Multi-omics integration to characterise mechanisms of molecular QTL from a sepsis cohort

Wellcome Sanger Institute

University of Cambridge
Churchill College

This thesis is submitted for the degree of Master of Philosophy

Nikhil Milind

August 2022

Abstract

Sepsis is a potentially lethal maladaptive host immune response to infection characterised by organ dysfunction. I used data from the UK Genomic Advances in Sepsis (GAinS) study to better understand the molecular mechanisms underlying heterogeneity in individual host immune response.

Weighted co-expression network analysis was used to decompose the transcriptome into modules. These modules identified pathways of pathological relevance to sepsis, including cell-type-specific modules associated with myeloid and lymphoid cells. The modules were used to perform module quantitative trait locus (QTL) mapping to identify genetic variants associated with variation in gene expression. These QTL, in addition to previously mapped *cis*-expression QTL (eQTL) and protein QTL (pQTL), were integrated and interpreted using Bayesian colocalisation and fine mapping methods. Finally, multiple functional enrichment methods integrating publicly available data sets were explored to predict the impact of QTL in various tissues and contexts.

These analyses provide biological insights into the genetic underpinnings of sepsis. In addition, the data generated from these analyses will be a useful resource for investigators exploring specific variants or sources of molecular heterogeneity in sepsis.

Declaration

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text. In accordance with the Statutes and Ordinances of the University of Cambridge, I declare that this thesis is not substantially the same as any work that I have submitted for a degree or diploma or similar qualification. I declare that my thesis does not exceed the word limit prescribed by the Biological Sciences Degree Committee. This thesis consists of 19,979 words, exclusive of tables, footnotes, bibliography, and appendices.

Collaboration

This thesis consists of analyses conducted by myself and groups involved in the Genomic Advances in Sepsis (GAinS) study. Investigators relevant to this thesis were present at either the Wellcome Sanger Institute (WSI) or the Wellcome Centre for Human Genetics (WHG) at the University of Oxford. Imputation of genotyping data, processing of RNA sequencing data, generation of gene expression data, and identification of expression quantitative trait loci (eQTL) was conducted by Katie Burnham and Wanseon Lee at the WSI. The processing of proteomics data, generation of protein expression data, and mapping of protein quantitative trait loci (pQTL) was conducted by Yuxin Mi from the WHG. Eddie Cano Gámez from the WHG developed the quantitative Sepsis Response Signature (SRSq) score and provided the processed microarray gene expression data. Probabilistic estimation of expression residuals (PEER) factors for the microarray gene expression data were computed by Katie Burnham. Andrew Kwok from the WHG processed the single-cell RNA sequencing data from the Sepsis Immunomics study and ran CIBERSORTx to estimate cell frequencies for patients with bulk RNA sequencing data. Other than these collaborative elements, all analyses presented in this thesis represent my own work.

Acknowledgements

I would like to thank Dr. Emma Davenport and her team at the Sanger Institute for allowing me to experience a year of exciting science. I want to especially thank Dr. Katie Burnham and Dr. Wanseon Lee for shaping my project and their willingness to provide feedback whenever I needed it. In addition, I would like to thank all the members of the research group for providing a friendly and intense intellectual environment that made every conversation exciting and every day something I looked forward to.

My path to the United Kingdom and beyond has been facilitated by the unconditional support of many people in my life. Thank you to my family - my parents and my sister - for their endless support. I am grateful to Natalie for her constant companionship. Thank you to all the friends that have made this year memorable. I would also be remiss not to mention some of the mentors that have enabled my passion for science, including Dr. David Aylor, Dr. Gregory Carter, Dr. Christoph Preuss, and Dr. Kyathanahalli Janardhan.

Finally, thank you to the Winston Churchill Foundation for funding this opportunity and to the University of Cambridge for providing a unique atmosphere for learning.

Contents

List of Figures **8**

List of Tables **10**

List of Acronyms **12**

1 Introduction **14**
1.1 Genetics of Complex Diseases . . . 14
1.1.1 Heritability . . . 16
1.1.2 Genome-Wide Association Studies . . . 16
1.1.3 Linkage Disequilibrium . . . 17
1.2 Multi-omics . . . 17
1.2.1 The Transcriptome and Proteome . . . 18
1.2.2 Regulation of Molecular Expression . . . 18
1.2.3 The Epigenome . . . 19
1.3 Functional Genomics . . . 19
1.3.1 Co-Expression Networks . . . 19
1.3.2 Molecular Quantitative Trait Loci . . . 20
1.3.3 Chromatin Accessibility . . . 21
1.4 Colocalisation and Fine Mapping . . . 21
1.4.1 Bayesian Fine Mapping . . . 21
1.4.2 Colocalisation . . . 23
1.5 Sepsis . . . 24
1.5.1 Immunological Response during Sepsis . . . 25
1.5.2 Sepsis Genomics . . . 29
1.5.3 Summary of *cis*-eQTL and pQTL . . . 31
1.6 Specific Aims . . . 32

2 Methods **34**
2.1 Description of Cohort . . . 34
2.2 Analysis of Gene Expression . . . 35
2.2.1 Weighted Network Correlation Analysis . . . 35
2.2.2 Module Annotation . . . 35
2.2.3 Module Association with Clinical Endophenotypes . . . 36
2.3 Molecular QTL . . . 37
2.3.1 Mapping of Module QTL . . . 38
2.3.2 Module QTL Replication . . . 38
2.4 Colocalisation . . . 39
2.5 Fine Mapping . . . 39
2.6 Publicly Available Data . . . 40
2.6.1 ATAC-seq Alignment . . . 40

2.6.2 ATAC-seq Sample Quality . . . 40
2.6.3 ATAC-seq Peaks . . . 40
2.6.4 Peak Annotation and Motif Enrichment . . . 41
2.7 Functional Interpretation . . . 42
2.7.1 Enrichment of eQTL in Functional Categories . . . 42
2.7.2 Partitioned Heritability . . . 43
2.7.3 Variant Effect Prediction . . . 44
2.8 Statistical Analysis . . . 44

3 Gene Co-Expression 45
3.1 Co-Expression Modules . . . 45
3.1.1 Signatures of Leukocytes . . . 46
3.1.2 Association with Endophenotypes . . . 53
3.1.3 Module Networks . . . 56
3.2 Module QTL . . . 57
3.2.1 Multiple Module Eigengenes . . . 59
3.2.2 Trait-Associated Variants . . . 61
3.2.3 Module QTL Replication . . . 63
3.3 Discussion . . . 65
3.3.1 Co-expression Modules . . . 65
3.3.2 Relationships between Modules and Clinical Variables . . . 66
3.3.3 Module QTL . . . 66

4 Colocalisation and Fine Mapping 68
4.1 Colocalisation of *cis*-eQTL . . . 68
4.2 Colocalisation of *cis*-eQTL and module QTL . . . 70
4.3 Colocalisation of *cis*-eQTL and *cis*-pQTL . . . 71
4.4 Colocalisation of *trans*-pQTL . . . 74
4.5 Colocalisation with GWAS Associations . . . 77
4.6 Statistical Fine Mapping . . . 78
4.6.1 Conditional *cis*-eQTL . . . 78
4.6.2 Module QTL . . . 80
4.6.3 pQTL . . . 81
4.7 Discussion . . . 83
4.7.1 Colocalisation of QTL across Omics Layers . . . 83
4.7.2 Colocalisation and Fine Mapping Methods . . . 84

5 Dysregulated Immune Cell Types 86
5.1 Reprocessing ATAC-seq Data . . . 86
5.2 Consensus and Cell-Type-Specific Peaks . . . 87
5.3 Enrichment of *cis*-eQTL in Genomic Annotation . . . 88
5.4 Partitioned Heritability . . . 94
5.5 Variant Effect Prediction . . . 96
5.6 Integration . . . 98
5.6.1 Module 92 . . . 98
5.6.2 Module 101 . . . 99
5.7 Discussion . . . 100
5.7.1 Enrichment of *cis*-eQTL . . . 100
5.7.2 Partitioned Heritability . . . 101
5.7.3 Variant Effect Prediction . . . 102
5.7.4 Integration . . . 102
5.7.5 Concluding Remarks . . . 102

Bibliography **102**

A Prior Work in GAinS **115**
A.1 Genotyping . . . 115
A.2 Genotype Imputation . . . 115
A.3 RNA Sequencing . . . 116
A.4 Microarray Gene Expression . . . 116
A.5 Mass Spectrometry . . . 116
A.6 Mapping of eQTL . . . 117
A.7 Mapping of pQTL . . . 118

B Summary Statistics **119**

C Publicly Available ATAC-seq Data **126**

D ATAC-seq Reprocessing **129**
D.1 Quality Control . . . 129
D.2 Comparison with Original Study . . . 129
D.3 Peak Sets . . . 132

E Roadmap Project Epigenomes **134**

F Partitioned Heritability **136**

G Variant Effect Prediction **138**

List of Figures

1.1 Sepsis as a complex disease . . . 15
1.2 Dysregulated response to infection during sepsis . . . 25
1.3 Coagulation and complement systems during sepsis . . . 26
1.4 Effects of sepsis on leukocyte phenotypes . . . 28
1.5 Mapping of *cis*-eQTL and pQTL . . . 32

3.1 Distribution of module sizes . . . 46
3.2 Cell-type-specific enrichment of modules . . . 47
3.3 Cell marker enrichment of modules . . . 48
3.4 Neutrophil subsets . . . 48
3.5 Associations between module eigengenes and clinical endophenotypes . . . 53
3.6 Association of module eigengenes and inferred cell frequencies . . . 55
3.7 Module 51 HIF-1 pathway . . . 56
3.8 Module 92 . . . 57
3.9 Module QTL from module eigengenes . . . 58
3.10 Composition of module QTL . . . 59
3.11 Module QTL from top five module eigengenes . . . 60
3.12 Composition of module QTL from the top five module eigengenes . . . 61
3.13 Replication of module eigengenes . . . 63
3.14 Forest plot of replicated effects . . . 64

4.1 Number of eGenes sharing lead conditional *cis*-eQTL . . . 69
4.2 Colocalising *cis*-eQTL of components of the TCR β chain . . . 70
4.3 Distribution of *cis*-eQTL colocalising with a module QTL . . . 71
4.4 FCGR3B locus *cis*-eQTL and *cis*-pQTL . . . 73
4.5 ORM2 locus *cis*-eQTL and *cis*-pQTL . . . 74
4.6 Chromosome 16 *trans*-pQTL . . . 75
4.7 Chromosome 14 *trans*-pQTL . . . 76
4.8 Credible set sizes . . . 79
4.9 Number of signals for module QTL . . . 80
4.10 Module QTL credible set sizes . . . 81
4.11 Number of signals for *cis*-pQTL . . . 81
4.12 *Cis*-pQTL credible set sizes . . . 82

5.1 Motif enrichment in group peak sets . . . 87
5.2 HOMER consensus peaks annotation . . . 88
5.3 Enrichment in ENCODE cCREs . . . 89
5.4 Enrichment in immune atlas peaks . . . 90
5.5 Enrichment in neutrophil atlas peaks . . . 91
5.6 CHEERS enrichment of *cis*-eQTL . . . 92
5.7 GoShifter overlap score matrix . . . 93

5.8 Partitioned heritability 95
5.9 VEP gene consequences 96
5.10 VEP change in motif score 97
5.11 *NLRC5* GoShifter overlap scores 99
5.12 Module 92 eigengene heritability 99
5.13 *RPS26* GoShifter overlap scores 100

D.1 TSS enrichment scores 129
D.2 Distribution of peak widths 130
D.3 Distribution of peaks across the genome 131
D.4 Correlation of read counts between peaks 131
D.5 Group peak sets from immune atlas 132
D.6 Cell type peak sets from immune atlas 133
D.7 Group peak sets from neutrophil atlas 133

E.1 Enrichment in ChromHMM states 135

G.1 VEP regulatory consequences 139
G.2 VEP module QTL gene consequences 139
G.3 VEP module QTL regulatory consequences 140

List of Tables

3.1 Key genes in modules 50
3.2 IMD-relevant traits in the EBI GWAS Catalog 62

4.1 Colocalisation of *cis*-eQTL with *cis*-pQTL 72
4.2 Proteins with *cis*-pQTL 72
4.3 Colocalisation of module QTL with GWAS variants 78

B.1 Summary statistics from GWAS analyses 119
B.2 EBI GWAS and module QTL overlap studies 119

C.1 Samples in immune atlas 127
C.2 Samples in neutrophil atlas 128

E.1 Roadmap Project epigenomes 134

G.1 VEP module QTL motifs 138

List of Acronyms

1000G	1000 Genomes
ANOVA	analysis of variance
APC	antigen-presenting cell
ATAC-seq	assay for transposase-accessible chromatin using sequencing
CAP	community-acquired pneumonia
cCRE	candidate cis-regulatory element
CD-CV	common disease-common variant
cDNA	complementary DNA
ChIP-seq	chromatin immunoprecipitation sequencing
CS	credible set
CTCF-only	(ENCODE cCRE Type) not TSS-overlapping and with high DNase and CTCF signals only
DAMP	damage-associated molecular pattern
DE	differentially expressed
dELS	(ENCODE cCRE Type) TSS-distal enhancer-like signature
DNA	deoxyribonucleic acid
DNase-H3K4me3	(ENCODE cCRE Type) not TSS-overlapping and with high DNase and H3K4me3 signals only
DNase-seq	DNase I hypersensitivity sites sequencing
EBI	European Bioinformatics Institute
eGene	gene with expression quantitative trait loci
ENCODE	Encyclopedia of DNA Elements
eQTL	expression quantitative trait locus
FDR	false discovery rate
FIMO	find individual motif occurrences
FP	faecal peritonitis
GAinS	Genomic Advances in Sepsis
GEO	Gene Expression Omnibus
GES	generalised eta squared
GRCh38	Genome Reference Consortium human build 38
GRM	genetic relationship matrix
GWAS	genome-wide association study
hg19	human genome build 19
HGI	Human Genetics Informatics
HLA	human leukocyte antigen
HRC	Haplotype Reference Consortium
IBD	identical by descent
IBSS	iterative Bayesian stepwise selection
ICU	intensive care unit
IFITM	interferon-induced transmembrane protein
IgG	immunoglobulin G
IMD	immune-mediated disease
kb	kilobase
LC-MS-MS	liquid chromatography with tandem mass spectrometry
LD	linkage disequilibrium
LMM	linear mixed model
logCPM	log-transformed counts per million
LPS	lipopolysaccharide
MAF	minor allele frequency
MAPQ	mapping quality
MARS	Molecular Diagnosis and Risk Stratification of Sepsis
Mb	megabase
MHC	major histocompatibility complex
miRNA	microRNA
MPRA	massively parallel reporter assay
MR	Mendelian randomisation
mRNA	messenger RNA
NCBI	National Center for Biotechnology Information

NET	neutrophil extracellular trap
NK	natural killer
NMD	nonsense-mediated decay
OCV	one causal variant
PAMP	pathogen-associated molecular pattern
PC	principal component
PEER	probabilistic estimation of expression residuals
pELS	(ENCODE cCRE Type) TSS-proximal enhancer-like signature
pGene	gene with protein quantitative trait loci
PIP	posterior inclusion probability
PLS	(ENCODE cCRE Type) promoter-like signature
pQTL	protein quantitative trait locus
PRR	pattern recognition receptor
PWM	position weight matrix
QTL	quantitative trait locus
REML	restricted maximum likelihood
RNA	ribonucleic acid
RNA-seq	RNA sequencing
SARS-CoV-2	severe acute respiratory syndrome coronavirus 2
scRNA-seq	single-cell RNA sequencing
SEA	simple enrichment analysis
SNP	single nucleotide polymorphism
SRA	Sequence Read Archive
SRS	sepsis response signature
SRS1	sepsis response signature 1
SRS2	sepsis response signature 2
SRSq	quantitative sepsis response signature
SuSiE	sum of single effects
SVD	singular value decomposition
T_{reg}	regulatory T cell
TAD	topologically associating domain
TCR	T cell receptor
TF	transcription factor
TLR	toll-like receptor
TOM	topological overlap metric
TSS	transcription start site
TTS	transcription termination site
TWMR	transcriptome-wide Mendelian randomisation
UTR	untranslated region
VEP	Variant Effect Predictor
WGCNA	weighted gene co-expression network analysis

1 | Introduction

The objective of this thesis is to explore the genetic mechanisms underlying variation in gene and protein expression in sepsis. Sepsis is a complex disease induced by infection that presents with broad clinical heterogeneity. A better understanding of this heterogeneity in the host immune response is critical for the identification of biomarkers and the development of novel therapeutic strategies. Although mortality from infection is heritable (Sørensen *et al.* 1988), genome-wide association studies (GWASs) have had limited success in sepsis. Functional genomics approaches, including the mapping of molecular quantitative trait loci (QTL), have proven to be more effective in dissecting clinical heterogeneity and are currently being explored as a part of the Genomic Advances in Sepsis (GAinS) study.

In this thesis, I will use functional genomics techniques to characterise the mechanisms through which QTL produce molecular and clinical heterogeneity. I will integrate genotype, transcriptomic, and proteomic layers from the GAinS cohort. In addition, I will use publicly available resources to characterise the molecular effects of QTL and the cell types through which they act. The goal of these investigations is to generate deeper biological insights into variation in gene and protein expression to inform therapeutic strategies for sepsis.

1.1 Genetics of Complex Diseases

Complex diseases have a polygenic basis and substantial influence from environmental factors, making them a biomedical and therapeutic challenge. The polygenic architecture of complex diseases is complicated by the presence of many common variants in the population that exert small effects on the disease phenotype. In addition, the role of epistatic interactions between causal loci remains poorly understood. Epistasis and environmental influence can produce variable penetrance and expressivity of disease-associated phenotypes. Thus, patients with complex diseases present with clinical, phenotypic, and molecular heterogeneity. One of the central challenges this heterogeneity presents is the inability to clearly define what specific set of requirements should be used to define the disease case and how therapeutic strategies should be

developed to target patient-specific biology.

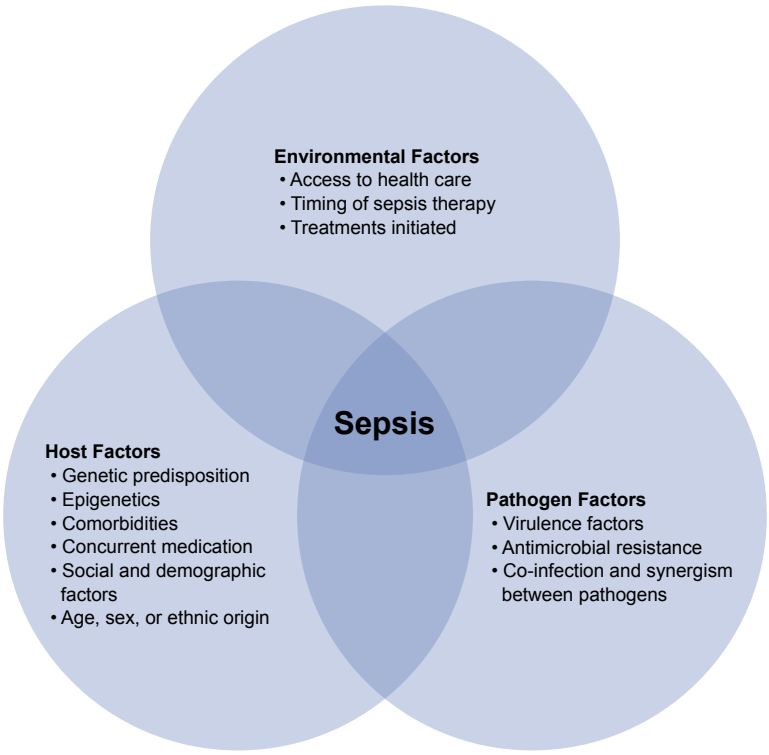

Figure 1.1: Sepsis as a complex disease. The dysregulated immune response to infection is associated with a confluence of host, environmental, and pathogen factors. Adapted from Goh *et al.* 2017

Sepsis is considered a complex disease (Figure 1.1). Sepsis is defined as a potentially lethal, maladaptive condition of organ dysfunction caused by a dysregulated host immune response to infection (Singer *et al.* 2016). The sepsis phenotype can be described through various modalities, including susceptibility to infection, the degree of organ dysfunction, disease severity, response to treatment, and outcome. In addition to host genetics, sepsis is modulated by multiple host, environmental, and pathogenic factors. Factors that tend to affect all complex diseases include comorbidity, concurrent medications, age, sex, social and demographic factors, access to treatments, and the treatments used during the course of the disease. In addition to these, sepsis presents a unique challenge to study because it is also influenced by factors associated with the infecting pathogen such as pathogen-specific immune response, co-infection, and the microbial genome (Goh *et al.* 2017). In contrast with commonly studied complex traits and diseases, the host sepsis phenotype is only observed after initial infection. Thus, even after factoring in family history, it is nearly impossible to demarcate true non-susceptible controls within the population. In addition, it is particularly challenging to assemble and recruit sepsis cohorts due to the requirement of an initial infection and the clinical challenges surrounding patient care during acute illness.

1.1.1 Heritability

When considering the genetic predisposition to a disease, it is important to address the concept of heritability. Heritability is not a measure of individual inheritability, which is a function of the genetic architecture and method of inheritance, but rather a population-level metric to quantify the amount of phenotypic variation that is explained by genetic variation.

Heritability in the broad sense H^2 is formally defined as the ratio of genotypic variance σ^2_G to phenotypic variance σ^2_P. An offspring generally shares up to one allele that is identical by descent (IBD) with a parent, which means that dominance and interactive genotypic effects do not play a role in phenotypic resemblance. As such, heritability generally refers to heritability in the narrow sense h^2, which is defined as the ratio of additive genotypic variance σ^2_A to phenotypic variance σ^2_P. Narrow-sense heritability can be estimated using regression-based approaches across the population. The simplest method is to take the slope of the regression between offspring phenotypic values and the midparent phenotypic values, although more sophisticated methods using linear mixed models (LMMs) are commonly used for efficient estimation in populations with mixed relatedness (Visscher *et al.* 2008). Recent methodological advances (Yang *et al.* 2011; Gusev *et al.* 2014; Finucane *et al.* 2015) allow for the estimation of the contribution from specific regions of the genome towards heritability. These methods partition the heritability of a trait based on genomic annotations using variance components models.

1.1.2 Genome-Wide Association Studies

The sequencing of the human genome and subsequent technological progress in genotyping human variation has allowed for the development of an analytical strategy called the GWAS. In contrast with traditional pedigree-based linkage analyses, GWASs utilise unrelated individuals in the population to associate genotypes with observed phenotypes. GWASs exploit linkage disequilibrium (LD), defined as the nonrandom association of alleles at two different loci within the population, to identify genomic regions associated with disease without the need to test all polymorphisms in the population. For statistical and genotyping simplicity, the most commonly used genetic variants are single nucleotide polymorphisms (SNPs).

The GWAS analytical strategy was based on the common disease-common variant (CD-CV) hypothesis that common diseases are caused by common variants with moderate effects. GWASs have successfully uncovered many of these trait-associated variants in complex diseases. However, the surprising discovery has been that trait-associated variants identified using GWASs do not explain the observed phenotypic variance. Specifically, narrow-sense heritabil-

ity estimates from regression-based approaches are typically much larger than the proportion of observed phenotypic variance explained by significant trait-associated variants from GWASs (Manolio *et al.* 2009).

The field has worked to genotype ever-larger cohorts to detect variants with even smaller effects based on the hypothesis that complex diseases may be caused due to thousands of common variants with very small effects. Others have proposed that common diseases may arise due to rare variation with large effects undetected in genotyping arrays. The generally accepted model is that complex traits arise from thousands of common variants with small effects, with rare variation also contributing to heritability. The omnigenic model proposes that a small number of variants with large effects are concentrated around core pathways that are biologically relevant to the expression of the phenotype, while a bulk of the heritability is spread across the genome in variants for peripheral genes that affect the core genes through *trans*-regulatory networks (Boyle *et al.* 2017).

1.1.3 Linkage Disequilibrium

For any pair of loci, LD can be quantified by comparing the observed co-occurrence of alleles against what is expected by chance based on the allele frequencies in the population, with larger deviations from expectation signifying more LD. The structure of LD is tied to the evolutionary and ancestral history of the population. Specifically, LD arises through selective sweeps, effects of genetic bottlenecks and random drift, and admixture that introduces novel variation into the population. Without the effects of selection, migration, and random drift, LD tends to decay through recombination (Slatkin 2008).

LD makes it challenging to identify the causal variants responsible for associated traits. The causal variant is often tagged by multiple variants in high LD, inducing spurious associations between non-causal variants and the trait of interest. Thus, although significant associations may identify causal genetic loci, the strength of association between variants in high LD cannot be reliably used to identify the causal variant within an associated region.

1.2 Multi-omics

Multi-omics strategies are concerned with the generation and integration of data from high-throughput assays for multiple “omes”, such as the genome, epigenome, transcriptome, proteome, and microbiome.

1.2.1 The Transcriptome and Proteome

The transcriptome refers to the entire set of RNA transcripts that may be expressed from the genome in a tissue of interest. Transcriptomes were initially assayed in a high-throughput manner using microarrays with probes designed to detect the quantity of a large number of transcripts. However, this technique relies heavily on *a priori* knowledge of transcripts. A more comprehensive approach is to use RNA sequencing (RNA-seq), which uses DNA sequencing technology to assay the transcriptome in an unbiased manner. In addition to quantifying transcripts, RNA-seq provides the opportunity to detect novel transcripts, quantify allele-specific expression, and identify splicing events.

Similar to the transcriptome, the proteome refers to the entire set of proteins that may be expressed from the genome in a tissue of interest. Techniques to assay the proteome can be divided into targeted and untargeted approaches. Similar to microarrays, targeted approaches are used to detect a preset library of proteins using affinity-based methods such as antibodies (Gold *et al.* 2010; Assarsson *et al.* 2014). Untargeted approaches are primarily based on mass spectrometry and face different challenges, such as the large variation in concentration of proteins in plasma.

In this thesis, I use gene expression data from RNA-seq and protein expression data from mass spectrometry. Due to the difficulty in obtaining samples from the critical illness setting, these data are derived from whole blood, which presents unique challenges when interpreting results.

1.2.2 Regulation of Molecular Expression

Quantification of the transcriptome and proteome provides a snapshot of the molecular state of a tissue. The quantity of molecules in a cell at any given time is tightly regulated through a diverse set of mechanisms that encode logic for basal tissue-specific functions and stimuli-specific responses. The regulatory code is itself encoded in the genome, and variation in genotype directly affects the regulation of expression.

Regulation of transcription occurs through functional elements such as promoters, enhancers, and silencers. Other functional elements in the genome such as topologically associating domains (TADs) and insulators can regulate local clusters of gene expression. Post-transcriptional modifications such as splicing and polyadenylation affect messenger RNA (mRNA) stability and function. mRNA can be degraded via nonsense-mediated decay (NMD) or the action of microRNAs (miRNAs). After translation into proteins, a variety of post-translational modifications

can alter protein function, localisation, and degradation.

Regulation of gene expression can be affected drastically by change in context. For instance, around 20% of the expressed leukocyte blood transcriptome is differentially expressed in sepsis patients compared to healthy subjects independent of the source of infection (Peters-Sengers *et al.* 2022).

1.2.3 The Epigenome

The epigenome refers to the set of chemical and steric configurations of chromatin that affect genome function. In this thesis, I use publicly available epigenomic data from various primary immune cell types to better understand which cell types may be affected by genotypic variation relevant to sepsis. Specifically, I focus on methods of detecting chromatin accessibility, which quantify how accessible regions of the genome are to factors that influence gene expression. Variation in chromatin accessibility is described in low resolution in the form of euchromatin and heterochromatin. More recently, high resolution characterisation of the epigenome using DNase I hypersensitivity sites sequencing (DNase-seq) and assay for transposase-accessible chromatin using sequencing (ATAC-seq) can nominate loci in the genome that are accessible to regulatory factors and primed for cell-type-specific responses (Boyle *et al.* 2008; Buenrostro *et al.* 2013). ATAC-seq is performed by using a hyperactive Tn5 transposase to simultaneously cut regions of accessible chromatin and ligate sequencing adapters. Regions of the genome that present less steric hindrance are more accessible to incorporation by transposase. Thus, the number of reads that align to a region of the genome is a readout of the level of accessibility (Buenrostro *et al.* 2013; Yan *et al.* 2020).

1.3 Functional Genomics

Functional genomics is concerned with describing the functions of genes and their molecular products. Below, I discuss some of the functional genomics tools used in this thesis.

1.3.1 Co-Expression Networks

Genes that are induced under similar conditions or under similar regulatory control tend to have correlated measures of expression. For example, genes that respond to a specific stimulus, are activated by a common *trans* factor, or belong to the same gene regulatory network are expected to be co-expressed. Since genes have multiple functions across various tissues and conditions,

co-expression patterns can resolve gene products that interact during specific responses, identify regulators of broad transcriptomic programs, and group genes by functional and biological relevance in a disease context (Dam *et al.* 2018).

Co-expression is described using networks. Genes represent vertices in the network and edges represent some measure of association between connected pairs of genes. Given n genes, an $n \times n$ similarity matrix $\mathbf{S} = [s_{ij}]$ represents the initial observed structure from the gene expression data. A similarity function such as $s_{ij} = |\text{cor}(\mathbf{X}_i, \mathbf{X}_j)|$ can be used to define similarity between the i-th and j-th genes. This matrix is transformed into the adjacency matrix for the network $\mathbf{A} = [a_{ij}]$ using some monotone adjacency function $a_{ij} = f(s_{ij})$. The motivation for this additional step is to impose constraints on the network. In weighted gene co-expression network analysis (WGCNA), a popular method for co-expression analysis, the adjacency function is used to approximate the scale-free topology observed in many biological and non-biological contexts (Barabási *et al.* 1999; Zhang *et al.* 2005). The co-expression network contains substructure, in that specific well-connected subgraphs capture different biological functions. These subgraphs, called network modules, are extracted from the network using various clustering approaches. Modules can be differentially co-expressed between disease states and conditions, which can assist in the identification of disease-relevant processes and regulators (Dam *et al.* 2018).

1.3.2 Molecular Quantitative Trait Loci

A QTL is a variant that is associated with a quantitative trait. That is, the genotype of the variant in an individual is predictive of some measured phenotypic quantity. In this thesis, I focus specifically on expression quantitative trait loci (eQTL) and protein quantitative trait loci (pQTL), which are variants associated with the quantity of mRNA and protein respectively. QTL are often present in functional elements in the genome involved in gene and protein regulation. The approach to mapping QTL in human populations is based on the GWAS. An additional challenge to QTL mapping compared to the GWAS, however, is unbiased multiple testing correction. False discovery rate (FDR) corrections traditionally used in GWASs are too stringent for eQTL discovery and do not account for biases between *cis* loci introduced by different LD structure, number of SNPs, and minor allele frequencies. Specialised methods such as permutation-based strategies and hierarchical gene-centric approaches are required to appropriately control the FDR while maximising power of discovery in an unbiased manner (Huang *et al.* 2018).

Due to the large multiple-testing burden and small effect sizes when testing all variants against all genes and proteins, studies with small cohorts tend to focus on mapping *cis*-QTL. In this analysis, variants near the transcription start site (TSS) are tested for association with the cog-

nate gene or protein, implying a *cis* mode of action. However, there is increasing interest in mapping *trans*-QTL, especially since a large proportion of heritability is explained in *trans* but large cohorts are required.

The common method for QTL mapping is to use linear models or LMMs. In either case, it is important to control for effects of population stratification and technical variation between samples. Principal components (PCs) from the genotype data or kinship matrices are the most popular methods to control for population stratification. Although technical covariates such as batch and sample quality metrics can be included, latent variable approaches such as probabilistic estimation of expression residuals (PEER) (Stegle *et al.* 2012) or PCs of expression can assist in controlling measured and unmeasured sources of technical variation when mapping molecular QTL.

1.3.3 Chromatin Accessibility

ATAC-seq peaks are regions of the genome that are enriched for ATAC-seq reads. Most ATAC-seq analyses utilise count-based methods and assume a Poisson background read distribution to call peaks and assign significance. The peaks are used for multiple downstream analyses. Different peaks are detected in different tissues and contexts, which can be used to infer regions of the genome that are important for context-specific regulation. Peaks can be annotated to characterise this context-specific accessibility profile. This includes simple genome metrics such as distance to the closest gene and motif enrichment tests that attempt to identify upstream *trans* factors that target detected peaks in a condition (Yan *et al.* 2020).

1.4 Colocalisation and Fine Mapping

1.4.1 Bayesian Fine Mapping

Association analyses such as the GWAS and QTL mapping tend to nominate multiple SNPs in LD as potentially causal. Statistical fine mapping methods have been developed to reduce the set of candidate causal SNPs at a locus. Early methods included using SNPs that tagged the lead variant at a certain heuristic LD threshold or joint SNP regression with shrinkage. Recently, a new suite of Bayesian fine mapping tools have been developed to identify smaller subsets of potentially causal variants. These models use the observed strength of association between SNPs and the trait of interest in addition to the underlying LD structure to quantify evidence for causal configurations at a locus. At a locus with k SNPs, a causal configuration is a binary vector

$\boldsymbol{\gamma} \in \{0,1\}^k$. For each causal configuration, $\gamma_i = 0$ indicates that the i-th SNP is not causal, while $\gamma_i = 1$ indicates that the i-th SNP is causal. Bayesian fine mapping methods generally have a prior distribution $\mathbb{P}(M_{\boldsymbol{\gamma}})$ for each causal configuration $\boldsymbol{\gamma}$. Using Bayes' rule, the evidence for a causal configuration given the association data D is

$$\mathbb{P}(M_{\boldsymbol{\gamma}} \mid D) = \frac{\mathbb{P}(D \mid M_{\boldsymbol{\gamma}})\mathbb{P}(M_{\boldsymbol{\gamma}})}{\sum_M \mathbb{P}(D \mid M)\mathbb{P}(M)}$$

The likelihood of the data given the causal configuration, $\mathbb{P}(D \mid M_{\boldsymbol{\gamma}})$, is based on the association summary statistics. Fine mapping methods assume standard association tests based on linear models. Thus, the likelihood for the data is often based on the vector of effects $\boldsymbol{\beta}$ from the regression such that

$$\mathbb{P}(D \mid M_{\boldsymbol{\gamma}}) = \int \mathbb{P}(D \mid \boldsymbol{\beta})\mathbb{P}(\boldsymbol{\beta} \mid M_{\boldsymbol{\gamma}})\,\mathrm{d}\boldsymbol{\beta}$$

An appropriate prior on $\boldsymbol{\beta}$ is also specified. The posterior inclusion probability (PIP) is often used to summarise the evidence for any given SNP being causal. The PIP for the i-th SNP is defined as

$$\mathbb{P}(\gamma_i = 1 \mid D) = \sum_{\boldsymbol{\gamma}:\gamma_i=1} \mathbb{P}(M_{\boldsymbol{\gamma}} \mid D)$$

which is the sum of the evidence for all causal configurations where the i-th SNP is causal. A common method to characterise the uncertainty surrounding the causal SNP is to generate a credible set (CS) of SNPs that, taken together, captures a set amount of posterior density. When we assume one causal variant (OCV) at a locus, the 95% CS is generated by ordering SNPs in decreasing order by PIP and taking the top m SNPs such that the sum of PIPs is greater than 0.95 (Schaid *et al.* 2018).

Recent models such as CAVIARBF and FINEMAP have relaxed the OCV assumption to search for at most L signals. CAVIARBF is an example of an exhaustive approach that attempts to enumerate all possible configurations with up to L signals. In contrast, FINEMAP and other algorithms such as DAPG and GUESSFM attempt to perform a smart search of the space of causal configurations to reduce computational time and increase the number of independent signals that can be jointly modelled at a locus (Hutchinson *et al.* 2020). FINEMAP, for instance, uses a shotgun stochastic search to efficiently explore causal configurations with up to L causal variants. In this approach, FINEMAP takes an initial configuration and performs a series of edits to identify neighbouring configurations. Using the unnormalised posterior density for each of these potential configurations as weights, FINEMAP then samples the edited configurations to identify the next configuration. This iterative procedure is repeated to identify a subset of all possible

configurations that are then used to approximate $\mathbb{P}(M_{\boldsymbol{\gamma}} \mid D)$ and the PIP for each SNP (Benner *et al.* 2016).

The sum of single effects (SuSiE) regression model is a novel approach to the fine mapping problem. In this formulation, the ℓ-th independent signal at a locus is represented by an effect vector $\boldsymbol{\beta}_\ell = \beta_\ell \boldsymbol{\gamma}_\ell$, where β_ℓ represents the effect size of the signal and $\boldsymbol{\gamma}_\ell$ represents the causal configuration of the signal. These single effect vectors are estimated using a new approach called iterative Bayesian stepwise selection (IBSS). Each iteration of IBSS involves fixing $L - 1$ effect vectors, deriving residuals of the trait after using a model with the fixed $L - 1$ effect vectors, and estimating the leftover effect vector using the analytical solution to the single effect regression model on the residuals (Wang *et al.* 2022a).

1.4.2 Colocalisation

Colocalisation can be considered an extension of the Bayesian fine mapping model. To demonstrate, I will use the COLOC method as an example of enumeration-based Bayesian colocalisation methods (Giambartolomei *et al.* 2014). At a locus with k SNPs that has been associated with two traits in different cohorts, a causal configuration is now a pair of binary vectors $\boldsymbol{\gamma}, \boldsymbol{\delta} \in \{0, 1\}^k$. A strong OCV assumption is used in COLOC, which reduces the model space to $(k + 1)^2$ possible configurations from 4^k. Each model $M_{\boldsymbol{\gamma\delta}}$ can be assigned to one of five mutually exclusive sets based on these hypotheses:

- $\mathbb{H}_0$: Neither trait is associated with the locus
- $\mathbb{H}_1$: Only the first trait is associated with the locus
- $\mathbb{H}_2$: Only the second trait is associated with the locus
- $\mathbb{H}_3$: Both traits are associated, but with different SNPs
- $\mathbb{H}_4$: Both traits are associated with the same SNP

Evidence for each hypothesis is then

$$\mathbb{P}(\mathbb{H}_i \mid D) \propto \sum_{M \in \mathbb{H}_i} \mathbb{P}(D \mid M)\mathbb{P}(M)$$

and the posterior odds for any hypothesis in reference to the null $\mathbb{H}_0$ is given by

$$\frac{\mathbb{P}(\mathbb{H}_i \mid D)}{\mathbb{P}(\mathbb{H}_0 \mid D)} = \sum_{M \in \mathbb{H}_i} \frac{\mathbb{P}(D \mid M)}{\mathbb{P}(D \mid M_0)} \times \frac{\mathbb{P}(M)}{\mathbb{P}(M_0)}$$

where $M_0 \in \mathbb{H}_0$ is the only configuration in the null set. Similar to the fine mapping methods, a prior is specified for each causal configuration $\mathbb{P}(M_{\boldsymbol{\gamma\delta}})$ and the vector of effects from the re-

gression β. In COLOC, Wakefield's method is used to calculate approximate Bayes' factors and the prior odds are constructed for each configuration from prior per-SNP probabilities (Giambartolomei *et al.* 2014).

The OCV assumption is unrealistic for most trait-associated loci, which are often composed of multiple independent causal variants (Yang *et al.* 2012). If iterative forward regression is used to identify sets of independently associated SNPs at a locus, signals can be conditioned on before using colocalisation methods. Bayesian fine mapping models are much more challenging to integrate into colocalisation methods. Although independent signals can be detected and the lead SNP from a credible set can be conditioned on, the joint inference over SNPs and uncertainty for each signal is lost in the process. The SuSiE reformulation, in comparison, can be directly integrated into the COLOC model, making it ideal for relaxing the OCV assumption in COLOC directly rather than following a two-step method to condition on independent signals (Wallace 2021).

1.5 Sepsis

Sepsis is defined as a potentially lethal, maladaptive condition of organ dysfunction caused by a dysregulated host immune response to infection (Singer *et al.* 2016). Sepsis poses a substantial worldwide burden, with an estimated 5.3 million deaths annually (Poll *et al.* 2017). The initial site of infection can vary between individuals, with the most common causes being respiratory infections followed by intra-abdominal and urinary tract infections (Angus *et al.* 2013). Although sepsis can be induced by a diverse range of pathogens interacting with a variety of pattern recognition receptors (PRRs) in the immune system (Takeuchi *et al.* 2010), the sepsis transcriptomic response in blood is largely independent of source and causative pathogen (Burnham *et al.* 2017). These transcriptomic responses are also similar to those induced by other non-infectious trauma, such as burn injuries (Xiao *et al.* 2011) or non-infectious respiratory distress (Scicluna *et al.* 2015). These observations indicate that in addition to infection, the septic response may be a more general immune response to extreme stress or trauma and that observed clinical heterogeneity may be due to a common set of underlying biological mechanisms.

1.5.1 Immunological Response during Sepsis

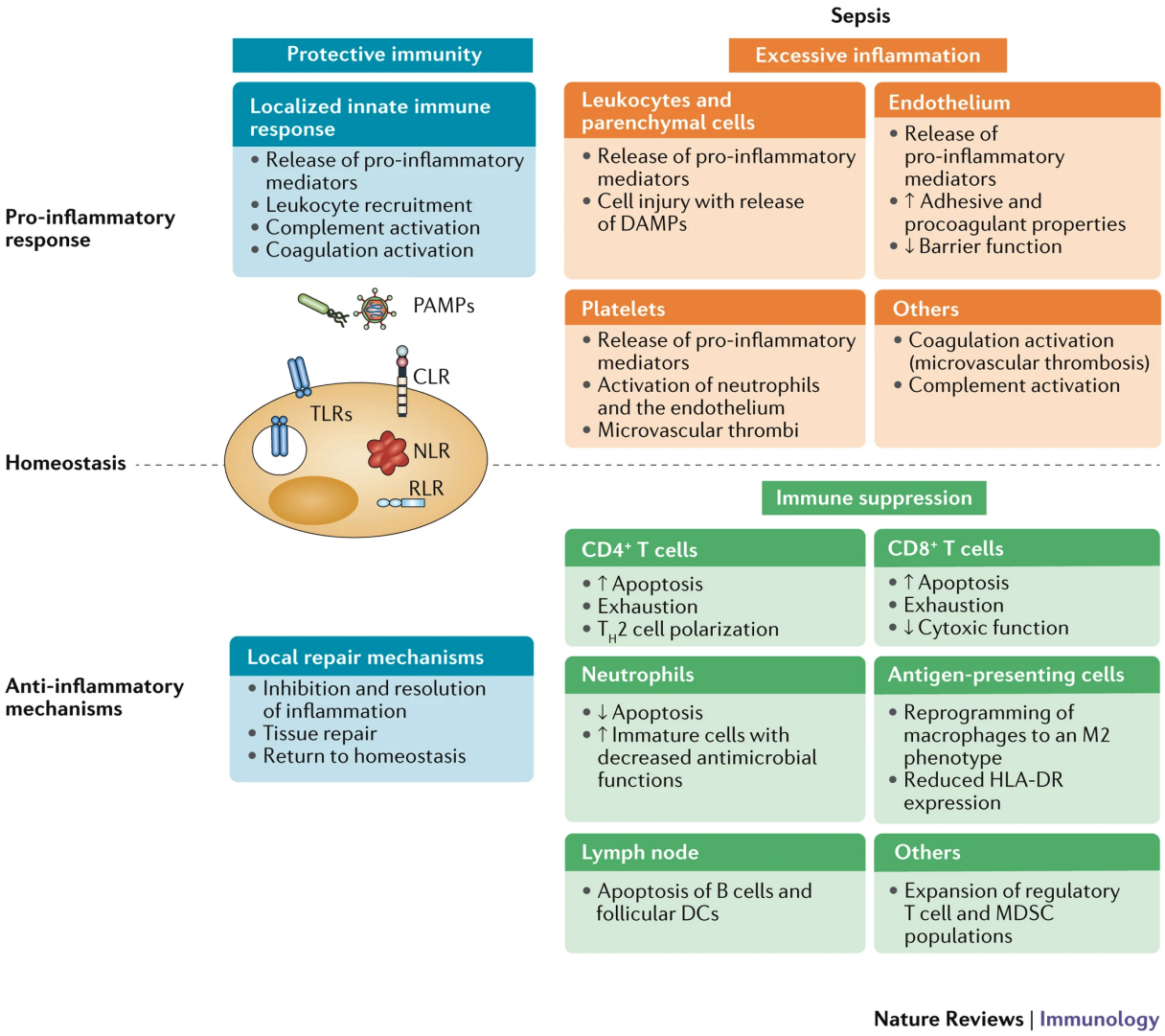

Figure 1.2: Dysregulated response to infection during sepsis. Sepsis is characterised by profound dysregulation of the normal immune response to infection. Certain components of the immune system induce excessive inflammation and cause tissue damage, while other components are pathologically suppressed. Image from Poll *et al.* 2017.

