 39, pp. 1278–84.

Kester, M. H., E. Kaptein, T. J. Roest, C. H. van Dijk, D. Tibboel, W. Meinl, H. Glatt, M. W. Coughtrie, and T. J. Visser (1999). "Characterization of human iodothyronine sulfotransferases". *J Clin Endocrinol Metab* 84, pp. 1357–64.

Khor, B., A. Gardet, and R. J. Xavier (2011). "Genetics and pathogenesis of inflammatory bowel disease". *Nature* 474, pp. 307–17.

Khuder, S. A. (2001). "Effect of cigarette smoking on major histological types of lung cancer: a meta-analysis". *Lung Cancer* 31, pp. 139–48.

Kim, M. S., X. Chang, K. Yamashita, J. K. Nagpal, J. H. Baek, G. Wu, B. Trink, E. A. Ratovitski, M. Mori, and D. Sidransky (2008). "Aberrant promoter methylation and tumor suppressive activity of the DFNA5 gene in colorectal carcinoma". *Oncogene* 27, pp. 3624–34.

Klein, J. (1986). "Seeds of time: fifty years ago Peter A. Gorer discovered the H-2 complex". *Immunogenetics* 24, pp. 331–8.

Knight, J. C. (2009). *Human Genetic Diversity: Functional Consequences for Health and Disease*. Oxford: Oxford University Press. ISBN: 9780191711015.

Knight, J. C., B. J. Keating, and D. P. Kwiatkowski (2004). "Allele-specific repression of lymphotoxin-alpha by activated B cell factor-1". *Nat Genet* 36, pp. 394–9.

Knowles, R. W., N. Flomenberg, K. Horibe, R. Winchester, S. F. Radka, and B. Dupont (1986). "Complexity of the supertypic HLA-DRw53 specificity: two distinct epitopes differentially expressed on one or all of the DR beta-chains depending on the HLA-DR allotype". *J Immunol* 137, pp. 2618–26.

Kobayashi, K., M. Atoh, A. Yagita, Y. Konoeda, H. Inoko, A. Ando, and S. Sekiguchi (1990). "Crohn's disease in the Japanese is associated with the HLA-DRw53". *Exp Clin Immunogenet* 7, pp. 101–8.

Kobayashi, K. S., M. Chamaillard, Y. Ogura, O. Henegariu, N. Inohara, G. Nunez, and R. A. Flavell (2005). "Nod2-dependent regulation of innate and adaptive immunity in the intestinal tract". *Science* 307, pp. 731–4.

Kochi, Y., Y. Okada, A. Suzuki, K. Ikari, C. Terao, A. Takahashi, K. Yamazaki, N. Hosono, K. Myouzen, T. Tsunoda, N. Kamatani, T. Furuichi, S. Ikegawa, K. Ohmura, T. Mimori, F. Matsuda, T. Iwamoto, S. Momohara, H. Yamanaka, R. Yamada, M. Kubo, Y. Nakamura, and K. Yamamoto (2010). "A regulatory variant in CCR6 is associated with rheumatoid arthritis susceptibility". *Nat Genet* 42, pp. 515–9.

Kolstad, A., K. Hannestad, B. Vandvik, and F. Vartdal (1989). "Multiple sclerosis patients have a high frequency of an HLA-DQ beta epitope defined by a human-human hybridoma antibody". *Tissue Antigens* 33, pp. 546–9.

Krausgruber, T., K. Blazek, T. Smallie, S. Alzabin, H. Lockstone, N. Sahgal, T. Hussell, M. Feldmann, and I. A. Udalova (2011). "IRF5 promotes inflammatory macrophage polarization and TH1-TH17 responses". *Nat Immunol* 12, pp. 231–8.

Krejsgaard, T., C. S. Vetter-Kauczok, A. Woetmann, H. Kneitz, K. W. Eriksen, P. Lovato, Q. Zhang, M. A. Wasik, C. Geisler, E. Ralfkiaer, J. C. Becker, and N. Odum (2009). "Ectopic expression of B-lymphoid kinase in cutaneous T-cell lymphoma". *Blood* 113, pp. 5896–904.

Kruglyak, L. (1999). "Prospects for whole-genome linkage disequilibrium mapping of common disease genes". *Nat Genet* 22, pp. 139–44.

Kruskal, J.B. (1964a). "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis." *Psychometrika* 29, pp. 1–27.

— (1964b). "Nonmetric multidimensional scaling: a numerical method." *Psychometrika* 29, pp. 115–129.

Kugathasan, S., R. N. Baldassano, J. P. Bradfield, P. M. Sleiman, M. Imielinski, S. L. Guthery, S. Cucchiara, C. E. Kim, E. C. Frackelton, K. Annaiah, J. T. Glessner, E. Santa, T. Willson, A. W. Eckert, E. Bonkowski, J. L. Shaner, R. M. Smith, F. G. Otieno, N. Peterson, D. J. Abrams, R. M. Chiavacci, R. Grundmeier, P. Mamula, G. Tomer, D. A. Piccoli, D. S. Monos, V. Annese, L. A. Denson, S. F. Grant, and H. Hakonarson (2008). "Loci on 20q13 and 21q22 are associated with pediatric-onset inflammatory bowel disease". *Nat Genet* 40, pp. 1211–5.

Kuhn, K., S. C. Baker, E. Chudin, M. H. Lieu, S. Oeser, H. Bennett, P. Rigault, D. Barker, T. K. McDaniel, and M. S. Chee (2004). "A novel, high-performance random array platform for quantitative gene expression profiling". *Genome Res* 14, pp. 2347–56.

Kumar, H., T. Kawai, and S. Akira (2009). "Toll-like receptors and innate immunity". *Biochem Biophys Res Commun* 388, pp. 621–5.

Kumasaka, N., M. Aoki, Y. Okada, A. Takahashi, K. Ozaki, T. Mushiroda, T. Hirota, M. Tamari, T. Tanaka, Y. Nakamura, N. Kamatani, and M. Kubo (2012). "Haplotypes with copy number and single nucleotide polymorphisms in CYP2A6 locus are associated with smoking quantity in a Japanese population". *PLoS One* 7, e44507.

Kwan, T., D. Benovoy, C. Dias, S. Gurd, D. Serre, H. Zuzan, T. A. Clark, A. Schweitzer, M. K. Staples, H. Wang, J. E. Blume, T. J. Hudson, R. Sladek, and J. Majewski (2007). "Heritability of alternative splicing in the human genome". *Genome Res* 17, pp. 1210–8.

Kwan, T., D. Benovoy, C. Dias, S. Gurd, C. Provencher, P. Beaulieu, T. J. Hudson, R. Sladek, and J. Majewski (2008). "Genome-wide analysis of transcript isoform variation in humans". *Nat Genet* 40, pp. 225–31.

Lahoud, M. H., A. I. Proietto, F. Ahmet, S. Kitsoulis, L. Eidsmo, L. Wu, P. Sathe, S. Pietersz, H. W. Chang, I. D. Walker, E. Maraskovsky, H. Braley, A. M. Lew, M. D. Wright, W. R. Heath, K. Shortman, and I. Caminschi (2009). "The C-type lectin Clec12A present on mouse and human dendritic cells can serve as a target for antigen delivery and enhancement of antibody responses". *J Immunol* 182, pp. 7587–94.

Lalonde, E., K. C. Ha, Z. Wang, A. Bemmo, C. L. Kleinman, T. Kwan, T. Pastinen, and J. Majewski (2011). "RNA sequencing reveals the role of splicing polymorphisms in regulating human gene expression". *Genome Res* 21, pp. 545–54.

Lander, E. S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland,

L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J. P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, N. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J. C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R. H. Waterston, R. K. Wilson, L. W. Hillier, J. D. McPherson, M. A. Marra, E. R. Mardis, L. A. Fulton, A. T. Chinwalla, K. H. Pepin, W. R. Gish, S. L. Chissoe, M. C. Wendl, K. D. Delehaunty, T. L. Miner, A. Delehaunty, J. B. Kramer, L. L. Cook, R. S. Fulton, D. L. Johnson, P. J. Minx, S. W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J. F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, *et al.* (2001). "Initial sequencing and analysis of the human genome". *Nature* 409, pp. 860–921.

Lee, J., C. Retamal, L. Cuitino, A. Caruano-Yzermans, J. E. Shin, P. van Kerkhof, M. P. Marzolo, and G. Bu (2008). "Adaptor protein sorting nexin 17 regulates amyloid precursor protein trafficking and processing in the early endosomes". *J Biol Chem* 283, pp. 11501–8.

Lee, K. H., K. W. Wucherpfennig, and D. C. Wiley (2001). "Structure of a human insulin peptide-HLA-DQ8 complex and susceptibility to type 1 diabetes". *Nat Immunol* 2, pp. 501–7.

Lee, W., P. Mitchell, and R. Tjian (1987). "Purified transcription factor AP-1 interacts with TPA-inducible enhancer elements". *Cell* 49, pp. 741–52.

Lee, Y. C., H. C. Kuo, J. S. Chang, L. Y. Chang, L. M. Huang, M. R. Chen, C. D. Liang, H. Chi, F. Y. Huang, M. L. Lee, Y. C. Huang, B. Hwang, N. C. Chiu, K. P. Hwang, P. C. Lee, L. C. Chang, Y. M. Liu, Y. J. Chen, C. H. Chen, Y. T. Chen, F. J. Tsai, and J. Y. Wu (2012). "Two new susceptibility loci for Kawasaki disease identified through genome-wide association analysis". *Nat Genet* 44, pp. 522–5.

Leek, J. T. and J. D. Storey (2007). "Capturing heterogeneity in gene expression studies by surrogate variable analysis". *PLoS Genet* 3, pp. 1724–35.

LeibundGut-Landmann, S., O. Gross, M. J. Robinson, F. Osorio, E. C. Slack, S. V. Tsoni, E. Schweighoffer, V. Tybulewicz, G. D. Brown, J. Ruland, and C. Reis e Sousa (2007). "Syk- and CARD9-dependent coupling of innate immunity to the induction of T helper cells that produce interleukin 17". *Nat Immunol* 8, pp. 630–8.

Leslie, S., P. Donnelly, and G. McVean (2008). "A statistical method for predicting classical HLA alleles from SNP data". *Am J Hum Genet* 82, pp. 48–56.

Lewontin, R.C. and K. Kojima (1960). "The evolutionary dynamics of complex polymorphisms." *Evolution* 14, pp. 458–72.

Ley, K., C. Laudanna, M. I. Cybulsky, and S. Nourshargh (2007). "Getting to the site of inflammation: the leukocyte adhesion cascade updated". *Nat Rev Immunol* 7, pp. 678–89.

Li, B., M. Carey, and J. L. Workman (2007). "The role of chromatin during transcription". *Cell* 128, pp. 707–19.

Li, N. and M. Stephens (2003). "Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data". *Genetics* 165, pp. 2213–33.

Li, Q., M. K. Spriggs, S. Kovats, S. M. Turk, M. R. Comeau, B. Nepom, and L. M. Hutt-Fletcher (1997). "Epstein-Barr virus uses HLA class II as a cofactor for infection of B lymphocytes". *J Virol* 71, pp. 4657–62.

Li, Y., C. Willer, S. Sanna, and G. Abecasis (2009). "Genotype imputation". *Annu Rev Genomics Hum Genet* 10, pp. 387–406.

Liang, L., N. Morar, A. L. Dixon, G. M. Lathrop, G. R. Abecasis, M. F. Moffatt, and W. O. Cookson (2013). "A cross-platform analysis of 14,177 expression quantitative trait loci derived from lymphoblastoid cell lines". *Genome Res* 23, pp. 716–26.

Livak, K. J. and T. D. Schmittgen (2001). "Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method". *Methods* 25, pp. 402–8.

Longini I. M., Jr., M. W. Higgins, P. C. Hinton, P. P. Moll, and J. B. Keller (1984). "Environmental and genetic sources of familial aggregation of blood pressure in Tecumseh, Michigan". *Am J Epidemiol* 120, pp. 131–44.

Lorenz, W. W., R. O. McCann, M. Longiaru, and M. J. Cormier (1991). "Isolation and expression of a cDNA encoding Renilla reniformis luciferase". *Proc Natl Acad Sci U S A* 88, pp. 4438–42.

Lozzio, C. B. and B. B. Lozzio (1975). "Human chronic myelogenous leukemia cell-line with positive Philadelphia chromosome". *Blood* 45, pp. 321–34.

Lu, M., M. Zhang, A. Takashima, J. Weiss, M. A. Apicella, X. H. Li, D. Yuan, and R. S. Munford (2005). "Lipopolysaccharide deacylation by an endogenous lipase controls innate antibody responses to Gram-negative bacteria". *Nat Immunol* 6, pp. 989–94.

Lucassen, A. M., C. Julier, J. P. Beressi, C. Boitard, P. Froguel, M. Lathrop, and J. I. Bell (1993). "Susceptibility to insulin dependent diabetes mellitus maps to a 4.1 kb segment of DNA spanning the insulin gene and associated VNTR". *Nat Genet* 4, pp. 305–10.

Lumeng, C. N., J. L. Bodzin, and A. R. Saltiel (2007). "Obesity induces a phenotypic switch in adipose tissue macrophage polarization". *J Clin Invest* 117, pp. 175–84.

Lundin, K. E., H. Scott, T. Hansen, G. Paulsen, T. S. Halstensen, O. Fausa, E. Thorsby, and L. M. Sollid (1993). "Gliadin-specific, HLA-DQ(alpha 1*0501,beta 1*0201) restricted T cells isolated from the small intestinal mucosa of celiac disease patients". *J Exp Med* 178, pp. 187–96.

Lusis, A. J., C. Heinzmann, R. S. Sparkes, J. Scott, T. J. Knott, R. Geller, M. C. Sparkes, and T. Mohandas (1986). "Regional mapping of human chromosome 19: organization of genes for plasma lipid transport (APOC1, -C2, and -E and LDLR) and the genes C3, PEPD, and GPI". *Proc Natl Acad Sci U S A* 83, pp. 3929–33.

Luu, N. T., G. E. Rainger, C. D. Buckley, and G. B. Nash (2003). "CD31 regulates direction and rate of neutrophil migration over and under endothelial cells". *J Vasc Res* 40, pp. 467–79.

Lyons, P. A., M. Koukoulaki, A. Hatton, K. Doggett, H. B. Woffendin, A. N. Chaudhry, and K. G. Smith (2007). "Microarray analysis of human leucocyte subsets: the advantages of positive selection and rapid purification". *BMC Genomics* 8, p. 64.

MacGregor, A. J., H. Snieder, A. S. Rigby, M. Koskenvuo, J. Kaprio, K. Aho, and A. J. Silman (2000). "Characterizing the quantitative genetic contribution to rheumatoid arthritis using data from twins". *Arthritis Rheum* 43, pp. 30–7.

Majewski, J. and T. Pastinen (2011). "The study of eQTL variations by RNA-seq: from SNPs to phenotypes". *Trends Genet* 27, pp. 72–9.

Mann, R., R. C. Mulligan, and D. Baltimore (1983). "Construction of a retrovirus packaging mutant and its use to produce helper-free defective retrovirus". *Cell* 33, pp. 153–9.

