Integrative Genomics to Understand Immune Function and Regulation

Artika Praveeta Nath

ORCID: 0000-0002-5688-2941

Doctor of Philosophy

October 2017

Department of Microbiology and Immunology

The Peter Doherty Institute for Infection and Immunity

The University of Melbourne, Parkville

Victoria, 3010, Australia

Submitted in Total Fulfillment of the Requirements of the Degree of Doctor of Philosophy

I would like to dedicate this thesis to my beloved parents
Lalita and Vishwa Nath
to my grandmother Rai Wati (Aji)
and to my darling sister Shomika Nath

Without whom none of my success, achievements, and dreams
would have been possible.

Thank you so much for your unconditional love, support, guidance, and prayers.

Abstract

Characterising the mechanistic principles underlying immune function and regulation will help us understand how the immune system provides effective host defence. However, the highly complex and multi-level nature of the immune system requires a systems-level analysis to gain multi-dimensional insight and unravel its complexity.

High-throughput profiling technologies allow quantitative measurement of various immunological parameters that capture system-wide information. This has led to the generation of large-scale multi-omics datasets from human populations, experimental set-ups, and a compendium of immune cell types. Developments in bioinformatics offer integrative approaches to explore the functional and regulatory relationships within and between various organisational levels of the immune system as well as across other biological systems.

For this thesis, multi-omic analysis was used to characterise immune processes in terms of genetics, transcriptional networks and interactions with metabolism. First, I mapped the genetics and interactions of immune gene networks with circulating metabolites in in a population-based study. I integrated blood transcriptomic, metabolomic, and genomic profiles from two population-based cohorts, including a subset with 7-year follow-up sampling. I identified topologically robust gene networks enriched for immune functions including cytotoxicity, viral response, B cell, platelet, neutrophil, and mast cell/basophil activity. These immune gene modules showed complex patterns of association with 158 circulating metabolites, including lipoprotein subclasses, lipids, fatty acids, amino acids, and CRP. Genome-wide scans for module expression quantitative trait loci (mQTLs) revealed five modules with mQTLs with both *cis* and *trans* effects.

Then, I explored the underlying shared genetic architecture between correlated cytokines, the regulatory agents of the immune system. Multivariate genome-wide association scan was performed to identify genetic variants regulating circulating cytokines in ~9,000 individuals from three independent population studies. Eight loci were identified as regulating this network, including two previously undetected loci. Expression quantitative loci (eQTL) analysis revealed that these loci harbour eQTLs. Further linking these loci with genetic variants associated with disease risk provided insight into the possible inflammatory pathways underlying these common diseases.

Thirdly, I explored a particular component of the immune system, immunological memory; with emphasis on tissue resident memory T-cells (TRM cells). I employed a network-based approach to identify a transcriptional sub-network related to the residency of murine TRM cells isolated from various tissues. Comparative analysis further revealed that expression profiles of tissue resident immune cells from different lineages share transcriptional similarity.

Finally, the role of TGF-β, an extrinsic tissue-derived factor in influencing the transcriptional signature of TRM cells was investigated. I compared the global transcription of T-cells induced *in vitro* by TGF-β with the residency-related transcription signature previously established in TRM cells. This demonstrated that the transcriptional signature of TRM cells is largely driven by TGF-β.

Findings from this thesis demonstrate the power of integrative bioinformatics analyses to gain novel insights into the immune system, which can assist in predicting its response to perturbations. It also may help explain how inter-individual variability in immune function contributes to differential disease susceptibility and treatment outcomes. This thesis offers a general framework to systematically integrate and analyse multi-omics data to answer important biological questions.

Declaration

This is to certify that this thesis conforms to relevant policies and procedures of the University of Melbourne and has been compiled with the following requirements:

i) This thesis, entitled "Integrative genomics to understand immune function and regulation", comprises only my original work towards the Doctor of Philosophy degree, except where otherwise indicated. The contribution of others towards this thesis and the proportion of the work that I have claimed as original has been acknowledged in the Preface of this thesis;

ii) Due acknowledgement has been made in the text to all other material used;

iii) This thesis is fewer than 100,000 words in length, exclusive of tables, maps, bibliographies, and appendices.

Artika Praveeta Nath, B.Sc (Hons) M.Sc
Department of Microbiology and Immunology,
The Peter Doherty Institute for Infection and Immunity,
The University of Melbourne, Parkville

Preface

This preface summarises the contents of each chapter, the proportion of work claimed as original, and the nature and extent of contribution to chapters from co-authors, collaborators, and supervisors. Chapters resulting in a manuscript that have been either published, submitted, or in preparation are listed.

Chapter 1: Introduction and literature review is an original work that provides an overview of the background to, research gap, and rational for the current research work done in this thesis. All original work discussed were appropriately cited. Comments on the structure and contents of this chapter were provided by my principal supervisor, Michael Inouye, and co-supervisor Francis Carbone.

Chapter 2: An interaction map of circulating metabolites, immune gene networks and their genetic regulation is an original work that resulted in a publication in *Genome Biology* with the same title. In this chapter, I systematically integrated multi-omic data from two population-based cohorts to identify interactions, both spatially and temporally, between circulating metabolites and immune gene networks and their genetic regulation. The article can be found online at: https://doi.org/10.1186/s13059-017-1279-y.

I am the first author and major contributor to the work presented in this co-authored manuscript. The following indicates the contributions of co-authors and myself:

- I was responsible for the majority of the bioinformatics analyses, interpretation of the results, and creating figures for the results presented in this chapter with input from Michael Inouye.

- I was responsible for planning, drafting and editing the manuscript, creating

figures, and responding to reviewer comments with help from Michael Inouye.

- Aki Havulinna, Anni Joensuu, Antti Kangas, Pasi Soininen, Annika Wennerström, Lili Milani, Andres Metspalu, Satu Männistö, Peter Würtz, Johannes Kettunen, Mika Kähönen, Markus Juonala, Mika Ala-Korpela, Samuli Ripatti, Terho Lehtimäki, Olli Raitakari, Veikko Salomaa, and Markus Perola were involved in the collection, measurement, preprocessing, and quality control of the multiple omics data in the population cohorts analysed in this chapter and provided feedback on the final version of the manuscript prior to submission.

- The genotype data analysed in this chapter had already been imputed and undergone post-imputation filtering. The gene expression data made available had also been processed. Details on data processing, quality control, and filtering are provided in the Methods section of this chapter.

- Scott Ritchie developed the NetRep software used in this analysis. He was also responsible for running NetRep on gene expression data to identify and replicate gene co-expression networks across cohorts.

- Liam Fearnley provided advice and feedback on the functional enrichment analysis.

- Michael Inouye, Gad Abraham, Scott Ritchie, and Sean Byars provided guidance and advice on data analyses, and the interpretation of the results and their visualisation.

Chapter 3: Multivariate genome-wide association analysis identifies eight loci associated with a network of circulating cytokines is an unpublished chapter; a co-authored manuscript, which is in preparation with the same title. I am the lead author on this manuscript. In this chapter, I performed genome-wide association analysis to identify and characterise genetic variants that regulated correlated cytokines in three population-based cohorts.

The following indicates the contributions of collaborators, colleagues, and myself towards the work presented in this chapter:

- I was responsible for the majority of the bioinformatics analyses, interpretation of the results, and creating figures for the results presented in this chapter with input from Michael Inouye.

- I was responsible for drafting and editing the manuscript in preparation, and creating figures with help from Michael Inouye.

- Ari Ahola-Olli, Peter Würtz, Aki Havulinna, Kristiina Aalto, Niina Pitkänen, Terho Lehtimäki, Mika Kähönen, Leo-Pekka Lyytikäinen, Emma Raitoharju, Ilkka Seppälä, Antti-Pekka Sarin, Samuli Ripatti, Aarno Palotie, Markus Perola, Jorma Viikari, Sirpa Jalkanen, Mikael Maksimow, Veikko Salomaa, Marko Salmi, Olli Raitakari, Johannes Kettunen were involved in the collection, measurement, preprocessing, and quality control of the omics data in the population cohorts analysed in this chapter.

- Michael Inouye, Gad Abraham, and Scott Ritchie provided guidance and advice on data analyses, and the interpretation of the results and their visualisation.

Chapter 4: Differential network analysis identifies a transcriptional network involved in tissue resident memory T-cell development is an original, unpublished work. In the first part of this chapter, I employed a network-based approach to identify and characterise a residency-related gene network associated with murine TRM cells isolated from various peripheral tissues. In the second part of this chapter, I performed comparative analysis to assess the shared similarities between transcriptional profiles of resident immune cells from different lineages. A manuscript for publication has been planned for this chapter.

The following indicates the contributions of collaborators, colleagues, and myself towards the work presented in this chapter:

- I was responsible for the majority of the bioinformatics analyses and creating figures for the results presented in this chapter with input from Michael Inouye and Francis Carbone.

- Scott Ritchie provided computational assistance with the network tool used to construct gene networks.

- I was responsible for interpreting the results with contributions and feedback provided by Michael Inouye, Francis Carbone, Thomas Gebhardt, and Laura Mackay.

Chapter 5: RNA-seq analysis reveals that the trancriptional signature of TRM cells is largely driven by TGF-β signalling is an original unpublished chapter. I performed comparative transcriptional analysis to assess the role of TGF-β signalling, a local extrinsic factor present at tissue sites, in influencing the transcriptional program of TRM cells. A manuscript for publication has been planned for this chapter.

The following indicates the contributions of collaborators, colleagues, and myself towards the work presented in this chapter:

- I was responsible for the majority of data analyses and creating figures for the results presented in this chapter with input from Michael Inouye, Thomas Gebhardt, and Asolina Braun.