The normal immune response to infection is composed of an acute pro-inflammatory phase, followed by a concomitant anti-inflammatory phase after the infection is resolved. The pro-inflammatory phase is concerned with eliminating the pathogen through recognition of the pathogen, recruitment of effector immune cells, and activation of supporting systems such as complement and coagulation. Non-effector tissues are also modulated to support the immune response, such as vasodilation to promote inflammation in the infected region. The goal of the anti-inflammatory phase is to attenuate the immune response after the pathogen is eliminated. Both the acute pro-inflammatory and later anti-inflammatory phases are dysregulated in patients with sepsis (Figure 1.2), who show signs of excessive inflammation and excessive immune suppression (Poll *et al.*

2017). The septic response compromises respiratory, cardiac, kidney, and brain function (Angus *et al.* 2013), demonstrating systemic effects that extend beyond the initial site of infection. The processes involved in excessive inflammation and excessive immune suppression in sepsis are discussed below.

§ Excessive Inflammation

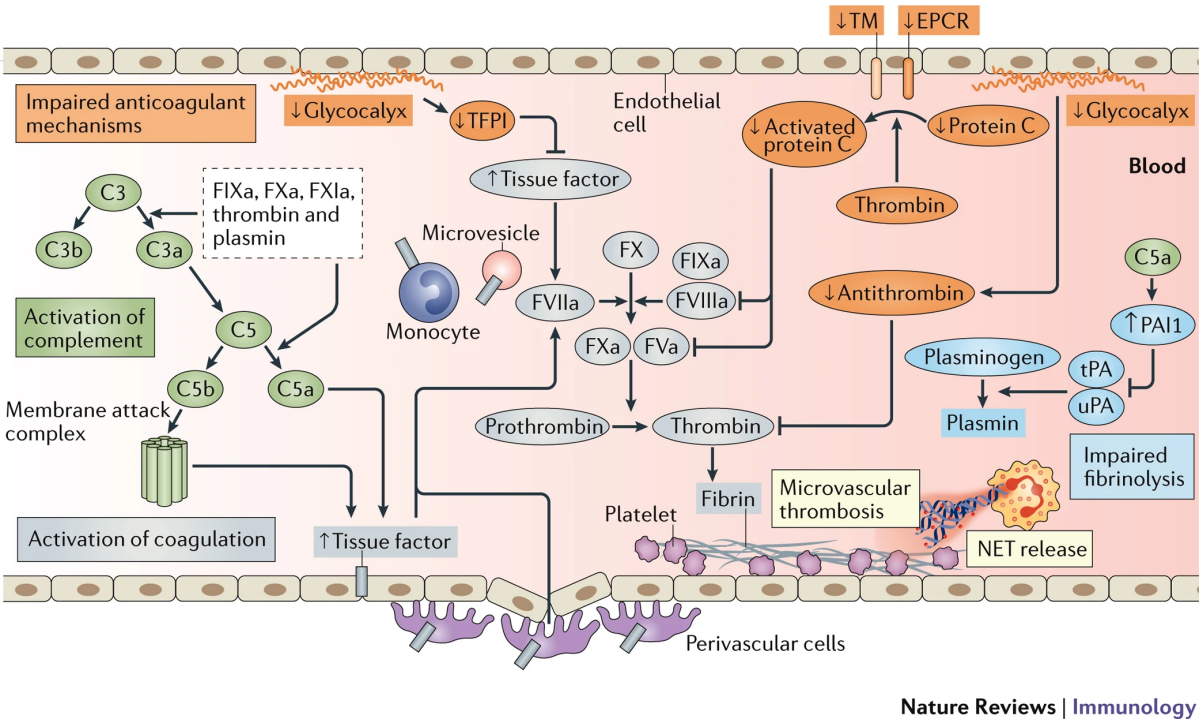

Figure 1.3: Coagulation and complement systems during sepsis. The serum proteome includes components of the coagulation cascade and complement system that are released by the liver and activated during the response to infection. Image from Poll *et al.* 2017.

Sepsis is characterised by excessive inflammation that occurs due to an interplay between effector cells of the immune system, the coagulation system, the complement system, and other non-effector tissues that support the course of inflammation.

The pathogen is detected prospectively through pathogen-associated molecular patterns (PAMPs) by PRRs and retrospectively through damage-associated molecular patterns (DAMPs) from tissue damage caused by the pathogen. The release of DAMPs due to damage caused by the immune system, in addition to or in the absence of the pathogen, can create a vicious cycle of sustained immune response (Chan *et al.* 2012).

The complement system consists of a repertoire of intravascular proteins with antimicrobial properties that are activated through protein-protein interactions during the innate immune response and are responsible for recruiting and activating effector cells, targeting microbes directly,

and assisting in the maintenance of an active immune response (Figure 1.3). Sepsis is characterised by elevated levels of complement proteins. The excessive activation of the complement system can cause organ damage and result in the release of DAMPs (Guo *et al.* 2005).

Although the coagulation system is generally associated with tissue repair following trauma, the coagulation system is activated by neutrophils during infection in a process called immunothrombosis (Figure 1.3) and has potent antimicrobial functions (Engelmann *et al.* 2013). The coagulation system is dysregulated during sepsis and results in microvascular thrombosis and haemorrhage due to the consumption of clotting factors (Levi *et al.* 2017). This process is simultaneously associated with excessive platelet activation and consumption, which has also been linked with organ injury through several mechanisms (Stoppelaar *et al.* 2014). Immunothrombosis is associated with the formation of neutrophil extracellular traps (NETs), which consist of DNA, histones, and serine proteases that are released by the neutrophil to entrap pathogens. NETs promote thrombosis by acting as scaffolds for platelet aggregation and have been associated with organ dysfunction in sepsis at elevated levels (Czaikoski *et al.* 2016). The dysregulation of immunothrombosis and NET formation may be driven by the rapid expansion of neutrophils during the innate immune response. Indeed, a recent study has implicated the expansion of a subset of immature neutrophils that are unique to sepsis and differ from neutrophilia observed in the normal immune response to infection (Kwok *et al.* 2022).

Taken together, excessive inflammation in sepsis is characterised by a profound dysregulation of the intertwined immune, coagulation, and complement response to infection.

§ Excessive Immune Suppression

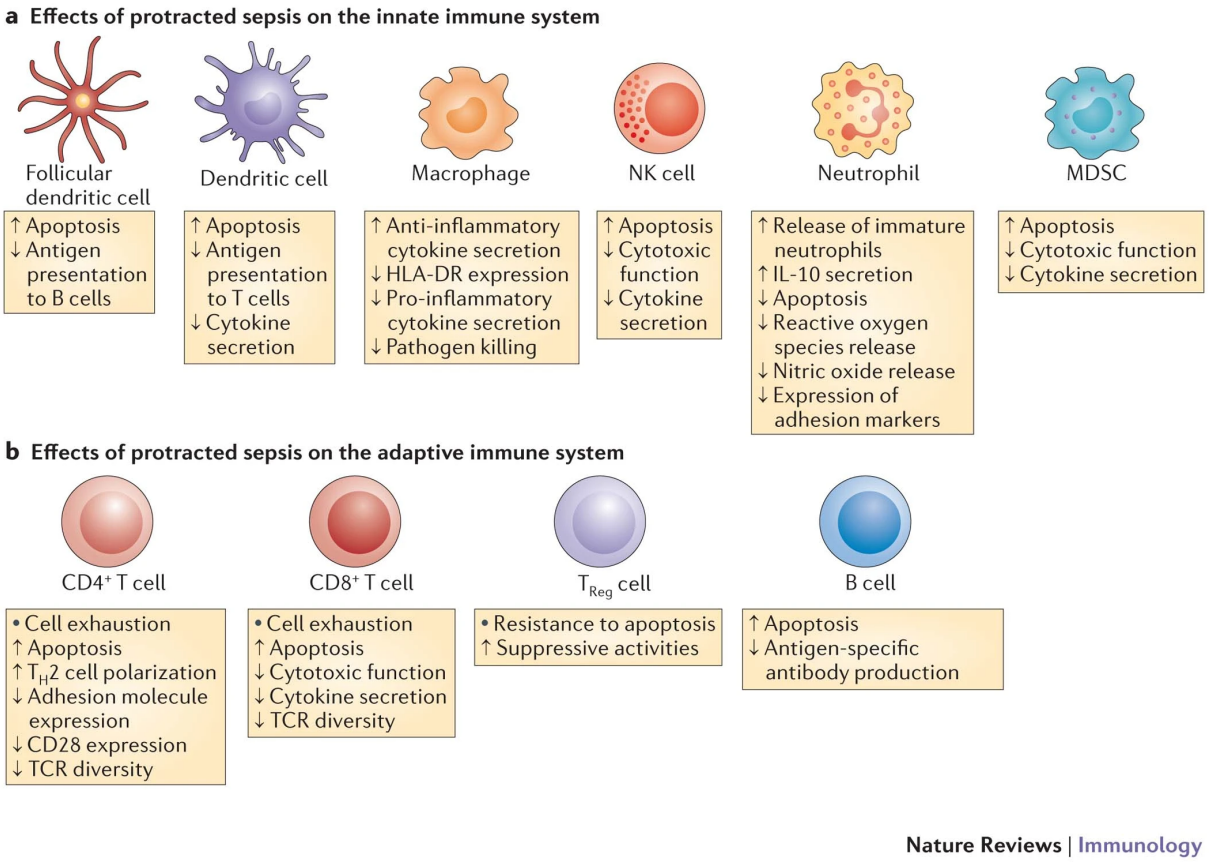

Figure 1.4: Effects of sepsis on leukocyte phenotypes. Leukocytes are affected differently by protracted sepsis, with some cellular functions being suppressed by anti-inflammatory signalling or via apoptosis. Image from Hotchkiss *et al.* 2013

At a cursory glance, excessive immune suppression during and following sepsis directly contradicts the exaggerated immune activation observed in sepsis. However, a key observation in early studies was an increased susceptibility to infections during and following survival from sepsis (Boomer *et al.* 2011). Patients experiencing sepsis and other non-infectious trauma experience increased incidence of viral reactivation and infection (Luyt *et al.* 2007; Goh *et al.* 2020). This immunosuppression is associated with lymphocyte exhaustion, which consists of decreased lymphocyte count and reduced lymphocyte activity (Figure 1.4). Immunosuppression in sepsis also involves reprogramming of professional antigen-presenting cells (APCs) such as macrophages and dendritic cells (Poll *et al.* 2017).

Apoptosis during sepsis drives a strong reduction in $CD4^+$ T cells, $CD8^+$ T cells, B cells, and dendritic cells. In addition, natural killer (NK) cells and $CD4^+$ T helper cell subsets demonstrate reduced activity that is consistent with T cell exhaustion. Regulatory T cells (T_{regs}), which are responsible for attenuating the immune response, are more resistant to sepsis-induced apoptosis

and also lead to the reduction of effector T cell function (Hotchkiss *et al.* 2013). The immature sepsis neutrophils identified recently also demonstrated the ability to suppress the activity and proliferation of $CD4^+$ T cells in a co-culture system (Kwok *et al.* 2022).

Professional APCs are reprogrammed during sepsis. Macrophages and dendritic cells have reduced *HLA-DR* expression, which is required for antigen presentation (Hotchkiss *et al.* 2013). In addition, macrophages enter an immunosuppressive phenotype similar to that observed during endotoxin tolerance (Poll *et al.* 2017).

1.5.2 Sepsis Genomics

Components of the immune system demonstrate substantial diversity between individuals. This diversity is deeply rooted in our evolutionary relationship with infectious agents. The evolutionary arms race between pathogens and humans necessarily runs into a generational time asymmetry — humans cannot outcompete pathogens using one antimicrobial strategy. Instead, evolution has favoured a huge diversity of immune activations and responses that vary in intensity between individuals. For a specific individual, immune diversity provides an evolutionary advantage because it increases the chance that a pathogen from a prior host will not have adapted to the immune response of the new host. Thus, although most humans have conserved immune responses maintained through strong purifying selection, they differ in the degree to which they are primed for different types of immune responses. These differences increase as humans encounter perturbations and new environments. The immune system is also under the genetic control of the most polymorphic genes in the genome (Liston *et al.* 2021). This immune diversity underpins the clinical and molecular heterogeneity observed in the response to infection and during sepsis.

The heritability of poor outcome during sepsis was strikingly observed in a study comparing risk of death by infection in adoptees to the risk of death by infection for biological and adoptive parents (Sørensen *et al.* 1988). In spite of evidence for heritability, the most recent analyses of 28-day outcome in sepsis patients failed to uncover any genome-wide significant associations (Rautanen *et al.* 2015; Scherag *et al.* 2016). One explanation for the failure of GWASs with 28-day endpoint is that the genetic component of sepsis may be explained by genetic predisposition for either susceptibility to infection or organ failure (Angus *et al.* 2013). For instance, GWASs have identified genetic associations with response to specific pathogenic agents such as severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) (Niemi *et al.* 2021), tuberculosis (Curtis *et al.* 2015), and some other common infections (Tian *et al.* 2017). Since sepsis is exacerbated by comorbidity with certain conditions, genetic predisposition for these related disorders may also

contribute to shared genetic mechanisms of immune dysregulation (Angus *et al.* 2013). Another explanation is that the initial estimate of heritability overestimates the current heritability of mortality from infection due to improved clinical care, in which case deeper phenotyping of sepsis cohorts is required to better understand the observed patient heterogeneity. Thus, recent inquiry into the genetic underpinnings of sepsis has focused on drivers of molecular variation in diseased patients.

The GAinS study is a collective effort to perform deep phenotyping of molecular variation in adult sepsis patients presenting to the intensive care unit (ICU). The study recruited adults with community-acquired pneumonia (CAP) or faecal peritonitis (FP) on admission to the ICU (Tridente *et al.* 2014; Walden *et al.* 2014). The data set generated from the study contains genotype data (Rautanen *et al.* 2015), gene expression from whole blood leukocytes (Davenport *et al.* 2016; Burnham *et al.* 2017), and protein expression from blood plasma. A GWAS of 28-day survival in patients with CAP identified no significant genome-wide associations, although 2 loci were associated with p-values lower than 1×10^{-5} and replicated in other cohorts (Rautanen *et al.* 2015).

Initial analysis of transcriptomic variation in this cohort identified transcriptomic signatures called sepsis response signature 1 (SRS1) and sepsis response signature 2 (SRS2) in patients with CAP (Davenport *et al.* 2016) and FP (Burnham *et al.* 2017). Individuals with SRS1 were associated with higher early mortality compared to SRS2 (Davenport *et al.* 2016; Burnham *et al.* 2017). Differentially expressed (DE) genes between SRS1 and SRS2 in CAP patients were enriched for pathways involved in T cell activation, cell death, apoptosis, necrosis, cytotoxicity, and phagocyte movement. Specifically, key mediators of endotoxin tolerance were present in these DE genes, with downregulation of human leukocyte antigen (HLA) class II genes and T cell activation genes in SRS1 samples. These pathways were similarly enriched in DE genes between sepsis response signature (SRS) groups in FP patients. The predominant source of variation in the transcriptome is associated with SRS assignment, with few differences arising due to source of infection. Recently, a quantitative sepsis response signature (SRSq) score was developed to position samples along a continuum, with higher values indicating a state close to SRS1 and lower values indicating a state close to healthy control patients (Cano-Gamez *et al.* 2022). Other efforts to perform transcriptomic stratification of patients in sepsis cohorts (Scicluna *et al.* 2017; Sweeney *et al.* 2018; Baghela *et al.* 2022) have similarly identified specific signatures associated with poorer outcomes. Patient stratification can also identify subsets of patients that may benefit from specific therapies (Marshall 2014). For instance, the use of hydrocortisone as an acute treatment in sepsis was associated with increased mortality in patients with an SRS2 phenotype (Antcliffe

et al. 2019). Taken together, these results demonstrate the clinical significance of transcriptomic heterogeneity in sepsis and the need for stratification to identify more precise point-of-care therapeutic strategies.

1.5.3 Summary of *cis*-eQTL and pQTL

The GAinS cohort now consists of 2,056 participants, with RNA-seq available for 667 patients and plasma protein data available for 1,182 patients. Variation in molecular expression is associated with genotypic variation between individuals. The promise of molecular heterogeneity for patient stratification and the potential to identify context-specific regulatory elements has motivated the identification of eQTL in whole blood leukocytes[1] and pQTL in plasma[2] in GAinS (Figure 1.5). Based on the RNA-seq data, 20,272 genes were considered to be expressed in the cohort of patients. For any given expressed gene, common SNPs in a 1 megabase (Mb) window around the TSS were tested as potential *cis*-eQTL. Of the expressed genes, 10,618 (52.4%) genes had *cis*-eQTL identified in this analysis. A conditional *cis*-eQTL analysis of the genes with expression quantitative trait loci (eGenes) identified evidence for multiple signals for 3,788 (35.7%) eGenes, resulting in 16,049 total eGene-signal pairs. The proteomics analysis of plasma, in contrast, detected 269 proteins. Since a *trans* analysis was powered with this number of proteins and restricting to *cis* windows would have removed many potentially interesting variants from the analysis, a genome-wide scan of protein expression was conducted and identified 29 (10.8%) genes with protein quantitative trait loci (pGenes). Of these, 23 (8.6%) pGenes had pQTL in *cis* and 6 (2.2%) pGenes had pQTL in *trans*, with no pGenes with both *cis*- and *trans*-pQTL.

[1] Described in Section A.6
[2] Described in Section A.7

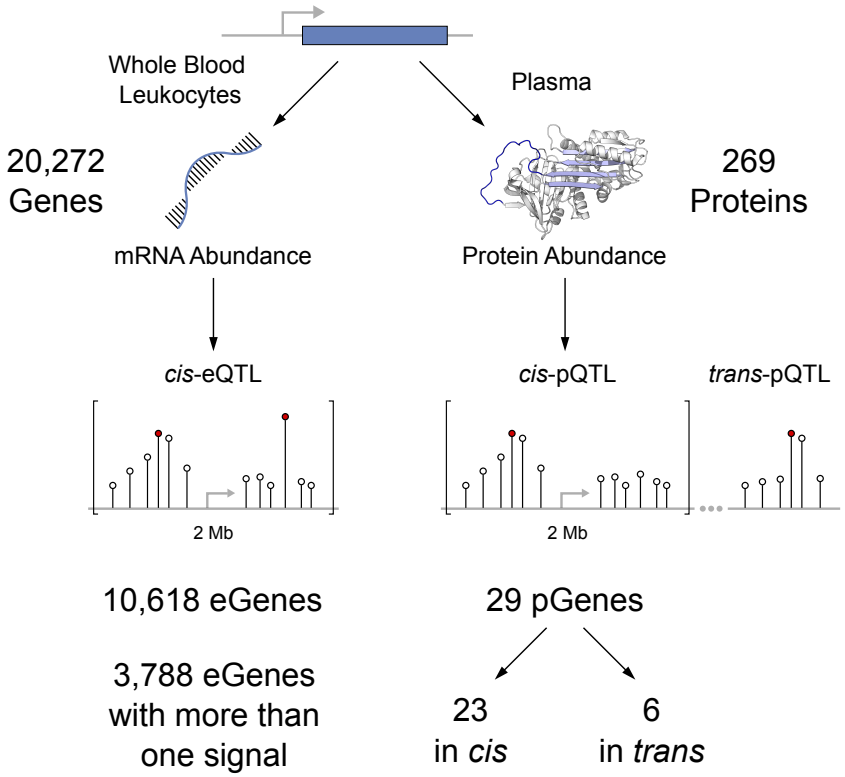

Figure 1.5: Mapping of *cis*-eQTL and pQTL. Expressed genes in whole blood leukocytes and detected proteins in plasma were tested for *cis*-eQTL and pQTL respectively. eGenes and pGenes were identified in each case. An initial 10,618 eGenes were identified, with 3,788 showing evidence of more than one signal based on a forward regression approach. The 29 pGenes were divided into 23 with only *cis*-pQTL and 6 with only *trans*-pQTL. There were no pGenes with both *cis*- and *trans*-pQTL.

In this thesis, I will explore the biological mechanisms that connect variation in genotype with molecular variation of clinical and therapeutic relevance in sepsis.

1.6 Specific Aims

§ Chapter 3: Characterise broad transcriptomic signatures in sepsis

The investigation of transcriptomic response to infection in GAinS has focused on the primary axis of variation in gene expression via SRS and SRSq. The molecular biology and heritability of SRS and its association with outcome remain to be explored. I aim to:

1. Decompose transcriptomic variation into co-expression modules of pathological interest
2. Identify the genetic basis of variation in modules
3. Determine the relevance of modules to outcome

§ Chapter 4: Investigate patterns of association with molecular expression

The eQTL and pQTL in GAinS provide initial evidence for the association between genotype and molecular traits. The flow of information between mRNA and proteins provides an opportunity to better understand how variation in gene expression in whole blood leukocytes is coupled with variation in protein abundance in plasma. Patterns of association can also be leveraged to connect molecular QTL with disease-associated variants. I aim to:

1. Colocalise eQTL and pQTL
2. Colocalise molecular QTL with disease-associated variants
3. Reduce uncertainty around causal variants due to LD using fine mapping

§ Chapter 5: Identify dysregulated cell types in sepsis

Molecular QTL in GAinS are not specific to one cell type or tissue. Whole blood leukocytes are a heterogeneous mixture of primary immune cells and proteins in plasma are produced and degraded by various tissues. Molecular QTL may be specific to certain cell types or may represent broad patterns of regulation across cell types. I aim to:

1. Characterise the accessibility landscape of stimulated immune cells
2. Identify cell types that manifest the effects of QTL
3. Predict the molecular effects of QTL

2 | Methods

2.1 Description of Cohort

The patients in this study consisted of adults recruited to the GAinS study from 34 ICUs in the United Kingdom. The patients were adults (aged greater than 18 years) diagnosed with either CAP or FP. Admission criteria for the study have been described previously (Walden *et al.* 2014; Tridente *et al.* 2014). For inclusion, CAP was defined as febrile illness associated with cough, sputum production, breathlessness, leukocytosis, and radiological features of pneumonia acquired in the community or within less than two days of hospital admission (Walden *et al.* 2014). Similarly, FP was defined as inflammation of the serosal membrane that lines the abdominal cavity, secondary to contamination by faeces, as diagnosed by laparotomy (Tridente *et al.* 2014).

Sample collection is described in prior studies (Davenport *et al.* 2016; Burnham *et al.* 2017). Briefly, whole blood samples from patients were collected approximately one, three, and/or five days after admission to the ICU. Most patients do not have samples from all three time points. Genotyping using SNP microarrays[1] and imputation[2] was performed previously. RNA-seq was performed on 864 samples from 667 patients[3]. Before RNA-seq was performed, gene expression in the initial subset of recruited patients was assayed using microarrays[4]. Mass spectrometry was used to quantify protein abundance in plasma for 1,680 samples from 1,068 patients[5].

[1] Described in Section A.1
[2] Described in Section A.2
[3] Described in Section A.3
[4] Described in Section A.4
[5] Described in Section A.5

2.2 Analysis of Gene Expression

2.2.1 Weighted Network Correlation Analysis

A recent simulation study showed that systematic variation in the transcriptome cannot be attributed to co-expression modules in a network when assuming a scale-free topology (Parsana *et al.* 2019). To control for this technical variation, the top 20 gene expression PCs were regressed out from the log-transformed counts per million (logCPM) gene expression matrix. The biweight midcorrelation matrix was then calculated for the residual gene expression to generate a similarity matrix using the `bicor` function from the WGCNA R package (Langfelder *et al.* 2008). For any given gene, the gene expression value of all samples from the same individual was replaced with the mean gene expression (Bland *et al.* 1995) to only measure between-individual correlation. Spatial quantile normalisation implemented in the spqn R package (Wang *et al.* 2022b) was used to account for the mean-correlation bias in the similarity matrix. The `normalize_correlation` function was used on the similarity matrix with 21 blocks of size 1000 and block 18 as the reference group.

Co-expression modules were identified in the gene expression data using the WGCNA R package (Langfelder *et al.* 2008). The `pickSoftThreshold` function determined a soft threshold value of 4 for the similarity matrix. This soft threshold was used to build an unsigned adjacency matrix using the `adjacency` function, which was used to calculate the topological overlap metric (TOM) matrix using the `TOMsimilarity` function. The dynamic tree cut algorithm included in the WGCNA package as the `cutreeDynamic` function was used to generate modules with default parameters and a minimum cluster size of 10. Similar modules were merged based on the similarity of their module eigengenes using the `mergeCloseModules` function with a cut height of 0.1. For a module, the eigengene was defined as the first PC of the gene expression data of the genes present in the module. The module eigengenes for the final set of modules were calculated using the `moduleEigengenes` function.

2.2.2 Module Annotation

The clusterProfiler R package was used to annotate modules with GO terms (Biological Processes, Cellular Components, and Molecular Functions) and KEGG pathways using the `enrichGO` and `enrichKEGG` functions respectively (Wu *et al.* 2021). The ReactomePA R package was used to annotate modules with Reactome pathways using the `enrichPathway` function (Yu *et al.* 2016). In each case, p-values were adjusted using Benjamini-Hochberg FDR correction. A p-value thresh-

old of 0.01 and a q-value threshold of 0.05 were used. The set of expressed genes was used as the background for enrichment.

Cell type signatures from the xCell R package (Aran *et al.* 2017) and marker genes for cell types detected in sepsis (Kwok *et al.* 2022) were used to identify cell-type-specific modules. The xCell signatures were derived based on differential gene expression from large transcriptomic studies of individual cell types and built to minimise classification error. Enrichment of gene signatures was performed using a hypergeometric test using the `phyper` function in R. The entire set of expressed genes was considered the background for enrichment. P-values were corrected using the Benjamini-Hochberg FDR procedure. Since multiple transcriptomic studies assayed the same cell types in xCell, one cell type often had multiple signatures. The median odds of enrichment per cell type in xCell were reported for any signatures that passed a q-value cutoff of 0.05. In contrast, each cell type in the Kwok *et al.* 2022 study had one set of markers. Signatures for cell types passing a q-value cutoff of 0.05 were reported. Enrichment was calculated as the ratio between the proportion of signature genes in the module and the proportion of signature genes in the entire set of expressed genes.

2.2.3 Module Association with Clinical Endophenotypes

For each endophenotype, FDR was controlled using the Benjamini-Hochberg procedure. Associations were considered to be significant if the adjusted p-value was less than 1×10^{-3}. Due to the large panel of cell frequencies, a lower adjusted p-value threshold of 1×10^{-4} was used.

Inverse normal transformed cell proportions for three broad leukocyte lineages (neutrophils, lymphocytes, and monocytes) were available for the majority of samples. To expand on these broad lineages, CIBERSORTx was used to estimate cell type frequencies from the bulk RNA-seq samples (Newman *et al.* 2019). Single-cell RNA sequencing (scRNA-seq) samples from the ongoing Sepsis Immunomics study were used as the panel for CIBERSORTx. An initial description of the scRNA-seq data is presented by Kwok *et al.* 2022. Spearman's Rho was used to measure association between eigengenes and quantitative endophenotypes (SRSq, cell proportions, cell frequencies, and xCell scores) and a two-sided significance test was performed using the `cor.test` function in R.

To identify eigengenes that were associated with changes over time or with diagnosis (CAP or FP), a repeated measures analysis of variance (ANOVA) was performed using the `anova_test` function implemented in the rstatix R package. The reported effect size was generalised eta squared (GES).

The 28-day survival of patients was measured in the GAinS cohort. Association between this patient outcome and each eigengene was tested using a Cox proportional hazards model as implemented in the survival R package using the `coxph` function. For each patient, the value of the eigengene at the last time point recorded was used as a predictor for the survival function.

2.3 Molecular QTL

A single-variant association analysis developed previously for repeated measurements (Davenport *et al.* 2018) was used to perform QTL mapping. In each QTL analysis, genotypes were filtered to only include biallelic SNPs on autosomes with a minor allele frequency (MAF) greater than 0.01. Genotypes were coded as 0, 1, or 2 based on the number of copies of the minor allele carried by each patient. The lme4 R package (Bates *et al.* 2015) was used to build the LMM for each variant using the `lmer` function. Patients were modelled as a random intercept. In each QTL analysis, a given variant was modelled as having a fixed effect on the trait. A likelihood ratio test was used to test the significance of the effect of the variant using the `anova` function implemented in the lme4 R package. This approach was previously used to identify *cis*-eQTL[1] and pQTL[2].

Let n be the number of samples, p be the number of covariates, and q be the number of patients. A LMM for a trait $\mathbf{Y} \in \mathbb{R}^{n\times 1}$ is modelled as

$$(\mathbf{Y} \mid \mathbf{B} = \mathbf{b}) \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b}, \sigma^2\mathbf{I}_n)$$

where $\mathbf{X} \in \mathbb{R}^{n\times p}$ is the design matrix of the fixed effects, $\boldsymbol{\beta} \in \mathbb{R}^{p\times 1}$ is the vector of fixed effects, $\mathbf{Z} \in \mathbb{R}^{n\times q}$ is the design matrix of the random effects, $\mathbf{B} \in \mathbb{R}^{q\times 1}$ is a random vector of patient-specific intercepts, $\mathbf{b}$ is a realisation of $\mathbf{B}$, and σ^2 is the residual variance. The random vector $\mathbf{B}$ is further assumed to be normally distributed as

$$\mathbf{B} \sim \mathcal{N}(\mathbf{0}, \sigma_R^2\mathbf{I}_q)$$

implying that the patient-specific random effects are independent. Restated for the i-th observation associated with the j-th patient, Y_i is modelled as

$$Y_i = \beta_0 + \beta_g g_j + \sum_{k=2}^{p} \beta_k X_{ik} + b_j$$

where β_0 is the intercept, β_g is the effect of an additional minor allele on the trait, $\mathbf{g}$ is the first

[1] Described in Section A.6
[2] Described in Section A.7

column vector of $\mathbf{X}$ and g_j is the genotype of the patient, β_k is the effect of the k-th covariate, and b_j is the patient-specific random effect.

2.3.1 Mapping of Module QTL

A set of 70,300 SNPs consisting of lead *cis*-eQTL, lead conditional *cis*-eQTL, and trait-associated SNPs from the European Bioinformatics Institute (EBI) GWAS Catalog (Buniello *et al.* 2019) were tested for associations with all module eigengenes. Similar to the *cis*-eQTL analysis, seven genotyping PCs, 20 PEER factors, SRS status (SRS1 versus non-SRS1), diagnosis (CAP versus FP), and cell proportions were used as fixed-effect covariates. A genome-wide threshold of 6.71×10^{-9} was used based on a Bonferroni FDR correction of 0.05, accounting for the number of SNPs and number of modules tested. Loci were defined for each module by constructing 1 Mb windows around each module QTL and merging intervals until a set of disjoint intervals was generated.

Four PCs in addition to the first module eigengene (together called the top 5 module eigengenes) were calculated using singular value decomposition (SVD) on gene expression Z scores as implemented in the `svd` R function. The genome-wide threshold was decreased to 1.34×10^{-9} based on the additional testing burden. Loci were identified using the same recursive merging strategy of 1 Mb intervals as that used for the top module eigengenes.

2.3.2 Module QTL Replication

The microarray gene expression data was used as a replication cohort for the module QTL. A module was only considered for replication if 5 or more genes from the original module were present on the microarray. These microarray modules were then used to calculate microarray module eigengenes using SVD on gene expression Z scores as implemented in the `svd` R function. The 134 overlapping samples between the microarray and RNA-seq data sets were used to assess the similarity of module eigengenes between the two methods using Spearman's Rho.

PEER factors for the microarray expression data were calculated using the same approach as the RNA-seq analysis. The 134 overlapping samples were held out before testing module QTL for replication. The same LMM was used to test lead module QTL from the RNA-seq analysis on microarray module eigengenes. Lead module QTL were tested only if they met a MAF cutoff of 0.01 in the entire GAinS cohort. P-values were adjusted using a Benjamini-Hochberg FDR correction as implemented in the `p.adjust` R function. The replicated genotypic effect was considered significant if the adjusted p-value was less than 0.05. Replication of the direction of effect was confirmed by comparing the direction of the original effect with the direction of the replicated

effect multiplied by the sign of Spearman's Rho comparing the module eigengenes between the microarray and RNA-seq methods.

2.4 Colocalisation

Colocalisation was performed using conditional *cis*-eQTL, pQTL, module QTL, and trait-associated SNPs from selected studies (Table B.1) in the EBI GWAS Catalog with matching ancestry. Compared to overlapping individual SNPs between association studies, statistical colocalisation is performed across a locus consisting of variants in a genomic interval. The 1 Mb window around the TSS presented a natural definition for the colocalisation locus for conditional *cis*-eQTL and *cis*-pQTL. For *trans*-pQTL and module QTL, colocalisation loci were defined by building 1 Mb windows around each QTL and recursively merging genome-wide until a disjoint set of intervals was generated. For trait-associated SNPs from the EBI GWAS Catalog, a 1 Mb window around the lead trait-associated SNP was used as the colocalisation locus.

The COLOC R package (Giambartolomei *et al.* 2014) was used to perform colocalisation between any two traits. A colocalisation event for two traits was defined to occur when $\mathrm{PP3}+\mathrm{PP4} > 0.25$ and $\mathrm{PP4}/(\mathrm{PP3}+\mathrm{PP4}) > 0.7$. Default priors in COLOC were used.

2.5 Fine Mapping

Fine mapping was performed on the conditional *cis*-eQTL, pQTL, and module QTL loci that were used for colocalisation. In the case of eGenes with multiple conditional *cis*-eQTL signals, conditional summary statistics were used to perform fine mapping. The SuSiE regression (Wang *et al.* 2020) implemented in the susieR R package and FINEMAP (Benner *et al.* 2016) were used to identify 95% CSs. For both methods, one causal variant was assumed to underlie the conditional *cis*-eQTL. Up to ten causal variants were assumed for the other loci. The LD matrix for each locus required by susieR and FINEMAP was retrieved from the genotype data of the cohort rather than an external LD panel. As a naive alternative to CSs, tagging variants in a 1 Mb window around the lead variant with $R^2 > 0.8$ (tagging SNP sets) were identified using PLINK (Purcell *et al.* 2007) on genotype data from the entire GAinS cohort.

2.6 Publicly Available Data

Paired-end ATAC-seq reads (Accessions: SRP066100, SRP156496, SRP265675) were retrieved from the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) (`https://www.ncbi.nlm.nih.gov/sra`). A total of 219 ATAC-seq samples (Tables C.1 and C.2) were retrieved using fasterq-dump (`https://github.com/ncbi/sra-tools`).

2.6.1 ATAC-seq Alignment

Adapter sequences were trimmed using TrimGalore (`https://github.com/FelixKrueger/TrimGalore`) with the `--paired` option. The reads were then aligned to the Genome Reference Consortium human build 38 (GRCh38) build of the human genome using Bowtie 2 (Langmead *et al.* 2012) with default parameters and `--no-mixed` and `--no-discordant` flags set. Duplicates were marked using the Picard (`https://github.com/broadinstitute/picard`) `MarkDuplicates` function. Reads that were unpaired, unmapped, duplicates, or mapped to multiple locations were filtered out. To ensure quality alignments, reads with a mapping quality (MAPQ) score of less than 30 were filtered out. Reads that mapped to the mitochondrial genome or to the blacklisted genomic regions reported by the Encyclopedia of DNA Elements (ENCODE) (Amemiya *et al.* 2019) were also filtered out. Technical replicates from the neutrophil atlas were merged. The alignment files were filtered, indexed, merged, and sorted using a combination of SAMtools (Danecek *et al.* 2021) and BEDTools (Quinlan *et al.* 2010).

2.6.2 ATAC-seq Sample Quality

TSS enrichment scores were used as a proxy for sample quality. TSS enrichment scores measure the signal-to-noise ratio for each sample based on the expectation that regions near the TSS of protein-coding genes are more accessible across the entire genome. The score was calculated by dividing the mean coverage of the 100 base pair regions centred at the TSS by the mean coverage of the 100 base pair regions that are 1 Mb away from the TSS. TSS regions of protein coding genes were retrieved from version 99 of the Ensembl human genome reference (Yates *et al.* 2020) and coverage was calculated using featureCounts (Liao et al. 2014) with `-p` and `-O`.

2.6.3 ATAC-seq Peaks

ATAC-seq peaks were called using MACS2 (Zhang *et al.* 2008) with `--keep-dup all`, `--nomodel`, `--nolambda`, and `-f BAMPE`. Three types of peak sets were defined: group, cell type, and consen-

sus peak sets. A group was defined as a cell-condition pair, where the condition was stimulation status (stimulated versus unstimulated) for the immune atlas and stimulation condition (ligand, *E. coli* + time point, or *S. aureus* + concentration) for the neutrophil atlas. For each group, sample peaks with MACS2-derived q-value less than 1×10^{-4} present in two or more samples were merged. If the group consisted of only one sample, only the q-value filter was applied. Any peaks larger than 3 kilobases (kb) were removed.

The group peak sets were used to create the cell type peak sets. For each group peak set, peaks that intersected with peaks from another group peak set from the same cell type with less than 90% overlap were filtered out. Group peak sets from the same cell type were then merged. Any peaks larger than 3 kb were removed.

All group peak sets in an atlas were used to create the consensus peak set. Peaks that intersected with peaks from another group with less than 90% overlap were filtered out before merging group peak sets. Any peaks larger than 3 kb were removed. Intersection and merging operations were performed using BEDTools (Quinlan *et al.* 2010).

2.6.4 Peak Annotation and Motif Enrichment

To characterise ATAC-seq peaks detected in various cell types and under various stimulations, peak sets were annotated using the `annotatePeaks.pl` script in HOMER (Heinz *et al.* 2010). The script was run with the GRCh38 build of the human genome, genome annotation from version 99 of the Ensembl human genome reference (Yates *et al.* 2020), and the `-organism human` option.

Motif enrichment analysis requires an appropriate background set of sequences to contrast with the query sequences of interest. Since ATAC-seq peaks are enriched near the TSS, background sequences were selected in a local region around the test peak set. For the cell type peaks, the flanking regions upstream and downstream of each peak were used as background sequences. Any flanking region that overlapped with another peak in the test set was removed from the background sequence set. For the group peaks discovered under stimulation, the background sequences were defined as any peaks that overlapped more than 90% between the control and the stimulated group peak sets. The test set was constructed by removing any peaks from the stimulated group peak set that overlapped with any peak from the control group peak set.

Simple enrichment analysis (SEA) implemented in the MEME suite (Bailey *et al.* 2015; Bailey *et al.* 2021) was used to identify enrichment of known motifs in peaks compared to the defined background set of sequences. Only motifs with an enrichment q-value less than 0.05 were retained. The find individual motif occurrences (FIMO) tool in the MEME suite (Grant *et al.* 2011; Bailey *et al.*

2015) was used to identify motif locations in peak sequences. Curated, non-redundant transcription factor binding motifs were retrieved from the vertebrate taxonomic group of the JASPAR 2022 CORE database (Sandelin *et al.* 2004; Castro-Mondragon *et al.* 2022).

2.7 Functional Interpretation

2.7.1 Enrichment of eQTL in Functional Categories

The SNPsnap web server (Pers *et al.* 2015) was used to generate 10,000 matched SNPs for each of the lead conditional *cis*-eQTL. Matching was performed based on LD from the European superpopulation within Phase 3 of the 1000 Genomes (1000G) Project (The 1000 Genomes Project Consortium *et al.* 2015). The recommended default parameters were used, but HLA SNPs were not excluded. The matched SNPs were used to test for enrichment of the conditional *cis*-eQTL in group peaks from the immune and neutrophil atlases, ENCODE candidate cis-regulatory elements (cCREs) (Moore *et al.* 2020), and ChromHMM (Ernst *et al.* 2012) states from the 18-state model for selected Roadmap Epigenomics Project (Kundaje *et al.* 2015) epigenomes (Table E.1). For each genome annotation, the measured statistic was the proportion of SNPs that overlapped an annotated region. The null distribution was estimated using the 10,000 samples of matched SNPs. A two-sided test was conducted using the null distribution with a significance threshold of $\alpha = 0.0001$.

The permutation-based analysis of GoShifter (Trynka *et al.* 2015) was used to test for enrichment of conditional *cis*-eQTL in group peak sets. The reimplementation can be found at `https://github.com/NMilind/LeanGoShifter`. Lead conditional *cis*-eQTL and any tagging SNPs within a 1 Mb window with $R^2 > 0.8$ were included for each eGene.

CHEERS is a method that is specifically developed to identify enrichment in peak count data from different stimulations of a cell type (Soskic *et al.* 2019). Enrichment of lead conditional *cis*-eQTL in specific neutrophil states was tested using CHEERS. A significance threshold of $\alpha = 0.001$ was used. Since *cis*-eQTL were identified in 1 Mb windows around TSSs, peaks were subset to only include those that fell in a 1 Mb window around any TSS. To reduce confounding caused by difference in peak sizes, a region the width of the median peak width centred at each peak was used to test for SNP overlap instead of the entire peak. The reimplementation can be found at `https://github.com/TrynkaLab/CHEERS/tree/python3`.

2.7.2 Partitioned Heritability

Consider a dichotomous genomic annotation that splits the complete set of m biallelic variants into two mutually exclusive sets, where α variants fall within the annotation and $\bar{\alpha}$ variants are not in the annotation. A LMM can be used to build a variance components model that jointly estimates the SNP heritability of a trait that is attributable to the annotation[1] (Yang *et al.* 2010; Gusev *et al.* 2014).