Manning, A. K., M. F. Hivert, R. A. Scott, J. L. Grimsby, N. Bouatia-Naji, H. Chen, D. Rybin, C. T. Liu, L. F. Bielak, I. Prokopenko, N. Amin, D. Barnes, G. Cadby, J. J. Hottenga, E. Ingelsson, A. U. Jackson, T. Johnson, S. Kanoni, C. Ladenvall, V. Lagou, J. Lahti, C. Lecoeur, Y. Liu, M. T. Martinez-Larrad, M. E. Montasser, P. Navarro, J. R. Perry, L. J. Rasmussen-Torvik, P. Salo, N. Sattar, D. Shungin, R. J. Strawbridge, T. Tanaka, C. M. van Duijn, P. An, M. de Andrade, J. S. Andrews, T. Aspelund, M. Atalay, Y. Aulchenko, B. Balkau, S. Bandinelli, J. S. Beckmann, J. P. Beilby, C. Bellis, R. N. Bergman, J. Blangero, M. Boban, M. Boehnke, E. Boerwinkle, L. L. Bonnycastle, D. I. Boomsma, I. B. Borecki, Y. Bottcher, C. Bouchard, E. Brunner, D. Budimir, H. Campbell, O. Carlson, P. S. Chines, R. Clarke, F. S. Collins, A. Corbaton-Anchuelo, D. Couper, U. de Faire, G. V. Dedoussis, P. Deloukas, M. Dimitriou, J. M. Egan, G. Eiriksdottir, M. R. Erdos, J. G. Eriksson, E. Eury, L. Ferrucci, I. Ford, N. G. Forouhi, C. S. Fox, M. G. Franzosi, P. W. Franks, T. M. Frayling, P. Froguel, P. Galan, E. de Geus, B. Gigante, N. L. Glazer, A. Goel, L. Groop, V. Gudnason, G. Hallmans, A. Hamsten, O. Hansson, T. B. Harris, C. Hayward, S. Heath, S. Hercberg, A. A. Hicks, A. Hingorani, A. Hofman, J. Hui, *et al.* (2012). "A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance". *Nat Genet* 44, pp. 659–69.

Manolio, T. A., F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, M. I. McCarthy, E. M. Ramos, L. R. Cardon, A. Chakravarti, J. H. Cho, A. E. Guttmacher, A. Kong, L. Kruglyak, E. Mardis, C. N. Rotimi, M. Slatkin, D. Valle, A. S. Whittemore, M. Boehnke, A. G. Clark, E. E. Eichler, G. Gibson, J. L. Haines, T. F. Mackay, S. A. McCarroll, and P. M. Visscher (2009). "Finding the missing heritability of complex diseases". *Nature* 461, pp. 747–53.

Mantovani, A., S. Sozzani, M. Locati, P. Allavena, and A. Sica (2002). "Macrophage polarization: tumor-associated macrophages as a paradigm for polarized M2 mononuclear phagocytes". *Trends Immunol* 23, pp. 549–55.

Marchini, J. and B. Howie (2010). "Genotype imputation for genome-wide association studies". *Nat Rev Genet* 11, pp. 499–511.

Mariathasan, S., D. S. Weiss, K. Newton, J. McBride, K. O'Rourke, M. Roose-Girma, W. P. Lee, Y. Weinrauch, D. M. Monack, and V. M. Dixit (2006). "Cryopyrin activates the inflammasome in response to toxins and ATP". *Nature* 440, pp. 228–32.

Marinissen, M. J. and J. S. Gutkind (2001). "G-protein-coupled receptors and signaling networks: emerging paradigms". *Trends Pharmacol Sci* 22, pp. 368–76.

Marsh, S. G., E. D. Albert, W. F. Bodmer, R. E. Bontrop, B. Dupont, H. A. Erlich, M. Fernandez-Vina, D. E. Geraghty, R. Holdsworth, C. K. Hurley, M. Lau, K. W. Lee, B. Mach, M. Maiers, W. R. Mayr, C. R. Muller, P. Parham, E. W. Petersdorf, T. Sasazuki, J. L. Strominger, A. Svejgaard, P. I. Terasaki, J. M. Tiercy, and J. Trowsdale (2010). "Nomenclature for factors of the HLA system, 2010". *Tissue Antigens* 75, pp. 291–455.

Martinez, F. O., S. Gordon, M. Locati, and A. Mantovani (2006). "Transcriptional profiling of the human monocyte-to-macrophage differentiation and polarization: new molecules and patterns of gene expression". *J Immunol* 177, pp. 7303–11.

Martinez, F. O., L. Helming, and S. Gordon (2009). "Alternative activation of macrophages: an immunologic functional perspective". *Annu Rev Immunol* 27, pp. 451–83.

Masuda, Y., M. Futamura, H. Kamino, Y. Nakamura, N. Kitamura, S. Ohnishi, Y. Miyamoto, H. Ichikawa, T. Ohta, M. Ohki, T. Kiyono, H. Egami, H. Baba, and H. Arakawa (2006). "The potential role of DFNA5, a hearing impairment gene, in p53-mediated cellular response to DNA damage". *J Hum Genet* 51, pp. 652–64.

Matthaei, J. H., O. W. Jones, R. G. Martin, and M. W. Nirenberg (1962). "Characteristics and composition of RNA coding units". *Proc Natl Acad Sci U S A* 48, pp. 666–77.

Mauri, C. (2010). "Regulation of immunity and autoimmunity by B cells". *Curr Opin Immunol* 22, pp. 761–7.

Mauri, C. and A. Bosma (2012). "Immune regulatory function of B cells". *Annu Rev Immunol* 30, pp. 221–41.

Maynard, N. D., J. Chen, R. K. Stuart, J. B. Fan, and B. Ren (2008). "Genome-wide mapping of allele-specific protein-DNA interactions in human cells". *Nat Methods* 5, pp. 307–9.

McCarroll, S. A., A. Huett, P. Kuballa, S. D. Chilewski, A. Landry, P. Goyette, M. C. Zody, J. L. Hall, S. R. Brant, J. H. Cho, R. H. Duerr, M. S. Silverberg, K. D. Taylor, J. D. Rioux, D. Altshuler, M. J. Daly, and R. J. Xavier (2008). "Deletion polymorphism upstream of IRGM associated with altered IRGM expression and Crohn's disease". *Nat Genet* 40, pp. 1107–12.

McCarthy, M. I., G. R. Abecasis, L. R. Cardon, D. B. Goldstein, J. Little, J. P. Ioannidis, and J. N. Hirschhorn (2008). "Genome-wide association studies for complex traits: consensus, uncertainty and challenges". *Nat Rev Genet* 9, pp. 356–69.

McDaniell, R., B. K. Lee, L. Song, Z. Liu, A. P. Boyle, M. R. Erdos, L. J. Scott, M. A. Morken, K. S. Kucera, A. Battenhouse, D. Keefe, F. S. Collins, H. F. Willard, J. D. Lieb, T. S. Furey, G. E. Crawford, V. R. Iyer, and E. Birney (2010). "Heritable individual-specific and allele-specific chromatin signatures in humans". *Science* 328, pp. 235–9.

McEver, R. P., K. L. Moore, and R. D. Cummings (1995). "Leukocyte trafficking mediated by selectin-carbohydrate interactions". *J Biol Chem* 270, pp. 11025–8.

McGinnis, W., M. S. Levine, E. Hafen, A. Kuroiwa, and W. J. Gehring (1984). "A conserved DNA sequence in homoeotic genes of the Drosophila Antennapedia and bithorax complexes". *Nature* 308, pp. 428–33.

McGovern, D. P., A. Gardet, L. Torkvist, P. Goyette, J. Essers, K. D. Taylor, B. M. Neale, R. T. Ong, C. Lagace, C. Li, T. Green, C. R. Stevens, C. Beauchamp, P. R. Fleshner,

M. Carlson, M. D'Amato, J. Halfvarson, M. L. Hibberd, M. Lordal, L. Padyukov, A. Andriulli, E. Colombo, A. Latiano, O. Palmieri, E. J. Bernard, C. Deslandres, D. W. Hommes, D. J. de Jong, P. C. Stokkers, R. K. Weersma, Y. Sharma, M. S. Silverberg, J. H. Cho, J. Wu, K. Roeder, S. R. Brant, L. P. Schumm, R. H. Duerr, M. C. Dubinsky, N. L. Glazer, T. Haritunians, A. Ippoliti, G. Y. Melmed, D. S. Siscovick, E. A. Vasiliauskas, S. R. Targan, V. Annese, C. Wijmenga, S. Pettersson, J. I. Rotter, R. J. Xavier, M. J. Daly, J. D. Rioux, and M. Seielstad (2010). "Genome-wide association identifies multiple ulcerative colitis susceptibility loci". *Nat Genet* 42, pp. 332–7.

McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, and M. A. DePristo (2010). "The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data". *Genome Res* 20, pp. 1297–303.

Medzhitov, R. (2007). "Recognition of microorganisms and activation of the immune response". *Nature* 449, pp. 819–26.

— (2008). "Origin and physiological roles of inflammation". *Nature* 454, pp. 428–35.

Medzhitov, R. and T. Horng (2009). "Transcriptional control of the inflammatory response". *Nat Rev Immunol* 9, pp. 692–703.

Mendel, G.J. (1866). "Versuche uber pflanzenhypriden (Experiments in plant hybridisation)." *Verhandlungen des Naturforschenden Veriens (The Proceedings of the Natural History Society)* 4, pp. 3–47.

Metzker, M. L. (2010). "Sequencing technologies - the next generation". *Nat Rev Genet* 11, pp. 31–46.

Miettinen, H. M. (2011). "Regulation of human formyl peptide receptor 1 synthesis: role of single nucleotide polymorphisms, transcription factors, and inflammatory mediators". *PLoS One* 6, e28712.

Migueles, S. A., M. S. Sabbaghian, W. L. Shupert, M. P. Bettinotti, F. M. Marincola, L. Martino, C. W. Hallahan, S. M. Selig, D. Schwartz, J. Sullivan, and M. Connors (2000). "HLA B*5701 is highly associated with restriction of virus replication in a subgroup of HIV-infected long term nonprogressors". *Proc Natl Acad Sci U S A* 97, pp. 2709–14.

Mikkelsen, T. S., M. Ku, D. B. Jaffe, B. Issac, E. Lieberman, G. Giannoukos, P. Alvarez, W. Brockman, T. K. Kim, R. P. Koche, W. Lee, E. Mendenhall, A. O'Donovan, A. Presser, C. Russ, X. Xie, A. Meissner, M. Wernig, R. Jaenisch, C. Nusbaum, E. S. Lander, and B. E. Bernstein (2007). "Genome-wide maps of chromatin state in pluripotent and lineage-committed cells". *Nature* 448, pp. 553–60.

Miller, L. H., S. J. Mason, J. A. Dvorak, M. H. McGinniss, and I. K. Rothman (1975). "Erythrocyte receptors for (Plasmodium knowlesi) malaria: Duffy blood group determinants". *Science* 189, pp. 561–3.

Miller, L. H., S. J. Mason, D. F. Clyde, and M. H. McGinniss (1976). "The resistance factor to Plasmodium vivax in blacks. The Duffy-blood-group genotype, FyFy". *N Engl J Med* 295, pp. 302–4.

Miretti, M. M., E. C. Walsh, X. Ke, M. Delgado, M. Griffiths, S. Hunt, J. Morrison, P. Whittaker, E. S. Lander, L. R. Cardon, D. R. Bentley, J. D. Rioux, S. Beck, and P. Deloukas (2005). "A high-resolution linkage-disequilibrium map of the human major histocompatibility complex and first generation of tag single-nucleotide polymorphisms". *Am J Hum Genet* 76, pp. 634–46.

Modrek, B., A. Resch, C. Grasso, and C. Lee (2001). "Genome-wide detection of alternative splicing in expressed sequences of human genes". *Nucleic Acids Res* 29, pp. 2850–9.

Moffatt, M. F., M. Kabesch, L. Liang, A. L. Dixon, D. Strachan, S. Heath, M. Depner, A. von Berg, A. Bufe, E. Rietschel, A. Heinzmann, B. Simma, T. Frischer, S. A. Willis-Owen, K. C. Wong, T. Illig, C. Vogelberg, S. K. Weiland, E. von Mutius, G. R. Abecasis, M. Farrall, I. G. Gut, G. M. Lathrop, and W. O. Cookson (2007). "Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma". *Nature* 448, pp. 470–3.

Monks, S. A., A. Leonardson, H. Zhu, P. Cundiff, P. Pietrusiak, S. Edwards, J. W. Phillips, A. Sachs, and E. E. Schadt (2004). "Genetic inheritance of gene expression in human cell lines". *Am J Hum Genet* 75, pp. 1094–105.

Montgomery, S. B., M. Sammeth, M. Gutierrez-Arcelus, R. P. Lach, C. Ingle, J. Nisbett, R. Guigo, and E. T. Dermitzakis (2010). "Transcriptome genetics using second generation sequencing in a Caucasian population". *Nature* 464, pp. 773–7.

Moore, C. B., M. John, I. R. James, F. T. Christiansen, C. S. Witt, and S. A. Mallal (2002). "Evidence of HIV-1 adaptation to HLA-restricted immune responses at a population level". *Science* 296, pp. 1439–43.

Morgan, T. H. (1910). "Sex Limited Inheritance in Drosophila". *Science* 32, pp. 120–2.

Morley, M., C. M. Molony, T. M. Weber, J. L. Devlin, K. G. Ewens, R. S. Spielman, and V. G. Cheung (2004). "Genetic analysis of genome-wide variation in human gene expression". *Nature* 430, pp. 743–7.

Morling, N., V. Andersen, L. Fugger, J. Georgsen, P. Halberg, P. Oxholm, N. Odum, and A. Svejgaard (1991). "Immunogenetics of rheumatoid arthritis and primary Sjogren's syndrome: DNA polymorphism of HLA class II genes". *Dis Markers* 9, pp. 289–96.

Morris, J. F., R. Hromas, and 3rd Rauscher F. J. (1994). "Characterization of the DNA-binding properties of the myeloid zinc finger protein MZF1: two independent DNA-binding domains recognize two DNA consensus sequences with a common G-rich core". *Mol Cell Biol* 14, pp. 1786–95.

Morrison, B. H., J. A. Bauer, J. Hu, R. W. Grane, A. M. Ozdemir, M. Chawla-Sarkar, B. Gong, A. Almasan, D. V. Kalvakolanu, and D. J. Lindner (2002). "Inositol hexakisphosphate kinase 2 sensitizes ovarian carcinoma cells to multiple cancer therapeutics". *Oncogene* 21, pp. 1882–9.

Morton, N. E. (1955). "Sequential tests for the detection of linkage". *Am J Hum Genet* 7, pp. 277–318.

Mosser, D. M. and J. P. Edwards (2008). "Exploring the full spectrum of macrophage activation". *Nat Rev Immunol* 8, pp. 958–69.

Mullis, K. B. and F. A. Faloona (1987). "Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction". *Methods Enzymol* 155, pp. 335–50.

Murphy, A., J. H. Chu, M. Xu, V. J. Carey, R. Lazarus, A. Liu, S. J. Szefler, R. Strunk, K. Demuth, M. Castro, N. N. Hansel, G. B. Diette, B. M. Vonakis, Jr. Adkinson N. F., B. J. Klanderman, J. Senter-Sylvia, J. Ziniti, C. Lange, T. Pastinen, and B. A. Raby (2010). "Mapping of numerous disease-associated expression polymorphisms in primary peripheral blood CD4+ lymphocytes". *Hum Mol Genet* 19, pp. 4745–57.

Murray, P. J. and T. A. Wynn (2011). "Protective and pathogenic functions of macrophage subsets". *Nat Rev Immunol* 11, pp. 723–37.

Musunuru, K., A. Strong, M. Frank-Kamenetsky, N. E. Lee, T. Ahfeldt, K. V. Sachs, X. Li, H. Li, N. Kuperwasser, V. M. Ruda, J. P. Pirruccello, B. Muchmore, L. Prokunina-Olsson, J. L. Hall, E. E. Schadt, C. R. Morales, S. Lund-Katz, M. C. Phillips, J. Wong, W. Cantley, T. Racie, K. G. Ejebe, M. Orho-Melander, O. Melander, V. Koteliansky, K. Fitzgerald, R. M. Krauss, C. A. Cowan, S. Kathiresan, and D. J. Rader (2010). "From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus". *Nature* 466, pp. 714–9.

Myers, A. J., J. R. Gibbs, J. A. Webster, K. Rohrer, A. Zhao, L. Marlowe, M. Kaleem, D. Leung, L. Bryden, P. Nath, V. L. Zismann, K. Joshipura, M. J. Huentelman, D. Hu-Lince, K. D. Coon, D. W. Craig, J. V. Pearson, P. Holmans, C. B. Heward, E. M. Reiman, D. Stephan, and J. Hardy (2007). "A survey of genetic human cortical gene expression". *Nat Genet* 39, pp. 1494–9.