- Asolina Braun was responsible for performing all the lab experiments, which included the *in vitro* TGF-β stimulation experiment and RNA extraction at the Peter Doherty Institute for Infection and Immunity.

- Matthew Tinning at the Australian Genome Research Facility was responsible for performing the RNA sequencing.

- I was responsible for interpreting the results with contributions and feedback provided by Michael Inouye, Thomas Gebhardt, and Asolina Braun.

Chapter 6: Conclusions is an original summary of the key findings for each chapter, and the implications and importance of work presented in this thesis.

Manuscripts arising from candidature

Peer-reviewed journal research article manuscripts, manuscripts that have been submitted for publication, and manuscripts in preparation arising from PhD candidature are:

Nath AP, Ritchie SC, Byars SG, Fearnley LG, Havulinna AS, Joensuu A, Kangas AJ, Soininen P, Wennerström A, Milani L, Metspalu A, Männistö A, Würtz P, Kettunen J, Kähönen M, Juonala M, Ala-Korpela M, Ripatti S, Lehtimäki T, Abraham G, Raitakari O, Salomaa V, Perola M, Inouye M. An interaction map of circulating metabolites, immune gene networks and their genetic regulation. *Genome Biol*. 2017;18(1):146.

Nath AP, Ritchie SC, Ahola-Olli A, Würtz P, Havulinna AS, Aalto K, Pitkänen N, Terho L, Kähönen M, Lyytikäinen LP, Raitoharju E, Seppälä I, Sarin AP, Ripatti S, Palotie A, Perola M, Viikari JS, Jalkanen S, Maksimow M, Salomaa V, Salmi M, Raitakari O, Abraham G, Kettunen J, Inouye M. Multivariate genome-wide association analysis identifies eight loci associated with a network of circulating cytokines. *Manuscript under preparation.* 2017.

Bhalala OG, **Nath AP**, UK Brain Expression Consortium, Inouye M, Sibley CR. Identification of expression quantitative trait loci associated with schizophrenia and affective disorders in normal brain tissue: A meta-analysis. *Manuscript submitted for publication.* 2017.

Ritchie SC, Würtz P, **Nath AP**, Abraham G, Havulinna AS, Fearnley LG, Sarin AP, Kangas AJ, Soininen P, Aalto K, Seppälä I, Raitoharju E, Salmi M, Maksimow M, Männistö S, Kähönen M, Juonala M, Ripatti S, Lehtimäki T, Jalkanene S, Perola M, Raitakari O, Salomaa V, Ala-Korpela M, Kettunen J, Inouye M. The biomarker glycA is associated with chronic inflammation and predicts long-term risk of severe infection. *Cell Systems.* 2015; 1(4): 293-301.

Acknowledgements

This PhD journey has been a unique and amazing experience, and this thesis would not have been possible without the support, guidance and encouragement of several people.

First and foremost, I would like to express my sincere thanks and appreciation to my principal supervisor A/Prof. Michael Inouye for giving me the valuable opportunity to work with him. I express my heartfelt gratitude for his patience, mentorship, and guidance that has enabled me to complete this PhD thesis. His leadership qualities and passion for science have greatly inspired my growth as a scientific researcher. His understanding nature and constant encouragement have always motivated me to perform to the best of my capabilities.

I would also like to thank my co – supervisor Prof. Francis Carbone for his presence and support during this journey,

I would like to thank my PhD committee members; Professor Terry Speed and Prof. Andrew Brooks for their expertise, guidance and support during my candidature.

The projects undertaken as part of this thesis would not have been possible without multiple collaborators who have assisted in data generation and provided their biological expertise during data analysis and interpretation of the results. In particular I would like to acknowledge A/Prof. Thomas Gebhard, Dr. Asolina Braun, and Dr. Laura Mackay.

It has been a blessing and privilege to be part of an extraordinary group of fellow lab mates. I am sincerely grateful for their immense support and help in my daily life at the university and helping me feel at home during this time. You have indeed become like family in the past four years. I would like to thank: Scott Ritchie, Dr. Liam Fearnley,

Dr. Marta Brożyńska, Dr. Lesley Gray, Dr. Oneil Bhalala, Dr. Howard Tang, Andrew Bakshi, Qinqin Huang, Yu Wan, Amy Hamilton, and Youwen Qin.

Special thanks to Dr. Gad Abraham, Dr. Shu Mei Teo, and Dr. Sean Byars for their assistance and valuable advice, both professional and personal.

I would like to express my appreciation to my fellow PhD colleagues, Claire Gorrie, Dr. Danielle Ingle, and Stephan Watts, for their moral support and encouragement during my PhD journey.

I would like to thank Rebecca Whitsed, the department postgraduate coordinator, for taking care of administrative matters regarding my candidature.

To all my friends around the globe, thank you for supporting me and lending me an ear when needed. I am grateful to Kriteshni Kaushal for her unconditional love and motivation throughout this journey.

To my family, your support, care and love during this time have been overwhelming. Without you in my life, this wouldn't have been possible. In particular, my parents and my sister for being the pillar of strength in my life.

Finally, many thanks to everyone who I could not mention personally, but have helped in various forms or ways to make this PhD a memorable experience and less stressful.

To Melbourne, for keeping me going through all those cold nights. Great whisky, coffee and lovely food.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

%	Percentage.
r^2	R-squared; a pairwise measure of linkage disequilibrium between two variants.
µg/ml	Microgram per millilitre.
^{1}H-NMR	Proton NMR.
BCM	B-cell activity module.
bp	Base pair.
CCLM	Cytotoxic cell-like module.
CD103	Integrin protein; cell surface marker expressed on TRM cells.
cis-eQTL	An eQTL that affects the expression of a nearby gene located within 1Mb.
DE	Differentially expressed.
DILGOM	The Dietary, Lifestyle, and Genetic Determinants of Obesity and Metabolic Syndrome Study.
DILGOM07	The 2007 collection of the DILGOM study.
DILGOM14	The 2014 follow-up collection of the DILGOM study.
eQTL	Expression quantitative trait loci.
FC	Fold change of differentially expressed genes.
FDR	False Discovery Rate.
GO	Gene Ontology.
GIMA	General immune module A.
GIMB	General immune module B.
GSEA	Gene Set Enrichment Analysis.
GTEx	The Genotype Tissue Expression consortium.
GWAS	Genome-wide association study.
ImmGen	The Immunological Genome Project

Kb	Kilobases; one thousand DNA base pairs.
LD	Linkage disequilibrium.
LLM	Lipid-leukocyte module.
log2FC	Log2 transformation of FC values
|log2FC|	Absolute value of log2FC
Mb	Megabases; one million DNA base pairs.
mQTL	Module quantitative Trait Loci.
NM	Neutrophil module.
NMR	Nuclear Magnetic Resonance spectroscopy.
PM	Platelet module.
QTL	Quantitative Trait Loci.
RNA-seq	RNA sequencing.
SNP	Single Nucleotide Polymorphism.
PC	Principal component.
PCA	Principal Components Analysis.
TCM	$CD8^+$ central memory T-cell.
TEM	$CD8^+$ effector memory T-cell.
TGF-β	Transforming growth factor beta.
TNF	Tumor necrosis factor alpha.
***Trans*-eQTL**	An eQTL that affects the expression of a distant gene located more than 5 Mb away.
TRM	$CD8^+$ tissue resident memory T-cell.
mol/L	Mole per litre
mg/ml	Milligram per millilitre.
mmol/L	Millimoles per litre
U/ml	Units per millilitre.
VRM	Virus response module.
WGCNA	Weighted Gene Coexpression Network Analysis.
YFS	The Cardiovascular Risk in Young Finns Study.
YFS07	The YFS follow-up study conducted in 2007.

Chapter 1

Introduction and literature review

1.1 Introduction

The immune system is a complex and dynamic network ordered in several hierarchical levels from simple molecules, cells to the whole organism. An effective host immune response against internal and external threats relies on the regulated interaction and integration between various levels of this hierarchy. Substantial progress has been made over the years to understand the mechanistic principles regulating the immune system and their relationship to diseases. However, this has mainly been done through a reductionist approach where the hierarchical sub-systems have been studied as individual entities with only a few selected candidate targets usually assessed.

A systems-level analysis is required to comprehensively evaluate and understand immune function and regulation. The Human Genome Project and the subsequent advancements in high-throughput profiling technologies have facilitated quantitative measurements of various aspects of the immune system at multiple levels. As a result, this has led to the generation of comprehensive information-rich datasets including genotypes, transcriptome, metabolome, cytokine profiles, and so forth. Besides, new developments in bioinformatics methods and tools enable us to integrate and represent the inter-relationship between components at each level of the immune system as functional or regulatory interaction networks.

This chapter is a review of the literature. First, I provide an overview of the immune system and response. Next, I summarise the different high-throughput profiling

technologies that probe the immune system, the bioinformatics approaches and tools for analysing large-scale data, and how the meaningful insights extracted from these data can be used to inform about immune-system.

1.2 The immune system and its components

The immune system is composed of a network of molecules, immune cells, and coordinated organ system that collectively function to guard the body against infection and foreign invaders (1,2).

One of the fundamental characteristics of the immune system is the ability to discriminate between self and non-self, which is essential to identify and eliminate invading microorganisms and abnormal cells from our bodies.