Let n be the number of samples, p be the number of covariates, and q be the number of patients. A LMM for a trait $\mathbf{Y} \in \mathbb{R}^{n\times 1}$ is modelled as

$$(\mathbf{Y} \mid \mathbf{B} = \mathbf{b}, \mathbf{B}_\alpha = \mathbf{b}_\alpha, \mathbf{B}_{\bar{\alpha}} = \mathbf{b}_{\bar{\alpha}}) \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}(\mathbf{b} + \mathbf{b}_\alpha + \mathbf{b}_{\bar{\alpha}}), \sigma^2\mathbf{I}_n)$$

where $\mathbf{X} \in \mathbb{R}^{n\times p}$ is the design matrix of the fixed effects, $\boldsymbol{\beta} \in \mathbb{R}^{p\times 1}$ is the vector of fixed effects, $\mathbf{Z} \in \mathbb{R}^{n\times q}$ is the design matrix of the random effects, $\mathbf{B} \in \mathbb{R}^{q\times 1}$ is a random vector of independent patient-specific intercepts, $\mathbf{B}_\alpha \in \mathbb{R}^{q\times 1}$ is a random vector of patient-specific intercepts with a covariance structure based on kinship estimated from SNPs within the annotation, and $\mathbf{B}_{\bar{\alpha}} \in \mathbb{R}^{q\times 1}$ is a random vector of patient-specific intercepts with a covariance structure based on kinship estimated from SNPs outside the annotation. The vectors $\mathbf{b}$, $\mathbf{b}_\alpha$, and $\mathbf{b}_{\bar{\alpha}}$ are realisations of these random variables. The random vectors are normally distributed as

$$\begin{aligned}
\mathbf{B} &\sim \mathcal{N}(\mathbf{0}, \sigma_R^2\mathbf{I}_q) \\
\mathbf{B}_\alpha &\sim \mathcal{N}(\mathbf{0}, \sigma_\alpha^2\boldsymbol{\Psi}_\alpha) \\
\mathbf{B}_{\bar{\alpha}} &\sim \mathcal{N}(\mathbf{0}, \sigma_{\bar{\alpha}}^2\boldsymbol{\Psi}_{\bar{\alpha}})
\end{aligned}$$

where $\boldsymbol{\Psi}_\alpha$ and $\boldsymbol{\Psi}_{\bar{\alpha}}$ are genetic relationship matrices (GRMs) derived from the variants within and outside the annotation respectively. The per-SNP heritability of the trait h^2_{SNP} and the annotations $h^2_{\text{SNP}\alpha}$ is

$$\begin{aligned}
h^2_{\text{SNP}} &= \frac{1}{m}\left[\frac{\sigma_\alpha^2 + \sigma_{\bar{\alpha}}^2}{\sigma_R^2 + \sigma_\alpha^2 + \sigma_{\bar{\alpha}}^2 + \sigma^2}\right] \\
h^2_{\text{SNP}\alpha} &= \frac{1}{\alpha}\left[\frac{\sigma_\alpha^2}{\sigma_R^2 + \sigma_\alpha^2 + \sigma_{\bar{\alpha}}^2 + \sigma^2}\right]
\end{aligned}$$

The enrichment of per-SNP heritability in the annotation is

$$\frac{h^2_{\text{SNP}\alpha}}{h^2_{\text{SNP}}} = \frac{m}{\alpha}\left[\frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_{\bar{\alpha}}^2}\right]$$

[1] Discussed in Section F

A variance components model was built using the `relmatLmer` function in the lme4qtl R package (Ziyatdinov *et al.* 2018) to estimate the heritability of module eigengenes partitioned by various functional annotations. Estimation was performed using restricted maximum likelihood (REML). For each annotation, biallelic variants were partitioned into mutually exclusive sets based on the annotation, which were used to estimate two separate GRMs using GCTA (Yang *et al.* 2010). Seven genotyping PCs, 20 PEER factors, SRS status (SRS1 versus non-SRS1), diagnosis (CAP versus FP), and cell proportions were used as fixed-effect covariates.

2.7.3 Variant Effect Prediction

The Ensembl Variant Effect Predictor (VEP) version 104 (McLaren *et al.* 2016) was used to annotate lead variants for conditional *cis*-eQTL, module QTL, and pQTL loci. The script to run VEP and its plugins is built and maintained by the Human Genetics Informatics (HGI) team at the Wellcome Sanger Institute.

2.8 Statistical Analysis

All statistical analyses were conducted in an Ubuntu environment. The code used for analysis can be found at `https://github.com/davenportlab/eQTL_pQTL_Characterization`. The complete list of software used is provided in the repository.

3 | Gene Co-Expression

The aim in this chapter is to decompose gene expression data from the GAinS study into co-expression modules. I will use these co-expression modules to better understand the relationship between genetic variation, transcriptomic variation, and outcome.

3.1 Co-Expression Modules

The gene expression data consists of RNA-seq of 864 whole blood leukocyte samples from 667 adult patients[1]. Samples were collected one, three, and/or five days after admission to the ICU whenever possible. 637 individuals with gene expression data were also present in the genotype data[2].

Co-expression modules are gene sets that show evidence of similar expression patterns within the tissue of interest. Correlation between gene expression profiles may be driven by upstream transcription factors, heterogeneity in cell proportions between individuals, or mechanistic interplay between gene products. Weighted correlation network analysis is a method to cluster the transcriptome into mutually exclusive modules of genes that are co-expressed. WGCNA was used to decompose the GAinS transcriptome into gene modules[3]. 106 modules were identified from the 20,272 expressed genes in the GAinS cohort. These modules ranged from 11 to 1,785 genes in size (Figure 3.1). Modules were labelled in decreasing order based on their size. Six genes were not assigned to any modules.

[1] Described in Sections 2.1 and A.3
[2] Described in Section A.1
[3] Described in Section 2.2.1

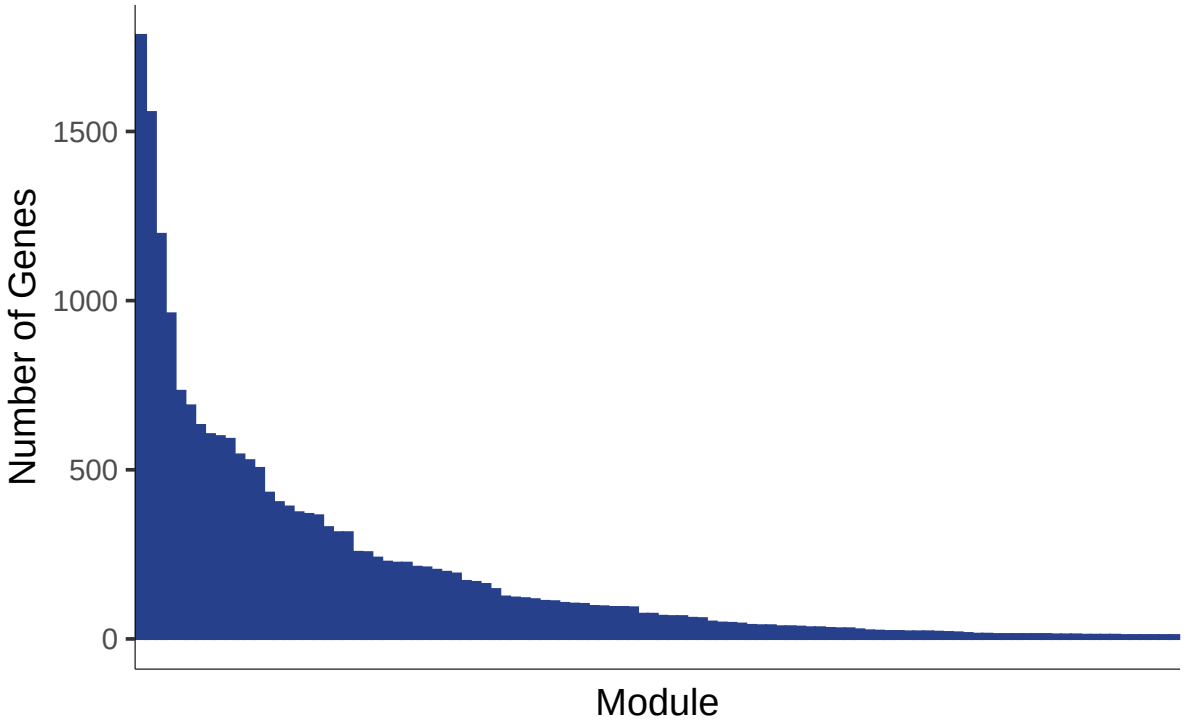

Figure 3.1: Distribution of module sizes. This figure depicts the distribution of module sizes for the 106 modules generated using WGCNA. The distribution had a long tail of smaller modules.

3.1.1 Signatures of Leukocytes

Enrichment for known pathways and cell-type-specific genes was used to characterise the biological pathways that were captured by the modules[1]. The modules demonstrated enrichment for xCell transcriptomic signatures derived from whole blood (Figure 3.2) and markers of cell types in blood derived from a sepsis cohort (Figure 3.3).

[1] Described in Section 2.2.2

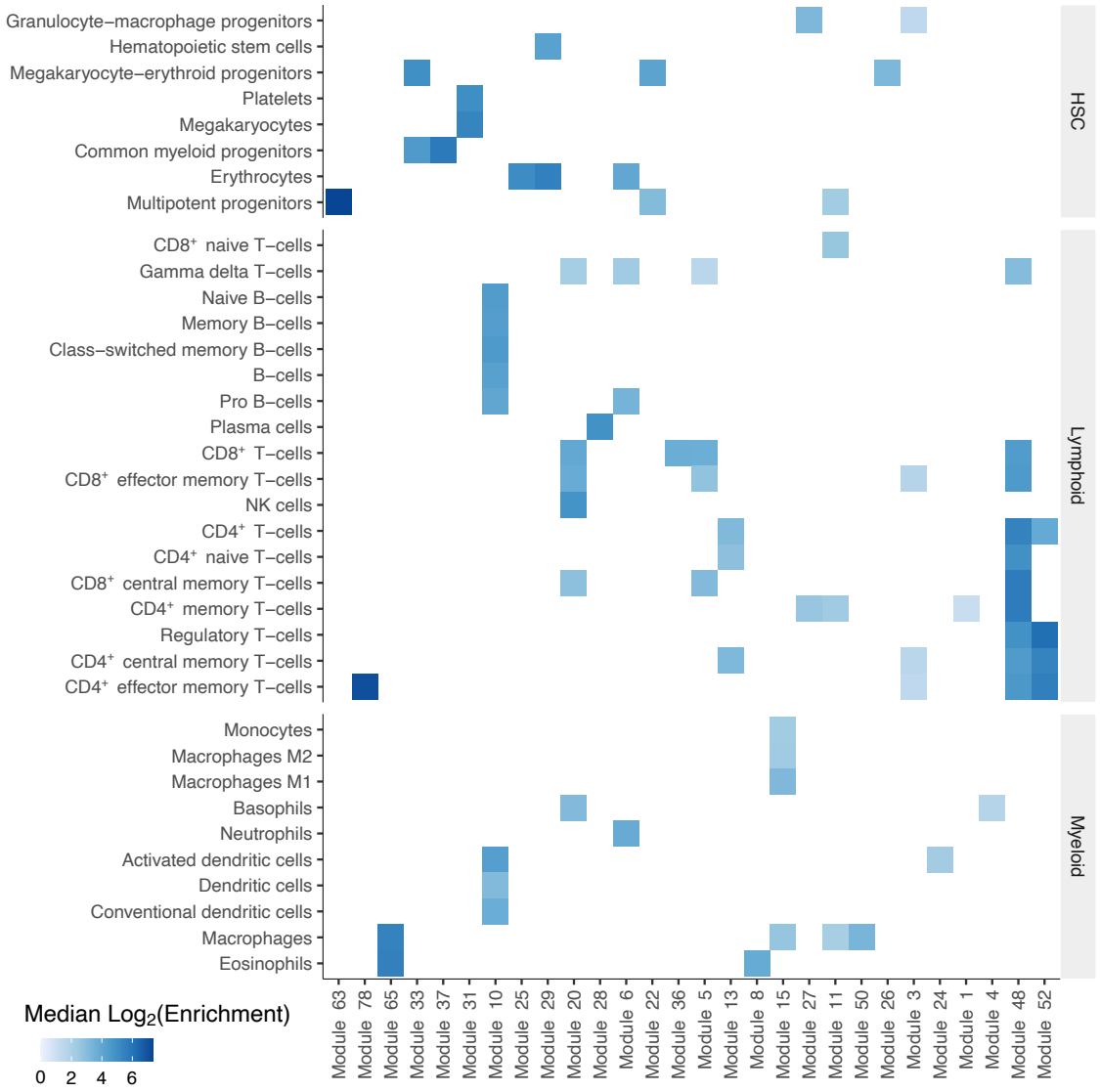

Figure 3.2: Cell-type-specific enrichment of modules. Modules were tested for enrichment of xCell gene signatures derived from large whole blood transcriptomic studies. Modules shown were enriched for one or a few signatures in related cell types, demonstrating cell-type specificity.

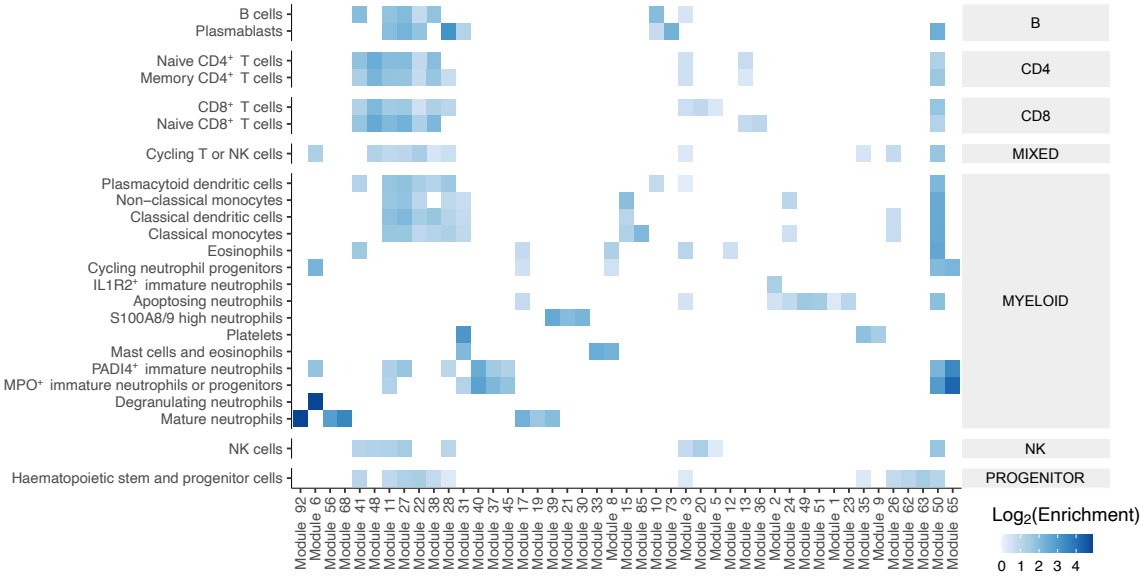

Figure 3.3: Cell marker enrichment of modules. Modules were tested for enrichment of leukocyte marker genes in sepsis identified by Kwok *et al.* 2022. Modules demonstrated cell-type specificity for markers for cell types detected in a sepsis context. This was especially apparent for various populations of neutrophils identified by Kwok *et al.* 2022.

A number of modules captured gene programs of particular relevance to the innate immune response, a process that induces excessive inflammation in sepsis. Neutrophils play a particularly important role in sepsis. Kwok *et al.* 2022 derived a trajectory for neutrophil subsets in sepsis from immature to mature (Figure 3.4).

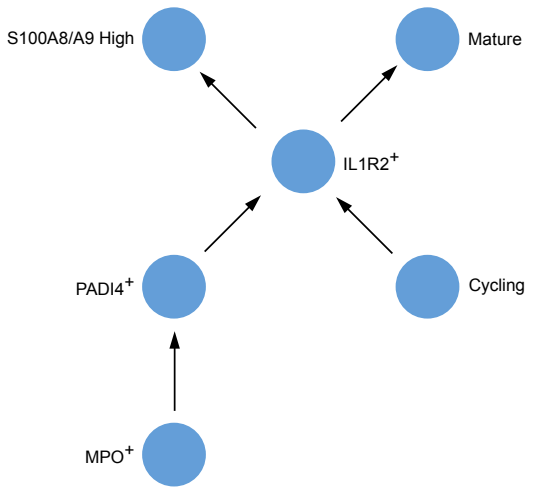

Figure 3.4: Neutrophil subsets. The differentiation pathway of neutrophil subsets identified by Kwok *et al.* 2022. Neutrophils move from MPO$^+$ immature neutrophils and cycling neutrophils towards mature neutrophils and S100A8/9 high neutrophils.

MPO$^+$ immature neutrophils differentiate into PADI4$^+$ immature neutrophils before eventually forming the large pool of IL1R2$^+$ immature neutrophils observed in sepsis. Modules uniquely en-

riched for the MPO^+ and $PADI4^+$ gene markers are different from those enriched for the $IL1R2^+$ immature neutrophils (Figure 3.3). This included module 37, which contains genes essential to neutrophil effector function (Table 3.1) and is enriched for neutrophil activation (GO:0042119) and Antimicrobial Peptides (R-HSA-6803157). These neutrophil subsets were also associated with module 65, which contains genes for multiple antimicrobial molecules (Table 3.1) and is enriched for neutrophil degranulation (GO:0043312) and Antimicrobial peptides (R-HSA-6803157). The other subset of immature neutrophils were the cycling neutrophils, which are neutrophils that are actively going through the cell cycle. Module 6 was enriched for the neutrophil xCell signature (Figure 3.2) and marker genes for degranulating neutrophils and cycling neutrophil progenitors (Figure 3.3). Consistent with this, module 6 was enriched for chromosome segregation (GO:0007059), cell cycle DNA replication (GO:0044786), and Cell Cycle Checkpoints (R-HSA-69620). Expansion of the $IL1R2^+$ immature neutrophil subset is specifically associated with sepsis. Module 2 was enriched for markers of $IL1R2^+$ immature neutrophils. S100A8/9 high neutrophils were one of the two terminal neutrophil subsets. S100A8/9 is released during infection by neutrophils and activates toll-like receptor (TLR) and RAGE signalling (Pruenster *et al.* 2016). Modules 21, 30, and 39 were enriched for S100A8/9 high neutrophil gene markers. Module 39 is also enriched for markers of mature neutrophils, which are a terminal population independent of the S100A8/A9 high neutrophils. Modules 56 and 92, enriched for mature neutrophil markers, contain NFκB pathway members and major histocompatibility complex (MHC) class I genes and associated regulators respectively (Table 3.1).

In addition to neutrophils, monocytes are known to be dysregulated during sepsis, especially through the process of endotoxin tolerance. Module 15 is specifically enriched for macrophage signatures and markers (Figures 3.2 and 3.3). It contains components associated with macrophage activation and chemotaxis (Table 3.1). *HIF1A* and associated genes in the hypoxia-induced glycolysis pathway that are implicated in endotoxin tolerance were captured separately in Module 51 (Table 3.1). Is is also enriched for the KEGG HIF-1 signalling pathway (hsa04066). Platelet activation and coagulation, which are activated alongside the cellular component of the humoral response, are dysregulated in sepsis. Module 31 was enriched for platelet and megakaryocyte signatures (Figure 3.2) as well as platelet and granulocyte gene markers (Figure 3.3). It contained receptors present on platelets that activate platelet aggregation (Table 3.1).

Table 3.1: Key genes in modules. Pathway and gene set enrichment analyses identified key genes in multiple modules. Some of these key genes and their relevance are listed in the table below.

Group	Module	Genes	Function
Innate Immune Response	Module 2	*MMP9, ADAM19, SELL*	Transendothelial migration
		TRAF3IP3	Toll-like receptor signaling
	Module 6	*CXCR1, CXCR2*	Neutrophil chemoattraction
		MMP25	Transendothelial migration
	Module 8	*CCR3, CLC, HRH4*	Eosinophil markers
		IL15RA, CEBPE, ADORA3	Granulocyte function
		ALOX15	Immunomodulation
	Module 15	*MARCO, FGR, TREM2, CSF1R, MSR1, CMKLR1*	Macrophage activation
		CCR2	Macrophage chemotaxis
		SFTPD, RNASE2, RNASE3	Antimicrobial molecules
		C1QA, C1QB, C1QC, CLU, CFP, C3	Complement system components
	Module 31	*GP1BA*	Interaction with von Willebrand factor
		GP6, CLEC1B, PEAR1, MPIG6B	Platelet receptors for platelet aggregation
		SERPINE1	Platelet granule content
	Module 37	*ELANE, SERPINB1*	Neutrophil elastase activity
		DEFA4, AZU, MPO, PRTN3, CTSG	Neutrophil granule components
	Module 65	*CAMP, BPI, LTF, HP*	Antimicrobial molecules
		CEACAM8, TCN1, LCN2	Granulocyte marker
Antiviral Response	Module 23	*PSMA2, PSMA3, PSMA5, PSMB4, PSMB8, PSMB9, PSMB10, PSME1, PSME2*	Proteasome components
		ERAP2, TAP1, TAP2	Antigen processing
		TAPBPL	Antigen presentation
	Module 39	*MX1, MX2*	Antiviral activity
		IFIT1, IFIT2, IFIT3, IFIT5	Viral RNA inhibition
		OASL, OAS1, OAS2, OAS3	dsRNA-induced antiviral response
		IRF7, RSAD2, IFI6	Other interferon-induced antiviral activity
	Module 81	*HLA-A, HLA-F, HLA-G*	MHC class I genes

Continued on next page

Group	Module	Genes	Function
	Module 92	*HLA-B, HLA-C, HLA-E, B2M* *NLRC5* *BTN2A1, BTN2A2, BTN2A3P, BTN3A1, BTN3A3, BTN3A2*	MHC class I genes Regulator of MHC class I expression Suppression of T cell interaction with APCs
Adaptive Immune Response	Module 10	*CD19, CD80* *PAX5, IKZF3, CCR6* *BLK, MS4A1, CD79A, CD79B* *CD180, TNFRSF13B* *CXCR5* *CIITA, HLA-DOA, HLA-DPB1*	B cell markers B cell maturation B cell activation NFκB activation B cell chemotaxis Professional antigen presentation
	Module 20	*NCR3, SH2D1B, CTSW, CD160* *GZMA, GZMM, GNLY, FASLG, PRF1* *KLRC1, KIR2DL4*	Activation of cytotoxic cells Cytotoxic response NK cell markers
	Module 48	*CRTAM, ITK, CXCR6, PDCD1, SIT1, UBASH3A, CTLA4* *SH2D1A, SLAMF1, LY9* *CD3D, CD3E, CD3G, CD4, CD5, CD6, CD28, CD40LG, CD96*	General T cell function SLAM transduction pathway T cell surface antigens
	Module 73	*IGHA1, IGHA2, IGKC* *JCHAIN, MZB1* *TNFRSF17*	Immunoglobulin components Immunoglobulin function NFκB and JNK activation in B cells
Sepsis Immune Response	Module 47	*ZNF268, ZNF227, ZNF606, ZNF226, ZNF585A, ZNF71, ZNF544, ZNF814, ZNF717, ZNF585B, ZNF345*	Transcriptional activators and repressors
	Module 51	*HIF1A* *PDK1, ALDOA, ENO1, GAPDH, PGK1*	Master regulator of HIF-1 signalling pathway Enzymes in anaerobic respiration
	Module 56	*TLR2, NFKBIA, DDIT4*	Mature neutrophil markers
	Module 63	*MPL* *PRSS57* *CD34, CYTL1, PROM1*	Platelet production Neutrophil granule component HSPC signature

Continued on next page

Group	Module	Genes	Function
		RBPMS, ZNF521	TGF-β/SMAD signaling
	Module 71	*ZNF737, ZNF273, ZNF595, ZNF675, ZNF680, ZNF506, ZNF107, ZNF253, ZNF14, ZNF90, ZNF91, ZNF93, ZNF486, ZNF682*	Transcriptional activators and repressors
	Module 84	*C4A, C4B* *SMAD7, MICA*	Complement factors Immune Signalling
	Module 85	*HADHA, ACADVL, ACAA2, ETFDH*	Fatty acid β-oxidation
	Module 99	*RNASET2, SLC15A4*	ssRNA detection
	Module 103	*LYZ*	Antimicrobial agent
	Module 106	*YEATS4* *IFITM1, IFITM2, IFITM3*	Transcriptional activation via acetylation Interferon-based response to pathogens

Adaptive immunity is expected to activate over the course of sepsis due to the prolonged nature of the immune response. Module 48 captured broad T cell signatures (Figure 3.2) and contained genes related to many T cell subsets (Table 3.1). Genes for cytotoxicity activation and response in both NK cells and $CD8^+$ T cells were captured in module 20 (Table 3.1). Modules 10 and 73 contained genes for various different functions of B cells (Table 3.1).

Compared to the larger modules, the smaller modules identified specific pathways of pathological interest to sepsis. This included specific immune functions such as single stranded RNA detection, antimicrobial agents, and transcriptional activators and repressors (Table 3.1). Module 106, for instance, contained three of the four members of the interferon-induced transmembrane protein (IFITM) family (Diamond *et al.* 2013), which restrict cellular entry of a diverse set of pathogens (Table 3.1).

3.1.2 Association with Endophenotypes

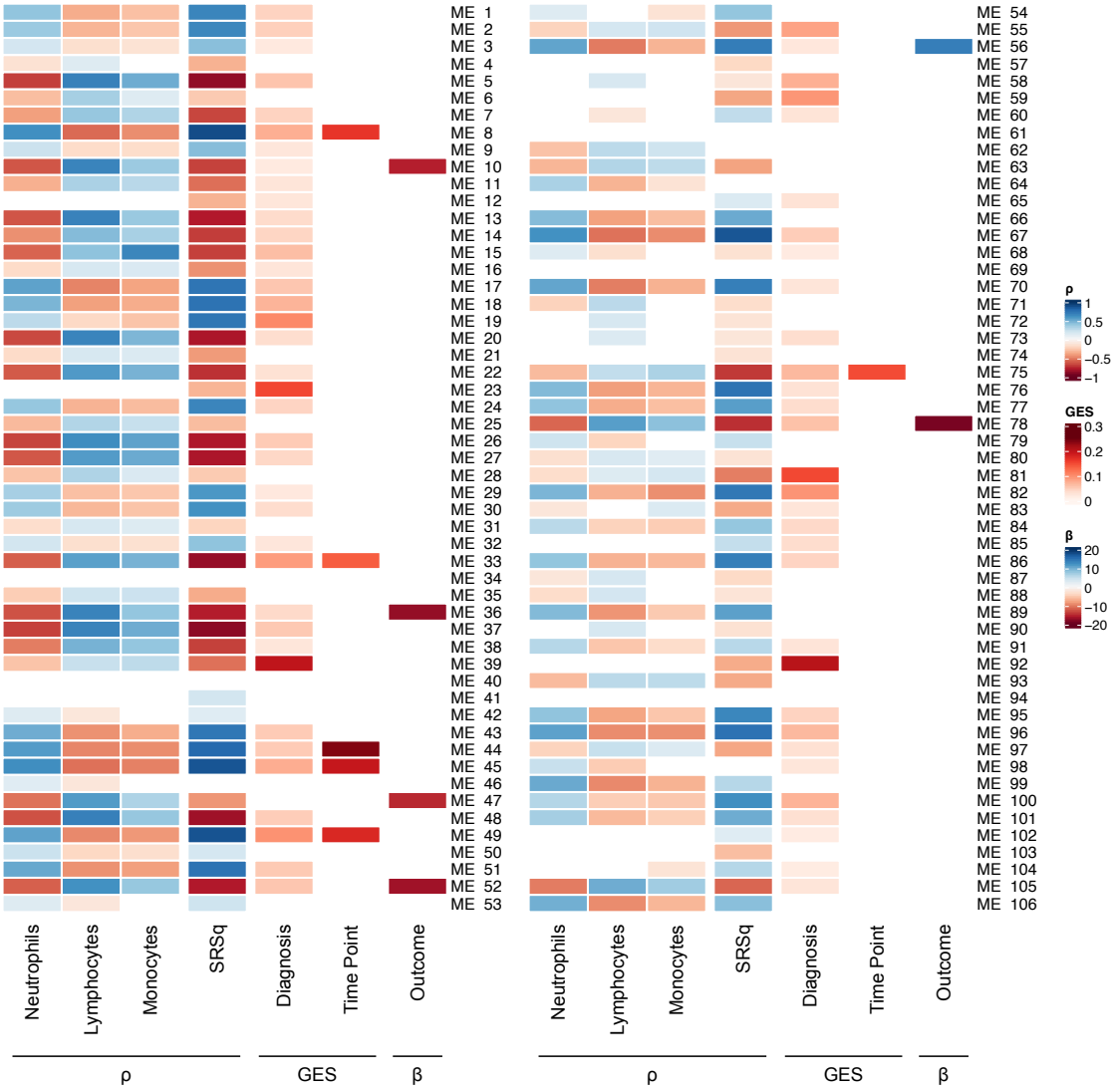

Figure 3.5: Associations between module eigengenes and clinical endophenotypes. Module eigengenes were tested for association with measured cell proportions (neutrophils, lymphocytes, and monocytes) and SRSq using Spearman's Rho. Association with diagnosis and time point was tested using repeated measures ANOVA. Finally, association with 28-day outcome was tested using a Cox proportional hazards model. Only associations passing the p-value threshold are displayed in this heatmap.

The module eigengene is the first principal component of the gene expression data associated with genes within the module. Since modules might resolve molecular programs underlying SRS and heterogeneity in other clinical parameters, module eigengenes were tested for association with clinical endophenotypes[1]. An association between the eigengene and an endophenotype suggests that variation in the gene network captured by the module is associated with observed

[1] Described in Section 2.2.3

phenotypic variation. The endophenotypes included the SRSq for each sample, the time point of sample collection, the diagnosis of the patient (CAP versus FP), the cell proportions for each sample, the patient outcome, and estimated cell frequencies from a reference scRNA-seq data set. 101 (95.3%) of the modules were associated with at least one clinical endophenotype (Figure 3.5) based on the module eigengene. Only 6 modules were associated with time point, and a different set of 6 modules were associated with outcome. These modules might represent pathways in whole blood that can act as potential biomarkers for disease progression or severity.

Several modules strongly associated with diagnosis (Figure 3.5) contained genes relevant to antiviral response. Module 23 was enriched for the antigen processing and presentation of peptide antigen via MHC class I (GO:0002474) term and contained genes of the proteasome complex that processes antigens for presentation via MHC class I molecules (Table 3.1) . Module 39 encoded immune responses induced by interferon specific to viral infections (Table 3.1). Module 39 was associated with type I interferon signalling pathway (GO:0060337) and Antiviral mechanism by IFN-stimulated genes (R-HSA-1169410). Modules 81 and 92, taken together, contained all the classical and non-classical MHC class I genes. Module 92 also contains *B2M*, which is another component of the MHC (Table 3.1).

Many of the modules were associated with either neutrophil, lymphocyte, or monocyte proportions. Since these proportions are necessarily anticorrelated, it was unclear which particular cell type drove the association with any given module. To address this challenge and to increase the panel of tested primary immune cell types, cell frequencies derived from CIBERSORTx were used as proxies (Figure 3.6). These scores are on an absolute scale and demonstrated that different modules were associated with variation in the frequency of different cell types.

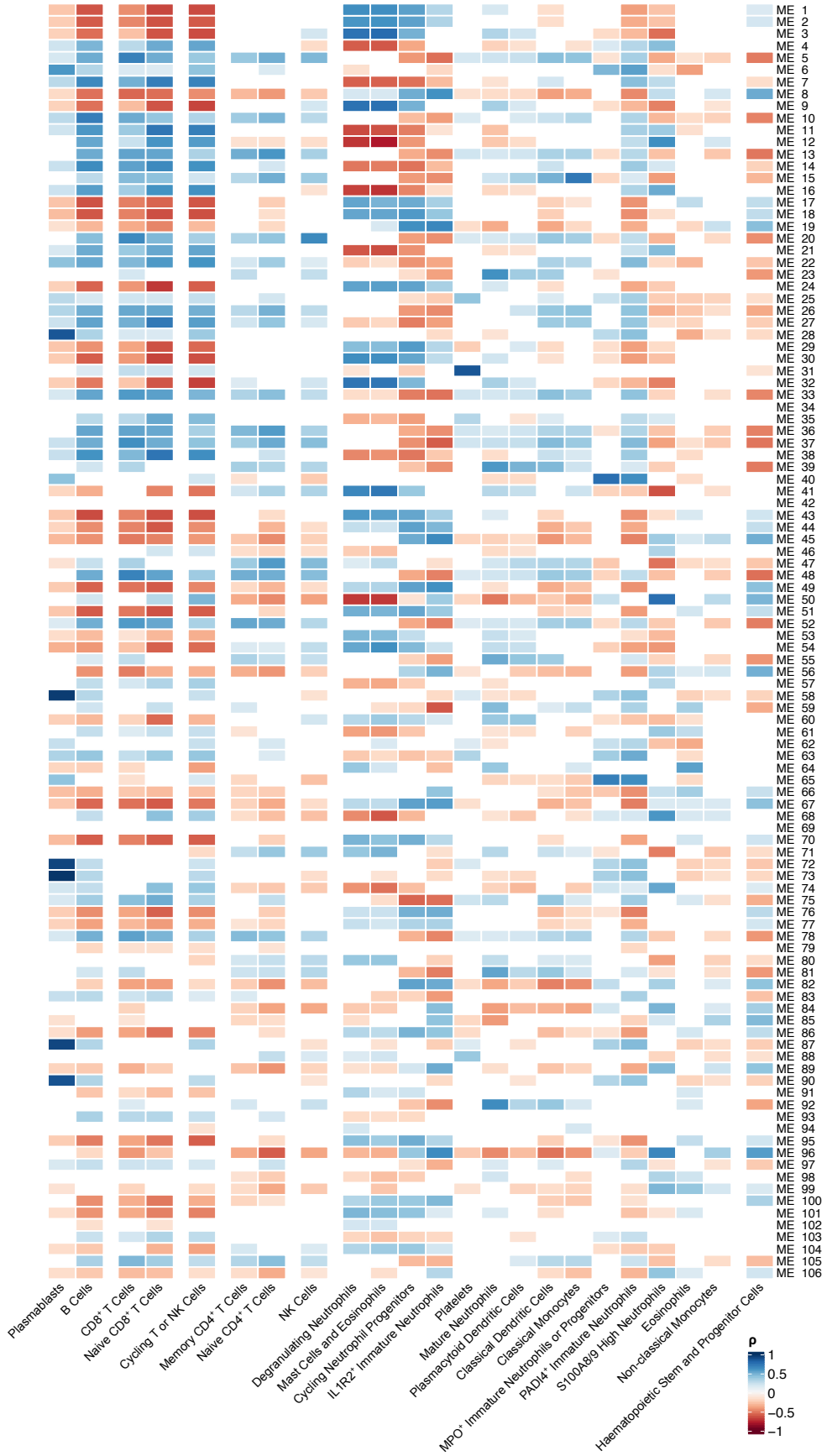

Figure 3.6: Association of module eigengenes and inferred cell frequencies. Cell frequencies inferred using CIBERSORTx were used as proxies for cell type proportion measurements in whole blood. Spearman's Rho was used to test for association. Only associations passing the p-value threshold are displayed in this heatmap.

3.1.3 Module Networks

Modules can be represented as networks to better understand the relationships between component genes. Module 51 is of particular interest because it contains members of the HIF-1 signalling pathway and is associated with endotoxin tolerance in monocytes. Selecting the genes associated with glycolysis and NFκB regulation present in this module (Figure 3.7) reveals that the glycolysis enzymes are positively correlated with each other and with *HIF1A* expression. The expression patterns between *HIF1A* and genes associated with NFκB signalling are less clear. The direction of association with *COMMD8*, *FPR1*, and *SMAD6* suggests a decrease in IκB and consequent increase in NFκB signalling (Migeotte *et al.* 2006; Starokadomskyy *et al.* 2013; Choi *et al.* 2006). In contrast, *HIF1A* is negatively correlated with *MIF*, which is a pro-inflammatory cytokine released into the bloodstream by leukocytes during infection and stress that is particularly responsible for reducing the immunosuppressive effects of glucocorticoids and increasing pro-inflammatory gene expression via NFκB by inhibiting the induction of IκB (Calandra *et al.* 2003).

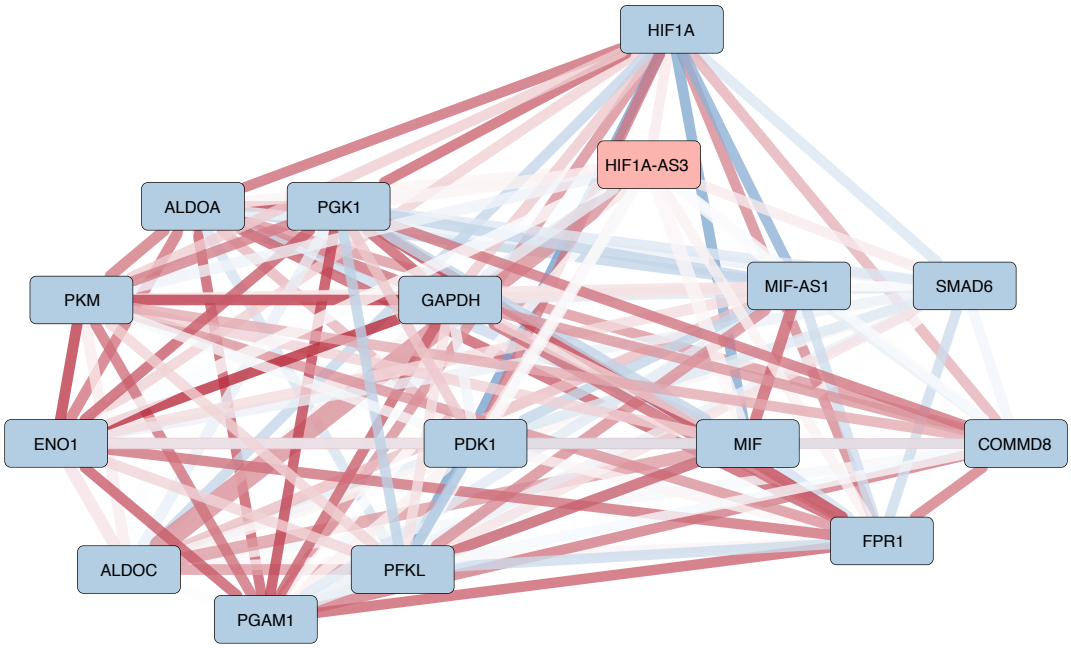

Figure 3.7: Module 51 HIF-1 pathway. This figure is a network diagram of selected genes from module 51. Edge colour represents correlation between gene expression profiles (blue being negative and red being positive). Edge size represents the edge weight in the adjacency matrix. Module 51 contains *HIF1A*. The hub gene in this module was *HIF1A-AS3*, an antisense RNA for *HIF1A*. The pathway contains *HIF1A* and downstream enzymes that are upregulated during glycolysis (Left). The module also contained markers of inflammation such as *MIF* that were negatively correlated with *HIF1A* expression.

Module 92, which was associated with diagnosis, contains both MHC class I molecules (*HLA-*

B, *HLA-C*, *HLA-E*, and *B2M*) and butyrophilin family proteins (Figure 3.8). The genes in these related families are strongly correlated with each other. Interestingly, the module contains *NLRC5*, which simultaneously inhibits NFκB activation (Cui *et al.* 2010) and activates MHC class I genes (Meissner *et al.* 2010). The butyrophilin genes are also present in the MHC class I region but are negatively correlated with *NLRC5*. Butyrophilins are expressed on many immune cell types and can act as either activators or inhibitors of T cells (Arnett *et al.* 2014).

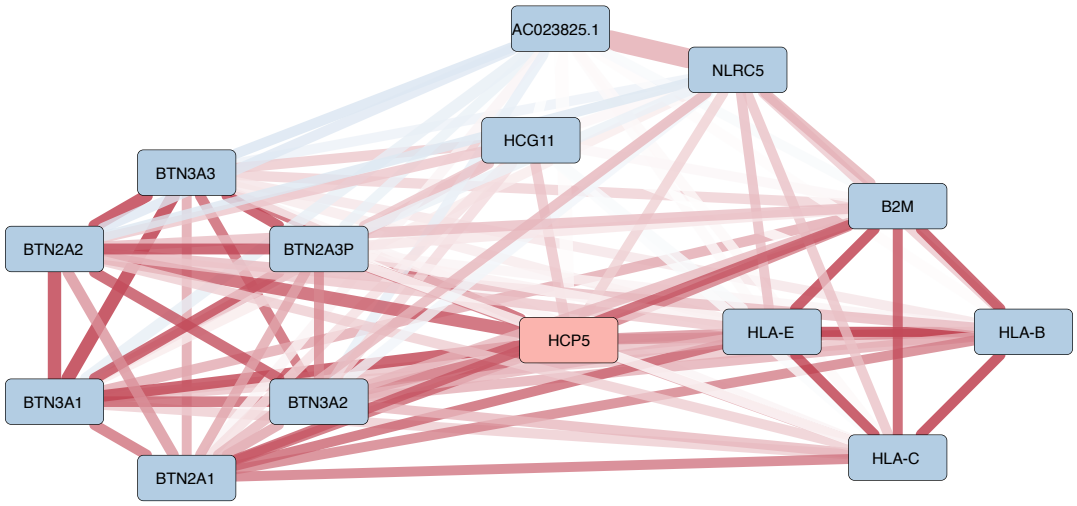

Figure 3.8: Module 92. This figure is a network diagram of selected genes from module 92. Edge colour represents correlation between gene expression profiles (blue being negative and red being positive). Edge size represents the edge weight in the adjacency matrix. Module 92 contains MHC class I members of the HLA complex (Right). It also contains members of the butyrophilin family that suppress interaction of T cells and APCs (Left). *HCP5* was the hub gene in this module.

3.2 Module QTL

Although a *trans*-eQTL analysis is an attractive approach to identify *trans* regulators of disease-associated molecular heterogeneity in sepsis, such a genome-wide analysis is underpowered in the GAinS cohort. As a proxy, the module eigengenes were mapped to identify drivers of broad transcriptomic programs that were called module QTL[1]. The specific hypothesis that a single SNP is associated with the expression of multiple distal genes via regulation of a shared upstream transcription factor in *cis* was tested by using a set of SNPs consisting of lead *cis*-eQTL, lead conditional *cis*-eQTL, and trait-associated SNPs from the EBI GWAS Catalog. The *cis*-eQTL were previously identified in the GAinS cohort[2].

[1] Described in Section 2.3.1
[2] Described in Section 1.5.3

Using the module eigengenes as the phenotype identified associations for 31 modules. This included 876 module QTL across 31 loci (Figure 3.9). Loci were defined as approximately 2 Mb disjoint intervals over the genome containing module QTL[1]. There was one locus on chromosome 3 associated with two modules, resulting in 32 module-locus pairs.