Myokai, F., S. Takashiba, R. Lebo, and S. Amar (1999). "A novel lipopolysaccharide-induced transcription factor regulating tumor necrosis factor alpha gene expression: molecular cloning, sequencing, characterization, and chromosomal assignment". *Proc Natl Acad Sci U S A* 96, pp. 4518–23.

Nagata, S., R. Hanayama, and K. Kawane (2010). "Autoimmunity and the clearance of dead cells". *Cell* 140, pp. 619–30.

Nair, R. P., P. E. Stuart, I. Nistor, R. Hiremagalore, N. V. Chia, S. Jenisch, M. Weichenthal, G. R. Abecasis, H. W. Lim, E. Christophers, J. J. Voorhees, and J. T. Elder (2006). "Sequence and haplotype analysis supports HLA-C as the psoriasis susceptibility 1 gene". *Am J Hum Genet* 78, pp. 827–51.

Nakamura, Y., M. Leppert, P. O'Connell, R. Wolff, T. Holm, M. Culver, C. Martin, E. Fujimoto, M. Hoff, E. Kumlin, and et al. (1987). "Variable number of tandem repeat (VNTR) markers for human gene mapping". *Science* 235, pp. 1616–22.

Nas, A. van, L. Ingram-Drake, J. S. Sinsheimer, S. S. Wang, E. E. Schadt, T. Drake, and A. J. Lusis (2010). "Expression quantitative trait loci: replication, tissue- and sex-specificity in mice". *Genetics* 185, pp. 1059–68.

Nathan, C. (2002). "Points of control in inflammation". *Nature* 420, pp. 846–52.

Nathan, C. and A. Ding (2010). "Nonresolving inflammation". *Cell* 140, pp. 871–82.

Nejentsev, S., J. M. Howson, N. M. Walker, J. Szeszko, S. F. Field, H. E. Stevens, P. Reynolds, M. Hardy, E. King, J. Masters, J. Hulme, L. M. Maier, D. Smyth, R. Bailey, J. D. Cooper, G. Ribas, R. D. Campbell, D. G. Clayton, and J. A. Todd (2007). "Localization of type 1 diabetes susceptibility to the MHC class I genes HLA-B and HLA-A". *Nature* 450, pp. 887–92.

Neph, S., J. Vierstra, A. B. Stergachis, A. P. Reynolds, E. Haugen, B. Vernot, R. E. Thurman, S. John, R. Sandstrom, A. K. Johnson, M. T. Maurano, R. Humbert, E. Rynes, H. Wang, S. Vong, K. Lee, D. Bates, M. Diegel, V. Roach, D. Dunn, J. Neri, A. Schafer, R. S. Hansen, T. Kutyavin, E. Giste, M. Weaver, T. Canfield, P. Sabo, M. Zhang, G. Balasundaram, R. Byron, M. J. MacCoss, J. M. Akey, M. A. Bender, M. Groudine, R. Kaul, and J. A. Stamatoyannopoulos (2012). "An expansive human regulatory lexicon encoded in transcription factor footprints". *Nature* 489, pp. 83–90.

Nestle, F. O., D. H. Kaplan, and J. Barker (2009). "Psoriasis". *N Engl J Med* 361, pp. 496–509.

Newman, P. J., M. C. Berndt, J. Gorski, 2nd White G. C., S. Lyman, C. Paddock, and W. A. Muller (1990). "PECAM-1 (CD31) cloning and relation to adhesion molecules of the immunoglobulin gene superfamily". *Science* 247, pp. 1219–22.

Nica, A. C., S. B. Montgomery, A. S. Dimas, B. E. Stranger, C. Beazley, I. Barroso, and E. T. Dermitzakis (2010). "Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations". *PLoS Genet* 6, e1000895.

Nica, A. C., L. Parts, D. Glass, J. Nisbet, A. Barrett, M. Sekowska, M. Travers, S. Potter, E. Grundberg, K. Small, A. K. Hedman, V. Bataille, J. Tzenova Bell, G. Surdulescu, A. S. Dimas, C. Ingle, F. O. Nestle, P. di Meglio, J. L. Min, A. Wilk, C. J. Hammond, N. Hassanali, T. P. Yang, S. B. Montgomery, S. O'Rahilly, C. M. Lindgren, K. T. Zondervan, N. Soranzo, I. Barroso, R. Durbin, K. Ahmadi, P. Deloukas, M. I. McCarthy, E. T. Dermitzakis, and T. D. Spector (2011). "The architecture of gene regulatory variation across multiple human tissues: the MuTHER study". *PLoS Genet* 7, e1002003.

Nicolae, D. L., E. Gamazon, W. Zhang, S. Duan, M. E. Dolan, and N. J. Cox (2010). "Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS". *PLoS Genet* 6, e1000888.

Nightingale, K. P., L. P. O'Neill, and B. M. Turner (2006). "Histone modifications: signalling receptors and potential elements of a heritable epigenetic code". *Curr Opin Genet Dev* 16, pp. 125–36.

Ober, C. and S. Hoffjan (2006). "Asthma genetics 2006: the long and winding road to gene discovery". *Genes Immun* 7, pp. 95–100.

Odegaard, J. I., R. R. Ricardo-Gonzalez, M. H. Goforth, C. R. Morel, V. Subramanian, L. Mukundan, A. Red Eagle, D. Vats, F. Brombacher, A. W. Ferrante, and A. Chawla (2007). "Macrophage-specific PPARgamma controls alternative activation and improves insulin resistance". *Nature* 447, pp. 1116–20.

Ogura, Y., D. K. Bonen, N. Inohara, D. L. Nicolae, F. F. Chen, R. Ramos, H. Britton, T. Moran, R. Karaliuskas, R. H. Duerr, J. P. Achkar, S. R. Brant, T. M. Bayless, B. S. Kirschner, S. B. Hanauer, G. Nunez, and J. H. Cho (2001a). "A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease". *Nature* 411, pp. 603–6.

Ogura, Y., N. Inohara, A. Benito, F. F. Chen, S. Yamaoka, and G. Nunez (2001b). "Nod2, a Nod1/Apaf-1 family member that is restricted to monocytes and activates NF-kappaB". *J Biol Chem* 276, pp. 4812–8.

Ohta, A. and M. Sitkovsky (2001). "Role of G-protein-coupled adenosine receptors in downregulation of inflammation and protection from tissue damage". *Nature* 414, pp. 916–20.

Okada, M., S. Nada, Y. Yamanashi, T. Yamamoto, and H. Nakagawa (1991). "CSK: a protein-tyrosine kinase involved in regulation of src family kinases". *J Biol Chem* 266, pp. 24249–52.

Olefsky, J. M. and C. K. Glass (2010). "Macrophages, inflammation, and insulin resistance". *Annu Rev Physiol* 72, pp. 219–46.

Olerup, O. and J. Hillert (1991). "HLA class II-associated genetic susceptibility in multiple sclerosis: a critical evaluation". *Tissue Antigens* 38, pp. 1–15.

Onouchi, Y., K. Ozaki, J. C. Burns, C. Shimizu, M. Terai, H. Hamada, T. Honda, H. Suzuki, T. Suenaga, T. Takeuchi, N. Yoshikawa, Y. Suzuki, K. Yasukawa, R. Ebata, K. Higashi, T. Saji, Y. Kemmotsu, S. Takatsuki, K. Ouchi, F. Kishi, T. Yoshikawa, T. Nagai, K. Hamamoto, Y. Sato, A. Honda, H. Kobayashi, J. Sato, S. Shibuta, M. Miyawaki, K. Oishi, H. Yamaga, N. Aoyagi, S. Iwahashi, R. Miyashita, Y. Murata, K. Sasago, A. Takahashi, N. Kamatani, M. Kubo, T. Tsunoda, A. Hata, Y. Nakamura, and T. Tanaka (2012). "A genome-wide association study identifies three new risk loci for Kawasaki disease". *Nat Genet* 44, pp. 517–21.

O'Sullivan, B. and R. Thomas (2003). "Recent advances on the role of CD40 and dendritic cells in immunity and tolerance". *Curr Opin Hematol* 10, pp. 272–8.

Pai, A. A., C. E. Cain, O. Mizrahi-Man, S. De Leon, N. Lewellen, J. B. Veyrieras, J. F. Degner, D. J. Gaffney, J. K. Pickrell, M. Stephens, J. K. Pritchard, and Y. Gilad (2012). "The contribution of RNA decay quantitative trait loci to inter-individual variation in steady-state gene expression levels". *PLoS Genet* 8, e1003000.

Park, P. J. (2009). "ChIP-seq: advantages and challenges of a maturing technology". *Nat Rev Genet* 10, pp. 669–80.

Parts, L., A. K. Hedman, S. Keildson, A. J. Knights, C. Abreu-Goodger, M. van de Bunt, J. A. Guerra-Assuncao, N. Bartonicek, S. van Dongen, R. Magi, J. Nisbet, A. Barrett, M. Rantalainen, A. C. Nica, M. A. Quail, K. S. Small, D. Glass, A. J. Enright, J. Winn, P. Deloukas, E. T. Dermitzakis, M. I. McCarthy, T. D. Spector, R. Durbin, and C. M. Lindgren (2012). "Extent, causes, and consequences of small RNA expression variation in human adipose tissue". *PLoS Genet* 8, e1002704.

Passlick, B., D. Flieger, and H. W. Ziegler-Heitbrock (1989). "Identification and characterization of a novel monocyte subpopulation in human peripheral blood". *Blood* 74, pp. 2527–34.

Pastinen, T. (2010). "Genome-wide allele-specific analysis: insights into regulatory variation". *Nat Rev Genet* 11, pp. 533–8.

Payne, R. and M. R. Rolfs (1958). "Fetomaternal leukocyte incompatibility". *J Clin Invest* 37, pp. 1756–63.

Pei, H., Y. Yao, Y. Yang, K. Liao, and J. R. Wu (2011). "Kruppel-like factor KLF9 regulates PPARgamma transactivation at the middle stage of adipogenesis". *Cell Death Differ* 18, pp. 315–27.

Peto, R., S. Darby, H. Deo, P. Silcocks, E. Whitley, and R. Doll (2000). "Smoking, smoking cessation, and lung cancer in the UK since 1950: combination of national statistics with two case-control studies". *BMJ* 321, pp. 323–9.

Petretto, E., J. Mangion, N. J. Dickens, S. A. Cook, M. K. Kumaran, H. Lu, J. Fischer, H. Maatz, V. Kren, M. Pravenec, N. Hubner, and T. J. Aitman (2006). "Heritability and tissue specificity of expression quantitative trait loci". *PLoS Genet* 2, e172.

Phillips, R. J., M. Lutz, and B. Premack (2005). "Differential signaling mechanisms regulate expression of CC chemokine receptor-2 during monocyte maturation". *J Inflamm (Lond)* 2, p. 14.

Pickrell, J. K., J. C. Marioni, A. A. Pai, J. F. Degner, B. E. Engelhardt, E. Nkadori, J. B. Veyrieras, M. Stephens, Y. Gilad, and J. K. Pritchard (2010). "Understanding mechanisms underlying human gene expression variation with RNA sequencing". *Nature* 464, pp. 768–72.

Plagnol, V., E. Uz, C. Wallace, H. Stevens, D. Clayton, T. Ozcelik, and J. A. Todd (2008). "Extreme clonality in lymphoblastoid cell lines with implications for allele specific expression analyses". *PLoS One* 3, e2966.

Plagnol, V., J. M. Howson, D. J. Smyth, N. Walker, J. P. Hafler, C. Wallace, H. Stevens, L. Jackson, M. J. Simmonds, P. J. Bingley, S. C. Gough, and J. A. Todd (2011). "Genome-

wide association analysis of autoantibody positivity in type 1 diabetes cases". *PLoS Genet* 7, e1002216.

Plenge, R. M., M. Seielstad, L. Padyukov, A. T. Lee, E. F. Remmers, B. Ding, A. Liew, H. Khalili, A. Chandrasekaran, L. R. Davies, W. Li, A. K. Tan, C. Bonnard, R. T. Ong, A. Thalamuthu, S. Pettersson, C. Liu, C. Tian, W. V. Chen, J. P. Carulli, E. M. Beckman, D. Altshuler, L. Alfredsson, L. A. Criswell, C. I. Amos, M. F. Seldin, D. L. Kastner, L. Klareskog, and P. K. Gregersen (2007). "TRAF1-C5 as a risk locus for rheumatoid arthritis–a genomewide study". *N Engl J Med* 357, pp. 1199–209.

Pollard, J. W. (2009). "Trophic macrophages in development and disease". *Nat Rev Immunol* 9, pp. 259–70.

Poltorak, A., I. Smirnova, X. He, M. Y. Liu, C. Van Huffel, O. McNally, D. Birdwell, E. Alejos, M. Silva, X. Du, P. Thompson, E. K. Chan, J. Ledesma, B. Roe, S. Clifton, S. N. Vogel, and B. Beutler (1998). "Genetic and physical mapping of the Lps locus: identification of the toll-4 receptor as a candidate gene in the critical region". *Blood Cells Mol Dis* 24, pp. 340–55.

Polychronakos, C. and Q. Li (2011). "Understanding type 1 diabetes through genetics: advances and prospects". *Nat Rev Genet* 12, pp. 781–92.

Powell, J. E., A. K. Henders, A. F. McRae, M. J. Wright, N. G. Martin, E. T. Dermitzakis, G. W. Montgomery, and P. M. Visscher (2012). "Genetic control of gene expression in whole blood and lymphoblastoid cell lines is largely independent". *Genome Res* 22, pp. 456–66.

Price, A. L., A. Helgason, G. Thorleifsson, S. A. McCarroll, A. Kong, and K. Stefansson (2011). "Single-tissue and cross-tissue heritability of gene expression via identity-by-descent in related or unrelated individuals". *PLoS Genet* 7, e1001317.

Pritchard, J. K. (2001). "Are rare variants responsible for susceptibility to complex diseases?" *Am J Hum Genet* 69, pp. 124–37.

Pugliese, A., M. Zeller, Jr. Fernandez A., L. J. Zalcberg, R. J. Bartlett, C. Ricordi, M. Pietropaolo, G. S. Eisenbarth, S. T. Bennett, and D. D. Patel (1997). "The insulin gene is transcribed in the human thymus and transcription levels correlated with allelic variation at the INS VNTR-IDDM2 susceptibility locus for type 1 diabetes". *Nat Genet* 15, pp. 293–7.

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. de Bakker, M. J. Daly, and P. C. Sham (2007). "PLINK: a tool set for whole-genome association and population-based linkage analyses". *Am J Hum Genet* 81, pp. 559–75.

Rabinovich, G. A., F. T. Liu, M. Hirashima, and A. Anderson (2007). "An emerging role for galectins in tuning the immune response: lessons from experimental models of inflammatory disease, autoimmunity and cancer". *Scand J Immunol* 66, pp. 143–58.

Rahman, A. and D. A. Isenberg (2008). "Systemic lupus erythematosus". *N Engl J Med* 358, pp. 929–39.

Randolph, G. J., G. Sanchez-Schmitz, R. M. Liebman, and K. Schakel (2002). "The CD16(+) (FcgammaRIII(+)) subset of human monocytes preferentially becomes migratory dendritic cells in a model tissue setting". *J Exp Med* 196, pp. 517–27.

Rathinam, V. A., S. K. Vanaja, and K. A. Fitzgerald (2012). "Regulation of inflammasome signaling". *Nat Immunol* 13, pp. 333–2.

Ray, D., R. Bosselut, J. Ghysdael, M. G. Mattei, A. Tavitian, and F. Moreau-Gachelin (1992). "Characterization of Spi-B, a transcription factor related to the putative oncoprotein Spi-1/PU.1". *Mol Cell Biol* 12, pp. 4297–304.