To be able to perform self/non-self discrimination, multiple proteins are required. This includes receptors that recognise pathogens such as pattern recognition receptors (PRRs) for innate immunity and receptors of T- and B-cells [T Cell Receptor (TCR) and B Cell Receptor (BCR)] for adaptive immunity (3). Signalling through these receptors trigger intracellular signalling cascades, for example the myeloid differentiation primary response gene 88 (MyD88)-dependent pathway in innate immunity. This in turn leads to the activation of several transcriptional factor families [e.g. nuclear factor kappa b (NF-κB) and interferon regulatory factors (IRFs) for innate immunity; NF-κB and nuclear factor of activated T cells (NFAT) for adaptive immunity], which subsequently drive the expression of effector cytokines and chemokines (3).

1.2.1 Organs of the immune system

Organs in the immune system can be divided into two groups. These are the primary lymphoid organs (bone marrow and thymus), sites where progenitor immune cells mature; and the secondary lymphoid organs (spleen, lymph nodes), where the mature immune cells further differentiate following an antigen encounter (4).

Bone Marrow is the soft tissue that lies inside the hollow cavities of bones. All the immune cells of the hematopoietic lineage are derived from precursor hematopoietic stem cells (HSCs) within the bone marrow (5).

The thymus is a bi-lobed organ located behind the breastbone in the upper chest where differentiation of precursor T-cells occurs (6).

The spleen is located in the upper left quadrant of the abdominal cavity, which functions as a blood filter by capturing foreign antigens and initiating an adaptive immune response (7). Aberrant and old cells are also detected and destroyed by the spleen.

The lymphatic system is an auxiliary circulatory system made up of lymph vessels, lymph nodes, and lymph (clear fluid rich in white blood cells). In the context of immunity, it functions as an internal filtration system. The microbes and toxins picked up by the lymph get cleared off in the lymph nodes (8).

Lymph nodes are dispersed throughout the body as concentrated regions of T-cells, B-cells, dendritic cells and macrophages, which function to filter the lymph (8).

Skin not only provides a physical barrier against invading pathogens but is also a matrix for a number of resident immune cell subtypes (e.g. resident memory T-cells).

1.2.2 Haematopoiesis gives rise to the cells of the immune system

Immune cells and other blood cell subpopulations are produced by a highly orchestrated developmental process, called haematopoiesis (9) (**Figure 1.1**). Adult HSCs, residing in the bone marrow, give rise to progeny, which progressively differentiate until they create distinct (unable to differentiate into other cell lineages) classes of blood and immune cells (10,11). The self-renewing HSCs first differentiate into multipotent progenitors (MPP), which lose their potential to self-renew. The MMPs further differentiate into two distinct branches – common myeloid progenitors (CMPs) and common lymphoid progenitors (CLPs) (11,12). The myeloid progenitors give rise to erythrocytes (red blood cells), platelets, granulocytes and macrophages, while T-, B-, and NK cells develop through the lymphoid path (9). Dendritic cells are derived from

both CLP and CMP lineages (13). The expression of a number of lineage-specific transcription factors is crucial in determining cell fate during haematopoiesis (12,14–19). In addition stem cell factor (SCF), thrombopoietin (TPO), erythropoietin (EPO), Fms-like tyrosine kinase 3 (FLT3) ligand, granulocyte–macrophage colony-stimulating factor (GM-CSF), and interleukins (IL-2, IL-3 and IL-7, IL-15) are growth factors and cytokines required for cell proliferation and survival at all points of the hematopoietic lineage (20,21). The functions of the blood elements formed during haematopoiesis are summarised in **Table 1.1**.

1.2.3 An immune response: crosstalk between innate and adaptive immunity

The immune system is divided into two distinctive, yet interconnected, subdivisions: the innate (natural) and adaptive (acquired) immunity (22). The **innate immune system** is constitutively present as the first line of defence and is activated within one hour of exposure to foreign invaders (23). This defence mechanism, which is mediated by various immune cells such as macrophages, neutrophils, dendritic cells, NK cells, and basophils, recognises pathogens in a non-specific manner via germ line-encoded pattern recognition receptors (24). The **adaptive immune system,** which involves T- and B-cells, is triggered after several days of infection in a more specific manner, by targeting antigens (foreign substance that triggers an immune response). This is achieved by the somatic rearrangement of genes encoding for receptors on these cells, thus generating a diverse repertoire of antigen-specific T- or B-cell receptors that are specific for unique foreign structures (3). This may allow it to clear pathogens that had somehow evaded innate immunity. Immunologic memory, the ability to recognise and rapidly clear reinfections, is another hallmark of adaptive immunity that is otherwise lacking in the innate immune response (24). However, there is intimate cross-talk between these two systems, and dendritic cells play a central role in this (25).

1.2.3.1 Innate immune response

An innate immune response is activated when a pathogen invades the physical barrier (e.g. skin) of a host. Pathogens interact with leukocyte receptors such as toll-like receptors (TLRs), nucleotide oligomerization domain (NOD) like receptors (NLR),

RIG-I like receptors (RLR), and other pattern recognition receptors (PRRs). These identify evolutionary conserved molecular structures (e.g. bacterial lipopolysaccharide, RNA from viruses) called pathogen-associated molecular patterns (PAMPs) that are exclusive to pathogens (26,27). PRRs can also detect endogenous host molecules that are released from damaged or dying cells and provoke an inflammatory response (28). PRRs trigger signalling pathways that converges on the activation of NFkB and IRF family of transcription factors, which mediate the expression of an array of cytokines including tumour necrosis factor alpha (TNF), interleukins (IL-1, IL-10, IL-12), type 1 and type II interferons [IFN-alpha (IFN-α), IFN-beta (IFN-β), IFN-gamma (IFN-γ)], and chemokines (26,29). Chemokines function as chemoattractants recruiting neutrophils, monocytes (precursors of macrophages), and other leukocytes to the infection site (29). Macrophages and neutrophils are phagocytes that ingest microbes and degrade them using digestive enzymes and toxic peroxides (30). Studies in translucent zebra fish showed that macrophages engulf fluid borne bacteria, while neutrophils engulf surface associated bacteria as they sweep over them using a "vacuum cleaner" technique (31). Cytokine production also activates the complement system, which bind to microbes, and mark with antibodies or proteins; thus “opsonising” them – making them more susceptible to ingestion by phagocytes (24). Innate-adaptive crosstalk occurs through cell–cell interaction upon antigen presentation by antigen presenting cells (APCs) to antigen specific receptors on the surface of T- and B-cells, hence, initiating an adaptive immune response (32–34).

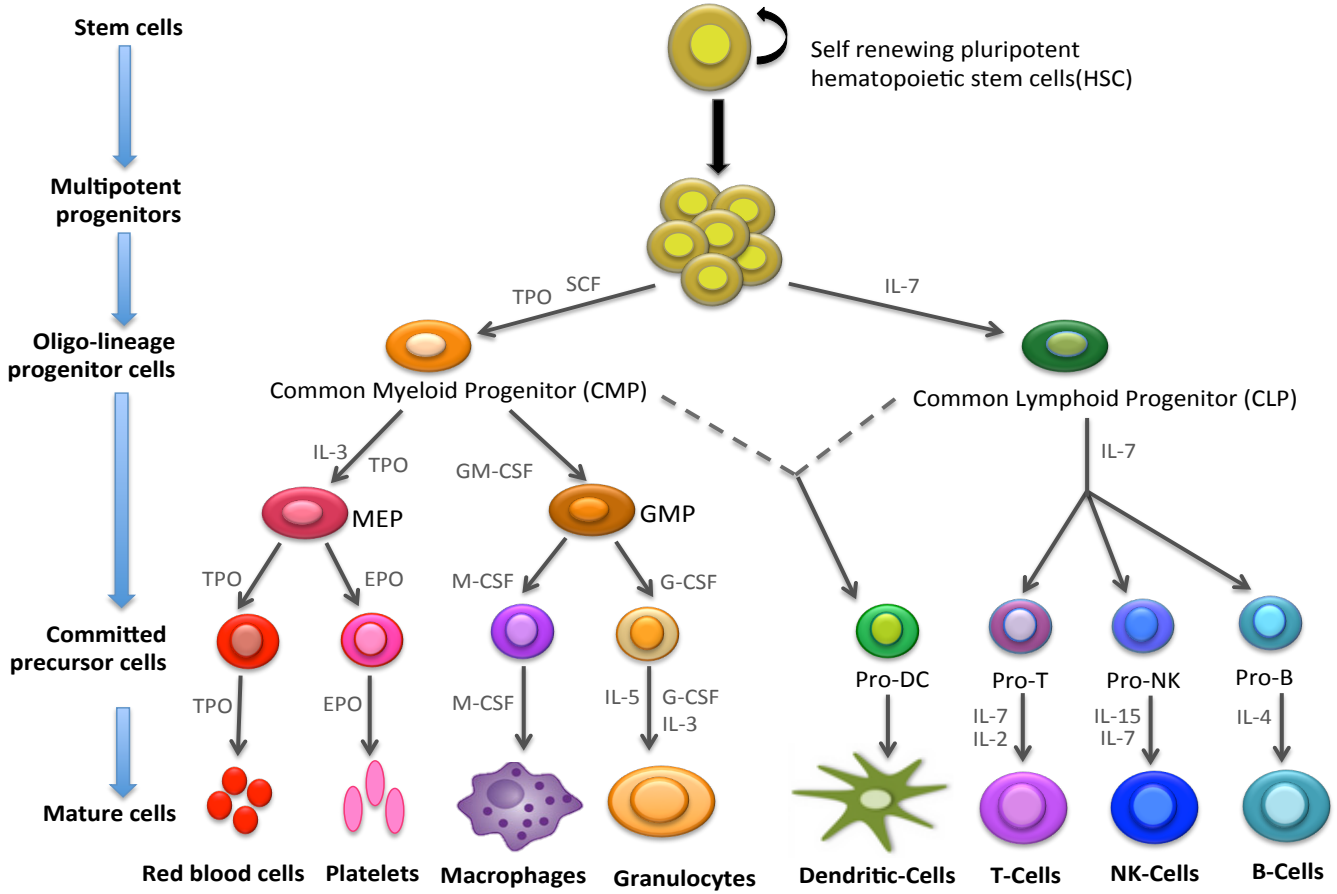


Figure 1.1: Simplified schematic of the hematopoietic lineage tree.