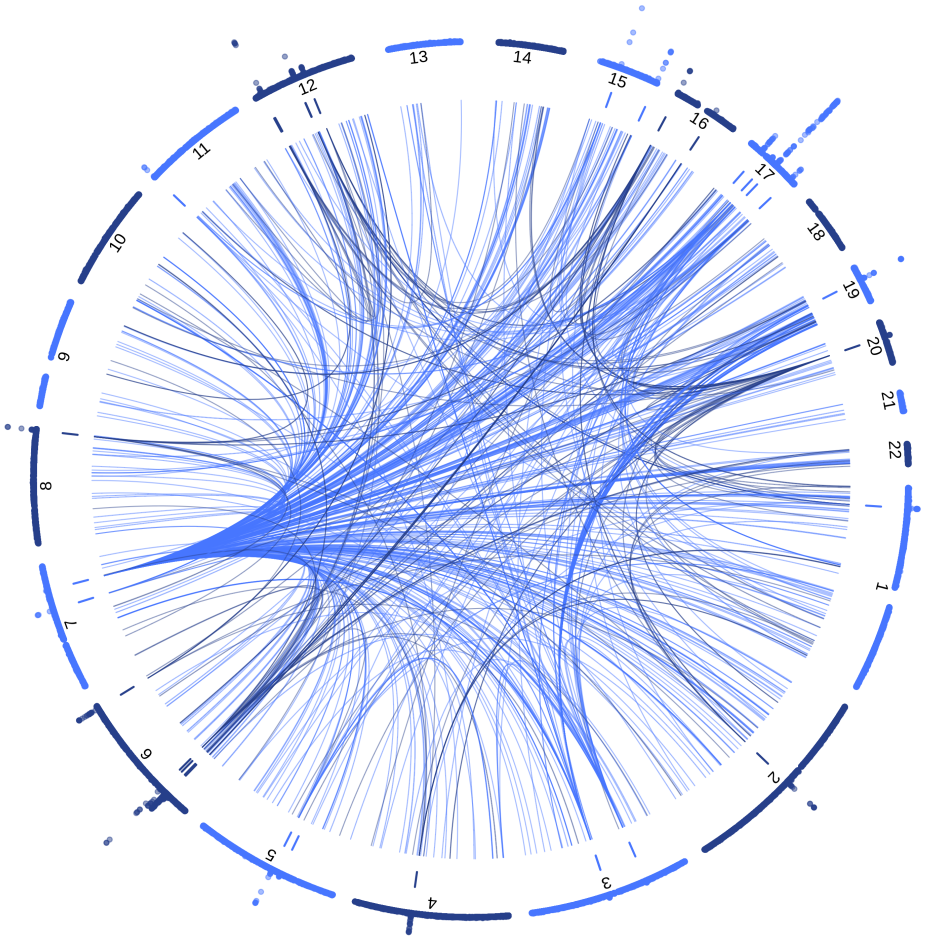

Figure 3.9: Module QTL from module eigengenes. Association with module eigengenes was tested genome-wide using lead *cis*-eQTL, lead conditional *cis*-eQTL, and trait-associated variants from the EBI GWAS Catalog. The analysis identified 31 module QTL loci (inner track), which consist of more than one module QTL in the same region. The outer track is the genome-wide Manhattan plot for each chromosome. Links originating from each module QTL locus represent genes in the module associated with the locus that are not on the same chromosome.

The input set of 70,300 SNPs contained 9,941 (14.1%) lead *cis*-eQTL, 14,937 (21.2%) lead conditional *cis*-eQTL, and 55,550 (79.0%) trait-associated variants. Of all the module QTL identified, 139 (15.9%) were previously-identified lead *cis*-eQTL, 236 (26.9%) were previously-identified lead

[1] Described in Section 2.3.1

conditional *cis*-eQTL, and 657 (75.0%) were SNPs from the EBI GWAS Catalog (Figure 3.10). Of the 31 modules with module QTL, 28 modules were associated with a *cis*-eQTL and also contained the corresponding eGene.

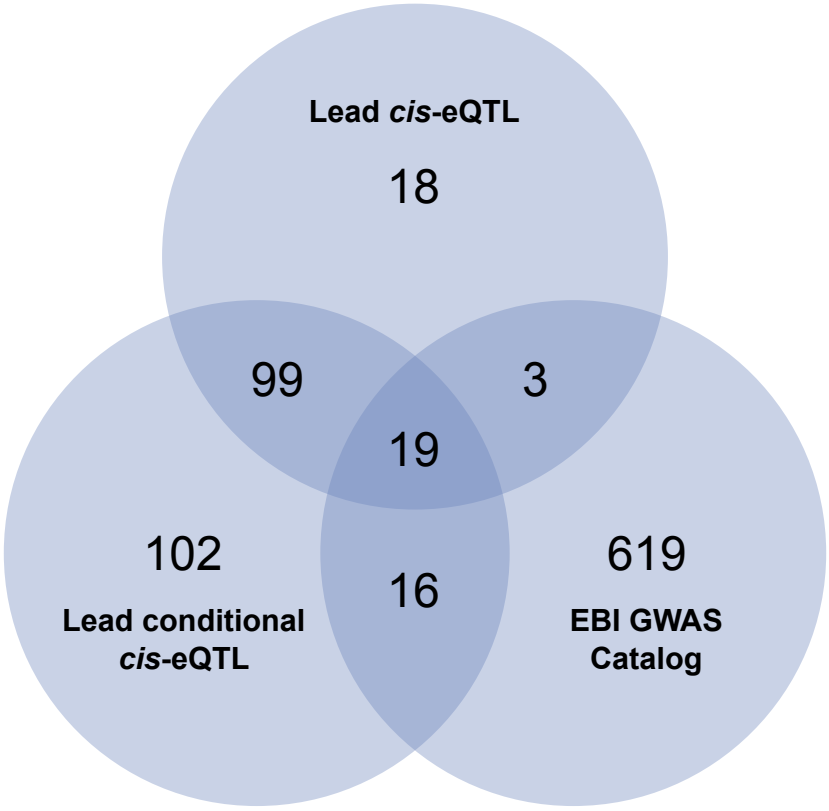

Figure 3.10: Composition of module QTL. Module QTL were identified from a test set of SNPs consisting of lead *cis*-eQTL, lead conditional *cis*-eQTL, and trait-associated variants from the EBI GWAS Catalog. Thus, module QTL have been previously associated with gene expression and/or a trait.

3.2.1 Multiple Module Eigengenes

Some have argued that the power to detect module QTL can be increased by testing against multiple gene expression PCs from each module (Wang *et al.* 2022b). To test this, 4 additional PCs were calculated for each module to create a set of 5 module eigengenes per module. Using the top 5 module eigengenes identified associations with a total of 48 modules. This analysis identified 1,935 module QTL across 76 loci (Figure 3.11). Due to the decreased p-value threshold, associations for 13 module QTL from the initial analysis were lost, including all module QTL for module 14. All loci were associated with one module, resulting in 76 module-locus pairs.

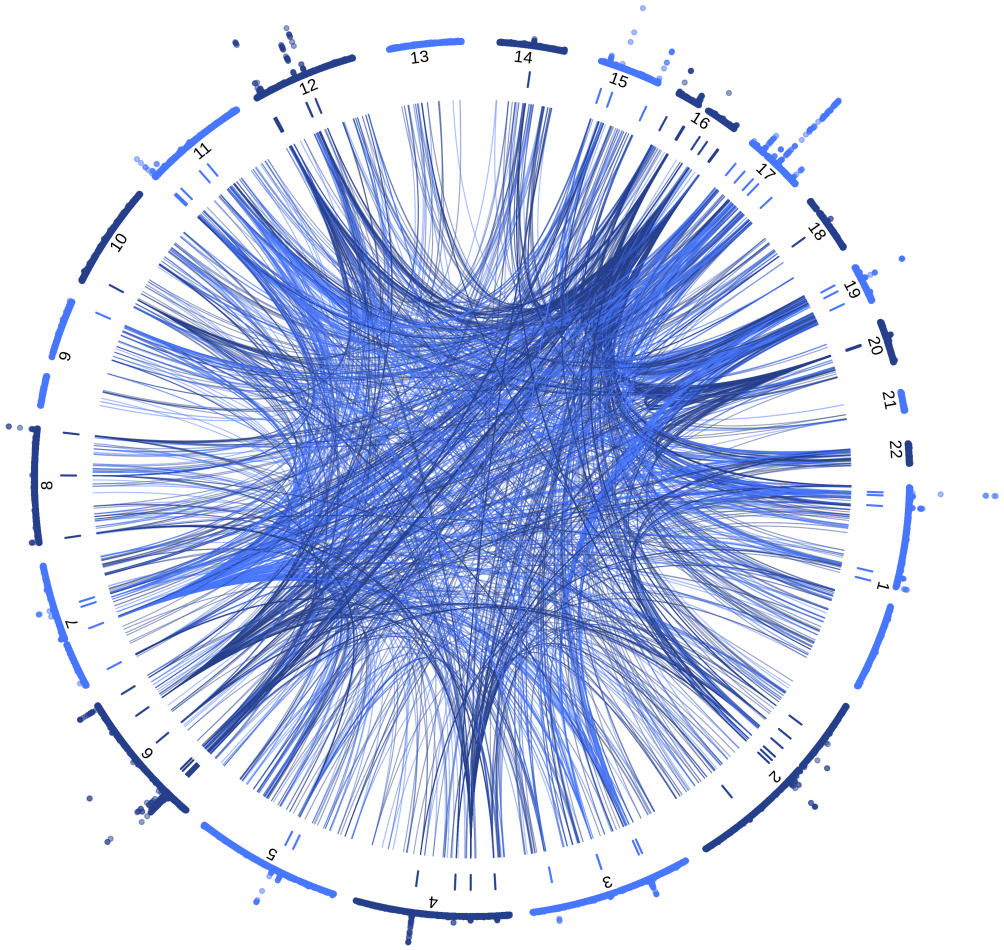

Figure 3.11: Module QTL from top five module eigengenes. To increase power, the top five module eigengenes for each module were tested genome-wide using lead *cis*-eQTL, lead conditional *cis*-eQTL, and trait-associated variants from the EBI GWAS Catalog. The analysis identified 76 module QTL loci (inner track), which consist of more than one module QTL in the same region. The outer track is the genome-wide Manhattan plot for each chromosome. Links originating from each module QTL locus represent genes in the module associated with the locus that are not on the same chromosome.

Of all the module QTL identified, 292 (15.1%) were previously-identified lead *cis*-eQTL, 486 (25.1%) were previously-identified lead conditional *cis*-eQTL, and 1,479 (76.4%) were SNPs from the EBI GWAS Catalog (Figure 3.12). Of the 48 modules with module QTL, 45 modules were associated with a *cis*-eQTL and also contained the corresponding eGene.

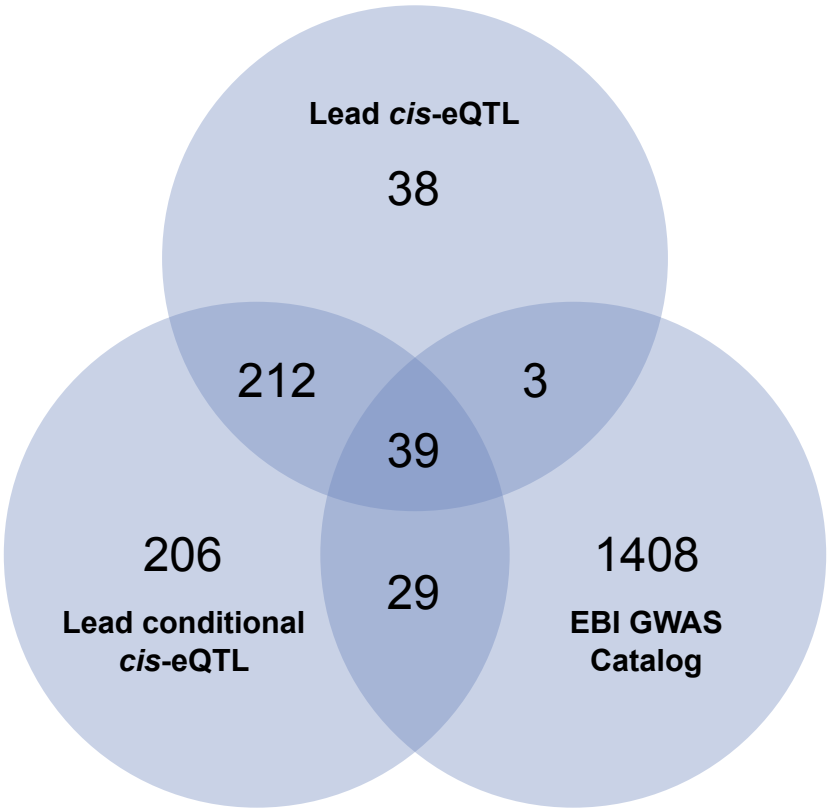

Figure 3.12: Composition of module QTL from the top five module eigengenes. Module QTL derived from the use of multiple module eigengenes were identified from a test set of SNPs consisting of lead *cis*-eQTL, lead conditional *cis*-eQTL, and trait-associated variants from the EBI GWAS Catalog. Thus, module QTL have been previously associated with gene expression and/or a trait.

3.2.2 Trait-Associated Variants

Certain traits relevant to sepsis and immune-mediated diseases (IMDs) were curated from the EBI GWAS Catalog (Tables 3.2 and B.2). Traits of interest were grouped into susceptibility to infection, serum proteins, leukocyte traits, and autoimmune diseases. 23 (47.9%) of the 48 modules with module QTL had module QTL that were previously associated with these EBI GWAS traits. Modules 81 and 92 contained HLA genes. QTL for these modules were previously associated with susceptibility to various infections (Table 3.2). QTL for module 84, which contained the genes for complement protein C4, were also associated with susceptibility to infections (Table 3.2). Serum biomarkers and autoimmune diseases have also been previously associated with module QTL (Table 3.2).

Table 3.2: IMD-relevant traits in the EBI GWAS Catalog. Some module QTL were SNPs from the EBI GWAS Catalog. Selected traits relevant to IMDs and their associated modules are listed in this table. Specific studies used are listed in Table B.2.

Trait Group	Trait	Modules
Susceptibility to Infection	HIV-1	84, 92, 97
	Chickenpox	81
	Shingles	81, 84
	Epstein-Barr Virus	84
	Hepatitis B Virus	84
	Hepatitis C Virus	84
	Mononucleosis	84
	Mycobacterium tuberculosis	84
	Pneumonia	84
	Scarlet Fever	84
Serum Proteins	C-reactive Protein (Inflammation Marker)	84, 94, 103
	Alanine Aminotransferase (Liver Function)	62, 101, 102
	Aspartate Aminotransferase (Liver Function)	63, 69, 81, 101, 102
	Albumin	69, 102
	Urate	69, 81, 84, 92, 94
	Alkaline Phosphatase	69, 106
	Creatinine	84, 94
	Cystatin C (Kidney Function)	64, 94, 102
	IgM	81
	IgA	62, 84
	IgE	81, 84
	Beta-2-microglobulin	84
	Complement C4	84
	IgG Glycosylation	88
	Interleukin 18	89
	Interleukin 1β	97
Primary Cell Trait	Lymphocyte Count / Proportion	69, 81, 84, 92, 97, 101, 103
	Neutrophil Count / Proportion	69, 81, 84, 92, 101, 103
	Monocyte Count / Proportion	75, 81, 84, 103, 106
	Eosinophil Count / Proportion	59, 69, 81, 84, 94, 99, 101, 106
	Basophil Count / Proportion	81, 84
	Platelet Count	59, 62, 63, 81, 84, 86, 88, 91, 97, 102, 104
	Erythrocyte Count	62, 69, 80, 81, 82, 91, 102

Continued on next page

Trait Group	Trait	Modules
Autoimmune Disease	Rheumatoid Arthritis	62, 84, 99
	Inflammatory Bowel Disease	71, 84, 91, 94, 99
	Coeliac Disease	84
	Psoriasis	84, 92
	Systemic Lupus Erythematosus	81, 84, 92
	Multiple Sclerosis	81, 84, 102
	Alopecia Areata	84, 101
	Vitiligo	81, 84, 99, 101
	Type 1 Diabetes	62, 69, 84, 91, 101
	Graves Disease	81, 84, 97, 99
	Myasthenia Gravis	84

3.2.3 Module QTL Replication

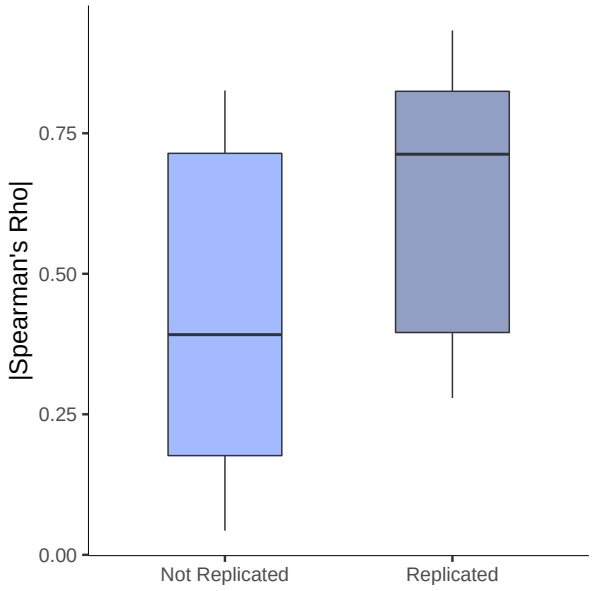

Figure 3.13: Replication of module eigengenes. Modules were reconstructed in an older microarray data set from the GAinS cohort with 134 overlapping samples. Similarity between modules from the two data sets was tested using Spearman's Rho for the 134 overlapping samples. Module QTL that were replicated in the microarray data set tended to have better correlated module eigengenes between the two data sets (Wilcoxon Rank Sum Test with Continuity Correction; p = 0.1118).

The original transcriptomic study in the GAinS cohort contained a subset of patients not included in the RNA-seq data set. Microarray gene expression data for this subset of patients was used

as an independent replication cohort for the module QTL[1]. Only module QTL associated with the first module eigengene were tested for replication. Since RNA-seq and microarrays are fundamentally different technologies and some expressed genes from RNA-seq are not assayed on the microarray, not all discovered modules could be replicated. Of the 31 lead module QTL, 26 could be tested for replication. Of these, 17 (65.4%) lead module QTL from the original analysis were significantly associated with the microarray module. A subset of 14 (53.8%) also matched in the direction of effect. Spearman's Rho was calculated as a measure of reproducibility between module eigengenes from the RNA-seq and microarray data based on shared samples. Comparing these values between module QTL that were replicated and not replicated (Figure 3.13) suggested that modules that were more consistent between data sets were more likely to also have replicable module QTL (Wilcoxon Rank Sum Test with Continuity Correction; p = 0.1118).

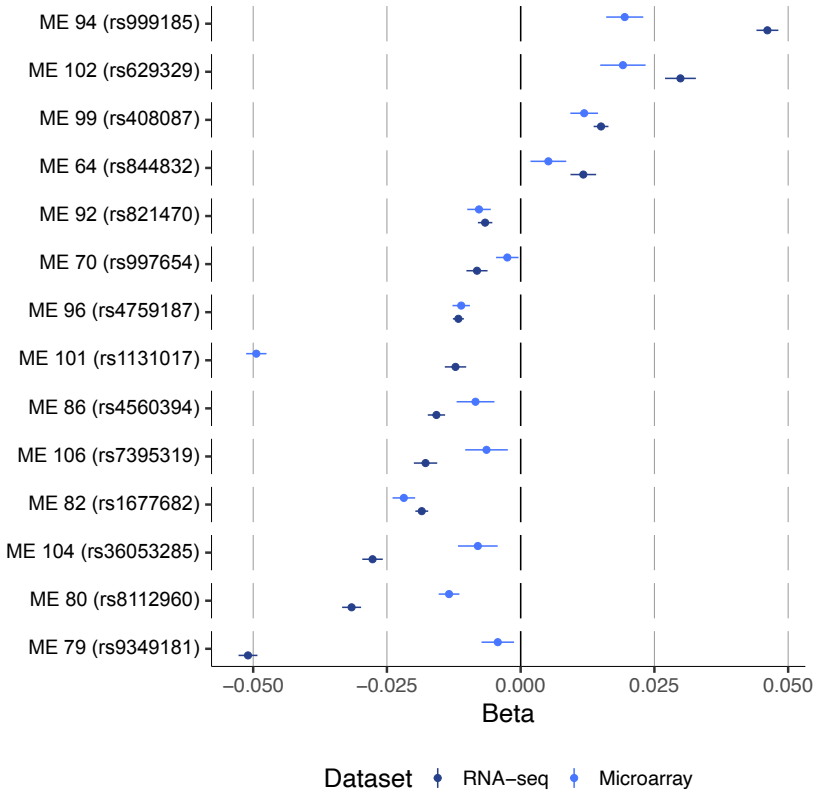

Figure 3.14: Forest plot of replicated effects. Of the 26 module QTL that could be tested, 14 replicated based on the lead module QTL. Effect sizes from the original RNA-seq data set and the replication microarray data set are displayed as points with 95% confidence intervals. The direction of the effect size from the replication analysis was multiplied by the sign of Spearman's Rho measured between the module eigengenes across the two data sets.

Due to regression towards the mean, replication effect sizes were expected to be smaller in magnitude than the original effect sizes. This pattern was indeed observed for the 14 replicated

[1] Described in Section 2.3.2

module QTL (Figure 3.14).

3.3 Discussion

In this chapter, I have identified co-expression modules from the gene expression data for the GAinS cohort. These modules identified gene networks relevant to the immune dysfunction that has been previously characterised in sepsis. I used these modules to identify variants associated with broad expression patterns that I called module QTL. These module QTL have been previously identified as trait-associated variants or *cis*-eQTL and are replicable in an earlier sepsis data set.

3.3.1 Co-expression Modules

The analysis identified module 31, which contained gene markers for platelets. The method used to isolate whole blood leukocytes[1] does not isolate platelets. The detection of RNA for platelet markers and known platelet-associated factors may be due to a high platelet load in septic patients or extracellular necrotic content that was not entirely filtered out. These markers and platelet factors may also be expressed in other leukocytes.

Module 51 contained *HIF1A* and many other factors involved in the hypoxic shift towards glycolysis that is characteristic of endotoxin tolerance. In macrophages, endotoxin tolerance is a brief state of hyporesponsiveness after sustained exposure to lipopolysaccharide (LPS). Samples in the SRS1 endotype showed enrichment of a gene expression signature associated with endotoxin tolerance, suggesting that macrophages in these patients may be immunosuppressed and in a hyporesponsive state (Davenport *et al.* 2016). The same study also identified HIF1α and the hypoxia pathway as differentially expressed between transcriptomic endotypes. The shift to glycolysis (The Warburg effect) and the hypoxic response driven by HIF1α in macrophages is important for the initial host immune response to infection and promotes pro-inflammatory gene expression programs (Tannahill *et al.* 2013). High levels of HIF1α, on the other hand, are associated with a immunosuppressive phenotype driven by suppression of TLRs via IRAKM and eventual endotoxin tolerance (Shalova *et al.* 2015). In addition to *HIF1A* and members of the glycolysis pathway, module 51 also contained multiple regulators of NFκB and IκB, suggesting that macrophage immunosuppression may be tied to the inhibitory activity of IκB.

[1] Described in Section A.3

3.3.2 Relationships between Modules and Clinical Variables

Prior work in sepsis has indicated that the underlying dysregulated immune response is largely common between sources of infections (Burnham *et al.* 2017) and may also be present in non-infectious sources of trauma (Xiao *et al.* 2011; Scicluna *et al.* 2015). However, in addition to this shared host response, there are source-specific responses that are also present in sepsis. While 78% of DE genes in whole blood leukocytes between sepsis and healthy subjects are common between pulmonary and abdominal infections, some enriched pathways differ between the two. Furthermore, mortality can also vary depending on the source of the infection (Peters-Sengers *et al.* 2022). Our transcriptomic analysis identified 4 modules that were strongly associated with source (CAP and FP). These modules encode pathways related to MHC class I antigen presentation and antiviral responses (Table 3.1). Thus, these modules may represent molecular variation due to different infecting pathogens between patients with CAP or FP. Although variation in mortality was expected, none of these modules were associated with outcome.

The modules are cell-type-specific, which is especially apparent when using gene markers from the sepsis scRNA-seq analysis (Kwok *et al.* 2022). Future work on these modules is to compare them directly with modules generated from the scRNA-seq data. Deconvolution methods such as CIBERSORTx are also capable of inferring gene expression profiles for individual cell types from bulk samples. These can be used to extend the module analysis to specific cell types in the larger bulk RNA-seq cohort. The scRNA-seq data was initially required to identify which modules captured variation from specific sepsis neutrophil subsets, since publicly available data do not capture pathological signatures of neutrophils. Now that these modules have been identified, they can be used to look for upregulation of particular sepsis neutrophil signatures in other bulk data sets.

3.3.3 Module QTL

By design, trait-associated variants were included in the module QTL analysis. We identified many QTL that were previously associated with traits related to IMDs. Although suggestive, a more rigorous analysis involves using statistical colocalisation methods to test if the pattern of association for these traits is consistent with the associations we identified. This analysis is conducted in chapter 4 of this thesis. Regardless, this initial analysis reveals that the module QTL have the potential to reveal interesting biology underlying pathological immune conditions.

The initial hypothesis when performing module QTL mapping was that they would identify *trans* factors that are regulated in *cis*. Thus, module QTL can be considered a form of *trans*-eQTL.

There are multiple mechanisms by which a *trans*-eQTL or module QTL may regulate associated genes. These include direct regulation via transcription factors (TFs), indirect regulation via TFs, co-regulation within the gene network, and protein-protein interactions (Võsa *et al.* 2021). Determining which mechanisms underlie specific module QTL requires functional interpretation of the associated variants and the gene network captured by the module. Some of this mechanistic interpretation is conducted in chapter 5 of this thesis.

53.8% of testable module QTL replicated in the original microarray data from GAinS. The method has some limitations. The microarray was only designed to assay a certain subset of genes, necessarily resulting in smaller modules that might not contain key genes required to reconstruct the signal of the original module eigengene. In addition, the method of quantifying transcript levels is on a relative scale in the microarray, compared to the absolute scale of the RNA-seq data. A future step for the replication analysis is to use an independent healthy cohort and sepsis cohort with paired genotype and RNA-seq data.

Next steps for the module QTL analysis are to explore if any specific category of traits is enriched in the module QTL compared to the input set of SNPs and to validate the approach in larger cohorts of gene expression in blood. The latter is especially relevant because the larger modules had fewer associations than the smaller modules, likely because the cohort was underpowered to detect small effects on a large set of genes. Other similar initiatives for sepsis, such as the Molecular Diagnosis and Risk Stratification of Sepsis (MARS) consortium (Scicluna *et al.* 2017), can also be used for further validation in the disease context. As with any QTL analysis in blood, these module QTL may be confounded by differences in cell proportion. This affects sepsis studies due to the large expansion of neutrophils during the acute phase of the infection. Identifying interaction module QTL (Zhernakova *et al.* 2017; Wijst *et al.* 2018), where the effect size is magnified or diminished based on cell proportion, can account for this confounding and identify cell types that are relevant to the QTL mechanism. Interaction module QTL for cell proportions may also identify module QTL that are actually cell proportion QTL that were detected because gene expression acted as a proxy for cell frequency across the samples.

4 | Colocalisation and Fine Mapping

The aim of this chapter is to leverage patterns of association across various molecular traits to gain mechanistic insight into variants associated with molecular expression. I colocalised eQTL with pQTL and module QTL to identify shared and distinct signals underlying these associations. I then used statistical fine mapping to refine the set of variants associated with each molecular trait.

4.1 Colocalisation of *cis*-eQTL

Colocalisation was performed between various associations using the COLOC R package with a predefined criteria for colocalisation[1]. Notably, 889 (6.0%) of the 14,938 independent *cis*-eQTL lead SNPs were associated with more than one eGene (Figure 4.1). To test the hypothesis that these loci represent a common functional element mediating gene expression of multiple nearby genes, colocalisation was performed between all 1,467 pairs of eGenes that shared a lead conditional *cis*-eQTL.

[1] Described in Section 2.4

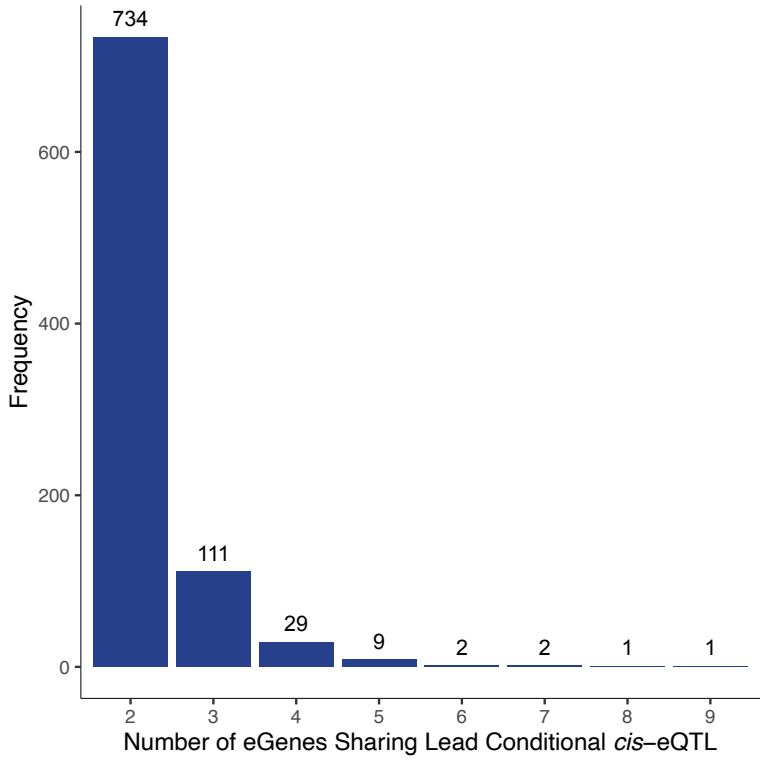

Figure 4.1: Number of eGenes sharing lead conditional *cis*-eQTL. 889 lead conditional *cis*-eQTL were shared between multiple eGenes. The distribution of shared eGenes showed that most (734) of the 889 cases involved sharing of the lead conditional *cis*-eQTL between two genes. Up to 9 eGenes were found to share a common lead conditional *cis*-eQTL.

Of the set of 889 lead conditional *cis*-eQTL that were shared between eGenes, 871 showed evidence of colocalisation between all pairs of shared eGenes and two additional *cis*-eQTL showed evidence of colocalisation between at least one pair of shared eGenes. The 871 lead conditional *cis*-eQTL, representing 1,435 pairs of colocalisations, were used as a confident set of *cis*-eQTL mediating the expression of multiple genes for further analysis. A notable example was the colocalisation of conditional *cis*-eQTL for members of the T cell receptor (TCR) β chain (Figure 4.2).

Under the hypothesis that these conditional *cis*-eQTL were genomic elements controlling expression of multiple genes, it was expected that certain modules may capture the co-expression patterns of these genes. Indeed, 77 of the 871 lead conditional *cis*-eQTL with shared eGenes were also module QTL. These module QTL were associated with 37 of the 48 modules with QTL.

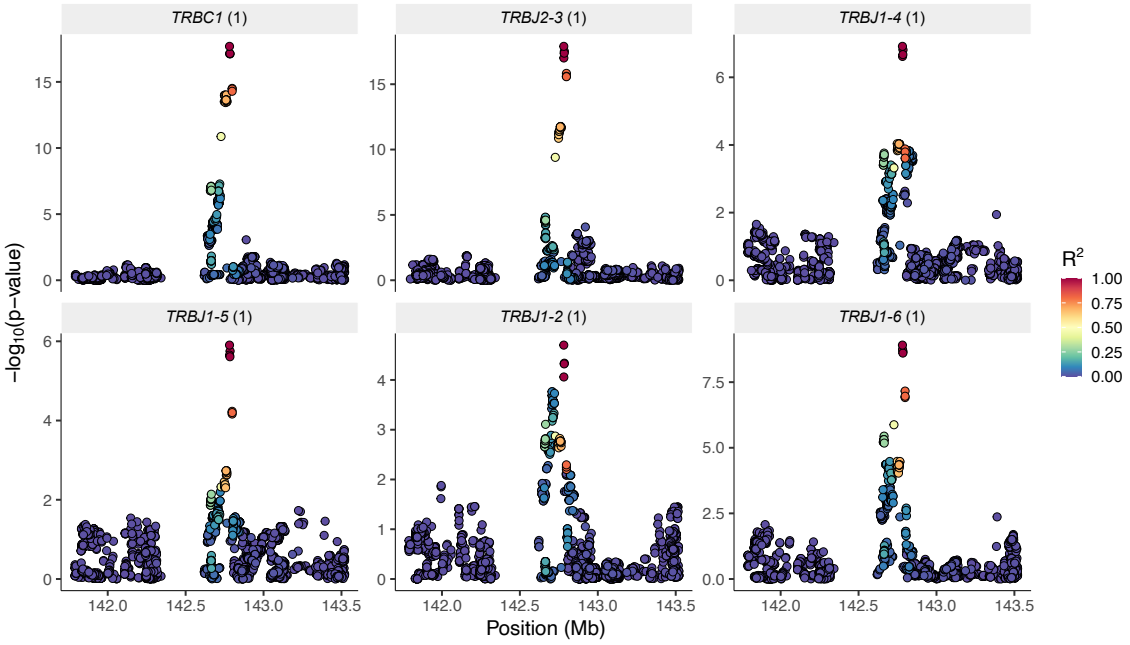

Figure 4.2: Colocalising *cis*-eQTL of components of the TCR β chain. These six eGenes shared the same lead conditional *cis*-eQTL and all colocalise with each other. They are members of the TCR β chain that undergo somatic recombination, including one constant region and five joining regions.

4.2 Colocalisation of *cis*-eQTL and module QTL

Many of the cases of colocalising *cis*-eQTL were found to be module QTL. To test if the same causal SNP was responsible for both, I colocalised all *cis*-eQTL overlapping any of the 76 module QTL loci detected from associations with the top 5 module eigengenes. Overall, this involved testing 12,135 pairs of module eigengenes and eGenes, of which 824 colocalised. This included a total of 361 eGenes and a total of 74 module QTL loci.

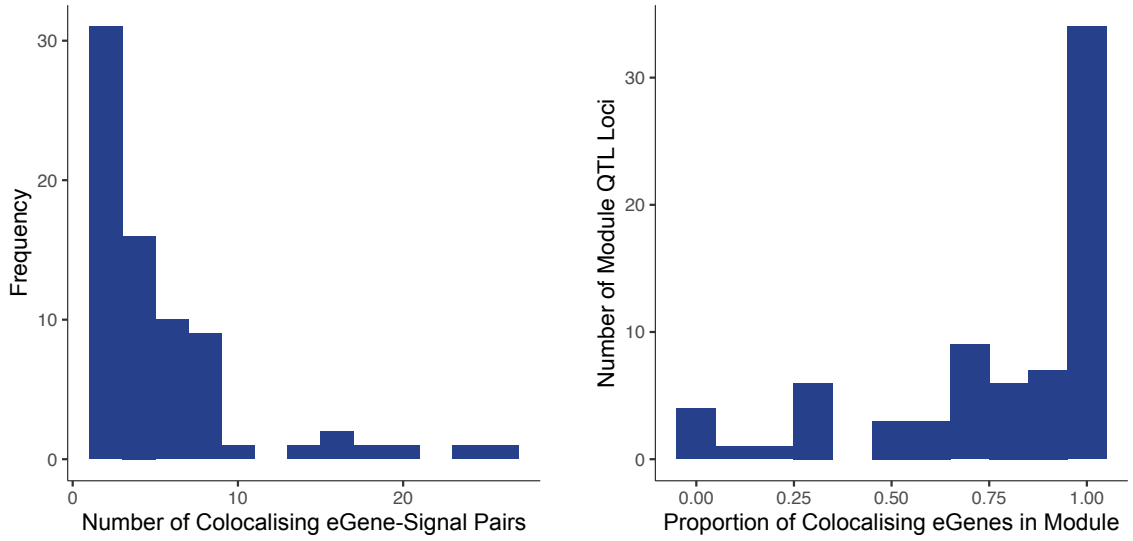

Figure 4.3: Distribution of *cis*-eQTL colocalising with a module QTL. (Left) The median number of eGene-signal pairs colocalising with a given module QTL locus was 4 (range 1 to 27). (Right) The median proportion of colocalising eGenes originating from the colocalising module was 88.9%.

The module QTL loci colocalised with a median of 4 eGene-signal pairs (Figure 4.3). At 70 of the 74 loci, at least one of the colocalising eGenes was in the module. The median proportion of eGenes colocalising with a particular module QTL that were also present in the module was 88.9% (Figure 4.3). These results indicate that module QTL loci that colocalised with eGenes were composed of common *cis*-eQTL shared by multiple co-expressed eGenes.

4.3 Colocalisation of *cis*-eQTL and *cis*-pQTL

To test for cases where the causal SNP may be shared between a pGene and its cognate eGene, which would suggest a common functional element regulating molecular expression in different tissues, colocalisation was performed between proteins and their cognate eGenes. Of the 269 proteins tested for pQTL, 258 were annotated for a unique gene and represented the proteins that could have *cis*-pQTL, with the remainder not annotated with a gene. One protein, neutrophil defensin 1 (DEFA1), was associated with two genes (*DEFA1* and *DEFA1B*). Of the 260 total genes for the 259 proteins, 97 were also eGenes. All 97 gene-protein pairs were tested for colocalisation, of which 14 had evidence of colocalisation. Within this set, 4 proteins were pGenes with genome-wide significant *cis*-pQTL (Table 4.1).

Category	Gene	Protein	Colocalisation Factor
eQTL and pQTL	*C4A*	Complement C4-A	0.992
Colocalisation	*IGLV6-57*	Immunoglobulin lambda variable 6-57	0.99
	MST1	Hepatocyte growth factor-like protein	0.722
	CFH	Complement factor H	0.707
eQTL and	*C1RL* (Signal 2)	Complement C1R subcomponent-like protein	0.985
Suggestive pQTL	*COL1A2*	Collagen alpha-2(I) chain	0.981
Colocalisation	*IGHG1*	Immunoglobulin heavy constant gamma 1	0.975
	C8G	Complement component C8 gamma chain	0.938
	HP	Haptoglobin	0.912
	F5	Coagulation factor V	0.897
	C4BPA	C4B-binding protein alpha chain	0.861
	PGLYRP2	N-acetylmuramoyl-L-alanine amidase	0.795
	A2M	Alpha-2-macroglobulin	0.759
	C1RL (Signal 1)	Complement C1R subcomponent-like protein	0.752
	PTGDS	Prostaglandin-H2 D-isomerase	0.72

Table 4.1: Colocalisation of *cis*-eQTL with *cis*-pQTL. All proteins with cognate eGenes were tested for colocalisation. Of the 14 proteins that colocalised, 4 were pGenes from the genome-wide scan for pQTL and 10 were potential pQTL that we were not powered to detect. The colocalisation factor was calculated as $\mathrm{PP4}/(\mathrm{PP3}+\mathrm{PP4})$ based on the posterior probabilities from COLOC.

The complementary analysis was to compare how many of the discovered genome-wide *cis*-pQTL colocalised with *cis*-eQTL. Of the 23 pGenes with *cis*-pQTL, 4 pGenes colocalised with their cognate eGenes as described above. 6 pGenes had a cognate eGene but did not colocalise. 2 pGenes had a cognate gene that was expressed but was not an eGene. Finally, 11 pGenes had a cognate gene that was not expressed in whole blood leukocytes (Table 4.2).

Category	pGenes with *cis*-pQTL
eQTL Present and Colocalisation	C4A, IGLV6-57, HGFL, CFH
eQTL Present but No Colocalisation	PLTP, FCGR3B, KLKB1, SERPINA1, C4B, ORM2
No eQTL but Expressed in Whole Blood Leukocytes	APOE, F12
Not Expressed in Whole Blood Leukocytes	PON1, FGL1, KNG1, HRG, CPN1, AGT, SERPINA10, AHSG, ITIH3, CLEC3B, APOC4

Table 4.2: Proteins with *cis*-pQTL. Of the 23 pGenes with *cis*-pQTL, 4 colocalised with their cognate eGenes. Of the rest, 6 did not colocalise, 2 did not have a cognate eGene, and 11 were not expressed in whole blood leukocytes.

The 4 pGenes that colocalised with their cognate eGenes (Table 4.2) represented instances where the same regulatory elements may control abundance of mRNA in whole blood leukocytes and protein in plasma. This may be due to secretion of proteins by leukocytes into the plasma, due to necrotic release of proteins into serum, or due to common regulation across tissues. Com-

plement factor 4 (C4) has two isotypes, encoded by either *C4A* or *C4B*. Interestingly, the C4A *cis*-pQTL colocalise with the *C4A* *cis*-eQTL but the C4B *cis*-pQTL do not colocalise with the *C4B* *cis*-eQTL, suggesting that the isotypes may be secreted or degraded via different pathways.

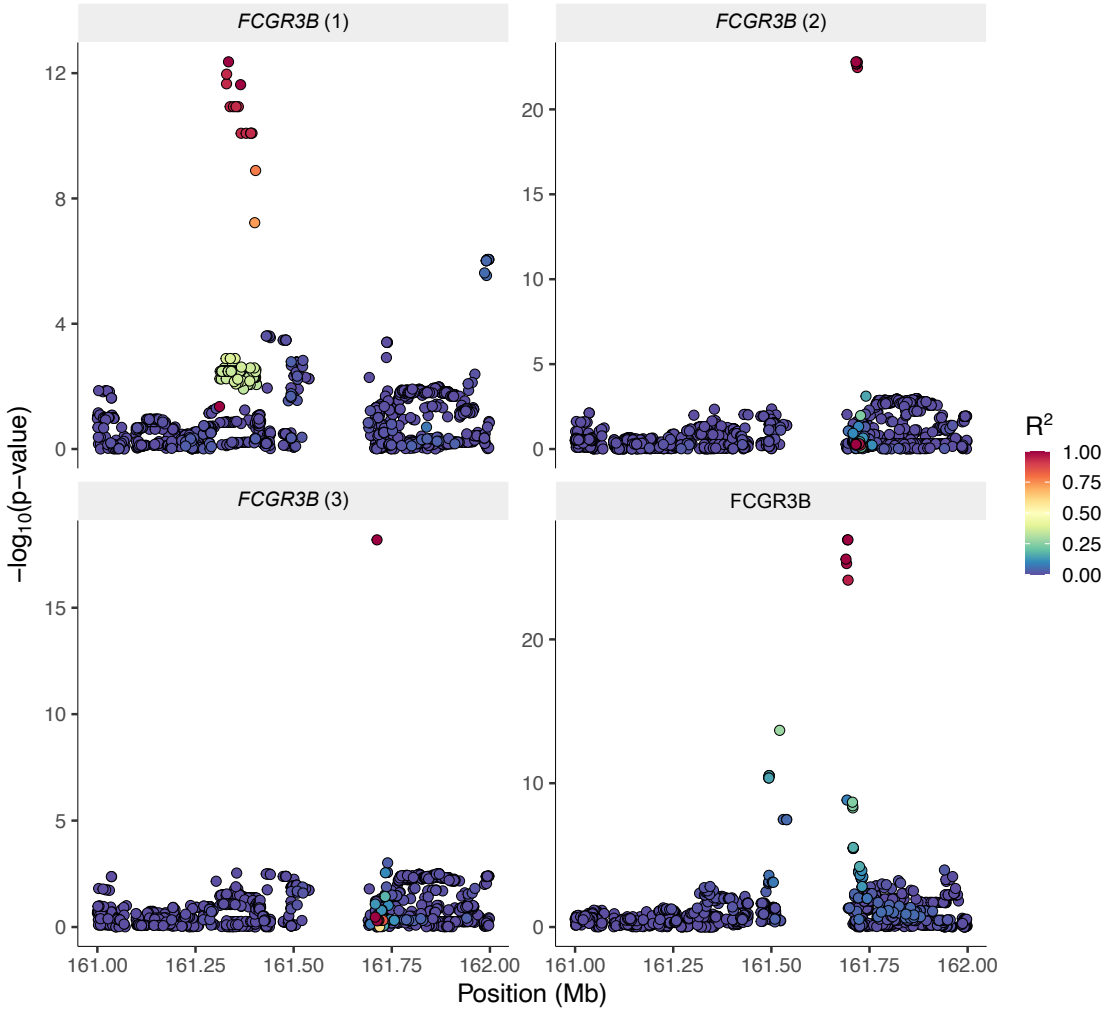

Figure 4.4: FCGR3B locus *cis*-eQTL and *cis*-pQTL. *FCGR3B* (1), *FCGR3B* (2), and *FCGR3B* (3) are the primary, secondary, and tertiary signals respectively detected from the conditional *cis*-eQTL analysis for *FCGR3B*. None of these *cis*-eQTL loci colocalise with the FCGR3B *cis*-pQTL locus.