Raychaudhuri, S., E. F. Remmers, A. T. Lee, R. Hackett, C. Guiducci, N. P. Burtt, L. Gianniny, B. D. Korman, L. Padyukov, F. A. Kurreeman, M. Chang, J. J. Catanese, B. Ding, S. Wong, A. H. van der Helm-van Mil, B. M. Neale, J. Coblyn, J. Cui, P. P. Tak, G. J. Wolbink, J. B. Crusius, I. E. van der Horst-Bruinsma, L. A. Criswell, C. I. Amos, M. F. Seldin, D. L. Kastner, K. G. Ardlie, L. Alfredsson, K. H. Costenbader, D. Altshuler, T. W. Huizinga, N. A. Shadick, M. E. Weinblatt, N. de Vries, J. Worthington, M. Seielstad, R. E. Toes, E. W. Karlson, A. B. Begovich, L. Klareskog, P. K. Gregersen, M. J. Daly, and R. M. Plenge (2008). "Common variants at CD40 and other loci confer risk of rheumatoid arthritis". *Nat Genet* 40, pp. 1216–23.

Raychaudhuri, S., C. Sandor, E. A. Stahl, J. Freudenberg, H. S. Lee, X. Jia, L. Alfredsson, L. Padyukov, L. Klareskog, J. Worthington, K. A. Siminovitch, S. C. Bae, R. M. Plenge, P. K. Gregersen, and P. I. de Bakker (2012). "Five amino acids in three HLA proteins explain most of the association between MHC and seropositive rheumatoid arthritis". *Nat Genet* 44, pp. 291–6.

Reddy, T. E., J. Gertz, F. Pauli, K. S. Kucera, K. E. Varley, K. M. Newberry, G. K. Marinov, A. Mortazavi, B. A. Williams, L. Song, G. E. Crawford, B. Wold, H. F. Willard, and R. M. Myers (2012). "Effects of sequence variation on differential allelic transcription factor occupancy and gene expression". *Genome Res* 22, pp. 860–9.

Redon, R., S. Ishikawa, K. R. Fitch, L. Feuk, G. H. Perry, T. D. Andrews, H. Fiegler, M. H. Shapero, A. R. Carson, W. Chen, E. K. Cho, S. Dallaire, J. L. Freeman, J. R. Gonzalez, M. Gratacos, J. Huang, D. Kalaitzopoulos, D. Komura, J. R. MacDonald, C. R. Marshall, R. Mei, L. Montgomery, K. Nishimura, K. Okamura, F. Shen, M. J. Somerville, J. Tchinda, A. Valsesia, C. Woodwark, F. Yang, J. Zhang, T. Zerjal, J. Zhang, L. Armengol, D. F. Conrad, X. Estivill, C. Tyler-Smith, N. P. Carter, H. Aburatani, C. Lee, K. W. Jones, S. W. Scherer, and M. E. Hurles (2006). "Global variation in copy number in the human genome". *Nature* 444, pp. 444–54.

Ricote, M., A. C. Li, T. M. Willson, C. J. Kelly, and C. K. Glass (1998). "The peroxisome proliferator-activated receptor-gamma is a negative regulator of macrophage activation". *Nature* 391, pp. 79–82.

Ridker, P. M., G. Pare, A. Parker, R. Y. Zee, J. S. Danik, J. E. Buring, D. Kwiatkowski, N. R. Cook, J. P. Miletich, and D. I. Chasman (2008). "Loci related to metabolic-syndrome pathways including LEPR,HNF1A, IL6R, and GCKR associate with plasma C-reactive protein: the Women's Genome Health Study". *Am J Hum Genet* 82, pp. 1185–92.

Ringner, M. (2008). "What is principal component analysis?" *Nat Biotechnol* 26, pp. 303–4.

Rioux, J. D., R. J. Xavier, K. D. Taylor, M. S. Silverberg, P. Goyette, A. Huett, T. Green, P. Kuballa, M. M. Barmada, L. W. Datta, Y. Y. Shugart, A. M. Griffiths, S. R. Targan, A. F. Ippoliti, E. J. Bernard, L. Mei, D. L. Nicolae, M. Regueiro, L. P. Schumm, A. H. Steinhart, J. I. Rotter, R. H. Duerr, J. H. Cho, M. J. Daly, and S. R. Brant (2007). "Genome-wide association study identifies new susceptibility loci for Crohn disease and implicates autophagy in disease pathogenesis". *Nat Genet* 39, pp. 596–604.

Rippe, V., N. Drieschner, M. Meiboom, H. Murua Escobar, U. Bonk, G. Belge, and J. Bullerdiek (2003). "Identification of a gene rearranged by 2p21 aberrations in thyroid adenomas". *Oncogene* 22, pp. 6111–4.

Risch, N. and K. Merikangas (1996). "The future of genetic studies of complex human diseases". *Science* 273, pp. 1516–7.

Ross, D. T., U. Scherf, M. B. Eisen, C. M. Perou, C. Rees, P. Spellman, V. Iyer, S. S. Jeffrey, M. Van de Rijn, M. Waltham, A. Pergamenschikov, J. C. Lee, D. Lashkari, D. Shalon, T. G. Myers, J. N. Weinstein, D. Botstein, and P. O. Brown (2000). "Systematic variation in gene expression patterns in human cancer cell lines". *Nat Genet* 24, pp. 227–35.

Ross, R. (1993). "The pathogenesis of atherosclerosis: a perspective for the 1990s". *Nature* 362, pp. 801–9.

Rotival, M., T. Zeller, P. S. Wild, S. Maouche, S. Szymczak, A. Schillert, R. Castagne, A. Deiseroth, C. Proust, J. Brocheton, T. Godefroy, C. Perret, M. Germain, M. Eleftheriadis, C. R. Sinning, R. B. Schnabel, E. Lubos, K. J. Lackner, H. Rossmann, T. Munzel, A. Rendon, J. Erdmann, P. Deloukas, C. Hengstenberg, P. Diemert, G. Montalescot, W. H. Ouwehand, N. J. Samani, H. Schunkert, D. A. Tregouet, A. Ziegler, A. H. Goodall, F. Cambien, L. Tiret, and S. Blankenberg (2011). "Integrating genome-wide genetic variations and monocyte expression data reveals trans-regulated gene modules in humans". *PLoS Genet* 7, e1002367.

Rowland, B. D., R. Bernards, and D. S. Peeper (2005). "The KLF4 tumour suppressor is a transcriptional repressor of p53 that acts as a context-dependent oncogene". *Nat Cell Biol* 7, pp. 1074–82.

Ruffell, B., N. I. Affara, and L. M. Coussens (2012). "Differential macrophage programming in the tumor microenvironment". *Trends Immunol* 33, pp. 119–26.

Ryan, E. J., A. J. Marshall, D. Magaletti, H. Floyd, K. E. Draves, N. E. Olson, and E. A. Clark (2002). "Dendritic cell-associated lectin-1: a novel dendritic cell-associated, C-type lectin-like molecule enhances T cell secretion of IL-4". *J Immunol* 169, pp. 5638–48.

Sachidanandam, R., D. Weissman, S. C. Schmidt, J. M. Kakol, L. D. Stein, G. Marth, S. Sherry, J. C. Mullikin, B. J. Mortimore, D. L. Willey, S. E. Hunt, C. G. Cole, P. C. Coggill, C. M. Rice, Z. Ning, J. Rogers, D. R. Bentley, P. Y. Kwok, E. R. Mardis, R. T. Yeh, B. Schultz, L. Cook, R. Davenport, M. Dante, L. Fulton, L. Hillier, R. H. Waterston, J. D. McPherson, B. Gilman, S. Schaffner, W. J. Van Etten, D. Reich, J. Higgins, M. J. Daly, B. Blumenstiel, J. Baldwin, N. Stange-Thomann, M. C. Zody, L. Linton, E. S. Lander, and D. Altshuler (2001). "A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms". *Nature* 409, pp. 928–33.

Saeki, N., D. H. Kim, T. Usui, K. Aoyagi, T. Tatsuta, K. Aoki, K. Yanagihara, M. Tamura, H. Mizushima, H. Sakamoto, K. Ogawa, M. Ohki, T. Shiroishi, T. Yoshida, and H. Sasaki (2007). "GASDERMIN, suppressed frequently in gastric cancer, is a target of LMO1 in TGF-beta-dependent apoptotic signalling". *Oncogene* 26, pp. 6488–98.

Saiki, R. K., S. Scharf, F. Faloona, K. B. Mullis, G. T. Horn, H. A. Erlich, and N. Arnheim (1985). "Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia". *Science* 230, pp. 1350–4.

Sandelin, A., W. Alkema, P. Engstrom, W. W. Wasserman, and B. Lenhard (2004). "JASPAR: an open-access database for eukaryotic transcription factor binding profiles". *Nucleic Acids Res* 32, pp. D91–4.

Sanger, F., S. Nicklen, and A. R. Coulson (1977). "DNA sequencing with chain-terminating inhibitors". *Proc Natl Acad Sci U S A* 74, pp. 5463–7.

Santiago-Schwarz, F., E. Belilos, B. Diamond, and S. E. Carsons (1992). "TNF in combination with GM-CSF enhances the differentiation of neonatal cord blood stem cells into dendritic cells and macrophages". *J Leukoc Biol* 52, pp. 274–81.

Sanyal, A., B. R. Lajoie, G. Jain, and J. Dekker (2012). "The long-range interaction landscape of gene promoters". *Nature* 489, pp. 109–13.

Satsangi, J., K. I. Welsh, M. Bunce, C. Julier, J. M. Farrant, J. I. Bell, and D. P. Jewell (1996). "Contribution of genes of the major histocompatibility complex to susceptibility and disease phenotype in inflammatory bowel disease". *Lancet* 347, pp. 1212–7.

Saveanu, L., O. Carroll, V. Lindo, M. Del Val, D. Lopez, Y. Lepelletier, F. Greer, L. Schomburg, D. Fruci, G. Niedermann, and P. M. van Endert (2005). "Concerted peptide trimming by human ERAP1 and ERAP2 aminopeptidase complexes in the endoplasmic reticulum". *Nat Immunol* 6, pp. 689–97.

Sawcer, S., G. Hellenthal, M. Pirinen, C. C. Spencer, N. A. Patsopoulos, L. Moutsianas, A. Dilthey, Z. Su, C. Freeman, S. E. Hunt, S. Edkins, E. Gray, D. R. Booth, S. C. Potter, A. Goris, G. Band, A. B. Oturai, A. Strange, J. Saarela, C. Bellenguez, B. Fontaine, M. Gillman, B. Hemmer, R. Gwilliam, F. Zipp, A. Jayakumar, R. Martin, S. Leslie, S. Hawkins, E. Giannoulatou, S. D'Alfonso, H. Blackburn, F. Martinelli Boneschi, J. Liddle, H. F. Harbo, M. L. Perez, A. Spurkland, M. J. Waller, M. P. Mycko, M. Ricketts, M. Comabella, N. Hammond, I. Kockum, O. T. McCann, M. Ban, P. Whittaker, A. Kemppinen, P. Weston, C. Hawkins, S. Widaa, J. Zajicek, S. Dronov, N. Robertson, S. J. Bumpstead, L. F. Barcellos, R. Ravindrarajah, R. Abraham, L. Alfredsson, K. Ardlie, C. Aubin, A. Baker, K. Baker, S. E. Baranzini, L. Bergamaschi, R. Bergamaschi, A. Bernstein, A. Berthele, M. Boggild, J. P. Bradfield, D. Brassat, S. A. Broadley, D. Buck, H. Butzkueven, R. Capra, W. M. Carroll, P. Cavalla, E. G. Celius, S. Cepok, R. Chiavacci, F. Clerget-Darpoux, K. Clysters, G. Comi, M. Cossburn, I. Cournu-Rebeix, M. B. Cox, W. Cozen, B. A. Cree, A. H. Cross, D. Cusi, M. J. Daly, E. Davis, P. I. de Bakker, M. Debouverie, B. D'Hooghe M, K. Dixon, R. Dobosi, B. Dubois, D. Ellinghaus, I. Elovaara, *et al.* (2011). "Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis". *Nature* 476, pp. 214–9.

Saxena, R., B. F. Voight, V. Lyssenko, N. P. Burtt, P. I. de Bakker, H. Chen, J. J. Roix, S. Kathiresan, J. N. Hirschhorn, M. J. Daly, T. E. Hughes, L. Groop, D. Altshuler, P. Almgren, J. C. Florez, J. Meyer, K. Ardlie, K. Bengtsson Bostrom, B. Isomaa, G. Lettre, U. Lindblad, H. N. Lyon, O. Melander, C. Newton-Cheh, P. Nilsson, M. Orho-Melander, L. Rastam, E. K. Speliotes, M. R. Taskinen, T. Tuomi, C. Guiducci,

A. Berglund, J. Carlson, L. Gianniny, R. Hackett, L. Hall, J. Holmkvist, E. Laurila, M. Sjogren, M. Sterner, A. Surti, M. Svensson, M. Svensson, R. Tewhey, B. Blumenstiel, M. Parkin, M. Defelice, R. Barry, W. Brodeur, J. Camarata, N. Chia, M. Fava, J. Gibbons, B. Handsaker, C. Healy, K. Nguyen, C. Gates, C. Sougnez, D. Gage, M. Nizzari, S. B. Gabriel, G. W. Chirn, Q. Ma, H. Parikh, D. Richardson, D. Ricke, and S. Purcell (2007). "Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels". *Science* 316, pp. 1331–6.

Schadt, E. E., S. A. Monks, T. A. Drake, A. J. Lusis, N. Che, V. Colinayo, T. G. Ruff, S. B. Milligan, J. R. Lamb, G. Cavet, P. S. Linsley, M. Mao, R. B. Stoughton, and S. H. Friend (2003). "Genetics of gene expression surveyed in maize, mouse and man". *Nature* 422, pp. 297–302.

Schadt, E. E., J. Lamb, X. Yang, J. Zhu, S. Edwards, D. Guhathakurta, S. K. Sieberts, S. Monks, M. Reitman, C. Zhang, P. Y. Lum, A. Leonardson, R. Thieringer, J. M. Metzger, L. Yang, J. Castle, H. Zhu, S. F. Kash, T. A. Drake, A. Sachs, and A. J. Lusis (2005). "An integrative genomics approach to infer causal associations between gene expression and disease". *Nat Genet* 37, pp. 710–7.

Schadt, E. E., C. Molony, E. Chudin, K. Hao, X. Yang, P. Y. Lum, A. Kasarskis, B. Zhang, S. Wang, C. Suver, J. Zhu, J. Millstein, S. Sieberts, J. Lamb, D. GuhaThakurta, J. Derry, J. D. Storey, I. Avila-Campillo, M. J. Kruger, J. M. Johnson, C. A. Rohl, A. van Nas, M. Mehrabian, T. A. Drake, A. J. Lusis, R. C. Smith, F. P. Guengerich, S. C. Strom, E. Schuetz, T. H. Rushmore, and R. Ulrich (2008). "Mapping the genetic architecture of gene expression in human liver". *PLoS Biol* 6, e107.

Schalkwyk, L. C., E. L. Meaburn, R. Smith, E. L. Dempster, A. R. Jeffries, M. N. Davies, R. Plomin, and J. Mill (2010). "Allelic skewing of DNA methylation is widespread across the genome". *Am J Hum Genet* 86, pp. 196–212.

Scherer, S. W., C. Lee, E. Birney, D. M. Altshuler, E. E. Eichler, N. P. Carter, M. E. Hurles, and L. Feuk (2007). "Challenges and standards in integrating surveys of structural variation". *Nat Genet* 39, S7–15.

Schilling, E., C. El Chartouni, and M. Rehli (2009). "Allele-specific DNA methylation in mouse strains is mainly determined by cis-acting sequences". *Genome Res* 19, pp. 2028–35.

Schlosstein, L., P. I. Terasaki, R. Bluestone, and C. M. Pearson (1973). "High association of an HL-A antigen, W27, with ankylosing spondylitis". *N Engl J Med* 288, pp. 704–6.

Schmid, R., P. Baum, C. Ittrich, K. Fundel-Clemens, W. Huber, B. Brors, R. Eils, A. Weith, D. Mennerich, and K. Quast (2010). "Comparison of normalization methods for Illumina BeadChip HumanHT-12 v3". *BMC Genomics* 11, p. 349.

Schneider, U., H. U. Schwenk, and G. Bornkamm (1977). "Characterization of EBV-genome negative "null" and "T" cell lines derived from children with acute lymphoblastic leukemia and leukemic transformed non-Hodgkin lymphoma". *Int J Cancer* 19, pp. 621–6.