Adapted from Moignard *et al*. (35). MEP, megakaryocyte-erythroid progenitor; GMP, granulocyte-macrophage progenitor. Key growth factors and cytokines are indicated (SCF, stem cell factor, TPO, thrombopoietin; IL, interleukin; GM-CSF, Granulocyte-macrophage colony-stimulating factor; M-CSF, macrophage colony-stimulating factor; G-CSF; Granulocyte-colony stimulating factor; EPO, erythropoietin.

Table 1.1: Major cells of the immune system and their functions.

Immune cell type	Functions
Erythrocytes (Red blood cells)	Transport oxygen around the body
Leukocytes	
Lymphocytes	
T-cells	**$CD4^+$ T-cells** different into various helper T-cells subsets during an adaptive immune response, which secretes specific cytokines to activate other immune cells as well as enhances their function. Also, suppresses immune overactivation (36,37). **$CD8^+$ T-cells** gain effector function following antigen challenge (adaptive immune response) to clear off infection. Few of the $CD4^+$ helper T cells and effector $CD8^+$ T cells go onto developing into memory subsets to protect against reinfections (38).
B-cells	Part of adaptive immunity and are responsible for clearing infections by producing antibodies, which bind to antigens found on invading pathogens (39).
Natural killer (NK) cells	Part of the innate immune system and possesses cytotoxic function to destroy abnormal or virus-infected cells (40).
Granulocytes	
Neutrophils	Functions in innate immunity as phagocytic cells, where engulfed bacteria of fungi are digested with the cytotoxic chemicals released from their granules (41).
Eosinophil	Provides defence against parasites and allergic diseases (42).
Basophil	Involved in inflammatory response to allergic reactions mainly by releasing histamine (43).
Monocytes Macrophages	Phagocytic cells that engulf debris, microbes and apoptotic cells (44). Also, function as antigen presenting cells. Monocytes are precursors of macrophages and dendritic cells (45).
Dendritic cells	Antigen presenting cells, which initiate an adaptive immune response (46).
Platelets	Involved in blood clotting (47).

Since two research chapters of this thesis will focus on $CD8^+$ T-cells (**Chapters 4 and 5**), $CD8^+$ T-cellmediated adaptive immune response is discussed in depth below.

1.3 $CD8^+$ T-cell mediated immunity

1.3.1 T-cell development in the thymus

T-cell progenitors initially arise from HSC in the bone marrow and then migrate through blood to the thymus where they complete their development (48). In the thymus, these progenitors, also known as thymocytes, undergo successive stages of differentiation, which are constrained to specific lineages, before maturing into distinct T-cell subsets and re-entering the circulation (48). The earliest thymocytes are devoid of surface expression of both co-receptors, CD4 and CD8, and are thus referred to as double-negative (DN; $CD4^-$ $CD8^-$) thymocytes (48,49). As these cells progressively differentiate during their development, they become double positives (DP; $CD4^+$ $CD8^+$) and then ultimately mature into single-positive naive $CD4^+$ or $CD8^+$ T-cells (48,49). Once released, these naive T-cells, which are specific for a unique antigen, constantly recirculate through secondary lymphoid organs scanning for their cognate antigens displayed on APCs (50).

1.3.2 Molecular mechanisms regulating effector and memory $CD8^+$ T-cell differentiation

The T-cell immune response following an infection or antigen exposure is comprised of three phases. This includes the initial priming and expansion of naive cells, followed by their contraction, and finally memory formation and maintenance (51). A simplified overview of $CD8^+$ T-cell differentiation is presented in **Figure 1.2**.

1.3.2.1 Priming and activation of naive T cells

A T-cell response is initiated when a naive T-cell recognises and binds to its cognate antigen on APCs through the T-cell antigen receptor (TCR) (52). Mature dendritic cells are the main APCs that activate naive T-cells, but stimulation by others, such as

macrophages, has also been shown (53,54). Dendritic cells process the ingested pathogen into small antigenic peptides (p) and display it on its cell surface by loading them onto the major histocompatibility complex (MHC) molecules (pMHC) (52,55).

Activated dendritic cells have increased expression of MHC and co-stimulatory molecules such as CD80 and CD86, together with altered chemokine receptor expression (up-regulation of CCR7 and CXCR4, and down-regulation of CCR1 and CCR5) (56,57). They migrate to lymph nodes where they effectively communicate with and activate T-cells via TCR-pMHC interactions, triggering an adaptive immune response. $CD4^+$ and $CD8^+$ T-cells interact with MHC class II and I molecules on the target cells, respectively. The CD4 and CD8 co-receptors on the T cells then further enhance TCR signalling by further stabilising the TCR-pMHC complex (58,59).

Besides binding to their respective MHC through TCR, the activation of $CD4^+$ and $CD8^+$ T-cells require accompanying co-stimulatory signals or adhesion molecules. Leukocyte function antigen 1 (LFA-1) and cluster of differentiation 28 (CD28) co-stimulatory receptors present on naive T-cells bind respectively to intercellular adhesion molecules 1 (ICAM-1) and CD80 found on dendritic cells (**Figure 1.3**), which further magnifies the signal received via TCR-pMHC interaction and as a result enhances naive T cell activation and survival (60). Also, pro-inflammatory cytokines and chemokines, particularly interleukin (IL)-12 and IFN-α, have also been shown to augment naive T-cell priming (61). Once activated these T-cells rapidly proliferate (~10^5 fold) and release more IL-2, producing a positive feedback on their proliferation and differentiation into effector T-cells (62,63). Activated $CD8^+$ T-cells mature and expand into effector cytotoxic T lymphocytes (CTLs) (64). The extent of synergy between three signals, which includes antigenic stimulation via TCR (signal 1) in combination with co-stimulatory (signal 2) and inflammatory signals (signal 3), has been suggested to play a role in determining effector fate (**Figure 1.3**) (65).

Activated $CD4^+$ T-cells differentiate into either one of the effector T helper (Th) lineages (Th1, Th2, Th17, and iTreg), which is characterised by the combination of cytokines they produce (62,66–68). Th1 cell differentiation is primarily induced by the cytokines IL-12 and IFN-γ, and the transcription factors STAT-5 (signal transducers and Activators of Transcription), STAT-4 and Tbet (69).

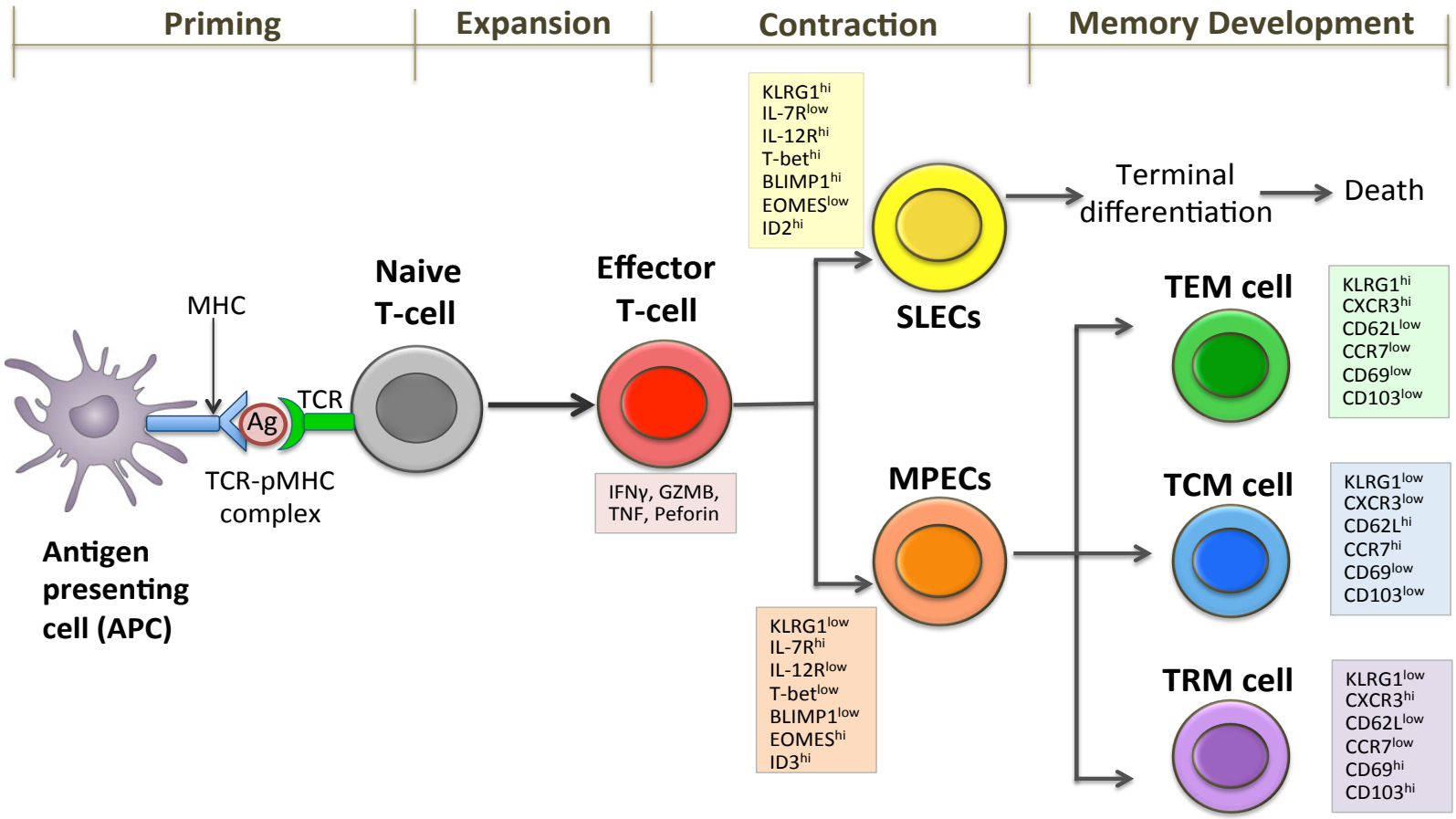


Figure 1.2: Simplified overview of the three main developmental phases of $CD8^+$ T-cell differentiation.