The lack of colocalisation between eGenes and pGenes was also informative, such as the lack of colocalisation between *FCGR3B* expression in leukocytes and FCGR3B abundance in plasma (Figure 4.4). FCGR3B is a receptor for immunoglobulin G (IgG) that may be involved in the sequestration of IgG complexes without activation of neutrophils. The protein is primarily expressed on the surface of neutrophils (Chen *et al.* 2012). A lack of colocalisation suggests that the production and/or degradation of FCGR3B may occur in another tissue, and tissue-specific functional elements in that tissue may affect FCGR3B abundance in plasma. Lack of colocalisation with eQTL may also be due to different mechanisms underlying the variation in protein levels, such as

splicing.

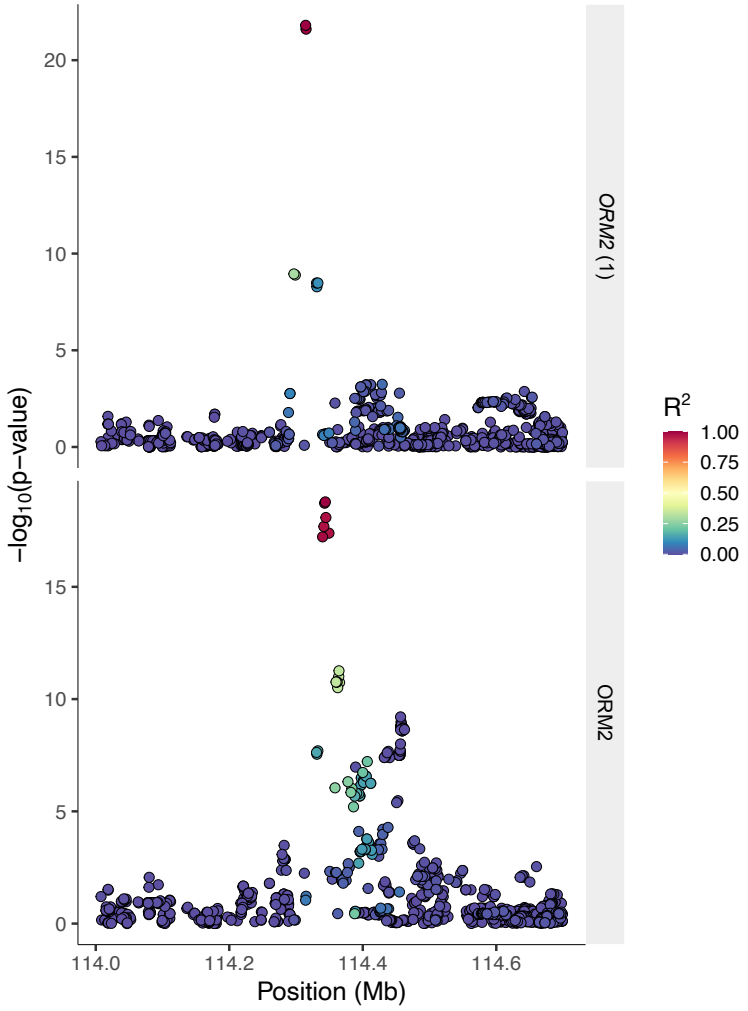

Figure 4.5: ORM2 locus *cis*-eQTL and *cis*-pQTL. (Top) Only one conditional *cis*-eQTL signal was detected for *ORM2*. (Bottom) This region did not colocalise with the ORM2 *cis*-pQTL.

ORM2 *cis*-pQTL are another example that did not colocalise with *ORM2 cis*-eQTL. ORM2 is released into the plasma during the acute phase of infection and is secreted by the liver, although there is also evidence of expression in endothelial cells and leukocytes (Sörensson *et al.* 1999). It is involved in maintaining capillary permeability (Haraldsson *et al.* 1987).

4.4 Colocalisation of *trans*-pQTL

Two *trans*-pQTL loci were present in the same genomic region on chromosome 16 and three *trans*-pQTL loci were present in the same genomic region on chromosome 14. To assess if an underlying effect in *cis* also accounted for the effect on these pGenes in *trans*, colocalisation was

performed between overlapping *trans*-pQTL and any *cis*-eQTL or *cis*-pQTL in the same region.

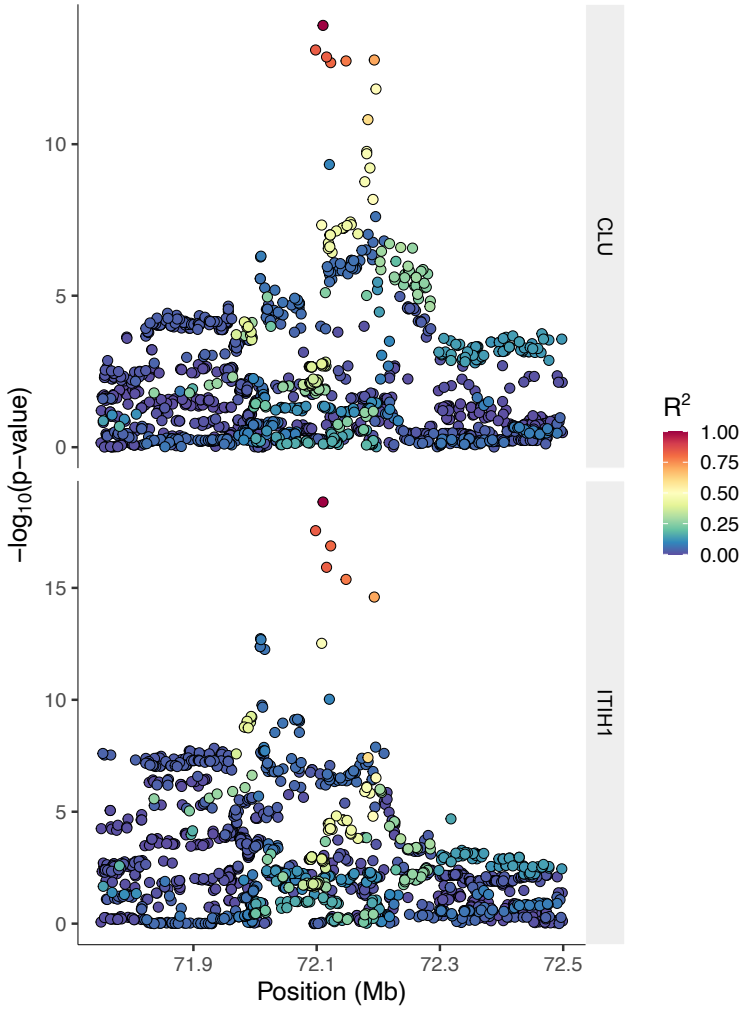

Figure 4.6: Chromosome 16 *trans*-pQTL. This figure contains association plots for the clusterin (CLU) and inter-alpha-trypsin inhibitor heavy chain H1 (ITIH1) *trans*-pQTL. These two loci colocalised with each other.

The chromosome 16 locus contained *trans*-pQTL for clusterin (CLU) and inter-alpha-trypsin inhibitor heavy chain H1 (ITIH1). The *trans*-pQTL for these two proteins colocalised (Figure 4.6). There were two *cis*-pQTL loci in the same region for haptoglobin (HP) and haptoglobin-related protein (HPR). Neither of the *trans*-pQTL loci colocalised with these *cis*-pQTL. There were 11 *cis*-eQTL loci in the same region. *EXOC6*, which encodes exocyst complex component 6, colocalised with CLU but not with ITIH1, making it an unlikely candidate for the underlying signal explaining both *trans*-pQTL loci.

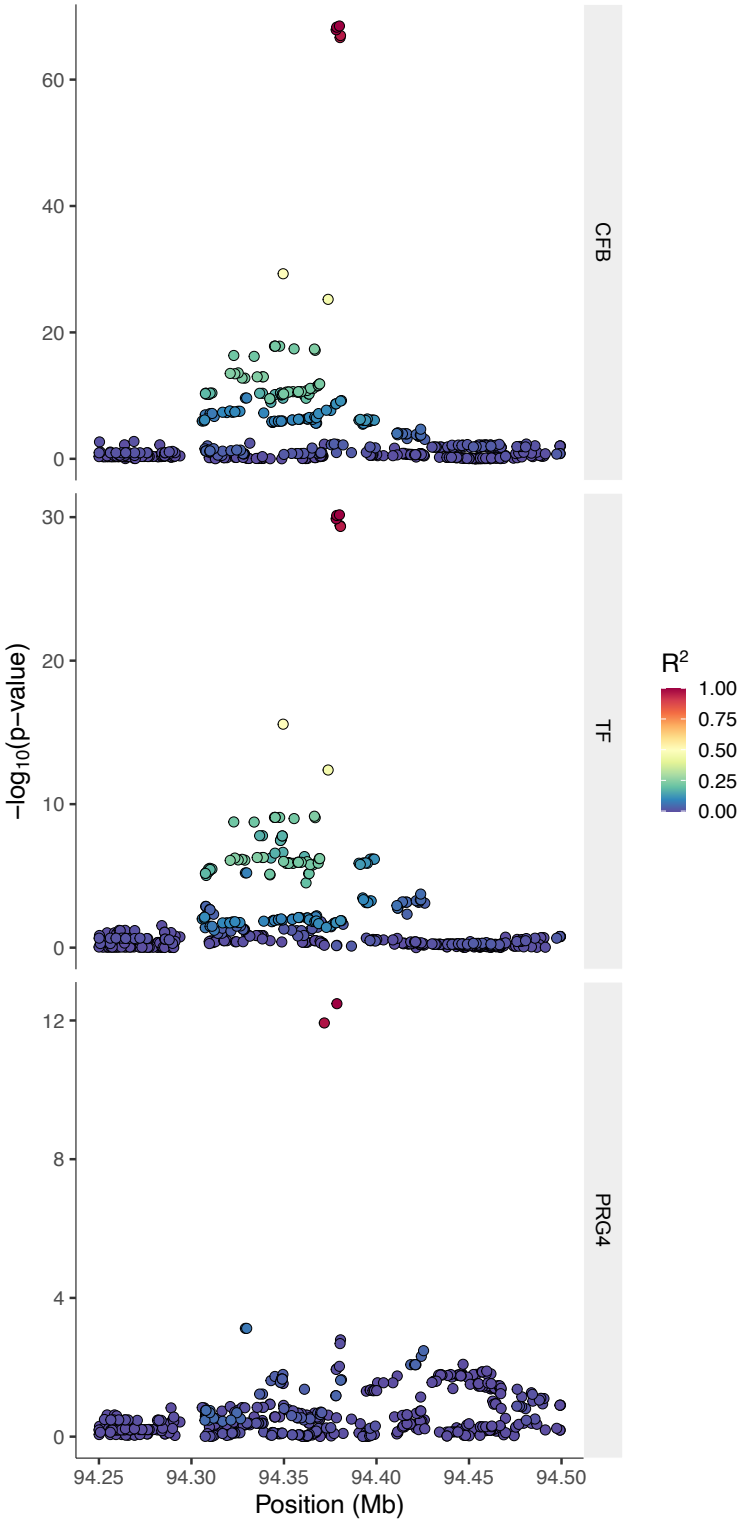

Figure 4.7: Chromosome 14 *trans*-pQTL. This figure contains association plots for the complement factor B (CFB), serotransferrin (TF), and proteoglycan 4 (PRG4) *trans*-pQTL. The CFB and TF loci colocalised with each other, but neither locus colocalised with PRG4.

The chromosome 14 locus contained *trans*-pQTL for complement factor B (CFB), serotransferrin (TF), and proteoglycan 4 (PRG4). The CFB and TF *trans*-pQTL regions colocalised with each

other, but neither colocalised with the PRG4 region (Figure 4.7). There were 6 *cis*-pQTL loci and 19 *cis*-eQTL loci that overlapped the *trans*-pQTL region, but none of them colocalised with any of the *trans*-pQTL. The *trans*-pQTL are in a region of chromosome 14 containing many genes of the extracellular serpin family, which are involved in immune functions. For instance, *SERPINA1* is an inhibitor of neutrophil elastase and *SERPINA5* is an inhibitor of active protein C (Law *et al.* 2006).

The last *trans*-pQTL locus was on chromosome 8, associated with insulin-like growth factor binding protein, acid labile subunit (IGFALS). There were no *cis*-pQTL loci and 16 *cis*-eQTL loci in the region, but none of them colocalised with the *trans*-pQTL. This suggests that another functional element, molecular QTL in a different tissue or context, or a different mechanism explains this *trans*-pQTL locus.

4.5 Colocalisation with GWAS Associations

Due to the inclusion of trait-associated variants from the EBI GWAS Catalog when mapping module QTL, multiple potential associations between module QTL and immune-related traits were identified[1]. However, association of the same SNP with two different traits does not necessarily imply the same underlying causal variant because of LD. To test for evidence of colocalisation, a subset of studies with summary statistics from the EBI GWAS Catalog were curated (Table B.1) based on matching ancestry and large sample size. The traits from these studies included four serum protein levels, seven cell frequency measures, and three autoimmune diseases. Of these 14 traits, seven colocalised with at least one module QTL locus (Table 4.3).

[1] Discussed in Section 3.2.2

Trait Group	Trait	Colocalising Module QTL
Serum Proteins	C-reactive Protein (Inflammation Marker)	None
	Alanine Aminotransferase (Liver Function)	Module 101 @ Chr 12 (55.0 Mb - 57.1 Mb)
	Alkaline Phosphatase	None
	Interleukin 18	None
Leukocyte Trait	Lymphocyte Count / Proportion	Module 101 @ Chr 12 (55.0 Mb - 57.1 Mb)
	Neutrophil Count / Proportion	Module 84 @ Chr 6 (28.8 Mb - 34.0 Mb) Module 103 @ Chr 12 (68.3 Mb - 70.4 Mb)
	Monocyte Count / Proportion	None
	Eosinophil Count / Proportion	Module 101 @ Chr 12 (55.0 Mb - 57.1 Mb)
	Basophil Count / Proportion	None
	Platelet Count	Module 63 @ 11 (46.5 Mb - 48.6 Mb) Module 88 @ Chr 14 (64.3 Mb - 66.4 Mb)
	Erythrocyte Count	None
Autoimmune Disease	Rheumatoid Arthritis	None
	Inflammatory Bowel Disease	Module 47 @ Chr 3 (100.4 Mb - 102.5 Mb) Module 71 @ Chr 3 (100.3 Mb - 102.5 Mb)
	Systemic Lupus Erythematosus	Module 84 @ Chr 6 (28.8 Mb - 34.0 Mb)

Table 4.3: Colocalisation of module QTL with GWAS variants. Some module QTL were variants from the EBI GWAS Catalog. A subset of studies based on matched ancestry were used to test these regions for colocalisation with the associated GWAS trait.

4.6 Statistical Fine Mapping

4.6.1 Conditional *cis*-eQTL

Statistical fine mapping takes into account both the association summary statistics and the underlying LD structure to identify CSs of variants that are likely to contain the causal SNP. Similar to the colocalisation framework, fine mapping is performed on a locus consisting of a set of variants in a genomic interval. The FINEMAP and SuSiE statistical fine mapping frameworks were used to identify CSs for each conditional *cis*-eQTL region[1]. FINEMAP identified a CS for all of the 16,054 eGene-signal pairs. In contrast, susieR uses a measure of purity to prune uninformative CSs and identified a CS for 11,055 (68.9%) eGene-signal pairs. Specifically, susieR prunes CSs if the minimum absolute correlation between SNP genotypes in the CS is less than 0.5. Performing

[1] Described in Section 2.5

the same pruning step on the FINEMAP results reduces the CSs to 10,332 (64.4%) eGene-signal pairs. For the 16,054 eGene-signal pairs, the lead SNP was used to identify tagging SNPs in LD. These tagging SNP sets represented a naive alternative to the CSs and served as a comparator to determine how much uncertainty was introduced or reduced using the Bayesian fine mapping approaches. The pruned FINEMAP and SuSiE CSs were smaller than the LD tagging SNP sets (Figure 4.8).

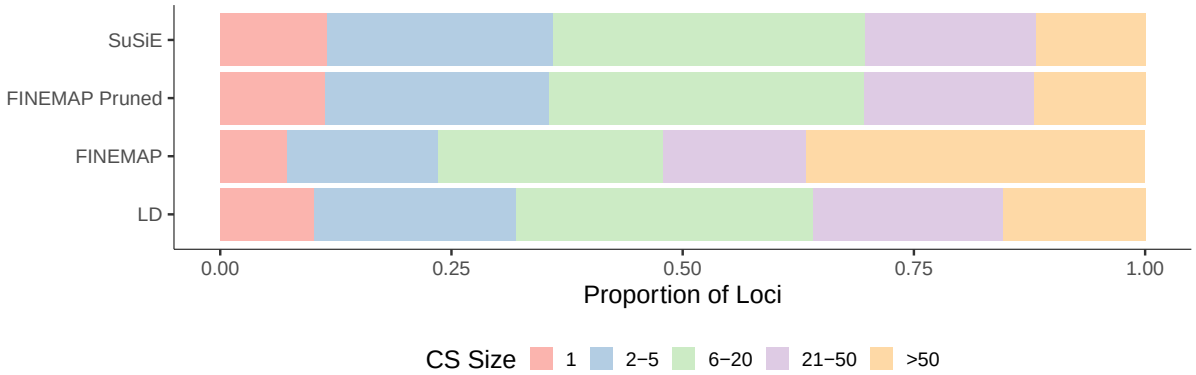

Figure 4.8: Credible set sizes. CSs were binned into five groups based on the number of variants present in the CS (1, 2-5, 6-20, 21-50, and >50). The data consists of all 16,054 eGene-signal pairs for the LD tagging SNP sets and FINEMAP CSs, the subset of 10,332 eGene-signal pairs for the pruned FINEMAP CSs, and the subset of 11,055 eGene-signal pairs with SuSiE CSs.

The 10,332 eGene-signal pairs that were present both in the pruned FINEMAP and SuSiE CSs were used to perform paired tests. Within this subset, pairs had a median of 14 SNPs in the tagging SNP sets. In comparison, FINEMAP and SuSiE CSs had a median of 10 and 9 SNPs respectively. Both FINEMAP and SuSiE generated CSs that were smaller than the tagging SNP sets (Wilcoxon Signed Rank Test with Continuity Correction; $p < 2.2 \times 10^{-16}$). In addition, the SuSiE CSs were smaller than the FINEMAP CSs (Wilcoxon Signed Rank Test with Continuity Correction; $p < 2.2 \times 10^{-16}$).

The 5,722 FINEMAP CSs that were removed by pruning (impure CSs) contained a median of 474.5 SNPs. Surprisingly, the tagging SNP sets of the associated eGene-signal pairs had a median of 8 SNPs. A paired test between the size of the CS and the size of the tagging SNP set demonstrated that the impure CSs were significantly larger than the tagging SNP sets (Wilcoxon Signed Rank Test with Continuity Correction; $p < 2.2 \times 10^{-16}$). This suggested that the diffusion of PIPs across a large number of SNPs in these CSs was not due to large LD blocks. The absolute Z scores of the lead SNP associated with each eGene-signal pair were lower in the impure CSs compared to the pruned FINEMAP CSs (Wilcoxon Rank Sum Test with Continuity Correction; $p < 2.2 \times 10^{-16}$), suggesting that the large CS size reflected uncertainty due to weak strength of association between SNPs in the region and the expression of the eGene.

4.6.2 Module QTL

There were 76 loci associated with module QTL. For each module, the locus was associated with at least one of the top five module eigengenes. All possible eigengene-locus pairs were used in fine mapping, resulting in a total of 380 pairs. Of these, SuSiE assigned CSs to 186 (48.9%) pairs. When using the same purity filter for FINEMAP, 191 (50.3%) pairs had CSs. Unlike the conditional *cis*-eQTL, the module QTL had not been refined using forward regression and thus potentially contained more than one signal. To identify these conditional signals, both FINEMAP and SuSiE were run assuming up to $L = 10$ signals. Thus, each eigengene-locus pair could have up to 10 CSs. The 174 eigengene-locus pairs that had CSs in both the SuSiE and pruned FINEMAP sets showed high concordance in the number of independent signals detected (Figure 4.9). 103 (59.2%) of the pairs had one signal detected by both methods.

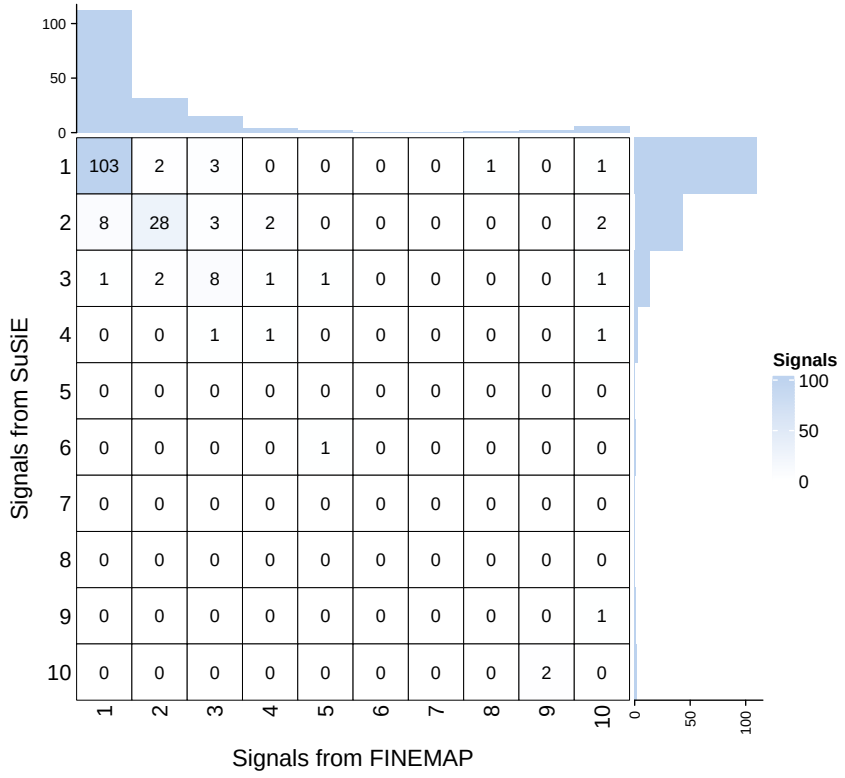

Figure 4.9: Number of signals for module QTL. Both SuSiE and FINEMAP were set to detect up to $L = 10$ signals at each module QTL locus. The number of signals detected at each locus by both frameworks were highly concordant, with one signal detected at most module QTL loci.

Compared to the conditional *cis*-eQTL, the CS sizes of the SuSiE and pruned FINEMAP approaches were comparable with the LD tagging SNP sets (Figure 4.10). The median number of tagging SNPs to the lead module QTL was 15, while the median number of SNPs in a CS was 14 for both the SuSiE and pruned FINEMAP approaches.

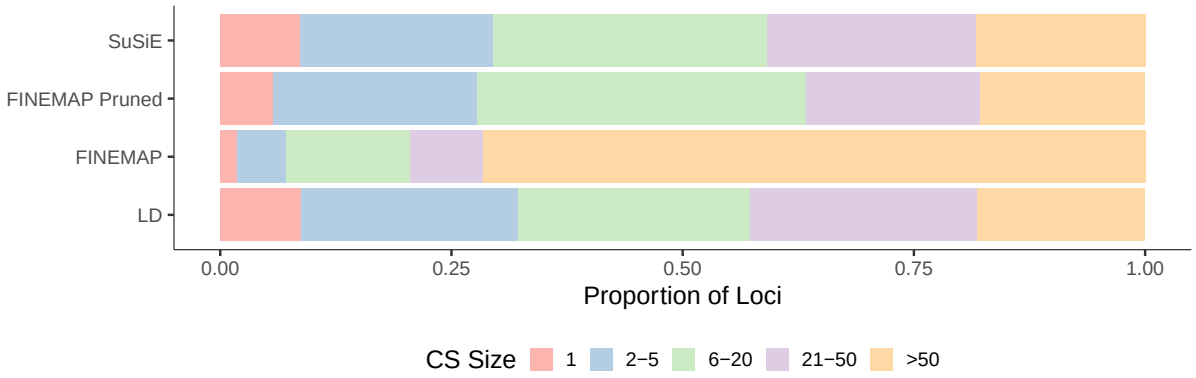

Figure 4.10: Module QTL credible set sizes. CSs were binned into five groups based on the number of variants present in the CS (1, 2-5, 6-20, 21-50, and >50). The data consists of all the LD tagging SNP sets and FINEMAP CSs, the subset of 191 eigengene-locus pairs for the pruned FINEMAP CSs, and the subset of 186 eigengene-locus pairs with SuSiE CSs.

4.6.3 pQTL

Of the 23 *cis*-pQTL loci, SuSiE assigned a CS to 13. FINEMAP, in contrast, assigned a CS to all loci, even after pruning based on purity. Similar to the module QTL, both approaches were run assuming up to $L = 10$ signals. Both FINEMAP and SuSiE detected one signal in 9 (69.2%) of the 13 pQTL loci present in results from both methods (Figure 4.11).

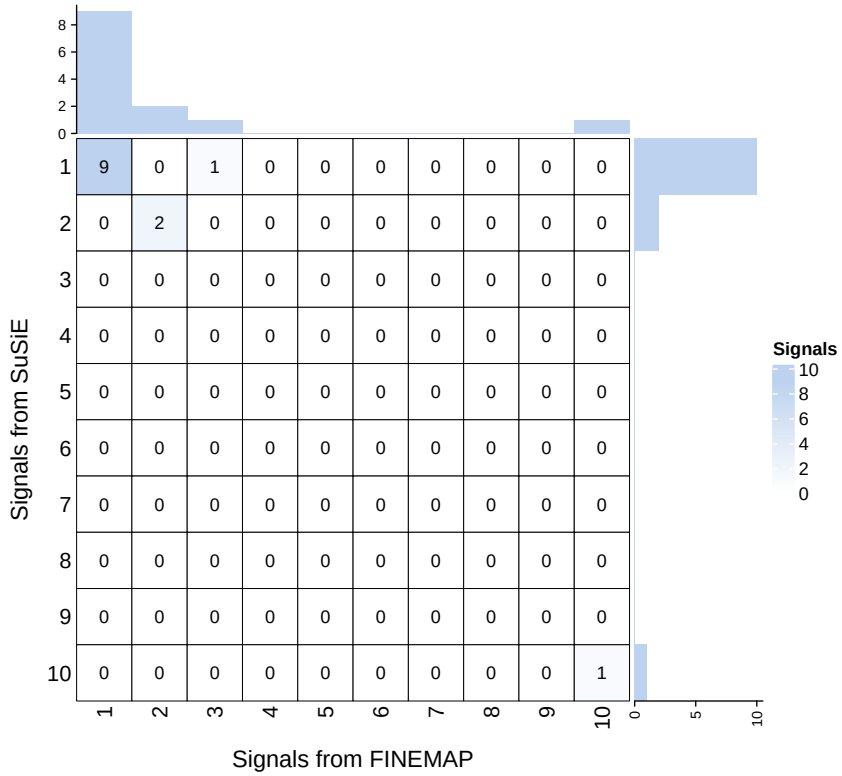

Figure 4.11: Number of signals for *cis*-pQTL. Both SuSiE and FINEMAP were set to detect up to $L = 10$ signals at each *cis*-pQTL locus. The number of signals detected at each locus by both frameworks were highly concordant, with one signal detected at most loci.

The FINEMAP and SuSiE CSs were larger than the tagging SNP sets for the 13 *cis*-pQTL loci (Figure 4.12). The median number of tagging SNPs was 18, while the median number of SNPs in a CS was 20 for both the SuSiE and FINEMAP approaches.

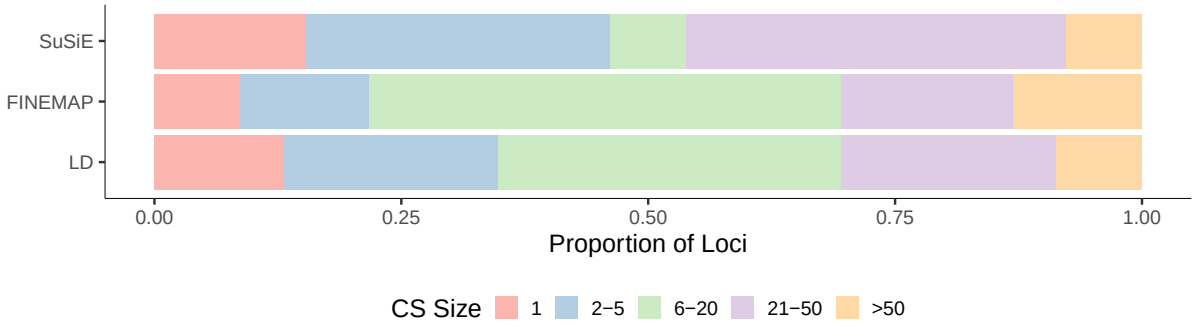

Figure 4.12: *Cis*-pQTL credible set sizes. CSs were binned into five groups based on the number of variants present in the CS (1, 2-5, 6-20, 21-50, and >50). The data consists of all the LD tagging SNP sets, all the FINEMAP CSs, and the subset of 13 *cis*-pQTL loci with SuSiE CSs.

The individual *trans*-pQTL loci represented interesting examples to test whether fine mapping could further resolve the underlying signal. The chromosome 8 *trans*-pQTL locus had 33 SNPs tagging the lead SNP at $R^2 > 0.8$. Both FINEMAP and SuSiE predicted only one signal at this locus. Both reported one CS consisting of the same 27 SNPs. Both also reported the same seven top SNPs with the highest PIPs. Unfortunately, the maximum PIP from either method was not substantial (0.052 from SuSiE and 0.049 from FINEMAP).

The three *trans*-pQTL loci on chromosome 14 were for CFB, TF, and PRG4. The colocalisation analysis revealed that the CFB and TF *trans*-pQTL colocalise with each other and neither colocalise with PRG4 *trans*-pQTL. SuSiE predicted 3 independent signals for CFB, while FINEMAP predicted 9 even after pruning based on purity. SuSiE predicted 1 independent signal for TF while pruned FINEMAP predicted 3 signals. However, both agreed on one independent signal for PRG4. Both approaches assigned a CS consisting of the same two SNPs (rs28929474 and rs112635299) to PRG4. SuSiE assigned a PIP of 0.777 to rs28929474. This SNP encodes a missense variant in *SERPINA1*, which codes for a serpin family antitrypsin that specifically targets and inhibits the activity of neutrophil elastase (Law *et al.* 2006). Both methods assigned a CS consisting of the same five SNPs to TF (rs11846959, rs1303, rs2073333, rs17090719, rs2070709), although FINEMAP had two other CSs that SuSiE did not report. SuSiE calculated the maximum PIP of 0.361 for rs11846959, which is a SNP in the intron of *SERPINA1*. The second highest PIP of 0.330 was calculated for rs1303, which is surprisingly another missense variant in *SERPINA1*. A very similar CS was assigned by both approaches to CFB, consisting of three SNPs by SuSiE (rs11846959, rs1303, rs2073333) and two SNPs by FINEMAP (rs11846959, rs1303). In this case,

SuSiE calculates a PIP of 0.555 for rs11846959 and FINEMAP calculates an even higher PIP of 0.867 for rs11846959.

The two *trans*-pQTL loci on chromosome 16 were for CLU and ITIH1. The *trans*-pQTL for these two proteins colocalised with each other. Interestingly, SuSiE fails to identify any signals at this locus and FINEMAP reports 10 independent signals for both loci after pruning by purity. FINEMAP assigned a CS of three SNPs (rs11647844, rs12925901, rs12708920) to both CLU and ITIH1, although the highest PIP for CLU (0.366) was for rs11647844 while ITIH1 had equal PIPs (0.333) for all three SNPs. All three SNPs are noncoding and in the intronic region of *PKD1L3*.

4.7 Discussion

In this chapter, I performed multiple colocalisation analyses between various molecular traits. I colocalised *cis*-eQTL with each other, *cis*-eQTL with module QTL, *cis*-eQTL with *cis*-pQTL, *trans*-pQTL with each other, and module QTL with GWAS associations. I also performed fine mapping of the molecular QTL in the GAinS cohort.

4.7.1 Colocalisation of QTL across Omics Layers

Colocalisation was used to identify regions of the genome that might be responsible for regulating multiple traits. However, these results do not directly implicate a single causal model for the effect of the locus on the measured traits. A natural hypothesis in the case of colocalising *cis*-eQTL and *cis*-pQTL, for instance, is that the effect of the variant on protein expression is mediated through a direct effect on the expression of the cognate gene. However, variants may also act on gene expression and protein expression through independent mechanisms, which is a case of horizontal pleiotropy (Sanderson *et al.* 2022). An additional complication is that the gene expression was measured in a heterogeneous tissue that is separate from the tissue in which protein was assayed, although closely linked via secretion from leukocytes and interactions with the coagulation and complement systems. Thus, a diverse set of mechanisms may explain the effect of the same variant on gene and protein expression.

A future step for this analysis is to use Mendelian randomisation (MR) to specifically test the causal relationships between traits that colocalise. Using *cis*-eQTL as instruments can be challenging because independent signals tend to be close to each other and can suffer from confounders associated with the same haplotype. Furthermore, as evidenced by the colocalisation analysis, *cis*-eQTL can be shared between neighbouring genes. New methods to account for

these concerns such as transcriptome-wide Mendelian randomisation (TWMR) are being developed to estimate the effect of gene expression on outcomes of interest (Porcu *et al.* 2021).

4.7.2 Colocalisation and Fine Mapping Methods

The goal of colocalisation and subsequent fine mapping was to increase evidence for the effect of genomic loci on molecular variation and to reduce uncertainty surrounding the causal variant. These results can be used to identify variants that can be tested further in functional assays. Evidence for colocalisation can be refined using fine mapping approaches. While conditionally independent *cis*-eQTL were used for colocalisation, fine mapping procedures for module QTL, pQTL, or GWAS associations were not used before performing colocalisation. Future work for this analysis is to use fine mapping to refine colocalisation, such as using COLOC integrated with the SuSiE framework (Wallace 2021).

Even after fine mapping, LD between mechanistically-independent causal molecular QTL limits the ability of statistical methods. A recent experimental fine mapping approach of *cis*-eQTL using a massively parallel reporter assay (MPRA) found that 17.7% of loci with strongly-linked variants had more than one allelic hit. Thus, once sepsis-relevant functional elements have been identified, experimental validation will be necessary. One method to design experiments for identified variants is to use variant effect prediction to identify potential mechanisms that can be tested. Another important consideration is to identify which cell types might be relevant to the variant, since cellular context is important to observe the effect in an experimental setup. Both of these computational approaches are explored in chapter 5 of this thesis. Finally, colocalisation was performed using the default priors and a liberal version of a threshold implemented previously (Nath *et al.* 2019). Future work is to test the sensitivity of the colocalisation results to this threshold and to the prior probabilities of association and colocalisation.

Colocalisation analysis assumes independent cohorts with the same LD structure. In this analysis, the QTL were derived from the same cohort and therefore necessarily share the same LD structure. However, *cis*-eQTL and *cis*-pQTL were mapped using overlapping samples. The *cis*-eQTL and module QTL were mapped on the same set of individuals. Thus, within-sample correlation may confound the colocalisation analysis and cause false inflation of the posterior probability of colocalisation. Overlapping cases are explicitly modelled in other approaches, such as HyPrColoc (Foley *et al.* 2021), which may be used in the future to determine the extent to which the assumption of independent samples affects colocalisation.

The module QTL remain challenging to interpret after colocalisation. A surprising number

of module QTL colocalise with multiple *cis*-eQTL, suggesting that these regions may represent functional elements that affect the transcription of multiple genes in a local genomic context. Indeed, a prior analysis generated local co-expression modules that consisted of genes in the same neighbourhood in the genome and observed that 45.6% of co-expressed eGene pairs had evidence for colocalisation between *cis*-eQTL (Ribeiro *et al.* 2021). A more unlikely scenario is that these regions represent true *cis*-eQTL for one gene in the module, which then regulates the other genes in *trans*. It is unclear why such factors would be restricted to affecting genes in *trans* in the local genomic context.

Taken together, these results suggest that it is possible to use modules as a method of aggregating the signal from multiple genes to identify genomic elements that act in *trans* in relatively small cohorts.

5 | Dysregulated Immune Cell Types

The aim of this chapter is to identify specific cell types that are dysregulated in sepsis. I characterised the accessibility landscape of stimulated immune cells using publicly available data, since chromatin accessibility profiles are cell-type-specific and reflect regions of the genome relevant to cell-type-specific functions. I also used variant effect prediction methods to nominate potential mechanisms for the QTL.

5.1 Reprocessing ATAC-seq Data

I used ATAC-seq data for stimulated and unstimulated leukocytes to identify functionally relevant regions of the genome[1]. In order to compare across all major cell types, I curated an immune atlas from multiple publicly available data sets comprising primary immune cell types in unstimulated and stimulated conditions (Corces *et al.* 2016; Calderon *et al.* 2019) and a neutrophil atlas containing neutrophils under various stimulations (Ram-Mohan *et al.* 2021).

Both atlases contained data that had been previously processed. However, samples were processed again for a few reasons. Both analyses used human genome build 19 (hg19) to align their reads, while all of the QTL analysis was conducted on GRCh38. While the original immune atlas contained a consensus peak set count matrix, the neutrophil atlas did not use a consensus peak set across all their samples. Finally, cell-type-specific and stimulation-specific peaks were not reported by either study.

The immune atlas consisted of 25 primary immune cell types from six broad lineages (Table C.1). Leukocytes were assayed either as unstimulated or after a cell-type-specific *ex vivo* stimulation. Although some cell types were tested with two different stimulations, all stimulated samples were grouped together in downstream analyses as done in the original analysis (Calderon *et al.* 2019) due to a high degree of concordance between effects. The neutrophil atlas consisted of three different experiments (Table C.2). In the first, neutrophils were stimulated *ex vivo* with six

[1] Described in Section 2.6

different sepsis-relevant ligands. In the second, neutrophils in whole blood were stimulated with *E. coli* for one or four hours. In the last, neutrophils in whole blood were stimulated with varying concentrations of *S. aureus*.

5.2 Consensus and Cell-Type-Specific Peaks

I aligned ATAC-seq data from both the immune atlas and neutrophil atlas and processed it to identify group peak sets[1] (Figures D.5 and D.7). These group peak sets were merged for each cell type to generate cell type peak sets in the immune atlas (Figure D.6).

Motif enrichment analysis[2] identified motifs that were enriched in each stimulated group compared to their respective control group (Figure 5.1).

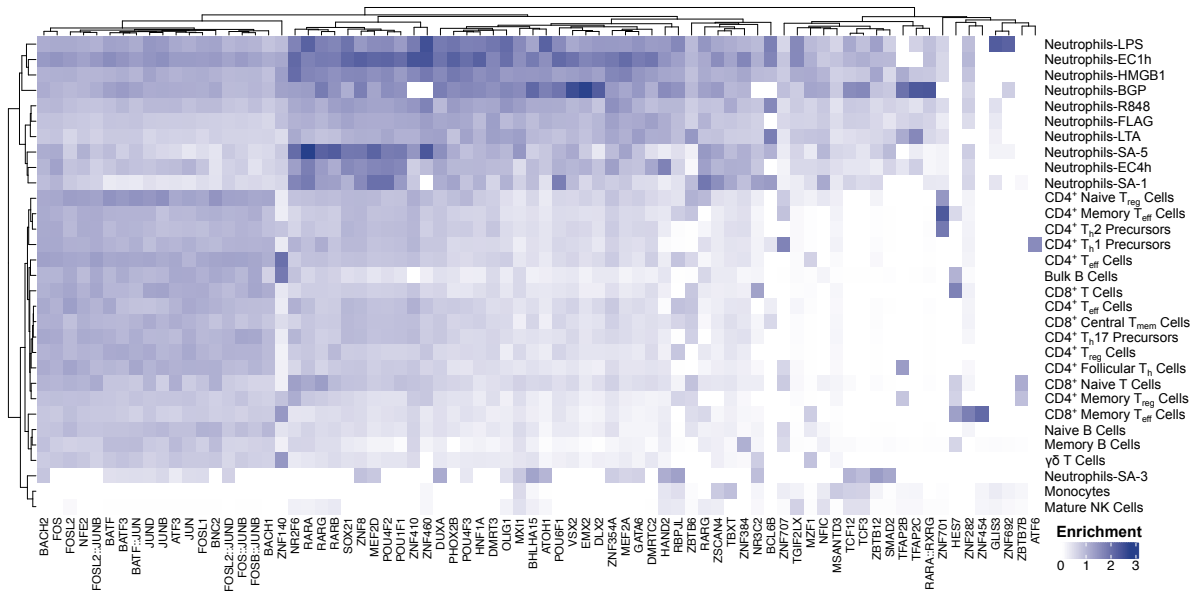

Figure 5.1: Motif enrichment in group peak sets. Motif enrichment was performed using SEA. Sequences in peaks present in the stimulated groups were compared to sequences in peaks from their respective control groups for enrichment of known motifs from JASPAR. Values shown here are $\log_2$ transformed enrichment ratios reported by SEA. The union of the top 5 motifs from each group was used for the heatmap. Any enrichment that did not pass the significance threshold were given a value of 0.

In both atlases, group peak sets were combined together to make a consensus peak set. The consensus peak set from the immune atlas and neutrophil atlas consisted of 296,994 peaks and 140,923 peaks respectively. The immune atlas had larger peaks, with a median size of 555 base pairs compared to the median peak size of 523 base pairs in the neutrophil atlas. A plurality of peaks (52.2% in the immune atlas and 48.7% in the neutrophil atlas) were intronic (Figure 5.2). 28.2% of the peaks in the immune atlas and 31.0% of the peaks in the neutrophil atlas were

[1] Described in Section 2.6.3
[2] Described in Section 2.6.4

intergenic. Fewer peaks were detected in exonic, promoter at TSS, and transcription termination site (TTS) sites.