Schoggins, J. W., S. J. Wilson, M. Panis, M. Y. Murphy, C. T. Jones, P. Bieniasz, and C. M. Rice (2011). "A diverse range of gene products are effectors of the type I interferon antiviral response". *Nature* 472, pp. 481–5.

Schork, N. J., S. S. Murray, K. A. Frazer, and E. J. Topol (2009). "Common vs. rare allele hypotheses for complex diseases". *Curr Opin Genet Dev* 19, pp. 212–9.

Schreiber, E., P. Matthias, M. M. Muller, and W. Schaffner (1989). "Rapid detection of octamer binding proteins with 'mini-extracts', prepared from a small number of cells". *Nucleic Acids Res* 17, p. 6419.

Scott, L. J., K. L. Mohlke, L. L. Bonnycastle, C. J. Willer, Y. Li, W. L. Duren, M. R. Erdos, H. M. Stringham, P. S. Chines, A. U. Jackson, L. Prokunina-Olsson, C. J. Ding, A. J. Swift, N. Narisu, T. Hu, R. Pruim, R. Xiao, X. Y. Li, K. N. Conneely, N. L. Riebow, A. G. Sprau, M. Tong, P. P. White, K. N. Hetrick, M. W. Barnhart, C. W. Bark, J. L. Goldstein, L. Watkins, F. Xiang, J. Saramies, T. A. Buchanan, R. M. Watanabe, T. T. Valle, L. Kinnunen, G. R. Abecasis, E. W. Pugh, K. F. Doheny, R. N. Bergman, J. Tuomilehto, F. S. Collins, and M. Boehnke (2007). "A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants". *Science* 316, pp. 1341–5.

Scott, W. A. and D. J. Wigmore (1978). "Sites in simian virus 40 chromatin which are preferentially cleaved by endonucleases". *Cell* 15, pp. 1511–8.
Serbina, N. V., T. Jia, T. M. Hohl, and E. G. Pamer (2008). "Monocyte-mediated defense against microbial pathogens". *Annu Rev Immunol* 26, pp. 421–52.
Serhan, C. N. and J. Savill (2005). "Resolution of inflammation: the beginning programs the end". *Nat Immunol* 6, pp. 1191–7.
Sethupathy, P. and F. S. Collins (2008). "MicroRNA target site polymorphisms and human disease". *Trends Genet* 24, pp. 489–97.
Shabalin, A. A. (2012). "Matrix eQTL: ultra fast eQTL analysis via large matrix operations". *Bioinformatics* 28, pp. 1353–8.
Shaulian, E. and M. Karin (2002). "AP-1 as a regulator of cell life and death". *Nat Cell Biol* 4, E131–6.
Sherry, S. T., M. H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski, and K. Sirotkin (2001). "dbSNP: the NCBI database of genetic variation". *Nucleic Acids Res* 29, pp. 308–11.
Shi, C. and E. G. Pamer (2011). "Monocyte recruitment during infection and inflammation". *Nat Rev Immunol* 11, pp. 762–74.
Sladek, R., G. Rocheleau, J. Rung, C. Dina, L. Shen, D. Serre, P. Boutin, D. Vincent, A. Belisle, S. Hadjadj, B. Balkau, B. Heude, G. Charpentier, T. J. Hudson, A. Montpetit, A. V. Pshezhetsky, M. Prentki, B. I. Posner, D. J. Balding, D. Meyre, C. Polychronakos, and P. Froguel (2007). "A genome-wide association study identifies novel risk loci for type 2 diabetes". *Nature* 445, pp. 881–5.
Small, K. S., A. K. Hedman, E. Grundberg, A. C. Nica, G. Thorleifsson, A. Kong, U. Thorsteindottir, S. Y. Shin, H. B. Richards, N. Soranzo, K. R. Ahmadi, C. M. Lindgren, K. Stefansson, E. T. Dermitzakis, P. Deloukas, T. D. Spector, and M. I. McCarthy (2011). "Identification of an imprinted master trans regulator at the KLF14 locus related to multiple metabolic phenotypes". *Nat Genet* 43, pp. 561–564.
Smirnov, D. A., M. Morley, E. Shin, R. S. Spielman, and V. G. Cheung (2009). "Genetic analysis of radiation-induced changes in human gene expression". *Nature* 459, pp. 587–91.
Smith, A. J. and S. E. Humphries (2009). "Characterization of DNA-binding proteins using multiplexed competitor EMSA". *J Mol Biol* 385, pp. 714–7.
Smith, H. O. and K. W. Wilcox (1970). "A restriction enzyme from Hemophilus influenzae. I. Purification and general properties". *J Mol Biol* 51, pp. 379–91.
Smith, M. W., M. Dean, M. Carrington, C. Winkler, G. A. Huttley, D. A. Lomb, J. J. Goedert, T. R. O'Brien, L. P. Jacobson, R. Kaslow, S. Buchbinder, E. Vittinghoff, D. Vlahov, K. Hoots, M. W. Hilgartner, and S. J. O'Brien (1997). "Contrasting genetic influence of CCR2 and CCR5 variants on HIV-1 infection and disease progression. Hemophilia Growth and Development Study (HGDS), Multicenter AIDS Cohort Study (MACS), Multicenter Hemophilia Cohort Study (MHCS), San Francisco City Cohort (SFCC), ALIVE Study". *Science* 277, pp. 959–65.
Smith, R. P., S. J. Riesenfeld, A. K. Holloway, Q. Li, K. K. Murphy, N. M. Feliciano, L. Orecchia, N. Oksenberg, K. S. Pollard, and N. Ahituv (2013a). "A compact, in vivo screen of all 6-mers reveals drivers of tissue-specific expression and guides synthetic regulatory element design". *Genome Biol* 14, R72.
Smith, R. P., L. Taher, R. P. Patwardhan, M. J. Kim, F. Inoue, J. Shendure, I. Ovcharenko, and N. Ahituv (2013b). "Massively parallel decoding of mammalian regulatory sequences supports a flexible organizational model". *Nat Genet* 45, pp. 1021–8.
Snell, G. D. and G. F. Higgins (1951). "Alleles at the histocompatibility-2 locus in the mouse as determined by tumor transplantation". *Genetics* 36, pp. 306–10.
Sollid, L. M., G. Markussen, J. Ek, H. Gjerde, F. Vartdal, and E. Thorsby (1989). "Evidence for a primary association of celiac disease to a particular HLA-DQ alpha/beta heterodimer". *J Exp Med* 169, pp. 345–50.
Solomon, M. J., P. L. Larsen, and A. Varshavsky (1988). "Mapping protein-DNA interactions in vivo with formaldehyde: evidence that histone H4 is retained on a highly transcribed gene". *Cell* 53, pp. 937–47.
Sotiriou, C. and L. Pusztai (2009). "Gene-expression signatures in breast cancer". *N Engl J Med* 360, pp. 790–800.

Speliotes, E. K., C. J. Willer, S. I. Berndt, K. L. Monda, G. Thorleifsson, A. U. Jackson, H. Lango Allen, C. M. Lindgren, J. Luan, R. Magi, J. C. Randall, S. Vedantam, T. W. Winkler, L. Qi, T. Workalemahu, I. M. Heid, V. Steinthorsdottir, H. M. Stringham, M. N. Weedon, E. Wheeler, A. R. Wood, T. Ferreira, R. J. Weyant, A. V. Segre, K. Estrada, L. Liang, J. Nemesh, J. H. Park, S. Gustafsson, T. O. Kilpelainen, J. Yang, N. Bouatia-Naji, T. Esko, M. F. Feitosa, Z. Kutalik, M. Mangino, S. Raychaudhuri, A. Scherag, A. V. Smith, R. Welch, J. H. Zhao, K. K. Aben, D. M. Absher, N. Amin, A. L. Dixon, E. Fisher, N. L. Glazer, M. E. Goddard, N. L. Heard-Costa, V. Hoesel, J. J. Hottenga, A. Johansson, T. Johnson, S. Ketkar, C. Lamina, S. Li, M. F. Moffatt, R. H. Myers, N. Narisu, J. R. Perry, M. J. Peters, M. Preuss, S. Ripatti, F. Rivadeneira, C. Sandholt, L. J. Scott, N. J. Timpson, J. P. Tyrer, S. van Wingerden, R. M. Watanabe, C. C. White, F. Wiklund, C. Barlassina, D. I. Chasman, M. N. Cooper, J. O. Jansson, R. W. Lawrence, N. Pellikka, I. Prokopenko, J. Shi, E. Thiering, H. Alavere, M. T. Alibrandi, P. Almgren, A. M. Arnold, T. Aspelund, L. D. Atwood, B. Balkau, A. J. Balmforth, A. J. Bennett, Y. Ben-Shlomo, R. N. Bergman, S. Bergmann, H. Biebermann, A. I. Blakemore, T. Boes, L. L. Bonnycastle, S. R. Bornstein, M. J. Brown, *et al.* (2010). "Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index". *Nat Genet* 42, pp. 937–48.

Spiegelman, B. M. (1998). "PPAR-gamma: adipogenic regulator and thiazolidinedione receptor". *Diabetes* 47, pp. 507–14.

Spielman, R. S., L. A. Bastone, J. T. Burdick, M. Morley, W. J. Ewens, and V. G. Cheung (2007). "Common genetic variants account for differences in gene expression among ethnic groups". *Nat Genet* 39, pp. 226–31.

Stahl, E. A., S. Raychaudhuri, E. F. Remmers, G. Xie, S. Eyre, B. P. Thomson, Y. Li, F. A. Kurreeman, A. Zhernakova, A. Hinks, C. Guiducci, R. Chen, L. Alfredsson, C. I. Amos, K. G. Ardlie, A. Barton, J. Bowes, E. Brouwer, N. P. Burtt, J. J. Catanese, J. Coblyn, M. J. Coenen, K. H. Costenbader, L. A. Criswell, J. B. Crusius, J. Cui, P. I. de Bakker, P. L. De Jager, B. Ding, P. Emery, E. Flynn, P. Harrison, L. J. Hocking, T. W. Huizinga, D. L. Kastner, X. Ke, A. T. Lee, X. Liu, P. Martin, A. W. Morgan, L. Padyukov, M. D. Posthumus, T. R. Radstake, D. M. Reid, M. Seielstad, M. F. Seldin, N. A. Shadick, S. Steer, P. P. Tak, W. Thomson, A. H. van der Helm-van Mil, I. E. van der Horst-Bruinsma, C. E. van der Schoot, P. L. van Riel, M. E. Weinblatt, A. G. Wilson, G. J. Wolbink, B. P. Wordsworth, C. Wijmenga, E. W. Karlson, R. E. Toes, N. de Vries, A. B. Begovich, J. Worthington, K. A. Siminovitch, P. K. Gregersen, L. Klareskog, and R. M. Plenge (2010). "Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci". *Nat Genet* 42, pp. 508–14.

Stanciu, L. A., J. Shute, S. T. Holgate, and R. Djukanovic (1996). "Production of IL-8 and IL-4 by positively and negatively selected CD4+ and CD8+ human T cells following a four-step cell separation method including magnetic cell sorting (MACS)". *J Immunol Methods* 189, pp. 107–15.

Stankiewicz, P. and J. R. Lupski (2010). "Structural variation in the human genome and its role in disease". *Annu Rev Med* 61, pp. 437–55.

Stastny, P. (1978). "Association of the B-cell alloantigen DRw4 with rheumatoid arthritis". *N Engl J Med* 298, pp. 869–71.

Stead, J. A. and K. J. McDowall (2007). "Two-dimensional gel electrophoresis for identifying proteins that bind DNA or RNA". *Nat Protoc* 2, pp. 1839–48.

Stead, J. A., J. N. Keen, and K. J. McDowall (2006). "The identification of nucleic acid-interacting proteins using a simple proteomics-based approach that directly incorporates the electrophoretic mobility shift assay". *Mol Cell Proteomics* 5, pp. 1697–702.

Steemers, F. J. and K. L. Gunderson (2007). "Whole genome genotyping technologies on the BeadArray platform". *Biotechnol J* 2, pp. 41–9.

Stefansson, H., D. Rujescu, S. Cichon, O. P. Pietilainen, A. Ingason, S. Steinberg, R. Fossdal, E. Sigurdsson, T. Sigmundsson, J. E. Buizer-Voskamp, T. Hansen, K. D. Jakobsen, P. Muglia, C. Francks, P. M. Matthews, A. Gylfason, B. V. Halldorsson, D. Gudbjartsson, T. E. Thorgeirsson, A. Sigurdsson, A. Jonasdottir, A. Jonasdottir, A. Bjornsson, S. Mattiasdottir, T. Blondal, M. Haraldsson, B. B. Magnusdottir, I. Giegling, H. J. Moller, A. Hartmann, K. V. Shianna, D. Ge, A. C. Need, C. Crombie, G. Fraser, N. Walker,

J. Lonnqvist, J. Suvisaari, A. Tuulio-Henriksson, T. Paunio, T. Toulopoulou, E. Bramon, M. Di Forti, R. Murray, M. Ruggeri, E. Vassos, S. Tosato, M. Walshe, T. Li, C. Vasilescu, T. W. Muhleisen, A. G. Wang, H. Ullum, S. Djurovic, I. Melle, J. Olesen, L. A. Kiemeney, B. Franke, C. Sabatti, N. B. Freimer, J. R. Gulcher, U. Thorsteinsdottir, A. Kong, O. A. Andreassen, R. A. Ophoff, A. Georgi, M. Rietschel, T. Werge, H. Petursson, D. B. Goldstein, M. M. Nothen, L. Peltonen, D. A. Collier, D. St Clair, and K. Stefansson (2008). "Large recurrent microdeletions associated with schizophrenia". *Nature* 455, pp. 232–6.

Steimle, V., C. A. Siegrist, A. Mottet, B. Lisowska-Grospierre, and B. Mach (1994). "Regulation of MHC class II expression by interferon-gamma mediated by the transactivator gene CIITA". *Science* 265, pp. 106–9.

Stewart, C. A., R. Horton, R. J. Allcock, J. L. Ashurst, A. M. Atrazhev, P. Coggill, I. Dunham, S. Forbes, K. Halls, J. M. Howson, S. J. Humphray, S. Hunt, A. J. Mungall, K. Osoegawa, S. Palmer, A. N. Roberts, J. Rogers, S. Sims, Y. Wang, L. G. Wilming, J. F. Elliott, P. J. de Jong, S. Sawcer, J. A. Todd, J. Trowsdale, and S. Beck (2004). "Complete MHC haplotype sequencing for common disease gene mapping". *Genome Res* 14, pp. 1176–87.

Stokes, P. L., P. Asquith, G. K. Holmes, P. Mackintosh, and W. T. Cooke (1972). "Histocompatibility antigens associated with adult coeliac disease". *Lancet* 2, pp. 162–4.

Stormo, G. D. (2000). "DNA binding sites: representation and discovery". *Bioinformatics* 16, pp. 16–23.

Strahl, B. D. and C. D. Allis (2000). "The language of covalent histone modifications". *Nature* 403, pp. 41–5.

Strange, A., F. Capon, C. C. Spencer, J. Knight, M. E. Weale, M. H. Allen, A. Barton, G. Band, C. Bellenguez, J. G. Bergboer, J. M. Blackwell, E. Bramon, S. J. Bumpstead, J. P. Casas, M. J. Cork, A. Corvin, P. Deloukas, A. Dilthey, A. Duncanson, S. Edkins, X. Estivill, O. Fitzgerald, C. Freeman, E. Giardina, E. Gray, A. Hofer, U. Huffmeier, S. E. Hunt, A. D. Irvine, J. Jankowski, B. Kirby, C. Langford, J. Lascorz, J. Leman, S. Leslie, L. Mallbris, H. S. Markus, C. G. Mathew, W. H. McLean, R. McManus, R. Mossner, L. Moutsianas, A. T. Naluai, F. O. Nestle, G. Novelli, A. Onoufriadis, C. N. Palmer, C. Perricone, M. Pirinen, R. Plomin, S. C. Potter, R. M. Pujol, A. Rautanen, E. Riveira-Munoz, A. W. Ryan, W. Salmhofer, L. Samuelsson, S. J. Sawcer, J. Schalkwijk, C. H. Smith, M. Stahle, Z. Su, R. Tazi-Ahnini, H. Traupe, A. C. Viswanathan, R. B. Warren, W. Weger, K. Wolk, N. Wood, J. Worthington, H. S. Young, P. L. Zeeuwen, A. Hayday, A. D. Burden, C. E. Griffiths, J. Kere, A. Reis, G. McVean, D. M. Evans, M. A. Brown, J. N. Barker, L. Peltonen, P. Donnelly, and R. C. Trembath (2010). "A genome-wide association study identifies new psoriasis susceptibility loci and an interaction between HLA-C and ERAP1". *Nat Genet* 42, pp. 985–90.