Figure 1.2: Simplified overview of the three main developmental phases of $CD8^+$ T-cell differentiation.

When naive $CD8^+$ T-cells encounter their pathogen-specific antigen (Ag) presented on antigen presenting cells (APCs), they undergo a differentiation program comprising of three steps: priming, expansion, and contraction followed by memory development. These naive cells rapidly proliferate and differentiate into cytotoxic effector cells, which secrete antiviral cytokines and cytolytic proteins (interferon gamma, IFNγ; granzyme, GZMB; and perforin). The early effector cells, which can adopt different fates, further differentiate into short-lived effector cells (SLEC) or memory precursor effector cells (MPEC). The terminally differentiated SLECs express increased amounts of killer cell lectin-like receptor G1 (KLRG1) and interleukin-2 receptor (IL-2R), and low amounts of interleukin-7 receptor (IL-7R). SLECs die during the contraction phase. By contrast, MPECs, which develop into long-lived, self-sustained memory T-cells, express high levels of IL-7R and low levels of KLRG1 and IL-2R. The memory population comprises of the circulating effector memory (TEM) cells and central memory (TCM) cells, and the tissue resident memory (TRM) cells. TCM cells express CD62L (L-selectin), and CCR7 homing molecules, which regulate their access to lymph node and localisation to secondary lymphoid organs. On the other hand, TEM cells lack lymph node targeting receptors, CCR7 and CD62L, but they express cytokine receptors such as CXCR3 allowing them to gain access to non-lymphoid peripheral tissues, spleen, and blood. TRM cells, which permanently reside in peripheral tissues, express the canonical CD103 and CD69 surface markers that facilitate their retention at tissue sites. This figure has been adapted from Kaech and Cui, 2012 (70).

IL-4 induced expression of transcription factors STAT-6 and GATA-3 (GATA Binding Protein 3) leads to polarization of Th2 cells (69). Th17 cell fate is determined by the cytokines IL-6 and TGF-β, and the transcription factors STAT-3 and RORγt (retinoic acid-related orphan nuclear receptor) (71). The TGF-β/FOXP3 signalling pathway plays a critical role in the polarization of regulatory T (Treg) cells (72). Each of the TH subsets have their own dominant cytokine profiles and lineage-specific transcription factors, which counter regulates the development of other subsets as well as self-reinforces their lineage commitment and maintenance (69).

1.3.2.2 Clonal expansion and differentiation of $CD8^+$ T-cells: Molecular mechanisms regulating effector and memory fates

$CD8^+$ effector CTLs mediate killing of infectious agents by secreting antiviral cytokines such as interferon-gamma (IFNγ), tumour necrosis factor (TNF), and cytolytic proteins including granzyme B (GZMB) and perforin (73).

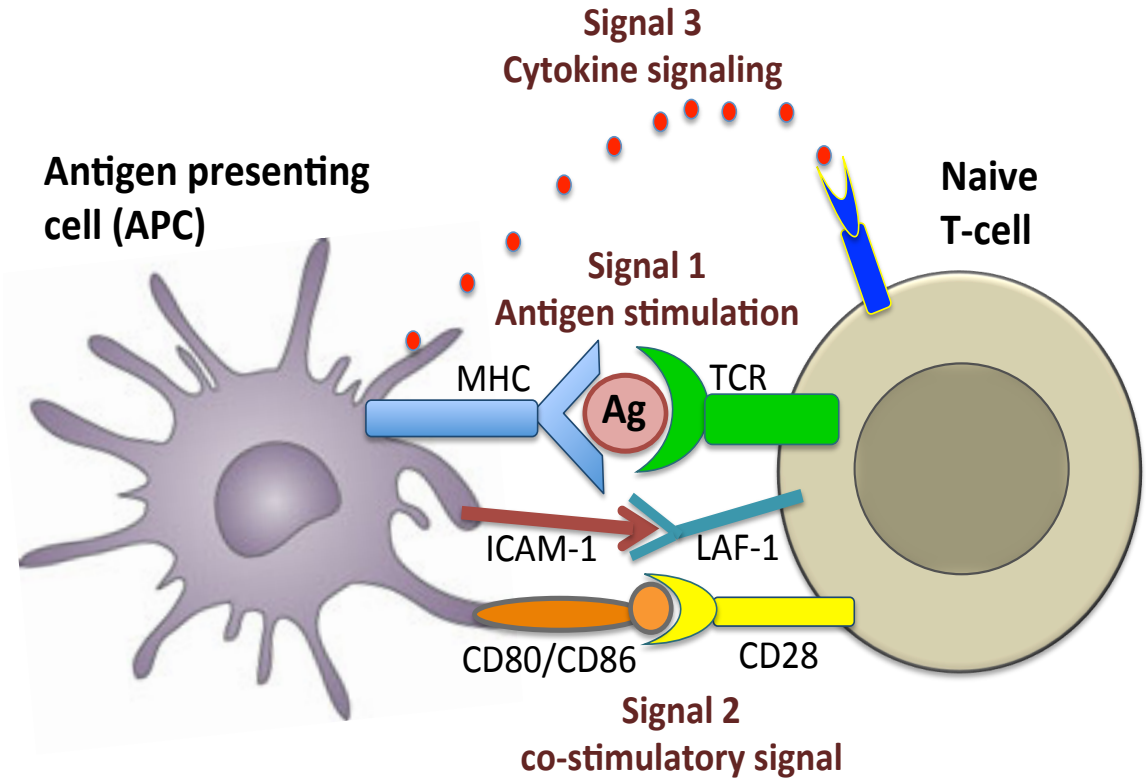


Figure 1.3: Naive T-cell stimulation and activation require the three signals provided by dendritic cells.

Figure 1.3: Naive T-cell stimulation and activation require the three signals provided by dendritic cells.

Signal 1 is the activation signal when the naive T-cell binds via the T-cell receptor to the MHC-associated peptides (Ag) presented on the surface of dendritic cells, forming a TCR-peptide-MHC (TCR-pMHC) complex. Signal 2 are the co-stimulatory signals triggered by co-stimulatory and adhesion molecules. Leukocyte function antigen 1 (LFA-1) and cluster of differentiation 28 (CD28) co-stimulatory receptors present on naive T-cells bind respectively to intercellular adhesion molecules 1 (ICAM-1), and CD80 found on dendritic cells. Signal 3 is cytokine and chemokine mediated signalling that leads to the functional polarisation and differentiation into effector T-cells. This figure has been adapted from Kapsenberg, 2003 (74).

The CTL effector group comprises two sub-populations that can be distinguished by the expression of surface markers interleukin-7 receptor (IL-7R) and killer cell lectin-like receptor G1 (KLRG1). These are: short-lived effector cells (SLECs; IL-7R^{low} and KLRG1^{high}) that are lost through apoptosis; and memory precursor effector cells (MPECs; IL-7R^{high} and KLRG1^{low}) which preserve long-lasting memory for a particular antigen, so that re-exposure to that antigen elicits a more rapid and enhanced response (75).

Although the role of effector CD8^{+} T-cells has long been well-known, the actual molecular mechanisms underlying their function and fate have only just begun to be elucidated. For instance, studies have indicated that the strength and nature of TCR signalling can regulate effector fate. Dampened TCR signalling skewed CD8^{+} T-cell differentiation towards a memory phenotype (76,77). Mice expressing a mutant TCR showed impaired TCR-mediated NF-kappaB signalling, which resulted in the loss of memory cell development but intact effector differentiation (78).

In addition, the combinatorial effects of the signals as mentioned earlier in Figure 1.3 may indirectly determine cell fate by altering transcription factor levels. Studies have identified a number of critical fate-determining transcription factors. These usually work in pairs and in a counter regulatory fashion to produce both SLECs and MPECs.

T-Bet* and *EOMES. It has been observed that high levels of IL-12 up-regulate *T-bet* expression inducing SLECs, while low levels of IL-12 coupled with low *T-bet* expression promoted accumulation of MPECs (75). Reciprocal expression of *T-bet* and *Eomes* has been noted in early effectors (T-Bet^{high} and EOMES^{low}) and memory cells (T-Bet^{low} and EOMES^{high}) (64,79).