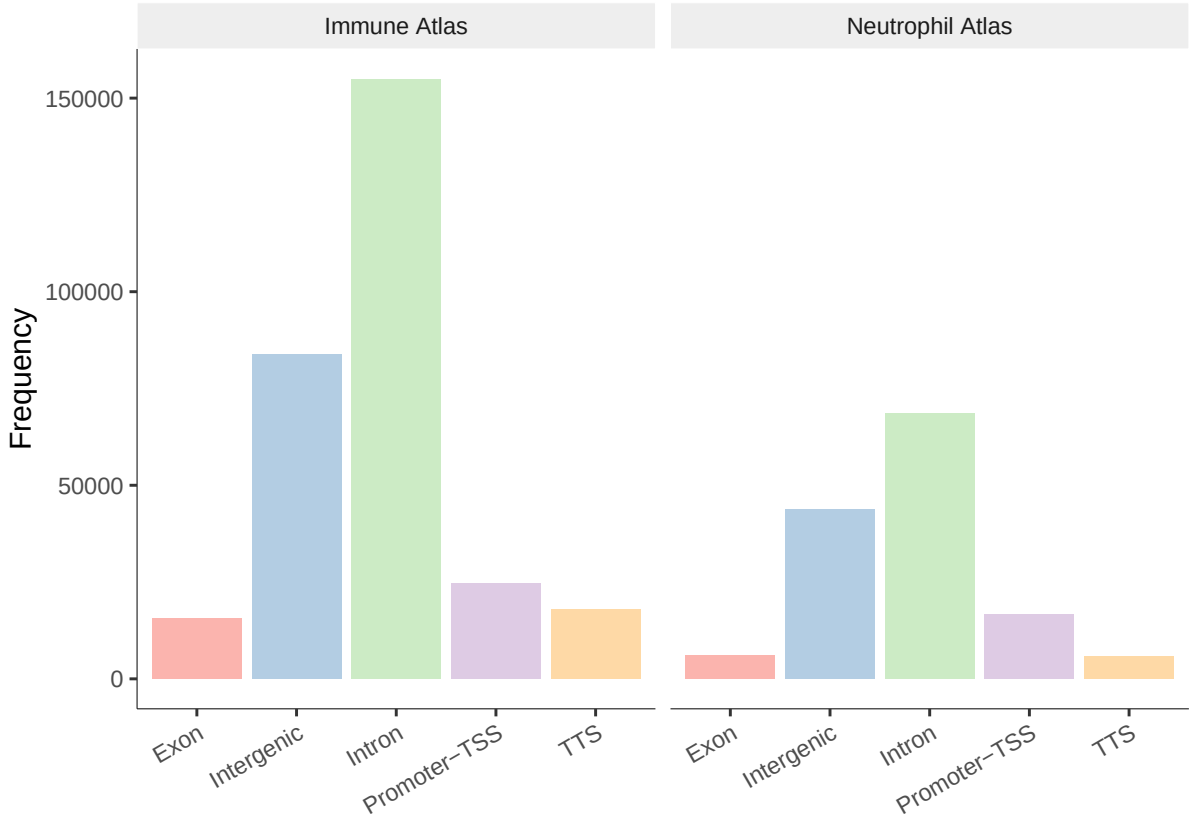

Figure 5.2: HOMER consensus peaks annotation. HOMER was used to annotate peaks based on their proximity to gene bodies. Peaks were annotated if they overlapped an exon, intron, promoter at a TSS, or the TTS. Any peaks not overlapping these elements were considered intergenic.

5.3 Enrichment of *cis*-eQTL in Genomic Annotation

Since the conditional *cis*-eQTL were derived from gene expression in a heterogeneous tissue, variants may exert their effects through various immune cell types. A matched SNP approach was used to test which cell-type-specific functional annotations across the genome were enriched for conditional *cis*-eQTL[1].

[1] Described in Section 2.7.1

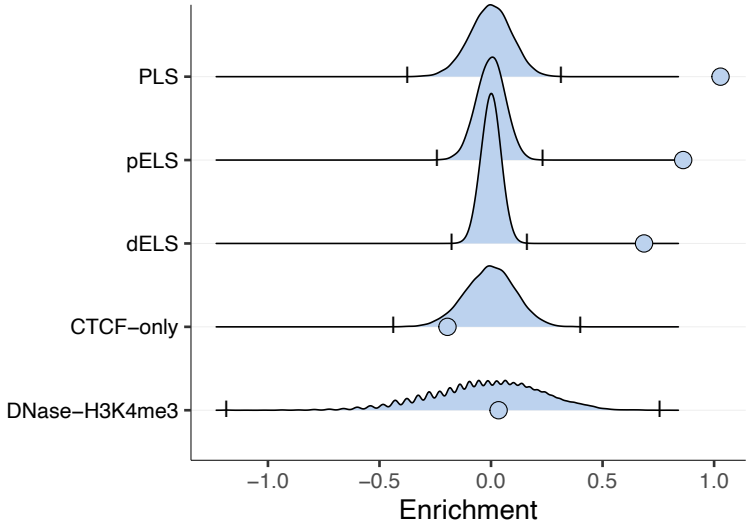

Figure 5.3: Enrichment in ENCODE cCREs. For each set of ENCODE cCREs, matched SNPs were used to generate null distributions for the proportion of overlapping variants. Each point represents the observed overlap of lead conditional *cis*-eQTL. The bars for each distribution represent the boundaries of the rejection region based on a significance threshold of $\alpha = 0.0001$. Enrichment was calculated as $\log_2$ of the ratio of the observed proportion of overlap to the mean null proportion of overlap.

Conditional *cis*-eQTL were expected to be enriched in proximal *cis* elements such as promoters and enhancers due to their association with gene expression. The cCREs in ENCODE are divided into promoter-like signatures (PLSs), TSS-proximal enhancer-like signatures (pELSs), TSS-distal enhancer-like signatures (dELSs), not TSS-overlapping and with high DNase and H3K4me3 signals only (DNase-H3K4me3), and not TSS-overlapping and with high DNase and CTCF signals only (CTCF-only). As expected, enrichment was observed in PLSs, pELSs, and dELSs (Figure 5.3).

ChromHMM states were used to further refine these enrichment results based on more granular genome annotations from an 18-state model and across epigenomes from specific primary immune cell types in blood (Figure E.1). Conditional *cis*-eQTL were enriched in genic enhancers, active enhancers, and weak enhancers across all the epigenomes. Conditional *cis*-eQTL were depleted in active TSS states, which are close to the TSS, and enriched in upstream and downstream flanking TSS states that are farther away. Unsurprisingly, conditional *cis*-eQTL were enriched in areas of transcription and depleted in inaccessible or transcriptionally repressed regions. There was no meaningful variation in the pattern of enrichment across cell types.

To test if the conditional *cis*-eQTL were specific to the chromatin accessibility profile of any cell type, the same approach was used to test for enrichment in group peaks from the immune and neutrophil atlases (Figures 5.4 and 5.5). Conditional *cis*-eQTL are enriched in all conditions, generally reflecting the observation that eQTL are enriched in accessible regions of the genome.

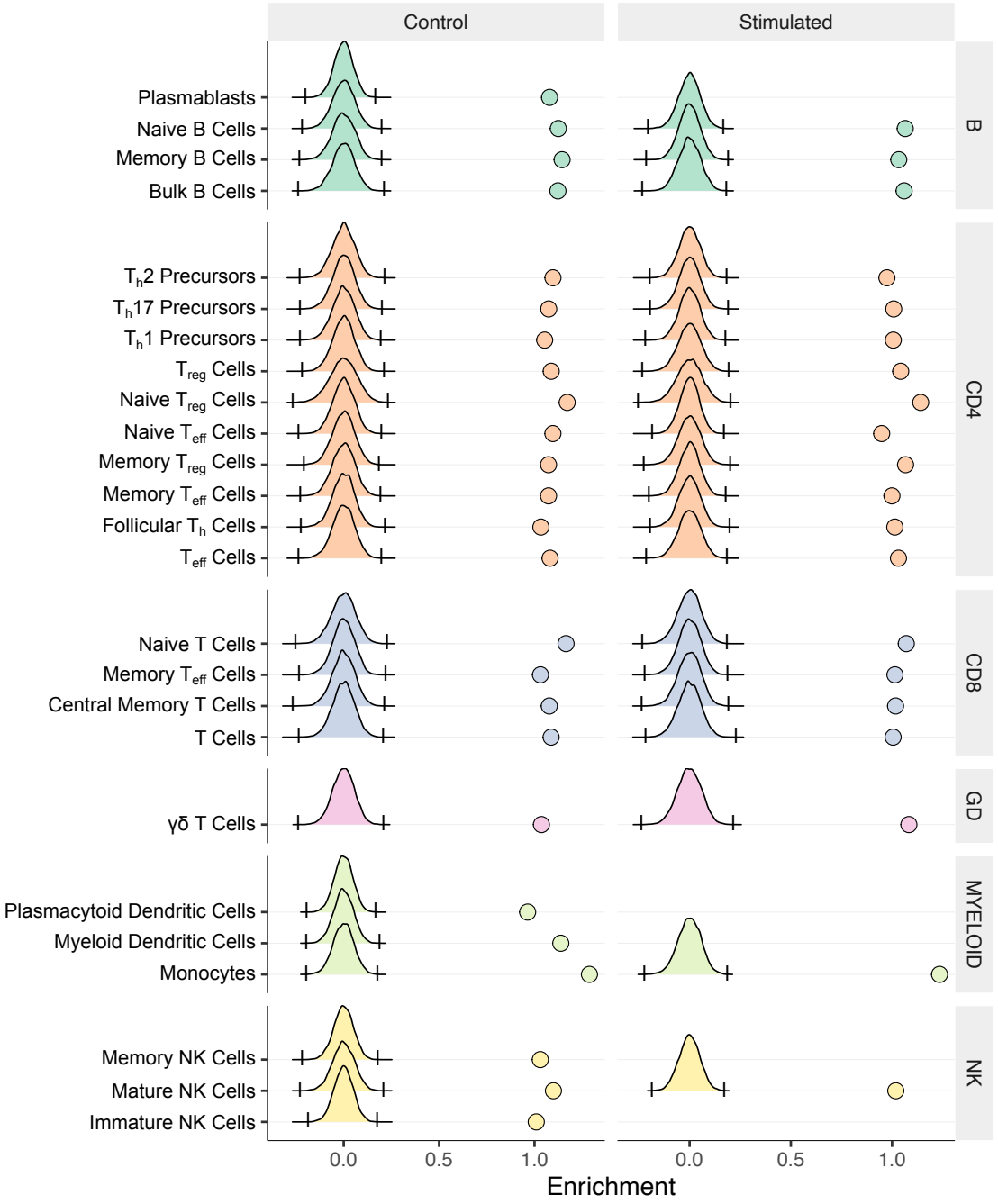

Figure 5.4: Enrichment in immune atlas peaks. For group peak sets in the immune atlas, matched SNPs were used to generate null distributions for the proportion of overlapping variants. Each point represents the observed overlap of lead conditional *cis*-eQTL. The bars for each distribution represent the boundaries of the rejection region based on a significance threshold of $\alpha = 0.0001$. Enrichment was calculated as $\log_2$ of the ratio of the observed proportion of overlap to the mean null proportion of overlap.

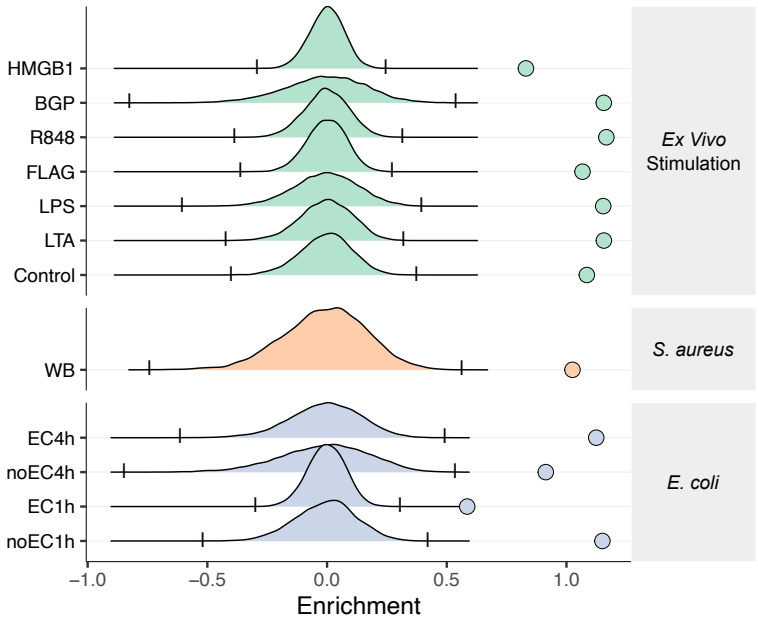

Figure 5.5: Enrichment in neutrophil atlas peaks. For group peak sets in the neutrophil atlas, matched SNPs were used to generate null distributions for the proportion of overlapping variants. Each point represents the observed overlap of lead conditional *cis*-eQTL. The bars for each distribution represent the boundaries of the rejection region based on a significance threshold of $\alpha = 0.0001$. Enrichment was calculated as $\log_2$ of the ratio of the observed proportion of overlap to the mean null proportion of overlap.

CHEERS integrates peak count information to detect small differences in the accessibility profiles of the same cell type under different stimulations. CHEERS was used to test for enrichment of conditional *cis*-eQTL in the neutrophil atlas (Figure 5.6). A strong enrichment was detected for neutrophil states induced by the ligands HMGB1, R848, and FLAG.

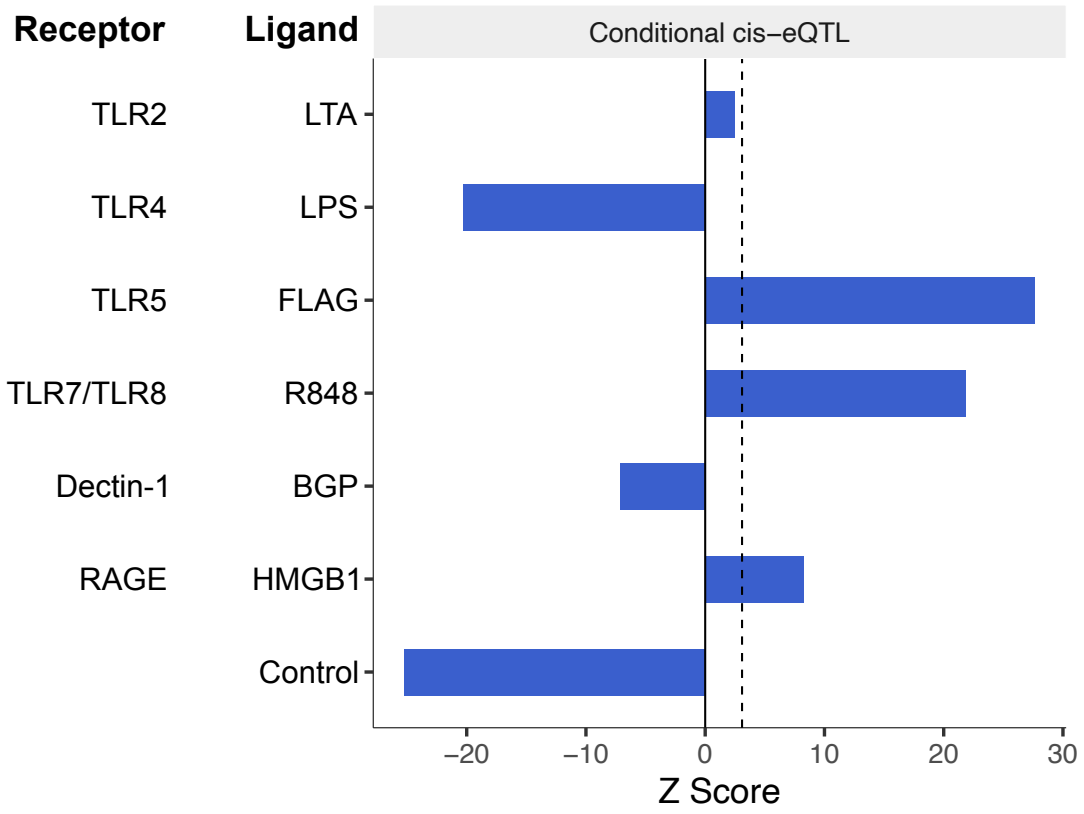

Figure 5.6: CHEERS enrichment of *cis*-eQTL. CHEERS was used to test for enrichment of lead conditional *cis*-eQTL and any tagging SNPs in stimulated neutrophil states.

A similar result to the matched SNP enrichment was obtained using GoShifter on the accessibility profiles from both atlases (empirical p-value of 1×10^{-4} for all group peak sets). GoShifter is a method that tests for enrichment of variants in genomic annotations using a local permutation strategy. GoShifter assigns an overlap score to each conditional *cis*-eQTL locus, which is the empirical probability that the locus overlaps an annotation by chance. Loci that have low overlap scores contribute strongly to the overall enrichment for the annotation computed by GoShifter. This locus by annotation score matrix demonstrates that some loci were enriched in all group peak sets (Figure 5.7). Some of the conditional *cis*-eQTL loci are also specific to the lineage or cell type. Thus, although the *cis*-eQTL are enriched in all of the group accessibility profiles, there are different cell and lineage specific loci that drive this enrichment in addition to a broad set of *cis*-eQTL that are generally in more accessible regions of the genome.

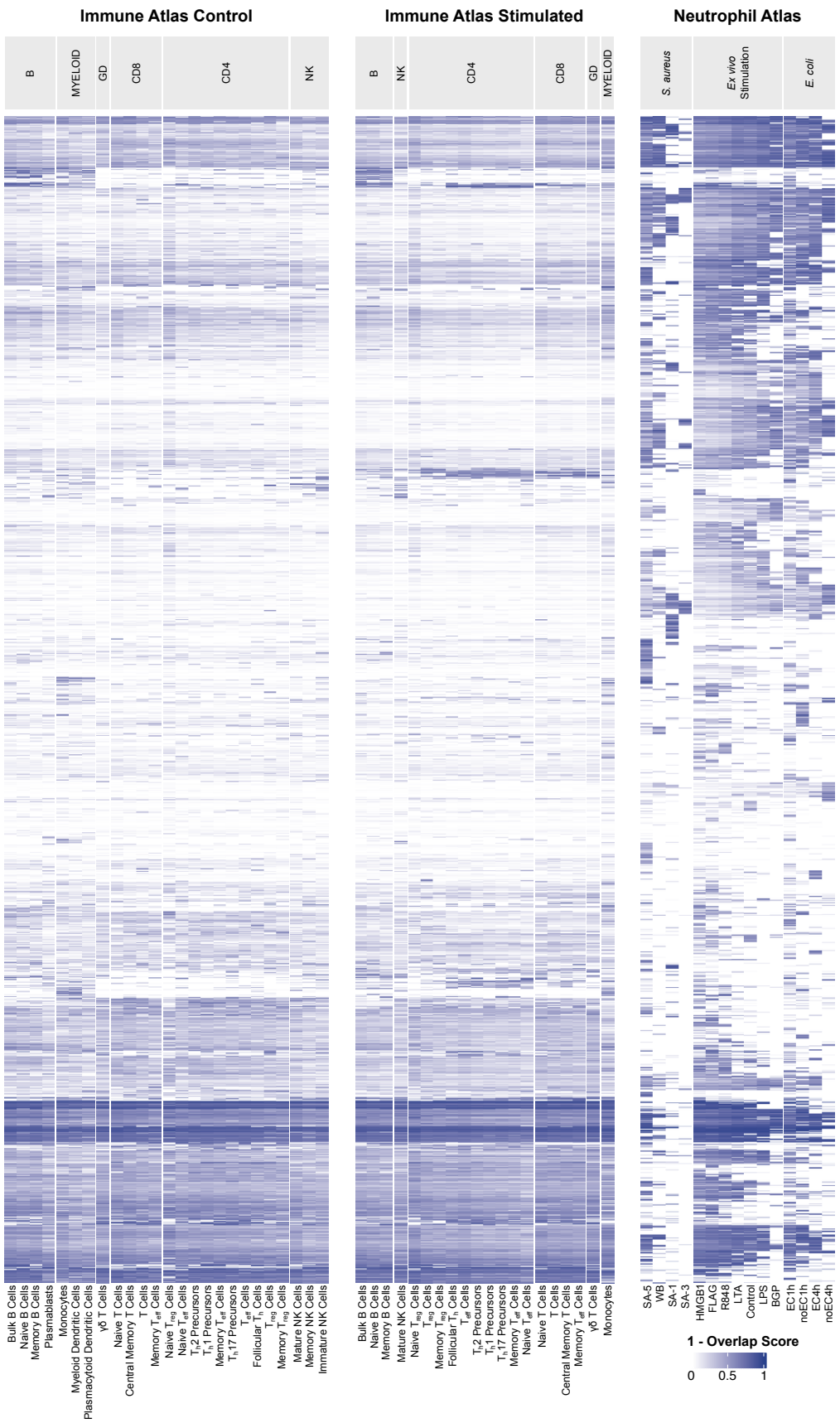

Figure 5.7: GoShifter overlap score matrix. This heatmap displays the matrix of locus by annotation overlap scores from GoShifter. Each row represents one conditional *cis*-eQTL locus. A locus includes the lead SNP and any tagging SNPs within 1 Mb with $R^2 > 0.8$. The score displayed here is 1 minus the probability that the locus overlaps the annotation by chance. Thus, a larger value implies a larger contribution to the overall enrichment p-value computed by GoShifter.

5.4 Partitioned Heritability

In comparison to enrichment methods that depend on overlap between annotations and trait-associated SNPs, testing for enrichment of per-SNP heritability in annotations leverages the entire polygenic architecture (Gusev *et al.* 2014). A variance component model was used to estimate the overall SNP heritability of module eigengenes and the proportion of heritability explained by SNPs within accessible regions from the immune and neutrophil atlases[1]. Patterns of enrichment and depletion of per-SNP heritability demonstrated that module eigengenes were enriched in different cell types (Figure 5.8). Broadly, the enrichment was stronger in the neutrophils compared to the other cell types. The eigengene for module 23, which contains genes for antigen processing and presentation (Table 3.1), was depleted in most cell types and specifically enriched in Naive B and T_{reg} cells. The eigengene for module 20, which contains genes associated with cytotoxic T cells and NK cells (Table 3.1), was enriched in stimulated NK and T cells. This preliminary analysis of partitioned heritability is a promising avenue to identify cell types associated with specific dysregulated molecular phenotypes.

[1] Described in Section 2.7.2

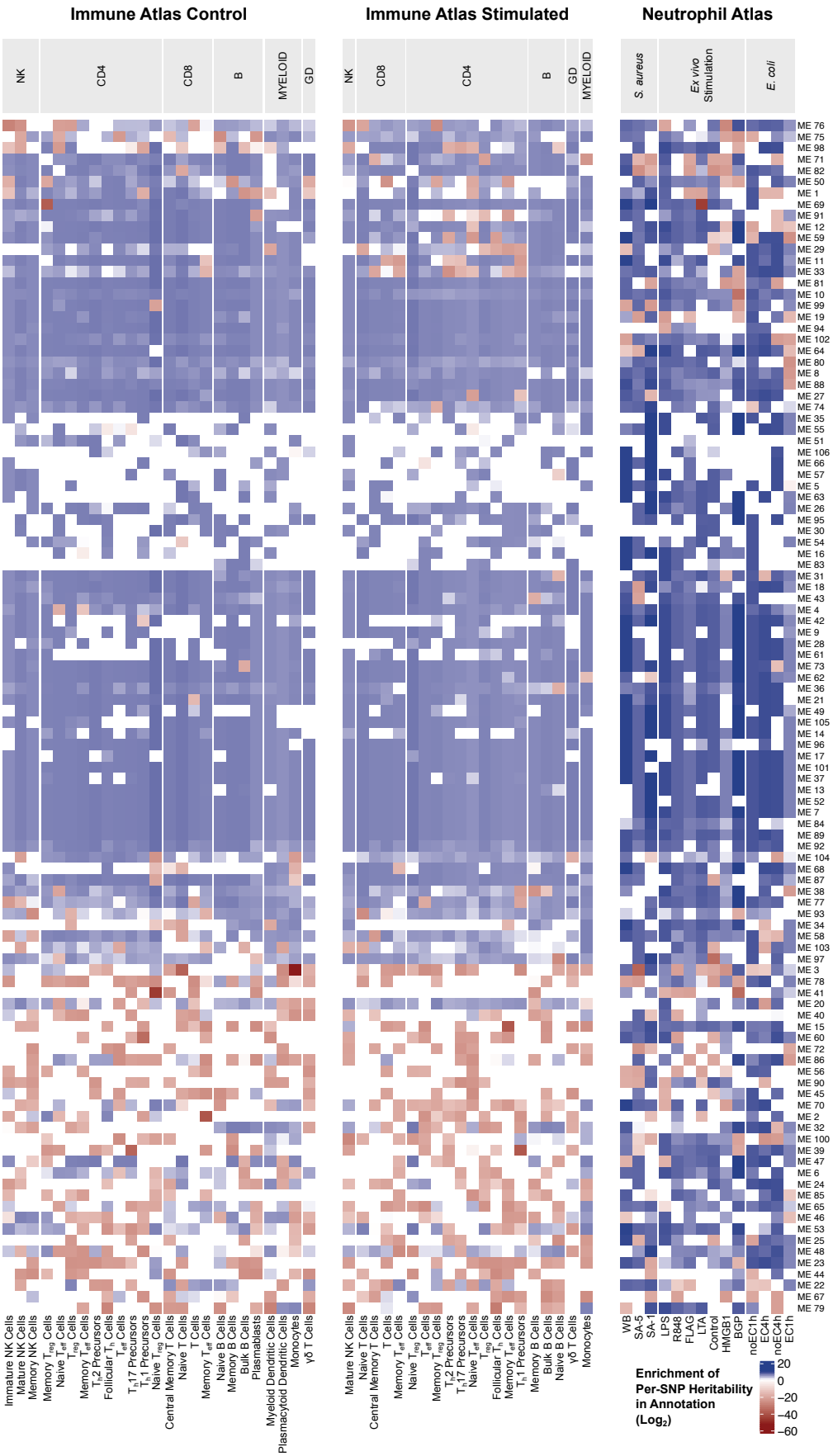

Figure 5.8: Partitioned heritability. This heatmap displays the $\log_2$ of the enrichment of per-SNP heritability ($h^2_{\mathrm{SNP}\alpha}/h^2_{\mathrm{SNP}}$) in each annotation for each module eigengene.

5.5 Variant Effect Prediction

My next aim was to determine any known or predicted functional consequences of the QTL to identify potential mechanisms underlying their effect on molecular traits. Using integrative genomics, data from other experiments and large consortium efforts can be used to predict the effect of variants in diverse contexts. Ensembl's VEP is a tool that annotates variants with consequences based on protein coding changes, proximity to genes and regulatory features, scores for various variant prioritisation schemes, and prior literature. VEP was used[1] to annotate lead conditional *cis*-eQTL and predicted a consequence for all 14,938 unique lead variants. 14,319 (89.2%) of the 16,054 unique *cis*-eQTL-eGene pairs were predicted to affect at least one transcript. Of these, 8,121 (56.7%) were predicted to affect only one gene, but others had more than one predicted consequence (Figure 5.9). Of all the pairs with gene consequences, only 6,351 (44.4%) were predicted to affect the associated eGene (Figure 5.9), suggesting that the conditional *cis*-eQTL captured regulatory biology that is not predictable from variant position alone.

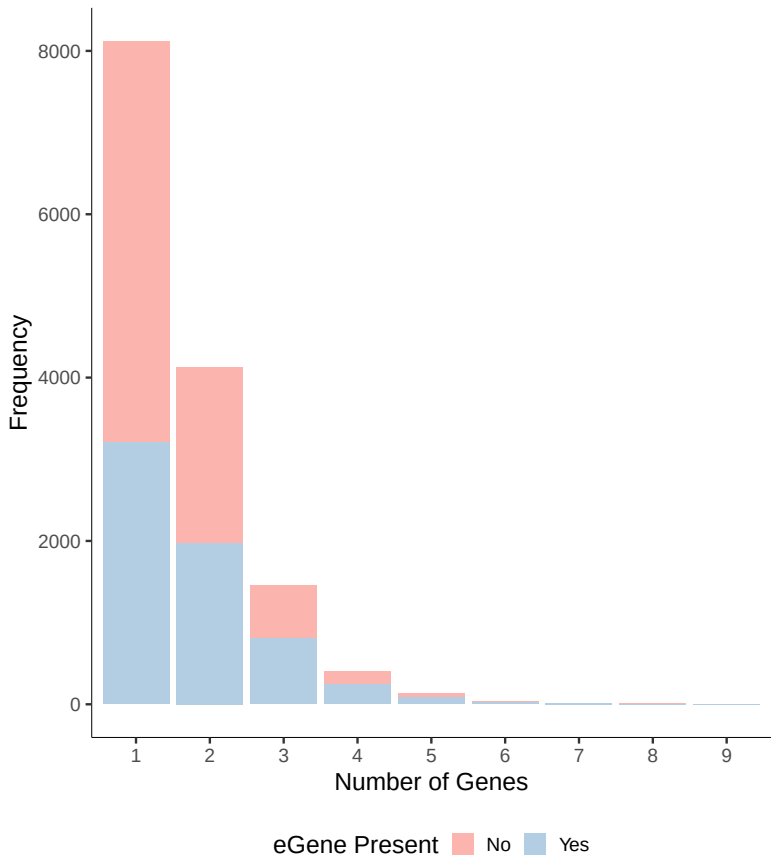

Figure 5.9: VEP gene consequences. VEP was used to identify predicted consequences of lead conditional *cis*-eQTL. While most variants were predicted to affect only one gene, some variants had multiple gene consequences. Conditional *cis*-eQTL that were predicted to affect their associated eGene are coloured in blue.

[1] Described in Section 2.7.3

The prior enrichment analysis revealed that the conditional *cis*-eQTL are enriched in regulatory elements across the genome. The VEP results identified specific regulatory consequences for 4,687 (31.4%) lead variants for 5,048 regulatory features across the genome. 4,326 (92.3%) variants with regulatory consequences were predicted to affect only one regulatory feature (Figure G.1). Most affected regulatory features were either promoter regions (35.4%) or flanking promoter regions (35.6%). Some specific enhancers (6.4%) and TF binding sites (1.7%) were also identified as potential consequences (Figure G.1).

Variation in genotype can affect gene expression by perturbing sequence motifs that encode TF binding. VEP reports motif sequences in the reference genome that overlap the variant. 639 (4.3%) unique lead conditional *cis*-eQTL overlapped a total of 1,716 motif features. The *cis*-eQTL that overlapped these motifs also changed the score of the motif based on the position weight matrix (PWM) (Figure 5.10).

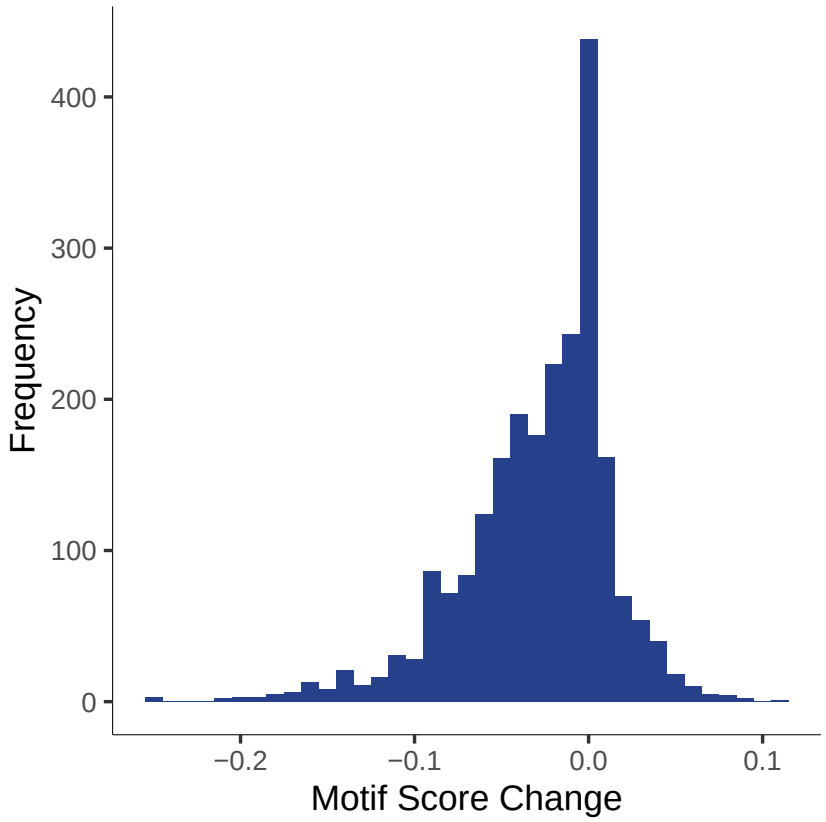

Figure 5.10: VEP change in motif score. VEP reports the change in the motif score based on the PWM for motifs in the reference genome that overlap a variant. This is the distribution of changes in motif scores of the lead conditional *cis*-eQTL that overlap a motif.

The same analysis was repeated for lead variants from module QTL and pQTL loci. 75 (98.7%) of the 76 lead module QTL had a predicted consequence. 64 (85.3%) lead variants were predicted to affect between 1 and 4 genes (Figure G.2). 27 (36.0%) lead variants fell in regulatory features,

which consisted primarily of promoters (Figure G.3). There were 32 regulatory features with an overlapping lead variant, of which 16 (50.0%) were promoter regions (Figure G.3). 8 (10.7%) of the lead variants were predicted to perturb motifs. A total of 12 motifs for 12 independent motif features overlapped a lead variant (Table G.1). Of particular interest are modules 101 and 103, which also colocalised with relevant immune traits (Table 4.3). The module 101 lead variant perturbed a GATA binding site, and the module 103 lead variant perturbed a HNF4A binding site.

All 23 lead *cis*-pQTL had a predicted functional consequence. 21 (91.3%) of the 23 were predicted to affect at least one transcript. 13 (61.9%) *cis*-pQTL were predicted to affect 1 gene, 7 (33.3%) were predicted to affect 2 genes, and 1 (4.8%) was predicted to affect 3 genes. Similar to the conditional *cis*-eQTL, only 15 (65.2%) of the 23 *cis-pqtl* were predicted to affect the pGene. 6 (26.1%) of the *cis*-pQTL were predicted to affect a total of 8 regulatory features, 2 of which were CTCF binding sites and 6 of which were promoter-flanking regions. None of the *cis*-pQTL were predicted to alter any motifs. VEP predicted consequences for 5 of the 6 *trans*-pQTL loci. All three of the chromosome 14 *trans*-pQTL were predicted to affect *SERPINA1*. As discussed previously, the PRG4 lead variant was a missense variant in *SERPINA1*, while the CFB and TF lead variant was an intronic variant. Both of the lead chromosome 16 *trans*-pQTL were predicted to affect *DHX38* and *PMFBP1* as noncoding variants. None of the *trans*-pQTL variants were predicted to affect regulatory regions or motifs.

5.6 Integration

5.6.1 Module 92

Module 92 was discussed previously as an interesting gene network (Figure 3.8) consisting of MHC class I molecules and butyrophilins (Table 3.1). Module 92 has module QTL on chromosome 6 and 16. Conditional *cis*-eQTL for *NLRC5*, a key regulator in this network, colocalise with the module QTL on chromosome 16. The module QTL association on chromosome 16 has 2 CSs from SuSiE consisting of one SNP each - rs821470 and rs12373120. The first variant (rs821470) was also the lead variant associated with *NLRC5*, with the minor allele rs821470^{G} being associated with increased expression. Overlap of the conditional *NLRC5 cis*-eQTL locus was observed in the group peak sets of a few subsets of T cells (Figure 5.11).

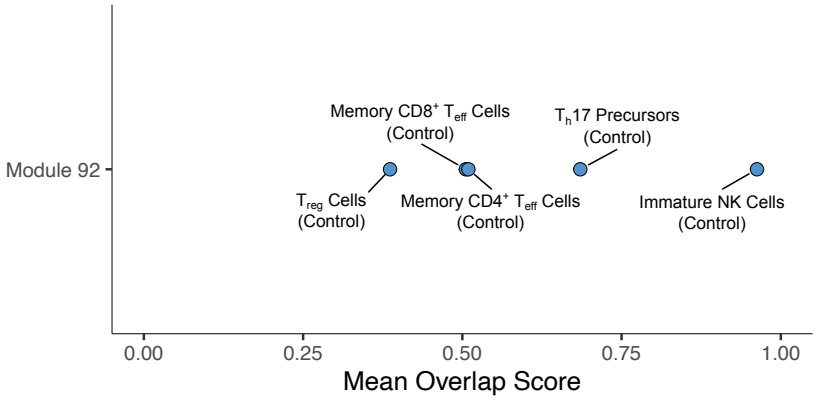

Figure 5.11: *NLRC5* GoShifter overlap scores. GoShifter overlap scores for the *NLRC5* locus are displayed, with lower scores indicating a lower probability of observing an overlap in the group peak set by chance.

The per-SNP heritability of the module 92 eigengene was enriched across all the group peak sets. The enrichment was higher in neutrophil accessible regions than other leukocytes, similar to patterns observed across the module eigengenes (Figure 5.8). The heritability was enriched the most in control and stimulated naive T_{reg} cells amongst the leukocytes in the immune atlas.

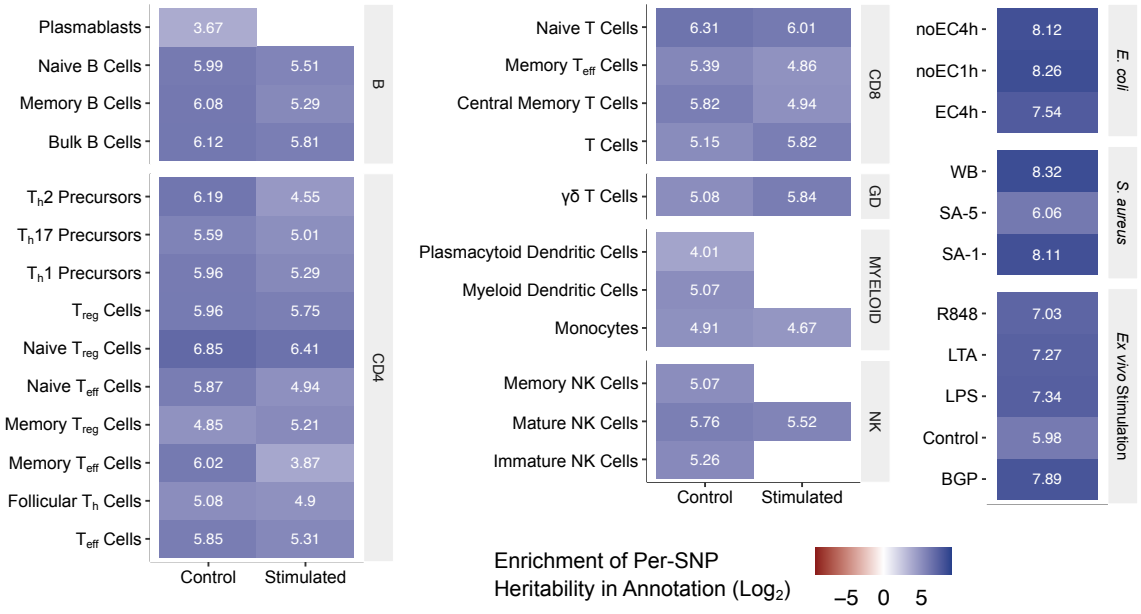

Figure 5.12: Module 92 eigengene heritability. This heatmap displays the $\log_2$ of the enrichment of per-SNP heritability ($h^2_{\mathrm{SNP}\alpha}/h^2_{\mathrm{SNP}}$) in each annotation for the module 92 eigengene.

5.6.2 Module 101

Module 101 is composed of 11 genes, 4 of which are protein coding genes - *SUOX*, *TMEM50A*, *RHD*, and *RPS26*. This module was of interest because it colocalised with GWAS associations for lymphocyte count, eosinophil count, and serum alanine aminotransferase level (Table 4.3).

The module QTL for module 101 on chromosome 12 colocalise with conditional *cis*-eQTL for two of the module members - *SUOX* and *RPS26*. In addition, the module QTL colocalise with conditional *cis*-eQTL for *GDF11*, which is not in the module. The top module eigengene had one CS from SuSiE with 6 SNPs. A 5' untranslated region (UTR) variant in *RPS26* (rs1131017) had the highest PIP of 0.62. The GoShifter results for the *RPS26* locus indicate that this locus is enriched specifically in neutrophil accessibility states (Figure 5.13).

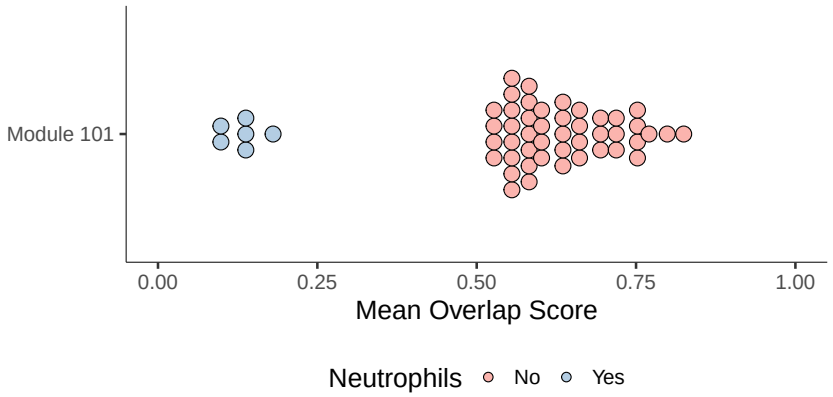

Figure 5.13: *RPS26* GoShifter overlap scores. GoShifter overlap scores for the *RPS26* locus are displayed, with lower scores indicating a lower probability of observing an overlap in the group peak set by chance. All the neutrophil peak sets have the lowest scores.

The minor allele in the QTL analysis was rs1131017^{C}. This allele is associated with increased expression of *RPS26*. Consistent with this, the variant falls in a binding site for GATA TFs (Table G.1) and rs1131017^{C} is predicted to increase the binding affinity of the TFs to the site. The rs1131017^{C} allele is also associated with reduced lymphocyte count, reduced alanine aminotransferase, and increased eosinophil count.

5.7 Discussion

In this chapter, I catalogued the results of reprocessing publicly available ATAC-seq data. In addition, I identified regions of the genome that are enriched for *cis*-eQTL, including cell-type-specific regions that can inform future inquiry of specific *cis*-eQTL. Finally, I have used a variant effect predictor to identify potential consequences of the molecular QTL explored in this thesis and integrated multiple analyses for a few modules of interest.

5.7.1 Enrichment of *cis*-eQTL

Enrichment tests for the various molecular QTL provide a method to compare these context-specific QTL to observations about QTL made in cohorts of healthy donors, particularly those fo-

cused on specific cell types. Similar to prior studies, the QTL were enriched in functionally-relevant regions such as enhancers and promoters. They were also enriched in accessible regions in all primary immune cell types. One of the goals of this analysis was to prioritise specific cell types that may be dysregulated in the sepsis response. However, the heterogeneous nature of the tissue meant that QTL were enriched uniformly in all primary immune cell types. The GoShifter results suggest that while some loci are enriched in accessible regions in all immune cell types and drive the overall observed enrichment, certain subsets of eQTL are particularly enriched in cell-type-specific accessibility profiles. The analysis of the heritability of module eigengenes similarly reveals that heritability for various molecular programs is distributed within specific accessibility profiles. Future efforts will focus on identifying subsets of context-specific QTL that are enhanced in sepsis compared to healthy cohorts to identify more meaningful cell type enrichment.

5.7.2 Partitioned Heritability

Methods that partition heritability provide insights into which functional regions of the genome contribute to the heritability of traits. Although well-established methods such as GCTA-GREML (Yang *et al.* 2010) and LD score regression (Bulik-Sullivan *et al.* 2015) exist, they do not provide the flexibility to account for repeat measurements from the same individuals. Thus, analysing the heritability of module eigengenes required a custom model. Partitioned heritability may identify enrichment without direct overlap of trait-associated variants with the annotation. For instance, the *NLRC5* conditional *cis*-eQTL locus overlapped accessible peaks in T_{reg} but not naive T_{reg} cells (Figure 5.11). The partitioned heritability analysis suggests that although overlap was not observed, the naive T_{reg} cells may be more relevant to *NLRC5* and module 92.