Stranger, B. E., A. C. Nica, M. S. Forrest, A. Dimas, C. P. Bird, C. Beazley, C. E. Ingle, M. Dunning, P. Flicek, D. Koller, S. Montgomery, S. Tavare, P. Deloukas, and E. T. Dermitzakis (2007a). "Population genomics of human gene expression". *Nat Genet* 39, pp. 1217–24.

Stranger, B. E., M. S. Forrest, M. Dunning, C. E. Ingle, C. Beazley, N. Thorne, R. Redon, C. P. Bird, A. de Grassi, C. Lee, C. Tyler-Smith, N. Carter, S. W. Scherer, S. Tavare, P. Deloukas, M. E. Hurles, and E. T. Dermitzakis (2007b). "Relative impact of nucleotide and copy number variation on gene expression phenotypes". *Science* 315, pp. 848–53.

Stranger, B. E., E. A. Stahl, and T. Raj (2011). "Progress and Promise of Genome-wide Association Studies for Human Complex Trait Genetics". *Genetics*.

Stranger, B. E., S. B. Montgomery, A. S. Dimas, L. Parts, O. Stegle, C. E. Ingle, M. Sekowska, G. D. Smith, D. Evans, M. Gutierrez-Arcelus, A. Price, T. Raj, J. Nisbett, A. C. Nica, C. Beazley, R. Durbin, P. Deloukas, and E. T. Dermitzakis (2012). "Patterns of cis regulatory variation in diverse human populations". *PLoS Genet* 8, e1002639.

Stumvoll, M., B. J. Goldstein, and T. W. van Haeften (2005). "Type 2 diabetes: principles of pathogenesis and therapy". *Lancet* 365, pp. 1333–46.

Stunz, L. L., R. W. Karr, and R. A. Anderson (1989). "HLA-DRB1 and -DRB4 genes are differentially regulated at the transcriptional level". *J Immunol* 143, pp. 3081–6.

Sturtevant, A.H. (1913). "The linear arrangement of six sex-linked factors in Drosophila as shown by their mode of association". *Journal of Experimental Zoology* 14, pp. 43–59.

Su, Z. J., C. N. Hahn, G. J. Goodall, N. M. Reck, A. F. Leske, A. Davy, G. Kremmidiotis, M. A. Vadas, and J. R. Gamble (2004). "A vascular cell-restricted RhoGAP, p73RhoGAP, is a key regulator of angiogenesis". *Proc Natl Acad Sci U S A* 101, pp. 12212–7.

Summers, J. and W. S. Mason (1982). "Replication of the genome of a hepatitis B–like virus by reverse transcription of an RNA intermediate". *Cell* 29, pp. 403–15.

Szanto, A., B. L. Balint, Z. S. Nagy, E. Barta, B. Dezso, A. Pap, L. Szeles, S. Poliska, M. Oros, R. M. Evans, Y. Barak, J. Schwabe, and L. Nagy (2010). "STAT6 transcription factor is a facilitator of the nuclear receptor PPARgamma-regulated gene expression in macrophages and dendritic cells". *Immunity* 33, pp. 699–712.

Tabas, I. (2010). "Macrophage death and defective inflammation resolution in atherosclerosis". *Nat Rev Immunol* 10, pp. 36–46.

Takaoka, A., H. Yanai, S. Kondo, G. Duncan, H. Negishi, T. Mizutani, S. Kano, K. Honda, Y. Ohba, T. W. Mak, and T. Taniguchi (2005). "Integral role of IRF-5 in the gene induction programme activated by Toll-like receptors". *Nature* 434, pp. 243–9.

Takeda, K., T. Tanaka, W. Shi, M. Matsumoto, M. Minami, S. Kashiwamura, K. Nakanishi, N. Yoshida, T. Kishimoto, and S. Akira (1996). "Essential role of Stat6 in IL-4 signalling". *Nature* 380, pp. 627–30.

Takeuchi, O. and S. Akira (2010). "Pattern recognition receptors and inflammation". *Cell* 140, pp. 805–20.

Tang, F., C. Barbacioru, E. Nordman, B. Li, N. Xu, V. I. Bashkirov, K. Lao, and M. A. Surani (2010). "RNA-Seq analysis to capture the transcriptome landscape of a single cell". *Nat Protoc* 5, pp. 516–35.

Tang, X., D. L. Marciano, S. E. Leeman, and S. Amar (2005). "LPS induces the interaction of a transcription factor, LPS-induced TNF-alpha factor, and STAT6(B) with effects on multiple cytokines". *Proc Natl Acad Sci U S A* 102, pp. 5132–7.

Tang, X. F., Z. Zhang, D. Y. Hu, A. E. Xu, H. S. Zhou, L. D. Sun, M. Gao, T. W. Gao, X. H. Gao, H. D. Chen, H. F. Xie, C. X. Tu, F. Hao, R. N. Wu, F. R. Zhang, L. Liang, X. M. Pu, J. Z. Zhang, J. W. Han, G. P. Pan, J. Q. Wu, K. Li, M. W. Su, W. D. Du, W. J. Zhang, J. J. Liu, L. H. Xiang, S. Yang, Y. W. Zhou, and X. J. Zhang (2013). "Association analyses identify three susceptibility Loci for vitiligo in the chinese han population". *J Invest Dermatol* 133, pp. 403–10.

Temin, H. M. and S. Mizutani (1970). "RNA-dependent DNA polymerase in virions of Rous sarcoma virus". *Nature* 226, pp. 1211–3.

Teslovich, T. M., K. Musunuru, A. V. Smith, A. C. Edmondson, I. M. Stylianou, M. Koseki, J. P. Pirruccello, S. Ripatti, D. I. Chasman, C. J. Willer, C. T. Johansen, S. W. Fouchier, A. Isaacs, G. M. Peloso, M. Barbalic, S. L. Ricketts, J. C. Bis, Y. S. Aulchenko, G. Thorleifsson, M. F. Feitosa, J. Chambers, M. Orho-Melander, O. Melander, T. Johnson, X. Li, X. Guo, M. Li, Y. Shin Cho, M. Jin Go, Y. Jin Kim, J. Y. Lee, T. Park, K. Kim, X. Sim, R. Twee-Hee Ong, D. C. Croteau-Chonka, L. A. Lange, J. D. Smith, K. Song, J. Hua Zhao, X. Yuan, J. Luan, C. Lamina, A. Ziegler, W. Zhang, R. Y. Zee, A. F. Wright, J. C. Witteman, J. F. Wilson, G. Willemsen, H. E. Wichmann, J. B. Whitfield, D. M. Waterworth, N. J. Wareham, G. Waeber, P. Vollenweider, B. F. Voight, V. Vitart, A. G. Uitterlinden, M. Uda, J. Tuomilehto, J. R. Thompson, T. Tanaka, I. Surakka, H. M. Stringham, T. D. Spector, N. Soranzo, J. H. Smit, J. Sinisalo, K. Silander, E. J. Sijbrands, A. Scuteri, J. Scott, D. Schlessinger, S. Sanna, V. Salomaa, J. Saharinen, C. Sabatti, A. Ruokonen, I. Rudan, L. M. Rose, R. Roberts, M. Rieder, B. M. Psaty, P. P. Pramstaller, I. Pichler, M. Perola, B. W. Penninx, N. L. Pedersen, C. Pattaro, A. N. Parker, G. Pare, B. A. Oostra, C. J. O'Donnell, M. S. Nieminen, D. A. Nickerson, G. W. Montgomery, T. Meitinger, R. McPherson, *et al.* (2010). "Biological, clinical and population relevance of 95 loci for blood lipids". *Nature* 466, pp. 707–13.

The ENCODE Project Consortium; Dunham, I., A. Kundaje, S. F. Aldred, P. J. Collins, C. A. Davis, F. Doyle, C. B. Epstein, S. Frietze, J. Harrow, R. Kaul, J. Khatun, B. R. Lajoie, S. G. Landt, B. K. Lee, F. Pauli, K. R. Rosenbloom, P. Sabo, A. Safi, A. Sanyal, N. Shoresh, J. M. Simon, L. Song, N. D. Trinklein, R. C. Altshuler,

E. Birney, J. B. Brown, C. Cheng, S. Djebali, X. Dong, I. Dunham, J. Ernst, T. S. Furey, M. Gerstein, B. Giardine, M. Greven, R. C. Hardison, R. S. Harris, J. Herrero, M. M. Hoffman, S. Iyer, M. Kelllis, J. Khatun, P. Kheradpour, A. Kundaje, T. Lassman, Q. Li, X. Lin, G. K. Marinov, A. Merkel, A. Mortazavi, S. C. Parker, T. E. Reddy, J. Rozowsky, F. Schlesinger, R. E. Thurman, J. Wang, L. D. Ward, T. W. Whitfield, S. P. Wilder, W. Wu, H. S. Xi, K. Y. Yip, J. Zhuang, B. E. Bernstein, E. Birney, I. Dunham, E. D. Green, C. Gunter, M. Snyder, M. J. Pazin, R. F. Lowdon, L. A. Dillon, L. B. Adams, C. J. Kelly, J. Zhang, J. R. Wexler, E. D. Green, P. J. Good, E. A. Feingold, B. E. Bernstein, E. Birney, G. E. Crawford, J. Dekker, L. Elinitski, P. J. Farnham, M. Gerstein, M. C. Giddings, T. R. Gingeras, E. D. Green, R. Guigo, R. C. Hardison, T. J. Hubbard, M. Kellis, W. J. Kent, J. D. Lieb, E. H. Margulies, R. M. Myers, M. Snyder, J. A. Starnatoyannopoulos, *et al.* (2012). "An integrated encyclopedia of DNA elements in the human genome". *Nature* 489, pp. 57–74.

The MicroArray Quality Control (MAQC) project Shi, L., L. H. Reid, W. D. Jones, R. Shippy, J. A. Warrington, S. C. Baker, P. J. Collins, F. de Longueville, E. S. Kawasaki, K. Y. Lee, Y. Luo, Y. A. Sun, J. C. Willey, R. A. Setterquist, G. M. Fischer, W. Tong, Y. P. Dragan, D. J. Dix, F. W. Frueh, F. M. Goodsaid, D. Herman, R. V. Jensen, C. D. Johnson, E. K. Lobenhofer, R. K. Puri, U. Schrf, J. Thierry-Mieg, C. Wang, M. Wilson, P. K. Wolber, L. Zhang, S. Amur, W. Bao, C. C. Barbacioru, A. B. Lucas, V. Bertholet, C. Boysen, B. Bromley, D. Brown, A. Brunner, R. Canales, X. M. Cao, T. A. Cebula, J. J. Chen, J. Cheng, T. M. Chu, E. Chudin, J. Corson, J. C. Corton, L. J. Croner, C. Davies, T. S. Davison, G. Delenstarr, X. Deng, D. Dorris, A. C. Eklund, X. H. Fan, H. Fang, S. Fulmer-Smentek, J. C. Fuscoe, K. Gallagher, W. Ge, L. Guo, X. Guo, J. Hager, P. K. Haje, J. Han, T. Han, H. C. Harbottle, S. C. Harris, E. Hatchwell, C. A. Hauser, S. Hester, H. Hong, P. Hurban, S. A. Jackson, H. Ji, C. R. Knight, W. P. Kuo, J. E. LeClerc, S. Levy, Q. Z. Li, C. Liu, Y. Liu, M. J. Lombardi, Y. Ma, S. R. Magnuson, B. Maqsodi, T. McDaniel, N. Mei, O. Myklebost, B. Ning, N. Novoradovskaya, M. S. Orr, T. W. Osborn, A. Papallo, T. A. Patterson, R. G. Perkins, E. H. Peters, *et al.* (2006). "The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements". *Nat Biotechnol* 24, pp. 1151–61.

Thelen, M. and J. V. Stein (2008). "How chemokines invite leukocytes to dance". *Nat Immunol* 9, pp. 953–9.

Thomas, D. (2010). "Gene–environment-wide association studies: emerging approaches". *Nat Rev Genet* 11, pp. 259–72.

Thorgeirsson, T. E., D. F. Gudbjartsson, I. Surakka, J. M. Vink, N. Amin, F. Geller, P. Sulem, T. Rafnar, T. Esko, S. Walter, C. Gieger, R. Rawal, M. Mangino, I. Prokopenko, R. Magi, K. Keskitalo, I. H. Gudjonsdottir, S. Gretarsdottir, H. Stefansson, J. R. Thompson, Y. S. Aulchenko, M. Nelis, K. K. Aben, M. den Heijer, A. Dirksen, H. Ashraf, N. Soranzo, A. M. Valdes, C. Steves, A. G. Uitterlinden, A. Hofman, A. Tonjes, P. Kovacs, J. J. Hottenga, G. Willemsen, N. Vogelzangs, A. Doring, N. Dahmen, B. Nitz, M. L. Pergadia, B. Saez, V. De Diego, V. Lezcano, M. D. Garcia-Prats, S. Ripatti, M. Perola, J. Kettunen, A. L. Hartikainen, A. Pouta, J. Laitinen, M. Isohanni, S. Huei-Yi, M. Allen, M. Krestyaninova, A. S. Hall, G. T. Jones, A. M. van Rij, T. Mueller, B. Dieplinger, M. Haltmayer, S. Jonsson, S. E. Matthiasson, H. Oskarsson, T. Tyrfingsson, L. A. Kiemeney, J. I. Mayordomo, J. S. Lindholt, J. H. Pedersen, W. A. Franklin, H. Wolf, G. W. Montgomery, A. C. Heath, N. G. Martin, P. A. Madden, I. Giegling, D. Rujescu, M. R. Jarvelin, V. Salomaa, M. Stumvoll, T. D. Spector, H. E. Wichmann, A. Metspalu, N. J. Samani, B. W. Penninx, B. A. Oostra, D. I. Boomsma, H. Tiemeier, C. M. van Duijn, J. Kaprio, J. R. Gulcher, M. I. McCarthy, L. Peltonen, U. Thorsteinsdottir, and K. Stefansson (2010). "Sequence variants at CHRNB3-CHRNA6 and CYP2A6 affect smoking behavior". *Nat Genet* 42, pp. 448–53.