BLIMP1* and *BCL6. B lymphocyte–induced maturation protein 1 (BLIMP1 or PRDM1) and BCL6 are reciprocally expressed exerting antagonist effects on each other to determine cell fate during differentiation of effector and memory T-cells (80,81). *Blimp1* is highly expressed in IL-7R^{low} KLRG1^{high} SLECs and enhances CTL functions through the production of effector molecules such as IFN-γ and granzyme. Murine T-cells lacking *Blimp1* differentiate into IL-7R^{high} KLRG1^{low} MPECs with attributes mirroring central memory T-cells (TCM). These cells are also deficient in granzyme B

and have increased *Bcl6* expression compared to *Blimp1*-expressing T-cells from wild-type mice (82,83). In memory T-cells, *Bcl6* expression levels are inversely correlated to *Blimp1* levels. The overexpression of *Bcl6* results in increased TCM cells (84) suggestive of its a role in TCM development.

***ID2* and *ID3*.** Inhibitor of DNA binding 2 (ID2) and 3 (ID3) are another important set of transcription regulators, which control effector and memory cell fates by inhibiting the DNA-binding activity of E-protein transcription factors (38,85). Studies with *Id2* and *Id3* reporter constructs in mice showed that *Id3*-GFP^{high} phenotype correlated with memory potential resulting in effector cells differentiating into MPECs; producing more IL-2. On the other hand, *Id2*-YF2 expression led to the generation of SLECs (38,85). Loss of ID2 and ID3 impaired the formation of memory, and short-lived effector subsets respectively (38). Moreover, BLIMP1-mediated repression of *Id3*, via direct binding to *Id3* promoters, is a key determinant of effector cell fate (86).

In conclusion, there is a complexity yet also an exquisite balance in the way that transcription factors regulate the differentiation and production of effector and memory immune cell. Hence, a genome-wide analysis is needed to comprehensively evaluate and deconvolute the complexity of the transcriptional regulation mediated by these transcription factors.

1.3.2.3 Contraction and development of memory T-cells

Following clearance of the primary infection, the expanded antigen-specific T-cell pool undergoes rapid contraction, with most effector cells dying by apoptosis. Only a small proportion (~10%) of memory cells remain to protect against reinfection (87). The contraction phase and subsequent generation of memory T-cells is mediated largely by the interactions between survival (BCL2, BCL-XL, BCL-W, MCL1) and apoptotic (BIM, BID, and PUMA) proteins.

Earlier studies have identified two distinct circulating groups of memory T-cells, based on their function and the expression of migratory homing receptors on their cell surface. **Central memory T-cells** (**TCM**) express CD62L (L-selectin), and CCR7 homing molecules, which regulate their access to lymph node and localisation to secondary lymphoid organs. TCM cells can produce interleukin (IL-2) following antigen

induction, stimulating them to proliferate expansively, but they have decreased effector capabilities (87–91). On the other hand, **effector memory T-cells** (**TEM**) lack lymph node-targeting receptors, CCR7 and CD62L, but they express other cytokine receptors such as CXCR3 allowing them to gain excess into non-lymphoid peripheral tissues, spleen, and blood (87–91). TEM cells have decreased proliferative capabilities, but have increased effector potential due to their ability to express effector cytokines like IFN-γ and IL-4 (87–91). In addition, TCMs have longer lifespans compared to TEM cells and can also themselves differentiate into TEM cells upon exposure to antigens (92).

The loss of TEM cells over time from circulation does not necessarily deprive extra-lymphoid tissues of T-cell immunity (93); a new subset of antigenic-specific long-lived memory T-cells, referred to as **tissue resident memory** (**TRM**) cells, have been identified to be permanently localised to peripheral tissues following an infection. TRM cells are disconnected from circulation and reside long-term in barrier and non-barrier tissues, where they have been shown to provide superior and rapid frontline defence against local reinfections (94–102). The expression of specific homing receptors and concurrent repression of genes involved in tissue egress facilitates the retention of TRM cells in tissues (95,103–105). Additionally, TRM cells display a transcriptional program that distinguishes them from their circulating TEM and TCM counterparts (88,103). Even though TRM cells are distinct from the circulating memory subsets, based on their phenotype, function and transcriptional profile, the mechanisms underlying tissue residency are still not fully understood.

As mentioned above memory T-cells express cytokine and chemokine receptors, which suggests that cytokines play a fundamental role in the development, migration and function of these memory cells.

1.4 Cytokines, chemokines, and growth factors

Cytokines, chemokines, and growth factors (hereafter referred to collectively as "cytokines", unless otherwise stated) are cell signalling molecules that play a key role in regulating immune cell differentiation, immune response and inflammation during host response to infection and injury (106).

1.4.1 Properties and function of cytokines

Cytokines are low molecular weight (usually ~10-40 kD), water-soluble proteins or glycoproteins, which are secreted by immune cells in response to a stimulus. These secreted cell signalling molecules are intimately involved in coordinating an effective immune response by facilitating communication between innate and adaptive responses. They exert their effect by binding to their cognate receptors on the surface of target cells (106–108) in an autocrine (same cell), paracrine (neighbouring cell) or endocrine (distant cells via circulation) fashion (109). Moreover, there is redundancy, in that a number of cytokines share similar biological functions. Several cells can produce the same cytokine, and each cytokine might be involved in multiple signalling pathways or act on multiple cells (110). In addition, cytokines also have the propensity to exhibit synergistic, antagonistic, and cascade-induction behaviour (110–112). Depending on the local inflammatory environment, cytokines may either promote (pro-inflammatory) or suppress (anti-inflammatory) inflammatory response, or both (113). The cytokine superfamily can be divided into seven broad classes including interferons, interleukins, colony-stimulating factors (CSF), angiogenic growth factors, tumour necrosis factors (TNF), transforming growth factors, and chemokines (**Table 1.2**).

1.4.2 Multiplexed cytokine profiling

Cytokines are traditionally measured using the antibody-based Enzyme-Linked Immuno-Sorbant Assay (ELISA) assays, which are the best-validated approach currently available. However, this approach generally requires large sample quantities, as multiple cytokines cannot be tested from one aliquoted sample; each cytokine requires its own aliquot for measurement (114). In recent years, several multiplexing technologies have emerged as an extension of ELISA to overcome these limitations. Multiplex cytokine profiling allows quantification of multiple cytokines (several dozens) simultaneously in serum or plasma. The two basic assay designs for multiplex cytokine profiling are beads-based multiplex assays and planar array assays. In the beads-based multiplex assay, relies on the capture and detection antibodies (**Figure 1.4**).

Table 1.2: Cytokine subgroups and their key functions

Group	Cytokines	Primary functions
Interferons (INFs)	INF-α, INF-β and INF-γ	Secreted in response to viral infections, exerting their effects by stimulating several downstream interferon-induced antiviral and immunomodulatory genes (115).
Interleukins (ILs)	IL-1, IL-2, IL-3, IL-4, IL-5, IL-6, IL-7, IL-9, IL-10, IL-11, IL-12, IL-13, and IL-17	Involved in immune cell growth, differentiation and activation (116).
Colony-stimulating factors (CSFs)	CSF-1, CSF-3	Exert their effect in a lineage-specific manner regulating the proliferation, differentiation and survival of cells in the macrophage and neutrophil lineages (117).
Angiogenic growth factors	Vascular endothelial growth factor A (VEGF-A), platelet-derived growth factor (PDGF), Hepatocyte growth factor (HGF), stromal cell-derived factor-1 (SDF-1)	Secreted by injured tissues, platelets, and immune cells to induce and regulate angiogenesis, a prerequisite for facilitating wound healing process (118).
Tumor necrosis factors (TNFs)	TNF-α, TNF-β	Exhibits an array of function including immune response, hematopoiesis, and initiation of cell survival, differentiation and apoptosis pathways (119).
Transforming growth factor beta (TGFB)	TGFB1, TGFB2, and TGFB3	Involved in regulating immune cell differentiation, proliferation, and immune homoeostasis (120).
Chemokines	Monocyte chemotactic protein (MCP)-1 MCP-2, IL-8, Macrophage inflammatory protein-1beta (MIP-1B)	Chemotactic migration and infiltration of macrophages, monocytes, neutrophils, and other immune cells (121).

The first laser light identifies the beads based on their fluorescence signature and the second laser light quantifies the cytokines by measuring the fluorescence intensity of the fluorescently labelled detection antibodies (114).

In the planar array assays, the two-dimensional array (96 wells) consists of different capture antibodies immobilised at different spots within each well, which are detected by chemiluminescence (114). The data generated from such multiplex platform needs to be treated with caution, as they tend to be confounded by technical and systematic noise. Hence, appropriate curve-fitting models and normalisation strategies should be employed when calculating concentrations (from fluorescence intensities) and preprocessing data, respectively.