The partitioned heritability model failed for some annotations due to singular solutions, indicating an issue of identifiability when estimating the variance components. This may occur because the regions of the genome in the annotation or background set show little relatedness ($\mathbf{\Psi}_\alpha \approx \mathbf{I}_q$ or $\mathbf{\Psi}_{\bar{\alpha}} \approx \mathbf{I}_q$) or if the relatedness estimated using the annotation and background sets is similar ($\mathbf{\Psi}_\alpha \approx \mathbf{\Psi}_{\bar{\alpha}}$). Thus, the heritability model must be interpreted carefully on a case-by-case basis. One method of testing which components are non-identifiable is to use a complementary Bayesian Markov Chain Monte Carlo approach for the hierarchical model. Future work for this analysis is to identify meaningful bounds on the heritability estimates using confidence intervals to aid interpretability and to use Bayesian hierarchical models to identify the conditions under which the LMM fails.

5.7.3 Variant Effect Prediction

The variant effect prediction highlighted the need for high-throughput molecular expression experiments to better understand context-specific regulation. Although VEP was able to identify functional regions of relevance, only 44% of the lead conditional *cis*-eQTL and 65% of the lead *cis*-pQTL were predicted to affect their cognate gene. Prediction of variant effects is also complicated by the algorithms used. For instance, VEP was able to identify motifs that may be affected by variants. However, these motifs were identified on the reference genome based on chromatin immunoprecipitation sequencing (ChIP-seq) peaks in healthy cohorts (Zerbino *et al.* 2015) and then tested for alteration by an overlapping variant. Thus, motifs that are generated by the alternate allele or are context-specific are not captured in the initial set of motifs. The motif score changes observed are likely biased to a reduction in motif strength because the gain-of-function variants are not represented.

5.7.4 Integration

Recently, it has been suggested that eQTL and GWAS variants are fundamentally different due to the differing method of discovery used for each (Mostafavi *et al.* 2022). One of the factors affecting discovery is that GWAS variants are detected in diseased cohorts where trait-associated variants are expected to have a higher allele frequency compared with healthy cohorts in which eQTL are often mapped. Thus, it is critical to profile molecular expression and QTL in disease cohorts to improve discovery of context-specific QTL. A future step for this analysis is to characterise the constraint and regulatory complexity of the *cis*-eQTL and module QTL to make direct comparisons with Mostafavi *et al.* 2022 and test if using disease-relevant cohorts identifies *cis*-eQTL that better resemble GWAS variants.

5.7.5 Concluding Remarks

The aim of this thesis was to generate biological insights into the molecular heterogeneity underlying sepsis. I used leukocyte transcriptomics and plasma proteomics data from the GAinS cohort. Using previously characterised *cis*-eQTL and pQTL in addition to the module QTL, I demonstrated multiple methods of nominating hypotheses for the effects of QTL on sepsis-relevant molecular traits. The methods explored and developed in this thesis will be used to characterise specific subsets of QTL variants that are enhanced in sepsis.

Bibliography

Amemiya, Haley M., Anshul Kundaje, and Alan P. Boyle (June 2019). The ENCODE Blacklist: Identification of Problematic Regions of the Genome. *Scientific Reports* 9 (1), 9354. DOI: 10.1038/s41598-019-45839-z.

Angus, Derek C. and Tom van der Poll (Aug. 2013). Severe Sepsis and Septic Shock. *The New England Journal of Medicine* 369, 840–851. DOI: 10.1056/NEJMra1208623.

Antcliffe, David B. *et al.* (Apr. 2019). Transcriptomic Signatures in Sepsis and a Differential Response to Steroids. From the VANISH Randomized Trial. *American Journal of Respiratory and Critical Care Medicine* 199 (8), 980–986. DOI: 10.1164/rccm.201807-1419OC.

Aran, Dvir, Zicheng Hu, and Atul J. Butte (Nov. 2017). xCell: digitally portraying the tissue cellular heterogeneity landscape. *Genome Biology* 18 (1), 220. DOI: 10.1186/s13059-017-1349-1.

Arnett, Heather A. and Joanne L. Viney (Aug. 2014). Immune modulation by butyrophilins. *Nature Reviews Immunology* 14 (8), 559–569. DOI: 10.1038/nri3715.

Assarsson, Erika *et al.* (Apr. 2014). Homogenous 96-Plex PEA Immunoassay Exhibiting High Sensitivity, Specificity, and Excellent Scalability. *PLOS ONE* 9 (4), e95192. DOI: 10.1371/journal.pone.0095192.

Baghela, Arjun *et al.* (Jan. 2022). Predicting sepsis severity at first clinical presentation: The role of endotypes and mechanistic signatures. *eBioMedicine* 75, 103776. DOI: 10.1016/j.ebiom.2021.103776.

Bailey, Timothy L. and Charles E. Grant (Aug. 2021). SEA: Simple Enrichment Analysis of motifs. *bioRxiv*. DOI: 10.1101/2021.08.23.457422.

Bailey, Timothy L. *et al.* (July 2015). The MEME Suite. *Nucleic Acids Research* 43 (W1), W39–W49. DOI: 10.1093/nar/gkv416.

Barabási, Albert-László and Réka Albert (Oct. 1999). Emergence of Scaling in Random Networks. *Science* 286 (5439), 509–512. DOI: 10.1126/science.286.5439.509.

Bates, Douglas *et al.* (Oct. 2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* 67, 1–48. DOI: 10.18637/jss.v067.i01.

Benner, Christian *et al.* (May 2016). FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* 32 (10), 1493–1501. DOI: 10.1093/bioinformatics/btw018.

Bland, J. Martin and Douglas G. Altman (Mar. 1995). Calculating correlation coefficients with repeated observations: Part 2—correlation between subjects. *BMJ* 310, 633. DOI: 10.1136/bmj.310.6980.633.

Boomer, Jonathan S. *et al.* (Dec. 2011). Immunosuppression in Patients Who Die of Sepsis and Multiple Organ Failure. *JAMA* 306 (23), 2594–2605. DOI: 10.1001/jama.2011.1829.

Boyle, Alan P. *et al.* (Jan. 2008). High-Resolution Mapping and Characterization of Open Chromatin across the Genome. *Cell* 132 (2), 311–322. DOI: 10.1016/j.cell.2007.12.014.

Boyle, Evan A., Yang I. Li, and Jonathan K. Pritchard (June 2017). An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell* 169 (7), 1177–1186. DOI: 10.1016/j.cell.2017.05.038.

Buenrostro, Jason D. *et al.* (Dec. 2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods* 10 (12), 1213–1218. DOI: 10.1038/nmeth.2688.

Bulik-Sullivan, Brendan K. *et al.* (Mar. 2015). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics* 47 (3), 291–295. DOI: 10.1038/ng.3211.

Buniello, Annalisa *et al.* (Jan. 2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research* 47 (D1), D1005–D1012. DOI: 10.1093/nar/gky1120.

Burnham, Katie L. *et al.* (Aug. 2017). Shared and Distinct Aspects of the Sepsis Transcriptomic Response to Fecal Peritonitis and Pneumonia. *American Journal of Respiratory and Critical Care Medicine* 196 (3), 328–339. DOI: 10.1164/rccm.201608-1685OC.

Calandra, Thierry and Thierry Roger (Oct. 2003). Macrophage migration inhibitory factor: a regulator of innate immunity. *Nature Reviews Immunology* 3 (10), 791–800. DOI: 10.1038/nri1200.

Calderon, Diego *et al.* (Oct. 2019). Landscape of stimulation-responsive chromatin across diverse human immune cells. *Nature Genetics* 51 (10), 1494–1505. DOI: 10.1038/s41588-019-0505-9.

Cano-Gamez, Eddie *et al.* (Mar. 2022). An immune dysfunction score for stratification of patients with acute infection based on whole blood gene expression. *medRxiv*. DOI: 10.1101/2022.03.17.22272427.

Castro-Mondragon, Jaime A *et al.* (Jan. 2022). JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. *Nucleic Acids Research* 50 (D1), D165–D173. DOI: 10.1093/nar/gkab1113.

Chan, James K. *et al.* (Aug. 2012). Alarmins: awaiting a clinical response. *The Journal of Clinical Investigation* 122 (8), 2711–2719. DOI: 10.1172/JCI62423.

Chen, Kan *et al.* (Nov. 2012). Endocytosis of soluble immune complexes leads to their clearance by FcγRIIIB but induces neutrophil extracellular traps via FcγRIIA in vivo. *Blood* 120 (22), 4421–4431. DOI: 10.1182/blood-2011-12-401133.

Choi, Kyung-Chul *et al.* (Oct. 2006). Smad6 negatively regulates interleukin 1-receptor–Toll-like receptor signaling through direct interaction with the adaptor Pellino-1. *Nature Immunology* 7 (10), 1057–1065. DOI: 10.1038/ni1383.

Corces, M. Ryan *et al.* (Oct. 2016). Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nature Genetics* 48 (10), 1193–1203. DOI: 10.1038/ng.3646.

COvid-19 Multi-omics Blood ATlas (COMBAT) Consortium *et al.* (Mar. 2022). A blood atlas of COVID-19 defines hallmarks of disease severity and specificity. *Cell* 185 (5), 916–938.e58. DOI: 10.1016/j.cell.2022.01.012.

Cui, Jun *et al.* (Apr. 2010). NLRC5 Negatively Regulates the NF-κB and Type I Interferon Signaling Pathways. *Cell* 141 (3), 483–496. DOI: 10.1016/j.cell.2010.03.040.

Curtis, James *et al.* (May 2015). Susceptibility to tuberculosis is associated with variants in the ASAP1 gene encoding a regulator of dendritic cell migration. *Nature Genetics* 47 (5), 523–527. DOI: 10.1038/ng.3248.

Czaikoski, Paula Giselle *et al.* (Feb. 2016). Neutrophil Extracellular Traps Induce Organ Damage during Experimental and Clinical Sepsis. *PLOS ONE* 11 (2), e0148142. DOI: 10.1371/journal.pone.0148142.

Dam, Sipko van *et al.* (July 2018). Gene co-expression analysis for functional classification and gene–disease predictions. *Briefings in Bioinformatics* 19 (4), 575–592. DOI: 10.1093/bib/bbw139.

Danecek, Petr *et al.* (Feb. 2021). Twelve years of SAMtools and BCFtools. *GigaScience* 10 (2), giab008. DOI: 10.1093/gigascience/giab008.

Davenport, Emma E *et al.* (Apr. 2016). Genomic landscape of the individual host response and outcomes in sepsis: a prospective cohort study. *The Lancet Respiratory Medicine* 4 (4), 259–271. DOI: 10.1016/S2213-2600(16)00046-1.

Davenport, Emma E *et al.* (Oct. 2018). Discovering in vivo cytokine-eQTL interactions from a lupus clinical trial. *Genome Biology* 19 (1), 168. DOI: 10.1186/s13059-018-1560-8.

Davis, Joe R. *et al.* (Jan. 2016). An Efficient Multiple-Testing Adjustment for eQTL Studies that Accounts for Linkage Disequilibrium between Variants. *The American Journal of Human Genetics* 98 (1), 216–224. DOI: 10.1016/j.ajhg.2015.11.021.

Diamond, Michael S. and Michael Farzan (Jan. 2013). The broad-spectrum antiviral functions of IFIT and IFITM proteins. *Nature Reviews Immunology* 13 (1), 46–57. DOI: 10.1038/nri3344.

Dobin, Alexander *et al.* (Jan. 2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29 (1), 15–21. DOI: 10.1093/bioinformatics/bts635.

Durbin, Richard (May 2014). Efficient haplotype matching and storage using the positional Burrows–Wheeler transform (PBWT). *Bioinformatics* 30 (9), 1266–1272. DOI: 10.1093/bioinformatics/btu014.

Engelmann, Bernd and Steffen Massberg (Jan. 2013). Thrombosis as an intravascular effector of innate immunity. *Nature Reviews Immunology* 13 (1), 34–45. DOI: 10.1038/nri3345.

Ernst, Jason and Manolis Kellis (Mar. 2012). ChromHMM: automating chromatin-state discovery and characterization. *Nature Methods* 9 (3), 215–216. DOI: 10.1038/nmeth.1906.

Finucane, Hilary K. *et al.* (Nov. 2015). Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nature Genetics* 47 (11), 1228–1235. DOI: 10.1038/ng.3404.

Foley, Christopher N. *et al.* (Feb. 2021). A fast and efficient colocalization algorithm for identifying shared genetic risk factors across multiple traits. *Nature Communications* 12 (1), 764. DOI: 10.1038/s41467-020-20885-8.

Fort, Alexandre *et al.* (June 2017). MBV: a method to solve sample mislabeling and detect technical bias in large combined genotype and sequencing assay datasets. *Bioinformatics* 33 (12), 1895–1897. DOI: 10.1093/bioinformatics/btx074.

Ge, Tian *et al.* (May 2017). Heritability analysis with repeat measurements and its application to resting-state functional connectivity. *Proceedings of the National Academy of Sciences* 114 (21), 5521–5526. DOI: 10.1073/pnas.1700765114.

Giambartolomei, Claudia *et al.* (May 2014). Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. *PLOS Genetics* 10 (5), e1004383. DOI: 10.1371/journal.pgen.1004383.

Goh, Cyndi and Julian C Knight (Mar. 2017). Enhanced understanding of the host–pathogen interaction in sepsis: new opportunities for omic approaches. *The Lancet Respiratory Medicine* 5 (3), 212–223. DOI: 10.1016/S2213-2600(17)30045-0.

Goh, Cyndi *et al.* (June 2020). Epstein-Barr virus reactivation in sepsis due to community-acquired pneumonia is associated with increased morbidity and an immunosuppressed host transcriptomic endotype. *Scientific Reports* 10 (1), 9838. DOI: 10.1038/s41598-020-66713-3.

Gold, Larry *et al.* (Dec. 2010). Aptamer-Based Multiplexed Proteomic Technology for Biomarker Discovery. *PLOS ONE* 5 (12), e15004. DOI: 10.1371/journal.pone.0015004.

Grant, Charles E., Timothy L. Bailey, and William Stafford Noble (Apr. 2011). FIMO: scanning for occurrences of a given motif. *Bioinformatics* 27 (7), 1017–1018. DOI: 10.1093/bioinformatics/btr064.

Guo, Ren-Feng and Peter A. Ward (2005). Role of C5A in Inflammatory Responses. *Annual Review of Immunology* 23 (1), 821–852. DOI: 10.1146/annurev.immunol.23.021704.115835.

Gusev, Alexander *et al.* (Nov. 2014). Partitioning Heritability of Regulatory and Cell-Type-Specific Variants across 11 Common Diseases. *The American Journal of Human Genetics* 95 (5), 535–552. DOI: 10.1016/j.ajhg.2014.10.004.

Haraldsson, B. and B. Rippe (1987). Orosomucoid as one of the serum components contributing to normal capillary permselectivity in rat skeletal muscle. *Acta Physiologica Scandinavica* 129 (1), 127–135. DOI: 10.1111/j.1748-1716.1987.tb08047.x.

Heinz, Sven *et al.* (May 2010). Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Molecular Cell* 38 (4), 576–589. DOI: 10.1016/j.molcel.2010.05.004.

Hotchkiss, Richard S., Guillaume Monneret, and Didier Payen (Dec. 2013). Sepsis-induced immunosuppression: from cellular dysfunctions to immunotherapy. *Nature Reviews Immunology* 13 (12), 862–874. DOI: 10.1038/nri3552.

Huang, Qin Qin *et al.* (Dec. 2018). Power, false discovery rate and Winner's Curse in eQTL studies. *Nucleic Acids Research* 46 (22), e133. DOI: 10.1093/nar/gky780.

Huber, Wolfgang *et al.* (July 2002). Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics* 18 (suppl_1), S96–S104. DOI: 10.1093/bioinformatics/18.suppl_1.S96.

Hutchinson, Anna, Jennifer Asimit, and Chris Wallace (Sept. 2020). Fine-mapping genetic associations. *Human Molecular Genetics* 29 (R1), R81–R88. DOI: 10.1093/hmg/ddaa148.

Johnson, W. Evan, Cheng Li, and Ariel Rabinovic (Jan. 2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* 8 (1), 118–127. DOI: 10.1093/biostatistics/kxj037.

Kong, Andy T. *et al.* (May 2017). MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry–based proteomics. *Nature Methods* 14 (5), 513–520. DOI: 10.1038/nmeth.4256.

Kuhn, Robert M., David Haussler, and W. James Kent (Mar. 2013). The UCSC genome browser and associated tools. *Briefings in Bioinformatics* 14 (2), 144–161. DOI: 10.1093/bib/bbs038.

Kundaje, Anshul *et al.* (Feb. 2015). Integrative analysis of 111 reference human epigenomes. *Nature* 518 (7539), 317–330. DOI: 10.1038/nature14248.

Kwok, Andrew J. *et al.* (Mar. 2022). Identification of deleterious neutrophil states and altered granulopoiesis in sepsis. *medRxiv*. DOI: 10.1101/2022.03.22.22272723.

Langfelder, Peter and Steve Horvath (Dec. 2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 9 (1), 559. DOI: 10.1186/1471-2105-9-559.

Langmead, Ben and Steven L. Salzberg (Apr. 2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods* 9 (4), 357–359. DOI: 10.1038/nmeth.1923.

Law, Ruby HP *et al.* (May 2006). An overview of the serpin superfamily. *Genome Biology* 7 (5), 216. DOI: 10.1186/gb-2006-7-5-216.

Lawrence, Michael, Robert Gentleman, and Vincent Carey (July 2009). rtracklayer: an R package for interfacing with genome browsers. *Bioinformatics* 25 (14), 1841–1842. DOI: 10.1093/bioinformatics/btp328.

Lawrence, Michael *et al.* (Aug. 2013). Software for Computing and Annotating Genomic Ranges. *PLOS Computational Biology* 9 (8), e1003118. DOI: 10.1371/journal.pcbi.1003118.

Levi, Marcel and Tom van der Poll (Jan. 2017). Coagulation and sepsis. *Thrombosis Research* 149, 38–44. DOI: 10.1016/j.thromres.2016.11.007.

Liao, Yang, Gordon K. Smyth, and Wei Shi (Apr. 2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 30 (7), 923–930. DOI: 10.1093/bioinformatics/btt656.

Liston, Adrian *et al.* (Nov. 2021). Human immune diversity: from evolution to modernity. *Nature Immunology* 22, 1479–1489. DOI: 10.1038/s41590-021-01058-1.

Loh, Po-Ru, Pier Francesco Palamara, and Alkes L. Price (July 2016). Fast and accurate long-range phasing in a UK Biobank cohort. *Nature Genetics* 48 (7), 811–816. DOI: 10.1038/ng.3571.

Luyt, Charles-Edouard *et al.* (May 2007). Herpes Simplex Virus Lung Infection in Patients Undergoing Prolonged Mechanical Ventilation. *American Journal of Respiratory and Critical Care Medicine* 175 (9), 935–942. DOI: 10.1164/rccm.200609-1322OC.

Manolio, Teri A. *et al.* (Oct. 2009). Finding the missing heritability of complex diseases. *Nature* 461 (7265), 747–753. DOI: 10.1038/nature08494.

Marshall, John C. (Apr. 2014). Why have clinical trials in sepsis failed? *Trends in Molecular Medicine*. Special issue : Sepsis 20 (4), 195–203. DOI: 10.1016/j.molmed.2014.01.007.

McCarthy, Shane *et al.* (Oct. 2016). A reference panel of 64,976 haplotypes for genotype imputation. *Nature Genetics* 48 (10), 1279–1283. DOI: 10.1038/ng.3643.

McLaren, William *et al.* (June 2016). The Ensembl Variant Effect Predictor. *Genome Biology* 17 (1), 122. DOI: 10.1186/s13059-016-0974-4.

Meissner, Torsten B. *et al.* (Aug. 2010). NLR family member NLRC5 is a transcriptional regulator of MHC class I genes. *Proceedings of the National Academy of Sciences* 107 (31), 13794–13799. DOI: 10.1073/pnas.1008684107.

Migeotte, Isabelle, David Communi, and Marc Parmentier (Dec. 2006). Formyl peptide receptors: A promiscuous subfamily of G protein-coupled receptors controlling immune responses. *Cytokine & Growth Factor Reviews* 17 (6), 501–519. DOI: 10.1016/j.cytogfr.2006.09.009.

Min, Alan, Elizabeth Thompson, and Saonli Basu (June 2022). Comparing Heritability Estimators under Alternative Structures of Linkage Disequilibrium. *G3 Genes|Genomes|Genetics*, jkac134. DOI: 10.1093/g3journal/jkac134.

Moore, Jill E. *et al.* (July 2020). Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* 583 (7818), 699–710. DOI: 10.1038/s41586-020-2493-4.

Mostafavi, Hakhamanesh *et al.* (May 2022). Limited overlap of eQTLs and GWAS hits due to systematic differences in discovery. *bioRxiv*. DOI: 10.1101/2022.05.07.491045.

Nath, Artika P. *et al.* (Dec. 2019). Multivariate Genome-wide Association Analysis of a Cytokine Network Reveals Variants with Widespread Immune, Haematological, and Cardiometabolic Pleiotropy. *The American Journal of Human Genetics* 105 (6), 1076–1090. DOI: 10.1016/j.ajhg.2019.10.001.

Newman, Aaron M. *et al.* (July 2019). Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nature Biotechnology* 37 (7), 773–782. DOI: 10.1038/s41587-019-0114-2.

Niemi, Mari E. K. *et al.* (Dec. 2021). Mapping the human genetic architecture of COVID-19. *Nature* 600 (7889), 472–477. DOI: 10.1038/s41586-021-03767-x.

Parsana, Princy *et al.* (May 2019). Addressing confounding artifacts in reconstruction of gene co-expression networks. *Genome Biology* 20 (1), 94. DOI: 10.1186/s13059-019-1700-9.

Pers, Tune H., Pascal Timshel, and Joel N. Hirschhorn (Feb. 2015). SNPsnap: a Web-based tool for identification and annotation of matched SNPs. *Bioinformatics* 31 (3), 418–420. DOI: 10.1093/bioinformatics/btu655.

Peters-Sengers, Hessel *et al.* (Jan. 2022). Source-specific host response and outcomes in critically ill patients with sepsis: a prospective cohort study. *Intensive Care Medicine* 48 (1), 92–102. DOI: 10.1007/s00134-021-06574-0.

Poll, Tom van der *et al.* (July 2017). The immunopathology of sepsis and potential therapeutic targets. *Nature Reviews Immunology* 17 (7), 407–420. DOI: 10.1038/nri.2017.36.

Porcu, Eleonora *et al.* (May 2021). Causal Inference Methods to Integrate Omics and Complex Traits. *Cold Spring Harbor Perspectives in Medicine* 11 (5), a040493. DOI: 10.1101/cshperspect.a040493.

Pruenster, Monika *et al.* (Nov. 2016). S100A8/A9: From basic science to clinical application. *Pharmacology & Therapeutics* 167, 120–131. DOI: 10.1016/j.pharmthera.2016.07.015.

Purcell, Shaun *et al.* (Sept. 2007). PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics* 81 (3), 559–575. DOI: 10.1086/519795.

Quinlan, Aaron R. and Ira M. Hall (Mar. 2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26 (6), 841–842. DOI: 10.1093/bioinformatics/btq033.

Ram-Mohan, Nikhil *et al.* (Aug. 2021). Profiling chromatin accessibility responses in human neutrophils with sensitive pathogen detection. *Life Science Alliance* 4 (8). DOI: 10.26508/lsa.202000976.

Rautanen, Anna *et al.* (Jan. 2015). Genome-wide association study of survival from sepsis due to pneumonia: an observational cohort study. *The Lancet Respiratory Medicine* 3 (1), 53–60. DOI: 10.1016/S2213-2600(14)70290-5.

Ribeiro, Diogo M. *et al.* (Aug. 2021). The molecular basis, genetic control and pleiotropic effects of local gene co-expression. *Nature Communications* 12 (1), 4842. DOI: 10.1038/s41467-021-25129-x.

Sandelin, Albin *et al.* (Jan. 2004). JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Research* 32 (suppl_1), D91–D94. DOI: 10.1093/nar/gkh012.

Sanderson, Eleanor *et al.* (Feb. 2022). Mendelian randomization. *Nature Reviews Methods Primers* 2 (1), 1–21. DOI: 10.1038/s43586-021-00092-5.

Schaid, Daniel J., Wenan Chen, and Nicholas B. Larson (Aug. 2018). From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nature Reviews Genetics* 19 (8), 491–504. DOI: 10.1038/s41576-018-0016-z.

Scherag, André *et al.* (Oct. 2016). Genetic Factors of the Disease Course after Sepsis: A Genome-Wide Study for 28Day Mortality. *EBioMedicine* 12, 239–246. DOI: 10.1016/j.ebiom.2016.08.043.

Scicluna, Brendon P *et al.* (Oct. 2017). Classification of patients with sepsis according to blood genomic endotype: a prospective cohort study. *The Lancet Respiratory Medicine* 5 (10), 816–826. DOI: 10.1016/S2213-2600(17)30294-1.

Scicluna, Brendon P. *et al.* (Oct. 2015). A Molecular Biomarker to Diagnose Community-acquired Pneumonia on Intensive Care Unit Admission. *American Journal of Respiratory and Critical Care Medicine* 192 (7), 826–835. DOI: 10.1164/rccm.201502-0355OC.

Shalova, Irina N. *et al.* (Mar. 2015). Human Monocytes Undergo Functional Re-programming during Sepsis Mediated by Hypoxia-Inducible Factor-1α. *Immunity* 42 (3), 484–498. DOI: 10.1016/j.immuni.2015.02.001.

Singer, Mervyn *et al.* (Feb. 2016). The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). *JAMA* 315 (8), 801–810. DOI: 10.1001/jama.2016.0287.

Slatkin, Montgomery (June 2008). Linkage disequilibrium — understanding the evolutionary past and mapping the medical future. *Nature Reviews Genetics* 9 (6), 477–485. DOI: 10.1038/nrg2361.

Sørensen, Thorkild I. A. *et al.* (Mar. 1988). Genetic and Environmental Influences on Premature Death in Adult Adoptees. *New England Journal of Medicine* 318, 727–732. DOI: 10.1056/NEJM198803243181202.

Sörensson, Jenny *et al.* (Feb. 1999). Human endothelial cells produce orosomucoid, an important component of the capillary barrier. *American Journal of Physiology-Heart and Circulatory Physiology* 276 (2), H530–H534. DOI: 10.1152/ajpheart.1999.276.2.H530.

Soskic, Blagoje *et al.* (Oct. 2019). Chromatin activity at GWAS loci identifies T cell states driving complex immune diseases. *Nature Genetics* 51 (10), 1486–1493. DOI: 10.1038/s41588-019-0493-9.

Starokadomskyy, Petro *et al.* (May 2013). CCDC22 deficiency in humans blunts activation of proinflammatory NF-κB signaling. *The Journal of Clinical Investigation* 123 (5), 2244–2256. DOI: 10.1172/JCI66466.

Stegle, Oliver *et al.* (Mar. 2012). Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nature Protocols* 7 (3), 500–507. DOI: 10.1038/nprot.2011.457.

Stoppelaar, Sacha F. de, Cornelis van 't Veer, and Tom van der Poll (2014). The role of platelets in sepsis. *Thrombosis and Haemostasis* 112 (10), 666–677. DOI: 10.1160/TH14-02-0126.

Sweeney, Timothy E. *et al.* (Feb. 2018). A community approach to mortality prediction in sepsis via gene expression analysis. *Nature Communications* 9 (1), 694. DOI: 10.1038/s41467-018-03078-2.

Takeuchi, Osamu and Shizuo Akira (Mar. 2010). Pattern Recognition Receptors and Inflammation. *Cell* 140 (6), 805–820. DOI: 10.1016/j.cell.2010.01.022.

Tannahill, G. M. *et al.* (Apr. 2013). Succinate is an inflammatory signal that induces IL-1β through HIF-1α. *Nature* 496 (7444), 238–242. DOI: 10.1038/nature11986.

The 1000 Genomes Project Consortium *et al.* (Oct. 2015). A global reference for human genetic variation. *Nature* 526, 68–74. DOI: 10.1038/nature15393.

Tian, Chao *et al.* (Sept. 2017). Genome-wide association and HLA region fine-mapping studies identify susceptibility loci for multiple common infections. *Nature Communications* 8 (1), 599. DOI: 10.1038/s41467-017-00257-5.

Tridente, Ascanio *et al.* (Feb. 2014). Patients with faecal peritonitis admitted to European intensive care units: an epidemiological survey of the GenOSept cohort. *Intensive Care Medicine* 40 (2), 202–210. DOI: 10.1007/s00134-013-3158-7.

Trynka, Gosia *et al.* (July 2015). Disentangling the Effects of Colocalizing Genomic Annotations to Functionally Prioritize Non-coding Variants within Complex-Trait Loci. *The American Journal of Human Genetics* 97 (1), 139–152. DOI: 10.1016/j.ajhg.2015.05.016.

Veiga Leprevost, Felipe da *et al.* (Sept. 2020). Philosopher: a versatile toolkit for shotgun proteomics data analysis. *Nature Methods* 17 (9), 869–870. DOI: 10.1038/s41592-020-0912-y.

Visscher, Peter M., William G. Hill, and Naomi R. Wray (Apr. 2008). Heritability in the genomics era — concepts and misconceptions. *Nature Reviews Genetics* 9 (4), 255–266. DOI: 10.1038/nrg2322.

Võsa, Urmo *et al.* (Sept. 2021). Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. *Nature Genetics* 53, 1300–1310. DOI: 10.1038/s41588-021-00913-z.

Walden, Andrew P. *et al.* (Apr. 2014). Patients with community acquired pneumonia admitted to European intensive care units: an epidemiological survey of the GenOSept cohort. *Critical Care* 18 (2), R58. DOI: 10.1186/cc13812.

Wallace, Chris (Sept. 2021). A more accurate method for colocalisation analysis allowing for multiple causal variants. *PLOS Genetics* 17 (9), e1009440. DOI: 10.1371/journal.pgen.1009440.

Wang, Gao *et al.* (2020). A simple new approach to variable selection in regression, with application to genetic fine mapping. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 82 (5), 1273–1300. DOI: 10.1111/rssb.12388.

Wang, Lili and Xuanyao Liu (May 2022a). Multivariate association testing on gene modules improves the power of trans-eQTLs detection. Poster. Cold Spring Harbor, New York.

Wang, Yi, Stephanie C. Hicks, and Kasper D. Hansen (Mar. 2022b). Addressing the mean-correlation relationship in co-expression analysis. *PLOS Computational Biology* 18 (3), e1009954. DOI: 10.1371/journal.pcbi.1009954.

Westra, Harm-Jan *et al.* (Aug. 2011). MixupMapper: correcting sample mix-ups in genome-wide datasets increases power to detect small genetic effects. *Bioinformatics* 27 (15), 2104–2111. DOI: 10.1093/bioinformatics/btr323.

Wijst, Monique G. P. van der *et al.* (Apr. 2018). Single-cell RNA sequencing identifies celltype-specific cis-eQTLs and co-expression QTLs. *Nature Genetics* 50 (4), 493–497. DOI: 10.1038/s41588-018-0089-9.

Wu, Tianzhi *et al.* (Aug. 2021). clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innovation* 2 (3), 100141. DOI: 10.1016/j.xinn.2021.100141.

Xiao, Wenzhong *et al.* (Nov. 2011). A genomic storm in critically injured humans. *Journal of Experimental Medicine* 208 (13), 2581–2590. DOI: 10.1084/jem.20111354.

Yan, Feng *et al.* (Feb. 2020). From reads to insight: a hitchhiker's guide to ATAC-seq data analysis. *Genome Biology* 21 (1), 22. DOI: 10.1186/s13059-020-1929-3.

Yang, Jian *et al.* (July 2010). Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics* 42 (7), 565–569. DOI: 10.1038/ng.608.

Yang, Jian *et al.* (June 2011). Genome partitioning of genetic variation for complex traits using common SNPs. *Nature Genetics* 43 (6), 519–525. DOI: 10.1038/ng.823.

Yang, Jian *et al.* (Apr. 2012). Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nature Genetics* 44 (4), 369–375. DOI: 10.1038/ng.2213.

Yates, Andrew D *et al.* (Jan. 2020). Ensembl 2020. *Nucleic Acids Research* 48 (D1), D682–D688. DOI: 10.1093/nar/gkz966.

Yu, Fengchao *et al.* (Sept. 2020). Fast Quantitative Analysis of timsTOF PASEF Data with MSFragger and IonQuant. *Molecular & Cellular Proteomics* 19 (9), 1575–1585. DOI: 10.1074/mcp.TIR120.002048.

Yu, Guangchuang and Qing-Yu He (Jan. 2016). ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization. *Molecular BioSystems* 12 (2), 477–479. DOI: 10.1039/C5MB00663E.

Zerbino, Daniel R. *et al.* (Mar. 2015). The Ensembl Regulatory Build. *Genome Biology* 16 (1), 56. DOI: 10.1186/s13059-015-0621-5.

Zhang, Bin and Steve Horvath (Aug. 2005). A General Framework for Weighted Gene Co-Expression Network Analysis. *Statistical Applications in Genetics and Molecular Biology* 4 (1). DOI: 10.2202/1544-6115.1128.

Zhang, Yong *et al.* (Sept. 2008). Model-based Analysis of ChIP-Seq (MACS). *Genome Biology* 9 (9), R137. DOI: 10.1186/gb-2008-9-9-r137.

Zhernakova, Daria V. *et al.* (Jan. 2017). Identification of context-dependent expression quantitative trait loci in whole blood. *Nature Genetics* 49 (1), 139–145. DOI: 10.1038/ng.3737.

Ziyatdinov, Andrey *et al.* (Feb. 2018). lme4qtl: linear mixed models with flexible covariance structure for genetic studies of related individuals. *BMC Bioinformatics* 19 (1), 68. DOI: 10.1186/s12859-018-2057-x.

A | Prior Work in GAinS

A.1 Genotyping

Genomic DNA was purified from whole blood or buffy coat and genotyped using three different Illumina arrays. The Illumina HumanOmniExpress BeadChip (730,525 SNPs; Illumina, San Diego, CA, USA) was used for 295 patients, the Infinium CoreExome BeadChip (551,839 SNPs; Illumina, San Diego, CA, USA) was used for 655 patients, and the Infinium Global Screening Array BeadChip (654,027 SNPs; Illumina, San Diego, CA, USA) was used for 307 patients.

A.2 Genotype Imputation

To increase the density of SNPs used in downstream analyses, genotypes were imputed based on the initial array data. Samples were excluded from each batch based on discordant sex information, genotyping missingness of greater than 2%, outlying heterozygosity rate, Hardy-Weinberg equilibrium p-value less than 1×10^{-5}, and high identity by descent ($\hat{\pi} \geq 0.1875$) using PLINK (Purcell *et al.* 2007). Sample mismatches with gene expression microarray data or RNA-seq were detected and excluded using MixupMapper (Westra *et al.* 2011) and MBV from QTLtools (Fort *et al.* 2017) respectively. Variants were also filtered in preparation for imputation using the McCarthy Group's pre-imputation check tool (`https://www.well.ox.ac.uk/~wrayner/tools/`). Genotypes from each batch were imputed into the Phase 1 Haplotype Reference Consortium (HRC) Panel (McCarthy *et al.* 2016) using the Sanger Imputation Service. Briefly, genotypes were phased using EAGLE2 (Loh *et al.* 2016) and imputed using PBWT (Durbin 2014). Variants with an imputation score less than 0.9 were removed. Data from the three different batches were then merged. After merging, variants with more than 2% missingness and MAF less than 1% were excluded. The SNP coordinates were then lifted over to the GRCh38 using liftOver (Kuhn *et al.* 2013).

A.3 RNA Sequencing

Whole blood leukocytes were isolated using LeukoLOCK (Thermo Fisher Scientific, Waltham, MA, USA) depletion filter technology. RNA was extracted using the Total RNA Isolation Protocol (Ambion, Thermo Fisher Scientific, Waltham, MA, USA). RNA-seq was performed on 864 samples from 667 patients from the GAinS study, as described previously (Cano-Gamez *et al.* 2022). NEB Ultra II Library Prep kits (Illumina, San Diego, CA, USA) were used to prepare complementary DNA (cDNA) libraries, which were sequenced on NovaSeq 6000 sequencers (Illumina, San Diego, CA, USA). Reads were aligned to Ensembl GRCh38 version 99 (Yates *et al.* 2020) using STAR v2.7.3 (Dobin *et al.* 2013) and quantified using featureCounts (Liao *et al.* 2014). Reads were reassigned based on individual patient imputed HLA types. Expressed genes were defined as those with at least 10 reads in 5% of samples. Count data for expressed genes was normalised and transformed into logCPM.

A.4 Microarray Gene Expression

Before RNA-seq was performed, the initial subset of recruited patients were assayed using microarrays (Davenport *et al.* 2016; Burnham *et al.* 2017; Cano-Gamez *et al.* 2022). This analysis was performed for 676 samples (514 patients), of which 134 samples overlapped with the RNA-seq study. Illumia HumanHT-12 v4 Expression BeadChip microarrays (Illumina, San Diego, CA, USA) were used to quantify transcript levels. Processing of the microarray data has been described previously (Cano-Gamez *et al.* 2022). Briefly, initial raw data processing was performed using GenomeStudio. The vsn package was used for background subtraction, quality control, transformation, and normalisation (Huber *et al.* 2002). Probes were filtered and measurements were averaged across all probes assigned to a gene. Batch effects were corrected using ComBat (Johnson *et al.* 2007).

A.5 Mass Spectrometry

The plasma proteome was assayed using high-throughput liquid chromatography with tandem mass spectrometry (LC-MS-MS) based on a previously described method (COvid-19 Multi-omics Blood ATlas (COMBAT) Consortium *et al.* 2022). This method was used to assay 1,680 samples from 1,068 patients. No affinity depletion was applied to the patient samples. Mass spectrometry data was acquired in PASEF mode from an Evosep One LC system connected to the TimsTOF

Pro mass spectrometer (Bruker Daltonics, Billerica, MA, USA). A Fragpipe pipeline consisting of Fragpipe 13.0, MSFragger 3.0 (Kong *et al.* 2017), and Philosopher 3.2.9 (Veiga Leprevost *et al.* 2020) was used to analyse the data. A library of human UniProt SwissProt sequences was used to generate the Philosopher database. Label-free quantification was conducted using IonQuant (Yu *et al.* 2020). Raw protein intensities were pre-processed through steps including protein and sample filtering, normalisation, and imputation.

A.6 Mapping of eQTL

For each expressed gene from the RNA-seq data, variants that were in a 1 Mb window around the TSS were tested. In addition to the genotype, seven genotyping PCs, 20 PEER factors (Stegle *et al.* 2012), SRS status (SRS1 versus non-SRS1), diagnosis (CAP versus FP), and cell proportions were used as fixed-effect covariates. The number of genotyping PCs was determined by identifying the elbow point in the scree plot. PEER factors were calculated based on all expressed genes, holding out the seven genotyping PCs, SRS status, diagnosis, and cell proportions. The number of PEER factors to use was determined by identifying the elbow point in the number of *cis*-eQTL detected to balance the addition of more explanatory covariates with maximising *cis*-eQTL discovery. Associated variants were identified using a hierarchical approach for multiple testing correction as described previously (Huang *et al.* 2018). For each gene, local FDR correction was performed using eigenMT (Davis *et al.* 2016), which uses local LD structure to estimate the effective number of tests performed. The corrected p-values of the lead SNPs were then adjusted using a Benjamini-Hochberg FDR correction. All genes with FDR-corrected lead SNPs with p-values less than 0.05 were considered to have *cis*-eQTL. The global FDR-corrected threshold was then used to determine the nominal p-value threshold at each gene locus. Finally, all variants passing the nominal p-value threshold at a locus were considered *cis*-eQTL for the gene.

Gene expression is often under combinatorial regulation by multiple *cis*-regulatory elements. Secondary *cis*-eQTL for a gene were discovered as previously described (Huang *et al.* 2018) for all eGenes from the initial pass. For each eGene, the most significant eQTL discovered in the initial mapping (lead eQTL) was added to the LMM as a fixed-effect covariate. All other SNPs in the *cis* window of the eGene were tested for association based on the local FDR threshold determined previously. Any top secondary eQTL discovered was added to the model as a covariate and the process was repeated until no new secondary eQTL were detected. For the set of *cis*-eQTL detected using this iterative forward regression approach, a backwards pass was performed by leaving one *cis*-eQTL out at a time in the model and scanning the *cis* window to ensure that an

association was detected and to identify the best signal SNP while accounting for all other signals. This final set of *cis*-eQTL were called conditional *cis*-eQTL. For all eGenes with conditional *cis*-eQTL, summary statistics were generated for each signal SNP by using models conditioning on all other signal SNPs for the eGene.

A.7 Mapping of pQTL

Similar to eQTL, pQTL are variants associated with protein expression. Due to the small number of proteins detected through LC-MS-MS, a *trans* approach was used when testing for pQTL by testing all 4,276,557 SNPs genome-wide. Seven genotyping PCs, 34 protein expression PCs, age, and sex were included as fixed-effect covariates. The number of protein expression PCs was determined by choosing the minimum required to explain at least 60% of the variation in protein expression. A genome-wide threshold of 1.86×10^{-10} was used based on a Bonferroni FDR correction of 5×10^{-8} for 269 proteins. If a pQTL was detected within 1 Mb of the TSS of the cognate gene of the protein, it was considered a *cis*-pQTL. Otherwise, the pQTL was considered a *trans*-pQTL. Loci were defined by constructing 1 Mb windows around each pQTL and merging intervals until a set of disjoint intervals was generated.

B | Summary Statistics

Trait Group	Trait	Study (EBI GWAS Catalog Accession)
Serum Proteins	C-reactive Protein (Inflammation Marker)	GCST009777
	Alanine Aminotransferase (Liver Function)	GCST90013405
	Alkaline Phosphatase	GCST90013406
	Interleukin 18	GCST90012024
Primary Cell Trait	Lymphocyte Count / Proportion	GCST90002388
	Neutrophil Count / Proportion	GCST90002398
	Monocyte Count / Proportion	GCST90002393
	Eosinophil Count / Proportion	GCST90002381
	Basophil Count / Proportion	GCST004618
	Platelet Count	GCST90002402
	Erythrocyte Count	GCST90002403
Autoimmune Disease	Rheumatoid Arthritis	GCST005569
	Inflammatory Bowel Disease	GCST004131
	Systemic Lupus Erythematosus	GCST003156

Table B.1: Summary statistics from GWAS analyses. I retrieved GWAS summary statistics from the following studies from the EBI GWAS Catalog to test for colocalisation between the module QTL and trait-associated SNPs.

Table B.2: EBI GWAS and module QTL overlap studies. These EBI GWAS studies that are relevant to IMDs contained lead variants that were also module QTL.