Thorleifsson, G., G. B. Walters, D. F. Gudbjartsson, V. Steinthorsdottir, P. Sulem, A. Helgadottir, U. Styrkarsdottir, S. Gretarsdottir, S. Thorlacius, I. Jonsdottir, T. Jonsdottir, E. J. Olafsdottir, G. H. Olafsdottir, T. Jonsson, F. Jonsson, K. Borch-Johnsen, T. Hansen, G. Andersen, T. Jorgensen, T. Lauritzen, K. K. Aben, A. L. Verbeek, N. Roeleveld, E. Kampman, L. R. Yanek, L. C. Becker, L. Tryggvadottir, T. Rafnar, D. M. Becker, J. Gulcher, L. A. Kiemeney, O. Pedersen, A. Kong, U. Thorsteinsdottir, and K. Stefansson

(2009). "Genome-wide association yields new sequence variants at seven loci that associate with measures of obesity". *Nat Genet* 41, pp. 18–24.
Thornberry, N. A. and Y. Lazebnik (1998). "Caspases: enemies within". *Science* 281, pp. 1312–6.
Thurman, R. E., E. Rynes, R. Humbert, J. Vierstra, M. T. Maurano, E. Haugen, N. C. Sheffield, A. B. Stergachis, H. Wang, B. Vernot, K. Garg, S. John, R. Sandstrom, D. Bates, L. Boatman, T. K. Canfield, M. Diegel, D. Dunn, A. K. Ebersol, T. Frum, E. Giste, A. K. Johnson, E. M. Johnson, T. Kutyavin, B. Lajoie, B. K. Lee, K. Lee, D. London, D. Lotakis, S. Neph, F. Neri, E. D. Nguyen, H. Qu, A. P. Reynolds, V. Roach, A. Safi, M. E. Sanchez, A. Sanyal, A. Shafer, J. M. Simon, L. Song, S. Vong, M. Weaver, Y. Yan, Z. Zhang, Z. Zhang, B. Lenhard, M. Tewari, M. O. Dorschner, R. S. Hansen, P. A. Navas, G. Stamatoyannopoulos, V. R. Iyer, J. D. Lieb, S. R. Sunyaev, J. M. Akey, P. J. Sabo, R. Kaul, T. S. Furey, J. Dekker, G. E. Crawford, and J. A. Stamatoyannopoulos (2012). "The accessible chromatin landscape of the human genome". *Nature* 489, pp. 75–82.
Thye, T., F. O. Vannberg, S. H. Wong, E. Owusu-Dabo, I. Osei, J. Gyapong, G. Sirugo, F. Sisay-Joof, A. Enimil, M. A. Chinbuah, S. Floyd, D. K. Warndorff, L. Sichali, S. Malema, A. C. Crampin, B. Ngwira, Y. Y. Teo, K. Small, K. Rockett, D. Kwiatkowski, P. E. Fine, P. C. Hill, M. Newport, C. Lienhardt, R. A. Adegbola, T. Corrah, A. Ziegler, A. P. Morris, C. G. Meyer, R. D. Horstmann, and A. V. Hill (2010). "Genome-wide association analyses identifies a susceptibility locus for tuberculosis on chromosome 18q11.2". *Nat Genet* 42, pp. 739–41.
Tobacco, the and Genetics Consortium (2010). "Genome-wide meta-analyses identify multiple loci associated with smoking behavior". *Nat Genet* 42, pp. 441–7.
Todd, J. A., J. I. Bell, and H. O. McDevitt (1987). "HLA-DQ beta gene contributes to susceptibility and resistance to insulin-dependent diabetes mellitus". *Nature* 329, pp. 599–604.
Todd, J. A., N. M. Walker, J. D. Cooper, D. J. Smyth, K. Downes, V. Plagnol, R. Bailey, S. Nejentsev, S. F. Field, F. Payne, C. E. Lowe, J. S. Szeszko, J. P. Hafler, L. Zeitels, J. H. Yang, A. Vella, S. Nutland, H. E. Stevens, H. Schuilenburg, G. Coleman, M. Maisuria, W. Meadows, L. J. Smink, B. Healy, O. S. Burren, A. A. Lam, N. R. Ovington, J. Allen, E. Adlem, H. T. Leung, C. Wallace, J. M. Howson, C. Guja, C. Ionescu-Tirgoviste, M. J. Simmonds, J. M. Heward, S. C. Gough, D. B. Dunger, L. S. Wicker, and D. G. Clayton (2007). "Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes". *Nat Genet* 39, pp. 857–64.
Tompa, M., N. Li, T. L. Bailey, G. M. Church, B. De Moor, E. Eskin, A. V. Favorov, M. C. Frith, Y. Fu, W. J. Kent, V. J. Makeev, A. A. Mironov, W. S. Noble, G. Pavesi, G. Pesole, M. Regnier, N. Simonis, S. Sinha, G. Thijs, J. van Helden, M. Vandenbogaert, Z. Weng, C. Workman, C. Ye, and Z. Zhu (2005). "Assessing computational tools for the discovery of transcription factor binding sites". *Nat Biotechnol* 23, pp. 137–44.
Tournamille, C., Y. Colin, J. P. Cartron, and C. Le Van Kim (1995). "Disruption of a GATA motif in the Duffy gene promoter abolishes erythroid gene expression in Duffy-negative individuals". *Nat Genet* 10, pp. 224–8.
Treisman, R., R. Marais, and J. Wynne (1992). "Spatial flexibility in ternary complexes between SRF and its accessory proteins". *EMBO J* 11, pp. 4631–40.
Trembath, R. C., R. L. Clough, J. L. Rosbotham, A. B. Jones, R. D. Camp, A. Frodsham, J. Browne, R. Barber, J. Terwilliger, G. M. Lathrop, and J. N. Barker (1997). "Identification of a major susceptibility locus on chromosome 6p and evidence for further disease loci revealed by a two stage genome-wide search in psoriasis". *Hum Mol Genet* 6, pp. 813–20.
Trowsdale, J. (2011). "The MHC, disease and selection". *Immunol Lett* 137, pp. 1–8.
Tsuchiya, S., M. Yamabe, Y. Yamaguchi, Y. Kobayashi, T. Konno, and K. Tada (1980). "Establishment and characterization of a human acute monocytic leukemia cell line (THP-1)". *Int J Cancer* 26, pp. 171–6.
Tysk, C., E. Lindberg, G. Jarnerot, and B. Floderus-Myrhed (1988). "Ulcerative colitis and Crohn's disease in an unselected population of monozygotic and dizygotic twins. A study of heritability and the influence of smoking". *Gut* 29, pp. 990–6.

Van Laer, L., E. H. Huizing, M. Verstreken, D. van Zuijlen, J. G. Wauters, P. J. Bossuyt, P. Van de Heyning, W. T. McGuirt, R. J. Smith, P. J. Willems, P. K. Legan, G. P. Richardson, and G. Van Camp (1998). "Nonsyndromic hearing impairment is associated with a mutation in DFNA5". *Nat Genet* 20, pp. 194–7.

Vandiedonck, C., M. S. Taylor, H. E. Lockstone, K. Plant, J. M. Taylor, C. Durrant, J. Broxholme, B. P. Fairfax, and J. C. Knight (2011). "Pervasive haplotypic variation in the spliceo-transcriptome of the human major histocompatibility complex". *Genome Res* 21, pp. 1042–54.

Veer, L. J. van 't, H. Dai, M. J. van de Vijver, Y. D. He, A. A. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards, and S. H. Friend (2002). "Gene expression profiling predicts clinical outcome of breast cancer". *Nature* 415, pp. 530–6.

Venter, J. C., M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, J. D. Gocayne, P. Amanatides, R. M. Ballew, D. H. Huson, J. R. Wortman, Q. Zhang, C. D. Kodira, X. H. Zheng, L. Chen, M. Skupski, G. Subramanian, P. D. Thomas, J. Zhang, G. L. Gabor Miklos, C. Nelson, S. Broder, A. G. Clark, J. Nadeau, V. A. McKusick, N. Zinder, A. J. Levine, R. J. Roberts, M. Simon, C. Slayman, M. Hunkapiller, R. Bolanos, A. Delcher, I. Dew, D. Fasulo, M. Flanigan, L. Florea, A. Halpern, S. Hannenhalli, S. Kravitz, S. Levy, C. Mobarry, K. Reinert, K. Remington, J. Abu-Threideh, E. Beasley, K. Biddick, V. Bonazzi, R. Brandon, M. Cargill, I. Chandramouliswaran, R. Charlab, K. Chaturvedi, Z. Deng, V. Di Francesco, P. Dunn, K. Eilbeck, C. Evangelista, A. E. Gabrielian, W. Gan, W. Ge, F. Gong, Z. Gu, P. Guan, T. J. Heiman, M. E. Higgins, R. R. Ji, Z. Ke, K. A. Ketchum, Z. Lai, Y. Lei, Z. Li, J. Li, Y. Liang, X. Lin, F. Lu, G. V. Merkulov, N. Milshina, H. M. Moore, A. K. Naik, V. A. Narayan, B. Neelam, D. Nusskern, D. B. Rusch, S. Salzberg, W. Shao, B. Shue, J. Sun, Z. Wang, A. Wang, X. Wang, J. Wang, M. Wei, R. Wides, C. Xiao, *et al.* (2001). "The sequence of the human genome". *Science* 291, pp. 1304–51.

Vercelli, D. (2008). "Discovering susceptibility genes for asthma and allergy". *Nat Rev Immunol* 8, pp. 169–82.

Veyrieras, J. B., S. Kudaravalli, S. Y. Kim, E. T. Dermitzakis, Y. Gilad, M. Stephens, and J. K. Pritchard (2008). "High-resolution mapping of expression-QTLs yields insight into human gene regulation". *PLoS Genet* 4, e1000214.

Viglianti, G. A., C. M. Lau, T. M. Hanley, B. A. Miko, M. J. Shlomchik, and A. Marshak-Rothstein (2003). "Activation of autoreactive B cells by CpG dsDNA". *Immunity* 19, pp. 837–47.

Visel, A., S. Prabhakar, J. A. Akiyama, M. Shoukry, K. D. Lewis, A. Holt, I. Plajzer-Frick, V. Afzal, E. M. Rubin, and L. A. Pennacchio (2008). "Ultraconservation identifies a small subset of extremely constrained developmental enhancers". *Nat Genet* 40, pp. 158–60.

Volinia, S., G. A. Calin, C. G. Liu, S. Ambs, A. Cimmino, F. Petrocca, R. Visone, M. Iorio, C. Roldo, M. Ferracin, R. L. Prueitt, N. Yanaihara, G. Lanza, A. Scarpa, A. Vecchione, M. Negrini, C. C. Harris, and C. M. Croce (2006). "A microRNA expression signature of human solid tumors defines cancer gene targets". *Proc Natl Acad Sci U S A* 103, pp. 2257–61.

Waddington, C. H. (1942). "Canalization of development and the inheritance of acquired characters". *Nature* 3811, pp. 563–565.

Walker, J. A., J. L. Barlow, and A. N. McKenzie (2013). "Innate lymphoid cells–how did we miss them?" *Nat Rev Immunol* 13, pp. 75–87.

Walsh, E. C., K. A. Mather, S. F. Schaffner, L. Farwell, M. J. Daly, N. Patterson, M. Cullen, M. Carrington, T. L. Bugawan, H. Erlich, J. Campbell, J. Barrett, K. Miller, G. Thomson, E. S. Lander, and J. D. Rioux (2003). "An integrated haplotype map of the human major histocompatibility complex". *Am J Hum Genet* 73, pp. 580–90.

Wan, Y. I., N. R. Shrine, M. Soler Artigas, L. V. Wain, J. D. Blakey, M. F. Moffatt, A. Bush, K. F. Chung, W. O. Cookson, D. P. Strachan, L. Heaney, B. A. Al-Momani, A. H. Mansur, S. Manney, N. C. Thomson, R. Chaudhuri, C. E. Brightling, M. Bafadhel, A. Singapuri, R. Niven, A. Simpson, J. W. Holloway, P. H. Howarth, J. Hui, A. W. Musk, A. L. James, M. A. Brown, S. Baltic, M. A. Ferreira, P. J. Thompson, M. D. Tobin, I. Sayers, and

I. P. Hall (2012). "Genome-wide association study to identify genetic determinants of severe asthma". *Thorax* 67, pp. 762–8.

Wang, E. T., R. Sandberg, S. Luo, I. Khrebtukova, L. Zhang, C. Mayr, S. F. Kingsmore, G. P. Schroth, and C. B. Burge (2008). "Alternative isoform regulation in human tissue transcriptomes". *Nature* 456, pp. 470–6.

Wang, H., L. Yang, M. S. Jamaluddin, and D. D. Boyd (2004). "The Kruppel-like KLF4 transcription factor, a novel regulator of urokinase receptor expression, drives synthesis of this binding site in colonic crypt luminal surface epithelial cells". *J Biol Chem* 279, pp. 22674–83.

Wasserman, W. W. and A. Sandelin (2004). "Applied bioinformatics for the identification of regulatory elements". *Nat Rev Genet* 5, pp. 276–87.

Watkins, H. and M. Farrall (2006). "Genetic susceptibility to coronary artery disease: from promise to progress". *Nat Rev Genet* 7, pp. 163–73.

Watson, J. D. and F. H. Crick (1953). "Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid". *Nature* 171, pp. 737–8.

Weiss, L. A., Y. Shen, J. M. Korn, D. E. Arking, D. T. Miller, R. Fossdal, E. Saemundsen, H. Stefansson, M. A. Ferreira, T. Green, O. S. Platt, D. M. Ruderfer, C. A. Walsh, D. Altshuler, A. Chakravarti, R. E. Tanzi, K. Stefansson, S. L. Santangelo, J. F. Gusella, P. Sklar, B. L. Wu, and M. J. Daly (2008). "Association between microdeletion and microduplication at 16p11.2 and autism". *N Engl J Med* 358, pp. 667–75.

Weiss, S. T., B. A. Raby, and A. Rogers (2009). "Asthma genetics and genomics 2009". *Curr Opin Genet Dev* 19, pp. 279–82.

Wellcome Trust Case Control Consortium, the (2007). "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls". *Nature* 447, pp. 661–78.

Wenger, R. H., D. P. Stiehl, and G. Camenisch (2005). "Integration of oxygen signaling at the consensus HRE". *Sci STKE* 2005, re12.

Westra, H. J., M. J. Peters, T. Esko, H. Yaghootkar, C. Schurmann, J. Kettunen, M. W. Christiansen, B. P. Fairfax, K. Schramm, J. E. Powell, A. Zhernakova, D. V. Zhernakova, J. H. Veldink, L. H. Van den Berg, J. Karjalainen, S. Withoff, A. G. Uitterlinden, A. Hofman, F. Rivadeneira, P. A. t Hoen, E. Reinmaa, K. Fischer, M. Nelis, L. Milani, D. Melzer, L. Ferrucci, A. B. Singleton, D. G. Hernandez, M. A. Nalls, G. Homuth, M. Nauck, D. Radke, U. Volker, M. Perola, V. Salomaa, J. Brody, A. Suchy-Dicey, S. A. Gharib, D. A. Enquobahrie, T. Lumley, G. W. Montgomery, S. Makino, H. Prokisch, C. Herder, M. Roden, H. Grallert, T. Meitinger, K. Strauch, Y. Li, R. C. Jansen, P. M. Visscher, J. C. Knight, B. M. Psaty, S. Ripatti, A. Teumer, T. M. Frayling, A. Metspalu, J. B. van Meurs, and L. Franke (2013). "Systematic identification of trans eQTLs as putative drivers of known disease associations". *Nat Genet* 45, pp. 1238–43.

Wet, J. R. de, K. V. Wood, M. DeLuca, D. R. Helinski, and S. Subramani (1987). "Firefly luciferase gene: structure and expression in mammalian cells". *Mol Cell Biol* 7, pp. 725–37.

Wigginton, J. E., D. J. Cutler, and G. R. Abecasis (2005). "A note on exact tests of Hardy-Weinberg equilibrium". *Am J Hum Genet* 76, pp. 887–93.

Wild, P. S., T. Zeller, A. Schillert, S. Szymczak, C. R. Sinning, A. Deiseroth, R. B. Schnabel, E. Lubos, T. Keller, M. S. Eleftheriadis, C. Bickel, H. J. Rupprecht, S. Wilde, H. Rossmann, P. Diemert, L. A. Cupples, C. Perret, J. Erdmann, K. Stark, M. E. Kleber, S. E. Epstein, B. F. Voight, K. Kuulasmaa, M. Li, A. S. Schafer, N. Klopp, P. S. Braund, H. B. Sager, S. Demissie, C. Proust, I. R. Konig, H. E. Wichmann, W. Reinhard, M. M. Hoffmann, J. Virtamo, M. S. Burnett, D. Siscovick, P. G. Wiklund, L. Qu, N. E. El Mokthari, J. R. Thompson, A. Peters, A. V. Smith, E. Yon, J. Baumert, C. Hengstenberg, W. Marz, P. Amouyel, J. Devaney, S. M. Schwartz, O. Saarela, N. N. Mehta, D. Rubin, K. Silander, A. S. Hall, J. Ferrieres, T. B. Harris, O. Melander, F. Kee, H. Hakonarson, J. Schrezenmeir, V. Gudnason, R. Elosua, D. Arveiler, A. Evans, D. J. Rader, T. Illig, S. Schreiber, J. C. Bis, D. Altshuler, M. Kavousi, J. C. Witteman, A. G. Uitterlinden, A. Hofman, A. R. Folsom, M. Barbalic, E. Boerwinkle, S. Kathiresan, M. P. Reilly, C. J. O'Donnell, N. J. Samani, H. Schunkert, F. Cambien, K. J. Lackner, L. Tiret, V. Salomaa, T. Munzel, A. Ziegler, and S. Blankenberg (2011). "A genome-

wide association study identifies LIPA as a susceptibility gene for coronary artery disease". *Circ Cardiovasc Genet* 4, pp. 403–12.