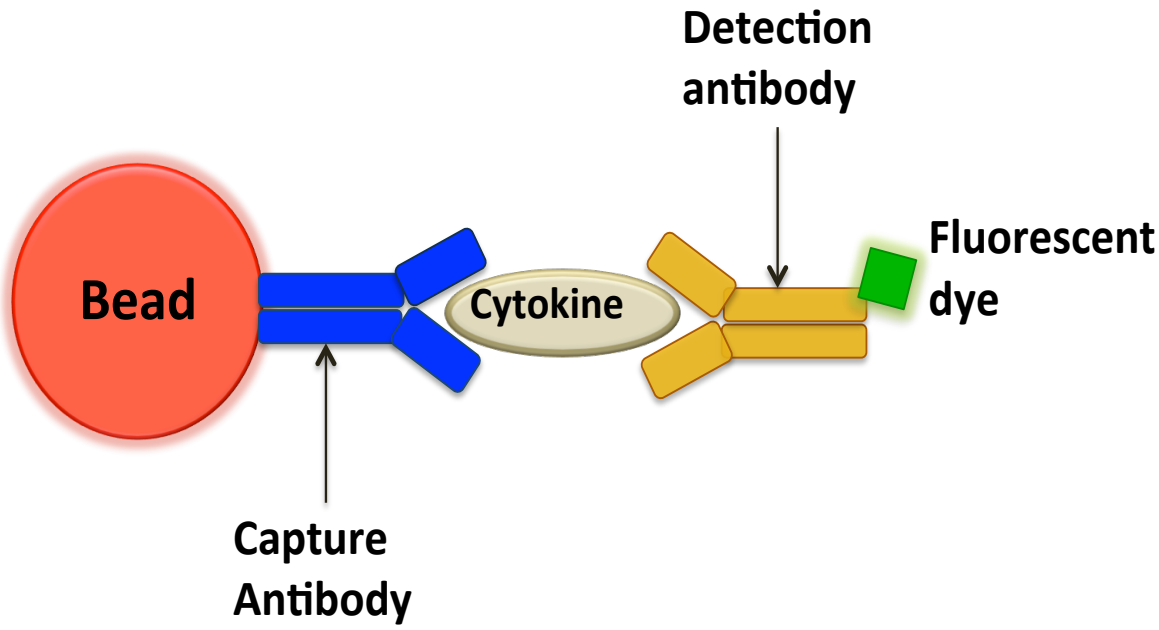


Figure 1.4: Schematic showing the bead-based sandwich immunoassay system used for cytokine detection in biological samples.

Figure has been adapted from Sachdeva and Asthana, 2007 (122).

1.4.3 Cytokine profiling to assess the immune system

The emergence of technologies such as multiplex cytokine profiling allows us to quantify a large number of cytokines across multiple samples simultaneously, which permits a systems-based investigation of complex immune and inflammatory responses. Several studies have utilised such technologies to assess cytokine patterns elicited following vaccination, infection or disease onset. For instance, cytokine profiling has been used in vaccination studies to establish the cytokine pattern induced in PBMCs, in response to vaccination by human papillomavirus (HPV) L1 virus-like particles (VLP) (123). It has also been used to quantify the difference in cytokine responses following the anti-tuberculosis Bacillus Calmette–Guérin (BCG) vaccination across infants from Malawi, UK, and Gambia (124,125). Finally, it has been used to characterise the response of $CD4^+$ T-cells in subjects who received the YFV-17D yellow fever vaccine (126). Multiplexed cytokine analysis has also been used in association with human diseases; cytokine profiles have been constructed to differentiate disease severity in patients with rheumatoid arthritis and sepsis (127,128); to examine the host immune response to variable compositions of gut microbiota in HIV patients (129); and to identify biomarkers signatures associated with other diseases (130–132).

The variation in immune response can be partially explained by heritable influences (133). Genetic variation in immune-related traits such as cytokine levels likely contributes to inter-individual differences in immune function and affects disease risk.

1.5 Genome-wide association studies (GWAS) to investigate the genetic architecture of complex diseases and traits

Over the last two decades, genome-wide association studies (GWAS) have emerged as a powerful approach to decipher the genetic basis of human diseases. GWAS has been used to identify common genetic variants (with minor allele frequency [MAF] > 5%) associated with disease risk (134), clinical phenotypes (135,136), and response to treatment and therapies (137).

Much progress has been made in creating a comprehensive catalogue of common variants occurring in diverse human populations. This includes the Human Genome

Project (138) and its extended International HapMap Project (138–140), the 1000 Genomes Project (141–143). Technological advances in genotyping have improved the viability of performing GWAS in both a time and fiscal sense. The two most popular genotyping platforms used in GWAS studies, Affymetrix GeneChip and Illumina BeadChip arrays, are capable of assaying thousands to millions of SNPs simultaneously across the genome (144). Genotype imputation, an approach to infer un-assayed SNPs using a reference panel derived from existing population data (e.g. 1000 Genomes), is commonly employed in GWASs to increase power, fine map regions, and enable summary statistics from separate studies to be combined via meta-analysis (145).
Additionally, the application of the genotyping platform has been expanded to detect structural variants such as copy number variations (CNVs). CNVs refer to insertions, duplications, or deletions of segments of a chromosome (> 1kb) that vary in number between individuals (146).

In performing a GWAS, one needs to account for linkage equilibrium (LD), the correlation between closely-neighbouring SNPs due to them being inherited together more frequently than SNPs that are further apart. A SNP identified via GWAS to be associated with a complex disease or quantitative trait may not be the causative SNP, but may instead be in LD with the causal allele (147–149).

GWASs are a popular method of analysis for genetic associations. The initial large-scale studies explored a large variety of human diseases, including age-related macular degeneration (150), Crohn's Disease (151), type 2 diabetes (152), inflammatory bowel disease (153). Of particular note is a landmark GWAS study conducted by the Wellcome Trust Case-Control Consortium (WTCCC), that jointly-investigated seven prominent diseases (bipolar disorder, coronary artery disease, Crohn's disease, rheumatoid arthritis, type 1 and type 2 diabetes, and hypertension) (154). Since then a number of studies have been conducted, leading to the identification of thousands of reproducible common genetic variants associated with complex traits. As of 17th April 2017, an online catalogue of all published GWAS studies, maintained by the he European Bioinformatics Institute (EMBL-EBI), contained 2,584 curated publications with 33,674 unique SNP-trait associations (155,156).

Inflammatory and autoimmune diseases have been at the front line of scrutiny using GWAS, which has led to the identification of an exceptional number of risk loci (157,158). Furthermore, several studies have provided insight into the genetic determinants underlying blood cell traits, which have been extensively used in clinical settings as indicators to assess human health. Genetic variants influencing a number of blood cell traits such as platelet counts and volumes (135,136), haemoglobin concentrations (159), and white blood cell counts (136) have been identified. As a result, the unprecedented number of susceptibility loci identified through these GWASs have considerably redefined our understanding of the genetic architecture (number, frequency and effect size of susceptibility alleles, and the way these alleles collectively interact) of common diseases or traits (148,160–162). Besides, studies are now focusing on characterising risk variants regulating intermediate immunological parameters (e.g. cytokines, gene expression) that are linked to disease development. The last few years has seen an increase in the availability of genetic data from large studies such as the UK Biobank, Electronic Medical Records and Genomics (eMERGE) network, and US National Institutes of Health Precision Medicine Initiative, in which individuals have been extensively phenotyped.

However, despite considerable progress in mapping the susceptibility loci of diseases, identification of casual variants at each of these loci, and the mechanism of their effect remains elusive. The two main reasons for this are as follows: firstly, the standard GWAS approach is often underpowered to identify causal SNPs with small effect sizes (163). Secondly, even if the GWAS is sufficiently powered, the majority of identified variants occur in intergenic, non-coding regions, with an uncertain mechanism of action. Some may exert their effect as transcriptional regulatory SNPs since they generally overlap with regulatory elements such as promoters (164).

A problem that often emerges in both population- and family-based GWA studies is confounding due to genetic relatedness between samples, which can lead to spurious (false positive) association signals if not appropriately accounted for in the analysis (165). The two types of relatedness that need to be identified and adjusted for are population structure and cryptic relatedness. Population structure occurs when there is a substantial difference in ancestry-specific allele frequencies between subgroups in a study population (165,166). In most cases, population structure is closely aligned with

self-reported ethnicity, race, language or geographic origin. Cryptic relatedness arises when close relatives (up to third-degree) are present in population-based studies (166). The inference of population structure and cryptic relatedness using genome-wide SNP data, and accounting for these effects accordingly in downstream association analysis, has become a common practice. PCA is one of the most widely used approaches to control for population structure (167). Usually PCA is performed on a small group of representative SNPs (with low pair-wise correlation) that can still extract the same structure of subpopulations within the dataset (168). The inferred PCs capturing the ancestry signal for each individual are then adjusted for as covariates in the association test for each SNP (169). In recent years, association analysis based on linear mixed models (LMM), which requires a matrix of pair-wise genetic relatedness computed between all individuals, has become a well accepted method to simultaneously account for population structure and cryptic relatedness (170). The quantile-quantile (QQ) plot of observed versus expects *P*-values and the genomic inflation factor (λ) have been routinely used to assess the presence of genetic relatedness (171). The inflation factor is estimated as the ratio of the median value of the empirically observed test statistics to its expected under the null hypothesis of no association, and it should be close to one in the absence of population stratification or cryptic relatedness (171).

1.6 Gene expression profiling- microarray and RNA-seq

The genetic knowledge gained from the completed sequence of the human genome has led to rapid developments in high-throughput technologies that make it possible to profile thousands of genes simultaneously within a cell or tissue. With regards to the immune system, global transcriptome analyses of specific immune cells and mixtures of immune cell populations (e.g. leukocytes in whole blood) have been used to characterise and understand the underlying molecular mechanisms regulating immunity in health and disease. Microarray-based profiling has been most widely used for such analyses, but in recent years, cDNA sequencing using next-generation sequencing (RNA-seq) has become an increasingly popular alternative (172).

1.6.1 Tools and methods for analysing gene expression data

1.6.1.1 Microarray and RNA-seq technology

Microarrays are based on the principle of hybridization (173) (adapted from Southern blotting) where fluorescently-labelled cDNA, derived from mRNA, hybridise to the oligonucleotide probes on the array chip. The chip is read by a laser scanner to generate a fluorescent image, where the fluorescence intensity correlates with the amount of mRNA isolated from the sample. Affymetrix GeneChips and Illumina BeadArrays are two platforms available for microarray-based expression studies, and they have been compared in Barnes *et al*.(174).