Trait Group	Trait	Accession	Modules
Susceptibility to Infection	HIV-1	GCST000549	84, 92, 97
		GCST000863	84, 92
		GCST002101	84
		GCST003183	84, 92
	Chickenpox	GCST004999	81
	Shingles	GCST005001	81, 84

Continued on next page

Trait Group	Trait	Accession	Modules
	Epstein-Barr Virus	GCST001812	84
		GCST003339	84
	Hepatitis B Virus	GCST001150	84
		GCST002068	84
		GCST002879	84
		GCST005004	84
	Hepatitis C Virus	GCST001815	84
		GCST001867	84
		GCST002316	84
		GCST007633	84
		GCST90018805	84
	Mononucleosis	GCST005002	84
	Mycobacterium tuberculosis	GCST005006	84
	Pneumonia	GCST005009	84
	Scarlet Fever	GCST005008	84
Serum Proteins	C-reactive Protein (Inflammation Marker)	GCST003680	84
		GCST007614	84, 94
		GCST007615	94
		GCST009777	84, 103
		GCST90018950	103
		GCST90019499	103
	Alanine Aminotransferase (Liver Function)	GCST90013405	101
		GCST90013663	102
		GCST90018943	62, 101
		GCST90020236	101
	Aspartate Aminotransferase (Liver Function)	GCST90011899	101
		GCST90013664	102
		GCST90019497	69
		GCST90020237	63, 81, 101
	Albumin	GCST90018945	102
		GCST90019493	69
	Urate	GCST008972	84, 92
		GCST011119	69, 81, 94
	Alkaline Phosphatase	GCST90013406	69
		GCST90019494	69, 106

Continued on next page

Trait Group	Trait	Accession	Modules
	Creatinine	GCST90018979	84
		GCST90019502	94
	Cystatin C (Kidney Function)	GCST90019504	64, 94, 102
	IgM	GCST008575	81
	IgA	GCST008568	62
		GCST011917	84
	IgE	GCST001316	81, 84
		GCST002302	84
	Beta-2-microglobulin	GCST001863	84
	Complement C4	GCST011833	84
	IgG Glycosylation	GCST001848	88
		GCST004925	88
		GCST005656	88
		GCST009860	88
	Interleukin 18	GCST90012024	89
	Interleukin 1β	GCST008209	97
Primary Cell Trait	Lymphocyte Count / Proportion	GCST004130	92
		GCST004627	81, 84
		GCST004632	69, 81, 97
		GCST90001554	92
		GCST90001560	84
		GCST90001568	84
		GCST90001684	92
		GCST90001687	92
		GCST90002316	97, 101
		GCST90002320	92, 97, 101
		GCST90002388	84, 92, 101
		GCST90002389	69, 81, 84, 97, 103
		GCST90085815	97
		GCST004130	81, 84, 92
	Neutrophil Count / Proportion	GCST002557	84
		GCST004623	81, 84, 101
		GCST004629	84, 103
		GCST004633	103
		GCST90002351	69, 84, 103
		GCST90002355	103

Continued on next page

Trait Group	Trait	Accession	Modules
		GCST90002398	81, 92, 103
		GCST90002399	81, 103
		GCST90018968	103
		GCST90056178	81, 84
	Monocyte Count / Proportion	GCST004609	103, 106
		GCST004625	84, 103
		GCST90002340	84, 103
		GCST90002344	84, 103
		GCST90002393	81, 84, 103
		GCST90002394	75, 103, 106
		GCST90018967	103
		GCST90056177	81, 84, 103
	Eosinophil Count / Proportion	GCST004600	81, 84, 101
		GCST004606	81, 84, 101
		GCST004617	81, 84, 101
		GCST007065	69
		GCST009457	81
		GCST90002298	84, 94, 99, 101
		GCST90002302	84, 94, 101
		GCST90002381	59, 81, 84, 101, 106
		GCST90002382	81, 84, 101, 106
		GCST90018733	84
		GCST90018953	84, 101
		GCST90056180	69, 81, 84
	Basophil Count / Proportion	GCST90056179	81, 84
	Platelet Count	GCST004599	59, 84
		GCST004603	63, 84
		GCST004607	84, 88
		GCST90002346	62
		GCST90002349	62
		GCST90002357	63, 86
		GCST90002358	81
		GCST90002361	63, 86
		GCST90002395	59, 84
		GCST90002400	59, 86, 88, 91
		GCST90002402	63, 84, 88

Continued on next page

Trait Group	Trait	Accession	Modules
		GCST90018969	97, 102, 104
		GCST90056183	84, 88
	Erythrocyte Count	GCST004008	69
		GCST004601	91
		GCST007069	69, 80, 82, 91, 102
		GCST90002363	62, 69, 80, 82
		GCST90002367	69, 82, 91, 102
		GCST90002403	81, 102
		GCST90018971	80, 102
Autoimmune Disease	Rheumatoid Arthritis	GCST000040	84
		GCST000677	84, 99
		GCST000679	99
		GCST000917	84
		GCST001042	84
		GCST002318	62, 84, 99
		GCST002323	84
		GCST002357	84
		GCST002433	84, 99
		GCST002434	99
		GCST005562	84
		GCST005568	99
		GCST005569	99
		GCST006048	99
		GCST006959	84, 99
		GCST90013534	99
		GCST90013684	99
		GCST90018690	99
		GCST90018910	99
	Inflammatory Bowel Disease	GCST000225	84
		GCST000531	91
		GCST001725	91, 94
		GCST003043	91, 94, 99
		GCST004131	71, 94
	Coeliac Disease	GCST000048	84
		GCST000612	84
		GCST002112	84
	Psoriasis	GCST000165	84
		GCST000173	84, 92

Continued on next page

Trait Group	Trait	Accession	Modules
		GCST000322	84, 92
		GCST000833	84, 92
		GCST000836	84, 92
		GCST005527	84, 92
		GCST008096	92
	Systemic Lupus Erythematosus	GCST000142	84
		GCST000144	84
		GCST000996	84
		GCST001795	84
		GCST002463	84
		GCST003103	84
		GCST003155	84
		GCST003156	84
		GCST003622	84
		GCST004867	84
		GCST005752	84
		GCST011426	92
		GCST011956	81, 84
		GCST90020042	84
	Multiple Sclerosis	GCST000062	84
		GCST000252	84
		GCST000424	81, 84
		GCST000425	84
		GCST000593	84
		GCST000716	84
		GCST001341	84
		GCST001459	84
		GCST001891	84
		GCST003566	84, 102
	Alopecia Areata	GCST000719	84, 101
		GCST004866	84, 101
	Vitiligo	GCST000662	81, 84
		GCST000692	99
		GCST001509	101
		GCST001670	101
		GCST004785	99
	Type 1 Diabetes	GCST000038	101
		GCST000043	101

Continued on next page

Trait Group	Trait	Accession	Modules
		GCST000054	84
		GCST000141	101
		GCST000258	101
		GCST000392	91, 101
		GCST001191	91, 101
		GCST007433	84
		GCST008377	101
		GCST009916	84
		GCST90013445	62, 69
		GCST90018925	101
	Graves Disease	GCST001200	84, 99
		GCST001219	81, 97
		GCST90018847	99
	Myasthenia Gravis	GCST90093465	84

C | Publicly Available ATAC-seq Data

Of the 213 ATAC-seq samples, 175 came from Calderon *et al.* 2019. The Illumina HiSeq 4000 was used to sequence 159 samples and the Illumina NovaSeq 6000 was used to sequence 16 samples. Both sequencers were used to generate 76 base pair reads. Cells were isolated from seven donors, of which the same four contributed a majority (90.9%) of samples. Six of the ATAC-seq samples were from the Corces *et al.* 2016 study, with three donors contributing two samples each. The Illumina NextSeq 500 sequencer was used to generate 150 base pair reads. The last 38 ATAC-seq samples of the 219 were from the neutrophil atlas. The Illumina X Ten sequencer was used for these samples to generate 150 base pair reads. For the six ligands for *ex vivo* stimulation, four donors were used. Two donors were used for the *S. aureus* stimulation, and only one donor was used for the *E. coli* stimulation.

Study	Lineage	Cell Type	ATAC-seq
Corces *et al.*	MYELOID	Monocytes	6
Calderon *et al.*	B	Bulk B cells	7
		Memory B cells	8
		Naive B cells	7
		Plasmablasts	3
	CD8	T cells	7
		Central Memory T cells	8
		Memory T_{eff} cells	8
		Naive T cells	8
	GD	γδ T cells	7
	CD4	T_{eff} cells	7
		Follicular T_h cells	9
		Memory T_{eff} cells	8
		Memory T_{reg} cells	8
		Naive T_{eff} cells	9
		T_{reg} cells	8
		T_h1 precursors	8
		T_h17 precursors	7
		T_h2 precursors	8
		Naive T_{reg} cells	4
	NK	Immature NK cells	5
		Mature NK cells	10
		Memory NK cells	6
	MYELOID	Monocytes	9
		Myeloid dendritic cells	3
		Plasmacytoid dendritic cells	3

Table C.1: Samples in immune atlas. Samples from various primary immune cell types were present across the two studies in the immune atlas. This table contains the number of samples from ATAC-seq experiments from each cell type.

Study	Cell Type	Experiment	Stimulation	ATAC-seq
Ram-Mohan *et al.*	Neutrophils	*Ex vivo* stimulation	LTA	4
			LPS	4
			FLAG	4
			R848	4
			BGP	4
			HMGB1	4
			Control	4
		S. aureus stimulation	*S. aureus* (10^1 Cells)	1
			S. aureus (10^3 Cells)	2
			S. aureus (10^5 Cells)	1
			Control	2
		E. coli stimulation	*E. coli* (1 hour)	1
			E. coli (4 hours)	1
			Control (1 hour)	1
			Control (4 hours)	1

Table C.2: Samples in neutrophil atlas. Samples of neutrophils under various simulations were present in the neutrophil atlas. This table contains the number of samples from ATAC-seq experiments from each stimulation.

D | ATAC-seq Reprocessing

D.1 Quality Control

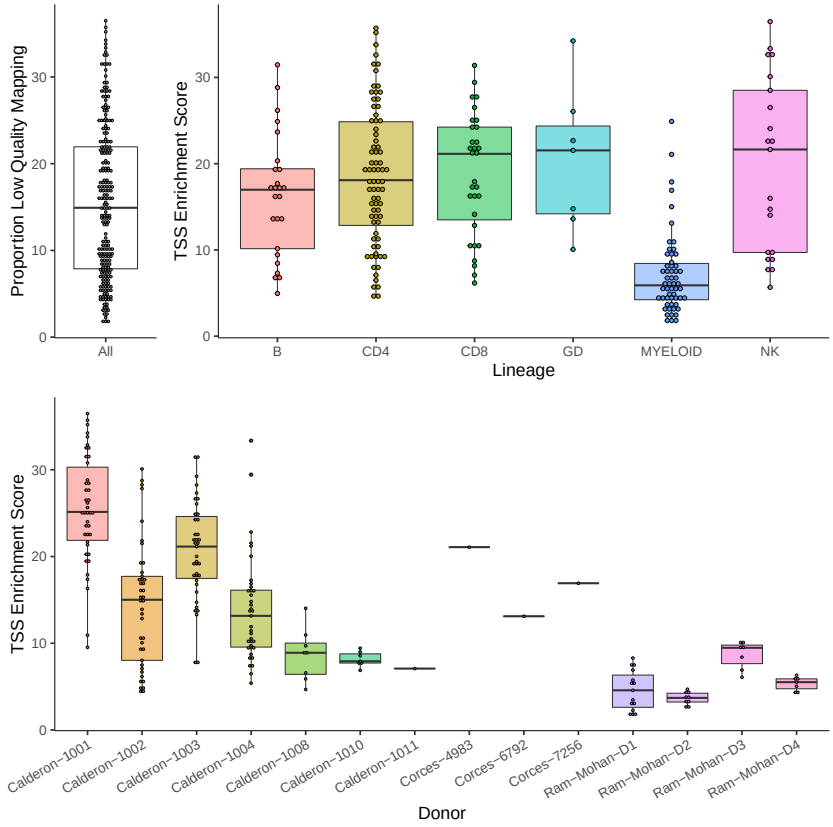

Figure D.1: TSS enrichment scores. The TSS enrichment score was used to filter samples that had a low signal-to-noise ratio.

D.2 Comparison with Original Study

The original count matrix from the Calderon *et al.* 2019 study was retrieved from the Gene Expression Omnibus (GEO) entry for the study (GSE118189). Peak intervals from the original study were converted from hg19 coordinates to GRCh38 coordinates using the `liftOver` function in the rtracklayer R package (Lawrence *et al.* 2009). Peaks that were split during the conversion

were discarded. The `findOverlaps` function in the GenomicRanges R package (Lawrence *et al.* 2013) was used to identify peaks from this analysis that overlapped peaks from the original study.

ATAC-seq alignment and peak calling was similar between this analysis and the initial analysis of the data (Calderon *et al.* 2019). The consensus peak set in this analysis was built by merging cell-type-specific peak sets and used a more stringent criteria for merging peaks across samples. Furthermore, this analysis utilised fewer cell types than the original analysis. Unsurprisingly, the consensus peak set in this analysis contained fewer peaks (296,994 peaks) than the original analysis (827,922 peaks). Although the median width of peaks in this analysis (554 base pairs) was slightly larger than the original analysis (490 base pairs), the overall distribution of peak sizes was comparable between the two analyses (Figure D.2). The distribution of peaks across the genome was also comparable between the two analyses (Figure D.3). Although reads were aligned to different peak sets and genome builds, read counts were highly concordant between overlapping peaks across the two analyses (Figure D.4).

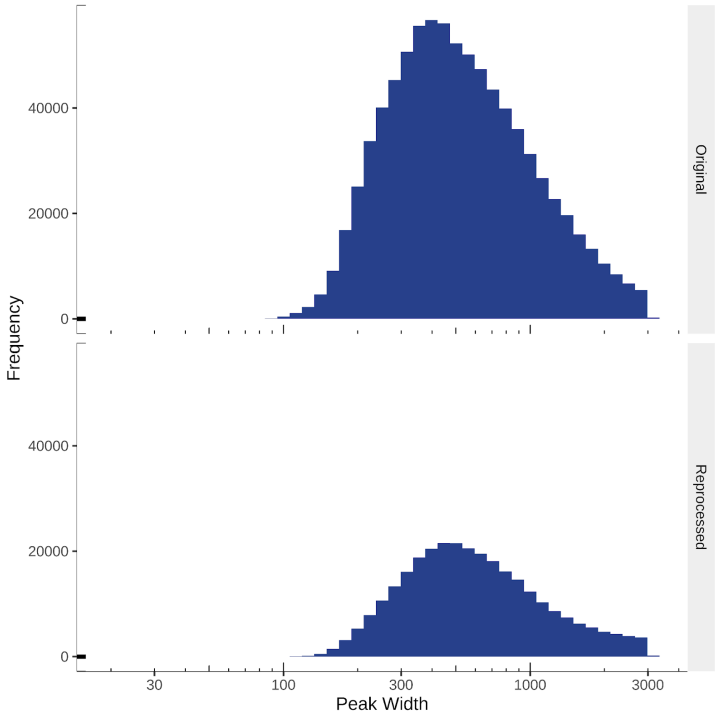

Figure D.2: Distribution of peak widths. Although fewer peaks were present in this analysis, the peak width distributions from the original analysis of the immune atlas (top) and this analysis (bottom) are comparable.

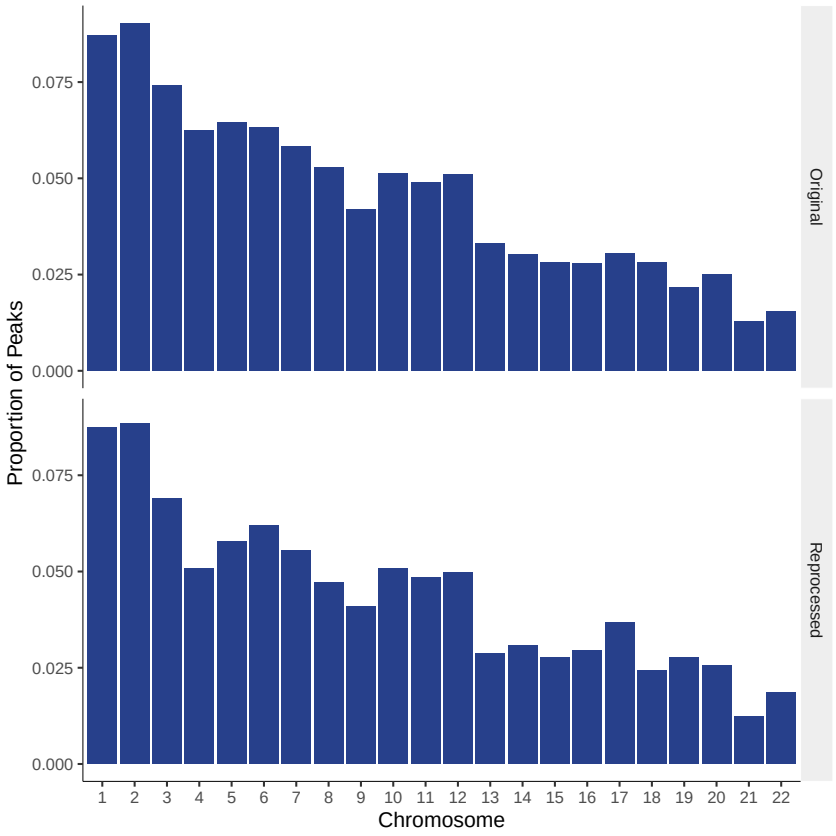

Figure D.3: Distribution of peaks across the genome. The distribution of peaks across the genome from the consensus peak set of the original study (top) and this analysis (bottom) are comparable.

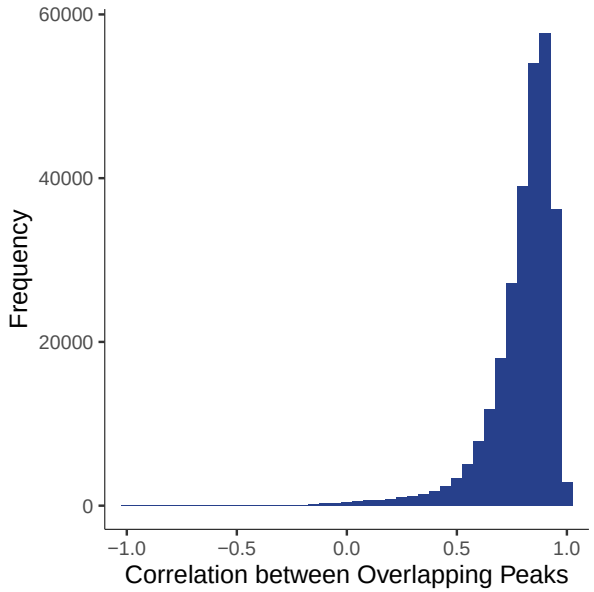

Figure D.4: Correlation of read counts between peaks. Read counts of peaks overlapping between the original study and this analysis are highly concordant. Spearman's Rho was used as a measure of similarity.

D.3 Peak Sets

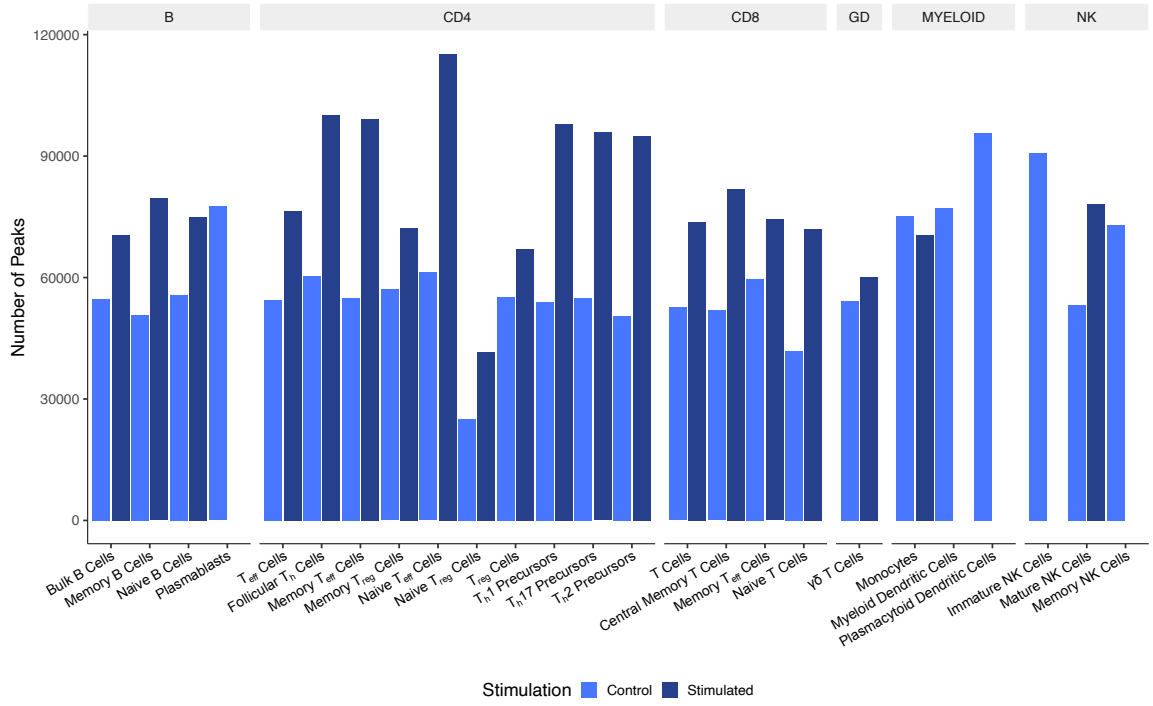

Figure D.5: Group peak sets from immune atlas. Group peak sets for each cell-condition pair from the immune atlas contained sets of peaks present in specific cell types in different conditions. Plasmablasts, myeloid dendritic cells, plasmacytoid dendritic cells, immature natural killer cells, and memory natural killer cells had no stimulated samples.

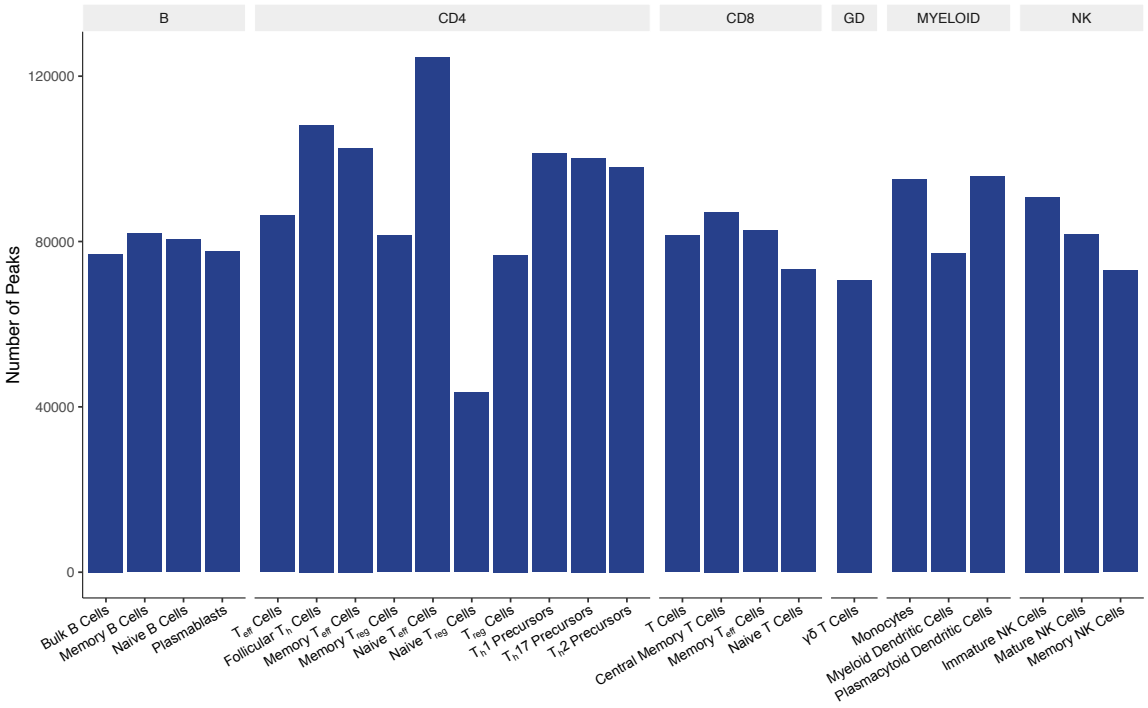

Figure D.6: Cell type peak sets from immune atlas. Cell type peak sets were generated by merging group peak sets from the same cell type.

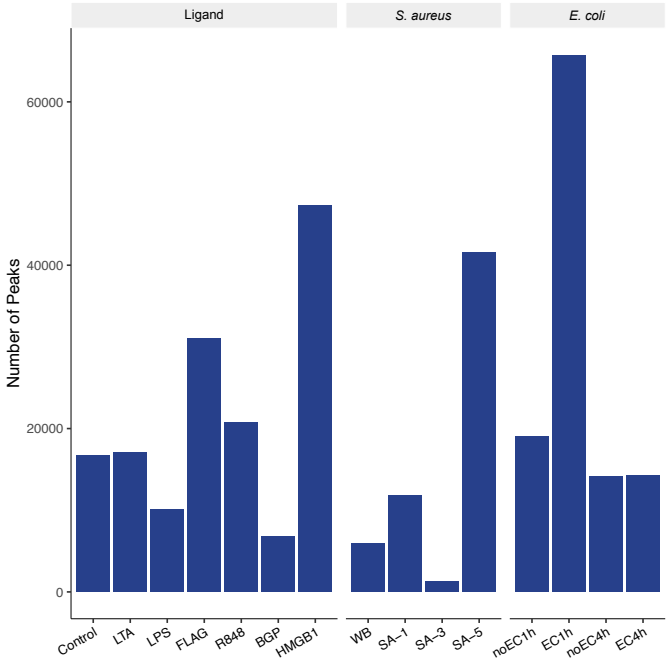

Figure D.7: Group peak sets from neutrophil atlas. Group peak sets for each cell-condition pair from the neutrophil atlas contained sets of peaks present in neutrophils under different conditions.

E | Roadmap Project Epigenomes

Group	Epigenome Name	Accession
HSC & B cell	Primary monocytes from peripheral blood	E029
HSC & B cell	Primary B cells from peripheral blood	E032
HSC & B cell	Primary Natural Killer cells from peripheral blood	E046
Blood & T cell	Primary T cells from peripheral blood	E034
Blood & T cell	Primary T helper naive cells from peripheral blood	E038
Blood & T cell	Primary T helper cells from peripheral blood	E043
Blood & T cell	Primary T regulatory cells from peripheral blood	E044
Blood & T cell	Primary T cells effector/memory enriched from peripheral blood	E045
Blood & T cell	Primary T $CD8^+$ naive cells from peripheral blood	E047
Blood & T cell	Primary T $CD8^+$ memory cells from peripheral blood	E048

Table E.1: Roadmap Project epigenomes. I retrieved the ChromHMM states from the 18-state models for the samples listed in this table.

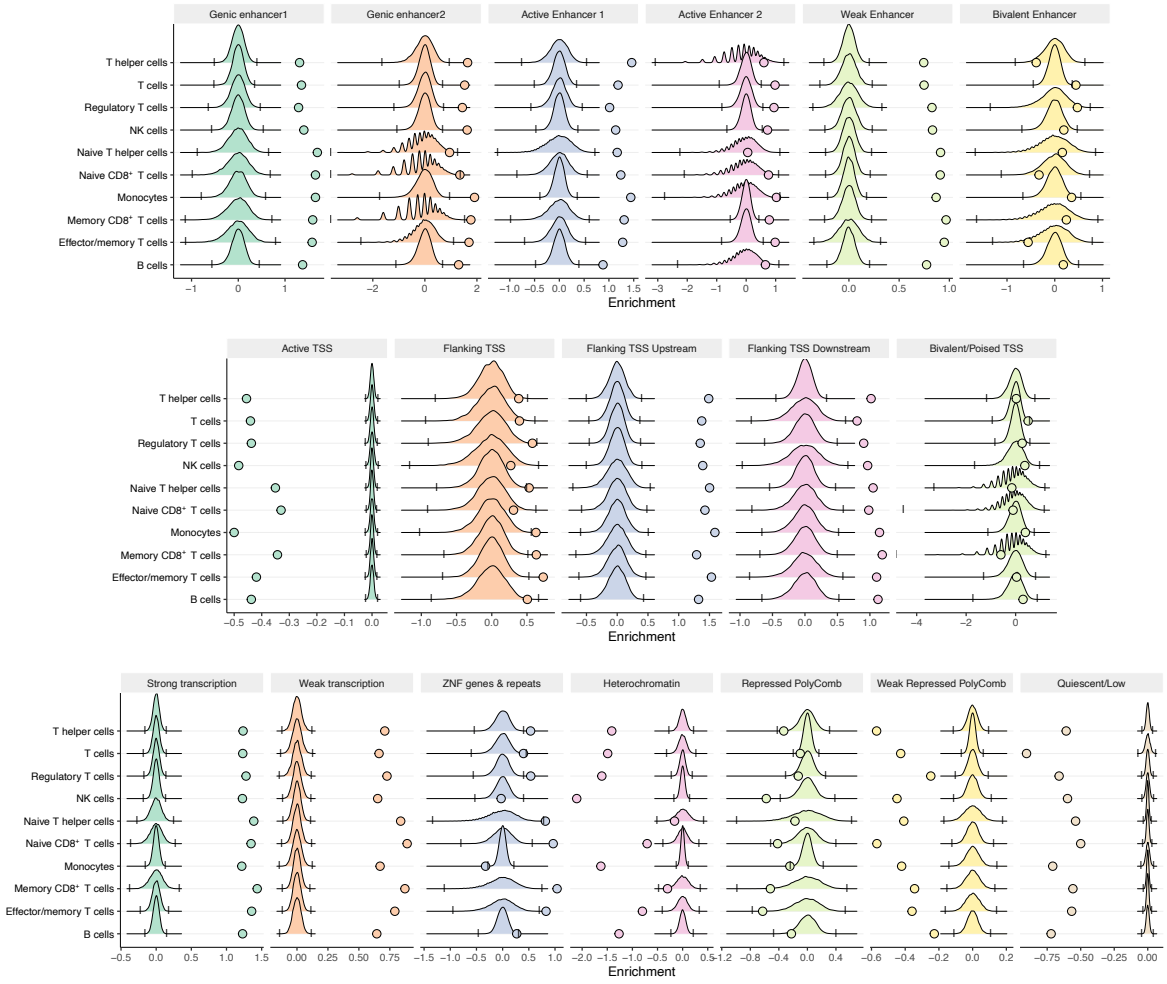

Figure E.1: Enrichment in ChromHMM states. For each ChromHMM state in various epigenomes, matched SNPs were used to generate null distributions for the proportion of overlapping variants. Each point represents the observed overlap of lead conditional *cis*-eQTL. The bars for each distribution represent the boundaries of the rejection region based on a significance threshold of $\alpha = 0.0001$. Enrichment was calculated as $\log_2$ of the ratio of the observed proportion of overlap to the mean null proportion of overlap. The states are divided into enhancer regions (top), TSS regions (middle), and transcription regions (bottom).

F | Partitioned Heritability

These derivations are presented in Yang *et al.* 2010, Ge *et al.* 2017, and Min *et al.* 2022. Let $\mathbf{\Gamma} \in \mathbb{R}^{q \times m}$ be the genotype matrix for q individuals and m biallelic SNPs. The genotypes for each SNP is centred and scaled so that Γ_{ij} has mean 0 and variance 1. Using this matrix, $\mathrm{Cov}(\Gamma_{ik}, \Gamma_{jk}) = \mathbb{E}[\Gamma_{ik}\Gamma_{jk}] - \mathbb{E}[\Gamma_{ik}]\mathbb{E}[\Gamma_{jk}] = \mathbb{E}[\Gamma_{ik}\Gamma_{jk}] = \phi_{ij}$ is the coefficient of relationship between the i-th and j-th individuals. The GRM is thus defined as

$$\mathbf{\Psi} = \frac{1}{m}\mathbf{\Gamma}\mathbf{\Gamma}^T$$

§ Assumptions for Additive Heritability Random SNP Effect Model

Let $\mathbf{\Gamma}^{(\alpha)} \in \mathbb{R}^{q \times \alpha}$ be the genotype matrix of SNPs that fall within the annotation. Assume that the first $m_\alpha < \alpha$ SNPs are causal SNPs for the trait. Similarly, let $\mathbf{\Gamma}^{(\bar{\alpha})} \in \mathbb{R}^{q \times \bar{\alpha}}$ be the genotype matrix of SNPs that are outside the annotation. Assume that the first $m_{\bar{\alpha}} < \bar{\alpha}$ SNPs are causal SNPs for the trait. For the i-th sample from the k-th individual, the model is assumed to be

$$Y_i = \sum_{j=1}^{m_\alpha} \Gamma_{kj}^{(\alpha)} \beta_j^{(\alpha)} + \sum_{j=1}^{m_{\bar{\alpha}}} \Gamma_{kj}^{(\bar{\alpha})} \beta_j^{(\bar{\alpha})} + b_k + \epsilon_i$$

Here, $\beta_j^{(\alpha)}$ and $\beta_j^{(\bar{\alpha})}$ represent the causal effect sizes, b_k represents the individual-level random intercept that captures within-individual variance, and ϵ_i represents the residual variance. It is assumed that

$$\begin{aligned}
\beta_j^{(\alpha)} &\sim \mathcal{N}\left(0, \frac{\sigma_\alpha^2}{m_\alpha}\right) \\
\beta_j^{(\bar{\alpha})} &\sim \mathcal{N}\left(0, \frac{\sigma_{\bar{\alpha}}^2}{m_{\bar{\alpha}}}\right) \\
b_k &\sim \mathcal{N}\left(0, \sigma_R^2\right) \\
\epsilon_i &\sim \mathcal{N}\left(0, \sigma^2\right)
\end{aligned}$$

Then the expected variance of Y_i is

$$\begin{aligned}\text{Var}[Y_i] &= \sum_{j=1}^{m_\alpha} \left(\Gamma_{kj}^{(\alpha)}\right)^2 \text{Var}\left[\beta_j^{(\alpha)}\right] + \sum_{j=1}^{m_{\bar{\alpha}}} \left(\Gamma_{kj}^{(\bar{\alpha})}\right)^2 \text{Var}\left[\beta_j^{(\bar{\alpha})}\right] + \text{Var}[b_k] + \text{Var}[\epsilon_i] \\ &= \frac{\sigma_\alpha^2}{m_\alpha} \sum_{j=1}^{m_\alpha} \left(\Gamma_{kj}^{(\alpha)}\right)^2 + \frac{\sigma_{\bar{\alpha}}}{m_{\bar{\alpha}}} \sum_{j=1}^{m_{\bar{\alpha}}} \left(\Gamma_{kj}^{(\bar{\alpha})}\right)^2 + \sigma_R^2 + \sigma^2\end{aligned}$$

Since $\mathbb{E}\left[\Gamma_{kj}^2\right] = \phi_{kk} = 1$, the expectation is

$$\text{Var}[Y_i] = \sigma_\alpha^2 + \sigma_{\bar{\alpha}}^2 + \sigma_R^2 + \sigma^2$$

§ Matrix Form of Heritability Model

The LMM proposed in Section 2.7.2 can be written in matrix form as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{ZB} + \mathbf{ZB}_\alpha + \mathbf{ZB}_{\bar{\alpha}} + \epsilon$$

Let n be the number of samples, $q < n$ be the number of individuals, and p be the number of fixed-effect covariates. $\mathbf{Y} \in \mathbb{R}^{n\times 1}$ is the measured trait of interest with n samples, some of which are repeated from the same individual. The design matrix $\mathbf{X} \in \mathbb{R}^{n\times p}$ and associated fixed-effect vector $\boldsymbol{\beta} \in \mathbb{R}^{p\times 1}$ encode the expected value of the trait $\mathbb{E}[\mathbf{Y}] = \mathbf{X}\boldsymbol{\beta}$. The incidence matrix for the random effects $\mathbf{Z} \in \mathbb{R}^{n\times q}$ is a block diagonal matrix that maps the individual-level random effects to the samples.

In this model, I assume that the individual-level random intercept can be partitioned into the sum of the additive genetic effect from SNPs in the annotation $\mathbf{B}_\alpha$, the additive genetic effect from SNPs outside the annotation $\mathbf{B}_{\bar{\alpha}}$, and the individual-specific environmental effect $\mathbf{B}$. The covariance structures take the form

$$\begin{aligned}\mathbf{B} &\sim \mathcal{N}(\mathbf{0}, \sigma_R^2 \mathbf{I}_q) \\ \mathbf{B}_\alpha &\sim \mathcal{N}(\mathbf{0}, \sigma_\alpha^2 \boldsymbol{\Psi}_\alpha) \\ \mathbf{B}_{\bar{\alpha}} &\sim \mathcal{N}(\mathbf{0}, \sigma_{\bar{\alpha}}^2 \boldsymbol{\Psi}_{\bar{\alpha}})\end{aligned}$$

which implies that the variance-covariance matrix of $\mathbf{Y}$ is partitioned as

$$\text{Var}[\mathbf{Y}] = \sigma_R^2 \mathbf{Z}\mathbf{I}_q\mathbf{Z}^T + \sigma_\alpha^2 \mathbf{Z}\boldsymbol{\Psi}_\alpha\mathbf{Z}^T + \sigma_{\bar{\alpha}}^2 \mathbf{Z}\boldsymbol{\Psi}_{\bar{\alpha}}\mathbf{Z}^T + \sigma^2\mathbf{I}_n$$

G | Variant Effect Prediction

SNP	Module QTL	Transcription Factors	Score Change
rs1131017	Module 101 @ Chr 12 (55.0 Mb - 57.1 Mb)	GATA3, GATA4, GATA5	-0.169
rs4761234	Module 103 @ Chr 12 (68.3 Mb - 70.4 Mb)	HNF4A	-0.059
rs1132812	Module 61 @ Chr 16 (29.2 Mb - 31.2 Mb)	ETV2::DLX3, HOXB2::ELF1, HOXB2::ELK3	-0.015
rs1023252	Module 62 @ Chr 1 (10.8 Mb - 12.8 Mb)	GCM1::MAX	-0.028
		ERF::MAX, FLI1::MAX	-0.025
		MYBL1::MAX	-0.018
		TEAD4::MAX	-0.027
rs7191618	Module 91 @ Chr 16 (27.3 Mb - 30.0 Mb)	ETV2::HOXA2, FLI1::DLX2, HOXB2::ELF1, HOXB2::ELK3, ETV2::DRGX, ELK1::HOXA1, FLI1::DRGX, ETV5::DRGX, ETV5::HOXA2, HOXB2::ELK1	0.016
rs11130192	Module 94 @ Chr 3 (47.7 Mb - 50.9 Mb)	POU2F1::ELK1	-0.094
		TEAD4::RFX5	0.001
rs4759187	Module 96 @ Chr 12 (54.7 Mb - 56.7 Mb)	TFAP2C::MAX	-0.008
rs13430254	Module 97 @ Chr 2 (71.2 Mb - 73.3 Mb)	ETV2::CEBPD, ETV2::TEF, ERF::CEBPD, ELK1::TEF, FLI1::CEBPB, FLI1::CEBPD, ETV5::CEBPD	0.071

Table G.1: VEP module QTL motifs. The results from the VEP motif analysis of the lead module QTL variants. Some motif features from Ensembl have multiple TFs.

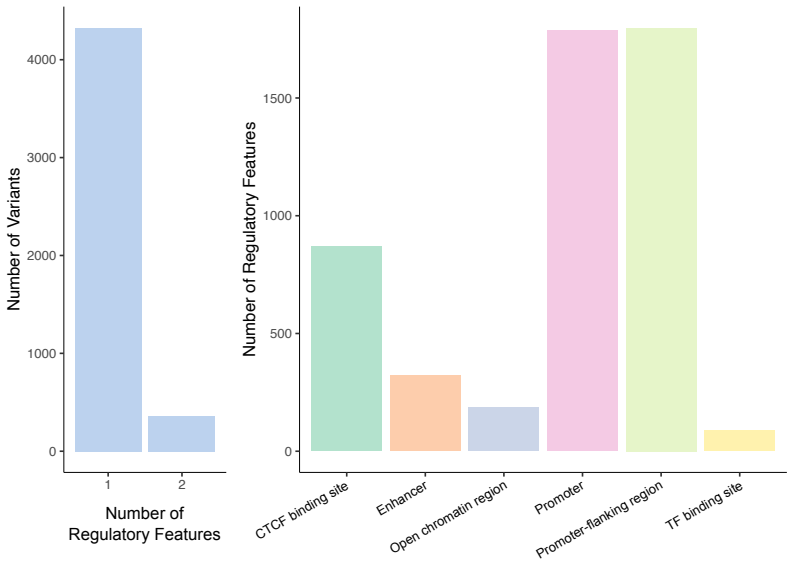

Figure G.1: VEP regulatory consequences. VEP was used to identify predicted consequences of lead conditional *cis*-eQTL. (Left) Most variants affected one specific regulatory feature, although some affected two. (Right) Affected regulatory features were divided by biotype based on their functional relevance.

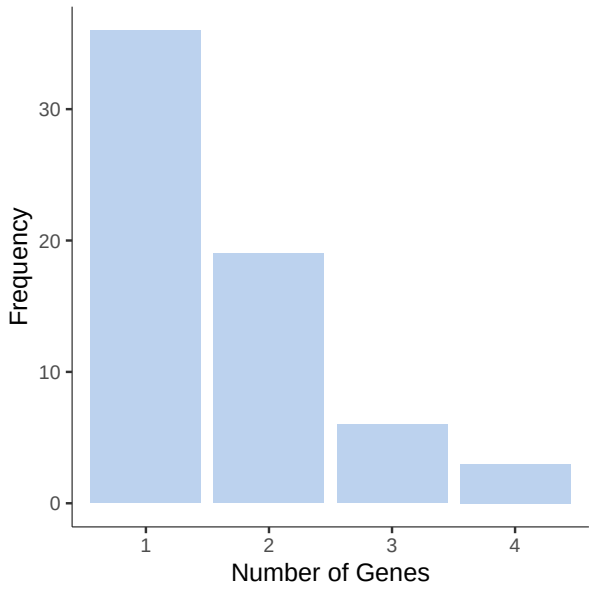

Figure G.2: VEP module QTL gene consequences. VEP was used to identify predicted consequences of lead module QTL. Lead variants were predicted to affect between 1 and 4 genes.

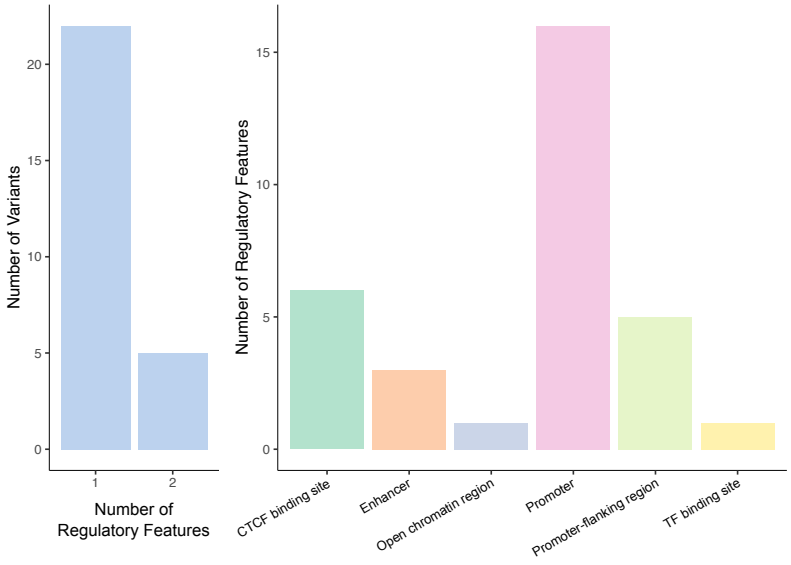

Figure G.3: VEP module QTL regulatory consequences. VEP was used to identify predicted consequences of lead module QTL. (Left) Most variants affected one specific regulatory feature, although some affected two. (Right) Affected regulatory features were divided by biotype based on their functional relevance.