Willer, C. J., S. Sanna, A. U. Jackson, A. Scuteri, L. L. Bonnycastle, R. Clarke, S. C. Heath, N. J. Timpson, S. S. Najjar, H. M. Stringham, J. Strait, W. L. Duren, A. Maschio, F. Busonero, A. Mulas, G. Albai, A. J. Swift, M. A. Morken, N. Narisu, D. Bennett, S. Parish, H. Shen, P. Galan, P. Meneton, S. Hercberg, D. Zelenika, W. M. Chen, Y. Li, L. J. Scott, P. A. Scheet, J. Sundvall, R. M. Watanabe, R. Nagaraja, S. Ebrahim, D. A. Lawlor, Y. Ben-Shlomo, G. Davey-Smith, A. R. Shuldiner, R. Collins, R. N. Bergman, M. Uda, J. Tuomilehto, A. Cao, F. S. Collins, E. Lakatta, G. M. Lathrop, M. Boehnke, D. Schlessinger, K. L. Mohlke, and G. R. Abecasis (2008). "Newly identified loci that influence lipid concentrations and risk of coronary artery disease". *Nat Genet* 40, pp. 161–9.

Willer, C. J., E. K. Speliotes, R. J. Loos, S. Li, C. M. Lindgren, I. M. Heid, S. I. Berndt, A. L. Elliott, A. U. Jackson, C. Lamina, G. Lettre, N. Lim, H. N. Lyon, S. A. McCarroll, K. Papadakis, L. Qi, J. C. Randall, R. M. Roccasecca, S. Sanna, P. Scheet, M. N. Weedon, E. Wheeler, J. H. Zhao, L. C. Jacobs, I. Prokopenko, N. Soranzo, T. Tanaka, N. J. Timpson, P. Almgren, A. Bennett, R. N. Bergman, S. A. Bingham, L. L. Bonnycastle, M. Brown, N. P. Burtt, P. Chines, L. Coin, F. S. Collins, J. M. Connell, C. Cooper, G. D. Smith, E. M. Dennison, P. Deodhar, P. Elliott, M. R. Erdos, K. Estrada, D. M. Evans, L. Gianniny, C. Gieger, C. J. Gillson, C. Guiducci, R. Hackett, D. Hadley, A. S. Hall, A. S. Havulinna, J. Hebebrand, A. Hofman, B. Isomaa, K. B. Jacobs, T. Johnson, P. Jousilahti, Z. Jovanovic, K. T. Khaw, P. Kraft, M. Kuokkanen, J. Kuusisto, J. Laitinen, E. G. Lakatta, J. Luan, R. N. Luben, M. Mangino, W. L. McArdle, T. Meitinger, A. Mulas, P. B. Munroe, N. Narisu, A. R. Ness, K. Northstone, S. O'Rahilly, C. Purmann, M. G. Rees, M. Ridderstrale, S. M. Ring, F. Rivadeneira, A. Ruokonen, M. S. Sandhu, J. Saramies, L. J. Scott, A. Scuteri, K. Silander, M. A. Sims, K. Song, J. Stephens, S. Stevens, H. M. Stringham, Y. C. Tung, T. T. Valle, C. M. Van Duijn, K. S. Vimaleswaran, *et al.* (2009). "Six new loci associated with body mass index highlight a neuronal influence on body weight regulation". *Nat Genet* 41, pp. 25–34.

Winter, J., S. Jung, S. Keller, R. I. Gregory, and S. Diederichs (2009). "Many roads to maturity: microRNA biogenesis pathways and their regulation". *Nat Cell Biol* 11, pp. 228–34.

Woo, A. J., J. S. Dods, E. Susanto, D. Ulgiati, and L. J. Abraham (2002). "A proteomics approach for the identification of DNA binding activities observed in the electrophoretic mobility shift assay". *Mol Cell Proteomics* 1, pp. 472–8.

Woods, D. B., J. Ghysdael, and M. J. Owen (1992). "Identification of nucleotide preferences in DNA sequences recognised specifically by c-Ets-1 protein". *Nucleic Acids Res* 20, pp. 699–704.

Wordsworth, B. P., J. S. Lanchbury, L. I. Sakkas, K. I. Welsh, G. S. Panayi, and J. I. Bell (1989). "HLA-DR4 subtype frequencies in rheumatoid arthritis indicate that DRB1 is the major susceptibility locus within the HLA class II region". *Proc Natl Acad Sci U S A* 86, pp. 10049–53.

Wright, M. D., G. W. Moseley, and A. B. van Spriel (2004). "Tetraspanin microdomains in immune cell signalling and malignant disease". *Tissue Antigens* 64, pp. 533–42.

Wright, S. D., R. A. Ramos, P. S. Tobias, R. J. Ulevitch, and J. C. Mathison (1990). "CD14, a receptor for complexes of lipopolysaccharide (LPS) and LPS binding protein". *Science* 249, pp. 1431–3.

Xie, W., C. L. Barr, A. Kim, F. Yue, A. Y. Lee, J. Eubanks, E. L. Dempster, and B. Ren (2012). "Base-resolution analyses of sequence and parent-of-origin dependent DNA methylation in the mouse genome". *Cell* 148, pp. 816–31.

Xu, H., A. Manivannan, I. Crane, R. Dawson, and J. Liversidge (2008). "Critical but divergent roles for CD62L and CD44 in directing blood monocyte trafficking in vivo during inflammation". *Blood* 112, pp. 1166–74.

Xu, J., Z. Mo, D. Ye, M. Wang, F. Liu, G. Jin, C. Xu, X. Wang, Q. Shao, Z. Chen, Z. Tao, J. Qi, F. Zhou, Z. Wang, Y. Fu, D. He, Q. Wei, J. Guo, D. Wu, X. Gao, J. Yuan, G. Wang, Y. Xu, G. Wang, H. Yao, P. Dong, Y. Jiao, M. Shen, J. Yang, J. Ou-Yang, H. Jiang, Y. Zhu, S. Ren, Z. Zhang, C. Yin, X. Gao, B. Dai, Z. Hu, Y. Yang, Q. Wu, H. Chen, P. Peng,

Y. Zheng, X. Zheng, Y. Xiang, J. Long, J. Gong, R. Na, X. Lin, H. Yu, Z. Wang, S. Tao, J. Feng, J. Sun, W. Liu, A. Hsing, J. Rao, Q. Ding, F. Wiklund, H. Gronberg, X. O. Shu, W. Zheng, H. Shen, L. Jin, R. Shi, D. Lu, X. Zhang, J. Sun, S. L. Zheng, and Y. Sun (2012). "Genome-wide association study in Chinese men identifies two new prostate cancer risk loci at 9q31.2 and 19q13.4". *Nat Genet* 44, pp. 1231–5.

Yamamoto, K., F. W. Quelle, W. E. Thierfelder, B. L. Kreider, D. J. Gilbert, N. A. Jenkins, N. G. Copeland, O. Silvennoinen, and J. N. Ihle (1994). "Stat4, a novel gamma interferon activation site-binding protein expressed in early myeloid differentiation". *Mol Cell Biol* 14, pp. 4342–9.

Yamazaki, K., J. Umeno, A. Takahashi, A. Hirano, T. A. Johnson, N. Kumasaka, T. Morizono, N. Hosono, T. Kawaguchi, M. Takazoe, T. Yamada, Y. Suzuki, H. Tanaka, S. Motoya, M. Hosokawa, Y. Arimura, Y. Shinomura, T. Matsui, T. Matsumoto, M. Iida, T. Tsunoda, Y. Nakamura, N. Kamatani, and M. Kubo (2013). "A genome-wide association study identifies 2 susceptibility Loci for Crohn's disease in a Japanese population". *Gastroenterology* 144, pp. 781–8.

Yang, W., H. Tang, Y. Zhang, X. Tang, J. Zhang, L. Sun, J. Yang, Y. Cui, L. Zhang, N. Hirankarn, H. Cheng, H. F. Pan, J. Gao, T. L. Lee, Y. Sheng, C. S. Lau, Y. Li, T. M. Chan, X. Yin, D. Ying, Q. Lu, A. M. Leung, X. Zuo, X. Chen, K. L. Tong, F. Zhou, Q. Diao, N. K. Tse, H. Xie, C. C. Mok, F. Hao, S. N. Wong, B. Shi, K. W. Lee, Y. Hui, M. H. Ho, B. Liang, P. P. Lee, H. Cui, Q. Guo, B. H. Chung, X. Pu, Q. Liu, X. Zhang, C. Zhang, C. Y. Chong, H. Fang, R. W. Wong, Y. Sun, M. Y. Mok, X. P. Li, Y. Avihingsanon, Z. Zhai, P. Rianthavorn, T. Deekajorndej, K. Suphapeetiporn, F. Gao, V. Shotelersuk, X. Kang, S. K. Ying, L. Zhang, W. H. Wong, D. Zhu, S. K. Fung, F. Zeng, W. M. Lai, C. M. Wong, I. O. Ng, M. M. Garcia-Barcelo, S. S. Cherny, N. Shen, P. K. Tam, P. C. Sham, D. Q. Ye, S. Yang, X. Zhang, and Y. L. Lau (2013). "Meta-analysis followed by replication identifies loci in or near CDKN1B, TET3, CD80, DRAM1, and ARID5B as associated with systemic lupus erythematosus in Asians". *Am J Hum Genet* 92, pp. 41–51.

Yao, T. P., S. P. Oh, M. Fuchs, N. D. Zhou, L. E. Ch'ng, D. Newsome, R. T. Bronson, E. Li, D. M. Livingston, and R. Eckner (1998). "Gene dosage-dependent embryonic development and proliferation defects in mice lacking the transcriptional integrator p300". *Cell* 93, pp. 361–72.

Zeggini, E., M. N. Weedon, C. M. Lindgren, T. M. Frayling, K. S. Elliott, H. Lango, N. J. Timpson, J. R. Perry, N. W. Rayner, R. M. Freathy, J. C. Barrett, B. Shields, A. P. Morris, S. Ellard, C. J. Groves, L. W. Harries, J. L. Marchini, K. R. Owen, B. Knight, L. R. Cardon, M. Walker, G. A. Hitman, A. D. Morris, A. S. Doney, M. I. McCarthy, and A. T. Hattersley (2007). "Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes". *Science* 316, pp. 1336–41.

Zeggini, E., L. J. Scott, R. Saxena, B. F. Voight, J. L. Marchini, T. Hu, P. I. de Bakker, G. R. Abecasis, P. Almgren, G. Andersen, K. Ardlie, K. B. Bostrom, R. N. Bergman, L. L. Bonnycastle, K. Borch-Johnsen, N. P. Burtt, H. Chen, P. S. Chines, M. J. Daly, P. Deodhar, C. J. Ding, A. S. Doney, W. L. Duren, K. S. Elliott, M. R. Erdos, T. M. Frayling, R. M. Freathy, L. Gianniny, H. Grallert, N. Grarup, C. J. Groves, C. Guiducci, T. Hansen, C. Herder, G. A. Hitman, T. E. Hughes, B. Isomaa, A. U. Jackson, T. Jorgensen, A. Kong, K. Kubalanza, F. G. Kuruvilla, J. Kuusisto, C. Langenberg, H. Lango, T. Lauritzen, Y. Li, C. M. Lindgren, V. Lyssenko, A. F. Marvelle, C. Meisinger, K. Midthjell, K. L. Mohlke, M. A. Morken, A. D. Morris, N. Narisu, P. Nilsson, K. R. Owen, C. N. Palmer, F. Payne, J. R. Perry, E. Pettersen, C. Platou, I. Prokopenko, L. Qi, L. Qin, N. W. Rayner, M. Rees, J. J. Roix, A. Sandbaek, B. Shields, M. Sjogren, V. Steinthorsdottir, H. M. Stringham, A. J. Swift, G. Thorleifsson, U. Thorsteinsdottir, N. J. Timpson, T. Tuomi, J. Tuomilehto, M. Walker, R. M. Watanabe, M. N. Weedon, C. J. Willer, T. Illig, K. Hveem, F. B. Hu, M. Laakso, K. Stefansson, O. Pedersen, N. J. Wareham, I. Barroso, A. T. Hattersley, F. S. Collins, L. Groop, M. I. McCarthy, M. Boehnke, and D. Altshuler (2008). "Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes". *Nat Genet* 40, pp. 638–45.

Zeller, T., P. Wild, S. Szymczak, M. Rotival, A. Schillert, R. Castagne, S. Maouche, M. Germain, K. Lackner, H. Rossmann, M. Eleftheriadis, C. R. Sinning, R. B. Schnabel,

E. Lubos, D. Mennerich, W. Rust, C. Perret, C. Proust, V. Nicaud, J. Loscalzo, N. Hubner, D. Tregouet, T. Munzel, A. Ziegler, L. Tiret, S. Blankenberg, and F. Cambien (2010). "Genetics and beyond–the transcriptome of human monocytes and disease susceptibility". *PLoS One* 5, e10693.

Zhang, D., L. Cheng, J. A. Badner, C. Chen, Q. Chen, W. Luo, D. W. Craig, M. Redman, E. S. Gershon, and C. Liu (2010). "Genetic control of individual differences in gene-specific methylation in human brain". *Am J Hum Genet* 86, pp. 411–9.

Zhang, F. R., W. Huang, S. M. Chen, L. D. Sun, H. Liu, Y. Li, Y. Cui, X. X. Yan, H. T. Yang, R. D. Yang, T. S. Chu, C. Zhang, L. Zhang, J. W. Han, G. Q. Yu, C. Quan, Y. X. Yu, Z. Zhang, B. Q. Shi, L. H. Zhang, H. Cheng, C. Y. Wang, Y. Lin, H. F. Zheng, X. A. Fu, X. B. Zuo, Q. Wang, H. Long, Y. P. Sun, Y. L. Cheng, H. Q. Tian, F. S. Zhou, H. X. Liu, W. S. Lu, S. M. He, W. L. Du, M. Shen, Q. Y. Jin, Y. Wang, H. Q. Low, T. Erwin, N. H. Yang, J. Y. Li, X. Zhao, Y. L. Jiao, L. G. Mao, G. Yin, Z. X. Jiang, X. D. Wang, J. P. Yu, Z. H. Hu, C. H. Gong, Y. Q. Liu, R. Y. Liu, D. M. Wang, D. Wei, J. X. Liu, W. K. Cao, H. Z. Cao, Y. P. Li, W. G. Yan, S. Y. Wei, K. J. Wang, M. L. Hibberd, S. Yang, X. J. Zhang, and J. J. Liu (2009). "Genomewide association study of leprosy". *N Engl J Med* 361, pp. 2609–18.

Zheng, W., H. Zhao, E. Mancera, L. M. Steinmetz, and M. Snyder (2010). "Genetic analysis of variation in transcription factor binding in yeast". *Nature* 464, pp. 1187–91.

Zhong, H., X. Yang, L. M. Kaplan, C. Molony, and E. E. Schadt (2010). "Integrating pathway analysis and genetics of gene expression for genome-wide association studies". *Am J Hum Genet* 86, pp. 581–91.

Ziegler-Heitbrock, H. W., G. Fingerle, M. Strobel, W. Schraut, F. Stelter, C. Schutt, B. Passlick, and A. Pforte (1993). "The novel subset of CD14+/CD16+ blood monocytes exhibits features of tissue macrophages". *Eur J Immunol* 23, pp. 2053–8.