In RNA-seq, libraries are prepared by ligating sequence adapters to cDNA, which are then subjected to PCR amplification, followed by sequencing. The millions of short sequence reads generated are aligned back to a known reference genome. Quantification by counting or estimating the number of reads overlapping within a genomic region (gene or exon) is used as an abundance measure (175–177). The sequence template can be sequenced from either one (single-end sequencing) or both ends (pair-end sequencing). Pair-end reads are particularly useful for detecting new transcripts and isoforms (178). Sequencing provides a number of advantages over microarrays, such as the ability to identify new transcribed regions and isoforms, gene fusions (translocations), allele-specific expression, differential splicing, and allows identification of single nucleotide variants (179,180). The recent applications of RNA-seq technology have expanded to single-cell sequencing, which has emerged as a powerful tool to profile the transcriptome of individual cells to investigate cell-to- cell heterogeneity within an individual (181).

1.6.1.2 Normalising gene expression data

Normalisation is an essential step in both microarray and RNA-seq data analysis to ensure that expression values are comparable across samples and reliable for downstream analysis. The choice of normalisation strategy may have a strong influence on the identification of differentially expressed genes, clustering, and the construction of gene networks (182).

Microarray. The common factors that usually contribute to unwanted systematic differences between samples include technical variation introduced during sample preparation and hybridization (183,184), and study design-related batch effects (185). The majority of published microarray studies have used Affymetrix GeneChips for profiling. The most popular normalising algorithms for Affymetrix gene arrays are Microarray Suite (MAS), Robust Multi-array Analysis methods including RMA (186), gcRMA (for correcting GC content) (187,188), and the Li-Wong model (189). However, since these methods do not take into account the study design, they often fail to remove batch effects (183,190–192). A number of methods such as ComBat (193), SVA (194), SNM (195), and RUV-2 (183), which operate in a supervised manner, have been proposed to filter batch effects. Supervised normalisation for microarrays (SNM) jointly fits a study-specific model to the expression data by defining two sets of variables: one that is of interest to the biological outcome (biological variables) and the other that is not (adjustment variables) (195).

RNA-seq. Specific technical biases inherent in RNA-seq data can also impact downstream analysis. To correct for the differences in library size or the total number of aligned reads across samples, several library size normalisation methods exist. The FPKM (Fragments Per Kilobase of exon per Million mapped reads) and RPKM (Reads Per Kilobase of exon per Million mapped reads) measures for paired-end and single-end reads, respectively, account for both the transcript length and library size within samples (196). However, since this approach only makes transcripts comparable within samples, it might not be appropriate in cases where a dissimilar distribution of transcripts is observed between samples. For example, in some samples, a few highly expressed transcripts may dominate the total reads counts resulting in a skewed distribution. Methods such as the trimmed mean of M-values and upper quartile scaling (UQ) are commonly used to normalise such between sample differences (197).

1.6.1.3 Differential gene expression analysis

The key goal of gene expression studies is to identify genes differentially expressed between two conditions (198).

Microarray. Fold change was the first crude method used in microarray analysis to rank differentially expressed genes. Fold change for each gene was calculated as the ratio of

mean expression intensities (log-transformed) between two groups. Often a list of differentially expressed genes was obtained using an arbitrary threshold of at least two-fold difference, with no statistical confidence given for the differential expression (199). This approach ignores the variance between replicates of the same condition (200). Hypothesis testing through statistical models has been developed to detect differentially expressed genes. A t-test is commonly used, and it assumes the data to be normally distributed with the groups having equal variance (198,199). However, the estimation of variance for each gene may not always be accurate since this method does not consider the heterogeneity in variance across all genes analysed (201,202). In cases where there are very few replicates, this can lead to false positives. To overcome this problem, Smyth *et al.* proposed a modified t-statistic approach, an empirical Bayes method, which estimates the variance of a gene while borrowing information from other genes (202).

RNA-seq. While analysis of microarray data assumes a continuous measurement of expression intensities, RNA-seq analysis requires separate statistical methods that take into account the discrete distribution of read counts. The commonly used models are Poisson distribution and Negative Binomial (NB) distribution (203,204). Poisson distribution is mainly used for its simplicity and has a single parameter where the variance of the model is identical to the mean (203). This model works well with technical replicates, but in the case of biological replicates, where the variability tends to be much larger, this model tends to underestimate the biological variation leading to the problem of overdispersion (203,205). The NB model is able to handle overdispersion by factoring in the relationship between the two parameters, mean and variance (204). Various tools, including Cuffdiff2 (206), DESeq (207), and edgeR (204), perform differential expression analysis using the NB model.

Multiple testing correction methods. A major problem encountered when analysing gene expression data is simultaneous testing of multiple null hypotheses (no association between the expression level of each gene with a condition or response) (208). Multiple hypothesis testing introduces two types of errors: a Type I error (false positive) results when a gene is declared differentially expressed when it is not; a Type II error (false negative) arises due to the failure to identify a truly differentially expressed gene (209). Statistical methods proposed for controlling false discovery rates includes the most

stringent Bonferroni correction method (210), and the less stringent and more appropriate methods for microarray analysis such as Holm (211), Storey's q-value (212), and Benjamini & Hochberg (213).

1.6.1.4 Cluster analysis and its limitations

Clustering is a valuable exploratory tool used as an initial dimension-reduction step in expression analysis, to partition genes with high expression-similarity into meaningful groups (214,215). Clustering of genes is based on the assumption that genes with similar expression have similar functions or share a common biological pathway, which can be useful to infer functions of unknown genes within the same cluster (216,217). The most popular clustering methods include K-means (218), hierarchical clustering (219), self-organizing maps (SOMs) (220), and principle component analysis (PCA) (221); these are all unsupervised, in that they do not rely on a priori knowledge of the data (222). Clustering of samples can lead to the identification of unknown sub-groups in the samples (223). A prerequisite for all clustering algorithms is that a measure of dissimilarity, or distance between two genes or samples, must be computed so that genes or samples placed in the same cluster are most similar (less dissimilar) to each other than to those from another cluster (224). Euclidean distance and Pearson correlation coefficient are two commonly used dissimilarity measures (224).

The traditional clustering methods mentioned above have a number of limitations. Firstly, since genes are grouped based on similar expression patterns across all samples, in cases where there is heterogeneity in experimental conditions, then these methods are no longer appropriate (225–228). Secondly, it misses out genes that belong to the same functional pathway if they have dissimilar expression patterns due to transactivation or transrepression (229). Thirdly, each gene is assigned to only one specific cluster, and most algorithms do not produce overlapping clusters; whereas in reality a gene may be involved in multiple pathways in different ways (activator, repressor) under varying experimental conditions (227,230). Finally, the clustering algorithm detects patterns in both noise and signal, so it ends up assigning all genes into a cluster, generating biologically irrelevant noisy clusters (230). The weakness of these traditional clustering algorithms have consequences on downstream biological interpretation and have motivated the development of co-expression network analysis approaches.

1.6.1.5 Gene co-expression network analysis

Gene co-expression networks, which are becoming increasingly popular in transcriptome analysis, facilitate the understanding of transcriptional programs governing specific biological processes or cellular traits (231,232). Since genes involved in the same pathway or biological processes are usually co-expressed, network analysis provides a basis to represent functionally related genes as interaction networks (233).

Gene network fundamentals. The idea of displaying gene expression data as a network was introduced by the work of Butte and Kohane on yeast *S. cerevisiae* expression data (234). They computed mutual information (MI), defined as the amount of information that one gene contains about another gene, for all pairwise gene-gene expression combinations (234). Then a threshold value was applied to screen for gene pairs that were biologically linked with similarity above the threshold (234). The resulting output was a co-expression network comprised of clusters of genes, an undirected graph where the nodes represent genes, and edges between gene pairs indicate the co-expression associations (216,229,234). Gene pairs exhibiting high correlation scores have a greater chance to be functionally related and are more likely to be co-regulated by common transcription factor(s) (235,236). MI has the following properties: its non-negative, symmetric, and additive for independent variables (234,237). Despite mutual information being robust, it can be computationally demanding to calculate. Hence, many network construction methods commonly use traditional Pearson correlation coefficients to quantify co-expression between two nodes. Alternatively, ranked-based correlation metrics such as Spearman and Kendell are applied when outliers are a concern.

To discriminate biologically relevant correlations from noise, an arbitrary minimum correlation threshold can be applied (hard thresholding) to generate an unweighted network (238). However, the optimal choice of a threshold can be challenging, as it can discard meaningful correlations resulting in information loss. On the contrary, a commonly-used algorithm, Weighted Gene Coexpression Network Analysis (WGCNA), deals with these limitations by giving the edges weights based on their correlation strength (soft thresholding) and only penalises weaker correlations (239).

In a co-expression network, a module refers to a subnetwork consisting of a subset of highly connected genes. Module detection with the WGCNA algorithm is based on using the topological overlap measure, a measure of connectedness between common neighbours shared by gene pairs, to hierarchically cluster densely-connected shared neighbours (240). Tree-cutting algorithms are used to cut the branches of the dendrogram based on its shape. Several cluster parameters can be tuned to identify nested and tightly connected modules (241).

Gene network structure. Several studies characterising the topology (arrangement of nodes in a network) of yeast, human, and mouse co-expression networks have shown that these networks exhibit small world and scale-free architecture (242–246) features that hold true for other cell biology (247) and real world networks (248). A scale-free co-expression network is heterogeneous in structure and exhibits properties where most nodes in the network will be connected to only a few other nodes (i.e. less connectivity), but some nodes will act as hubs connected to many nodes (i.e. high connectivity) (242,244,247). Co-expression networks have small-world architecture whereby non-neighbouring nodes can be reached from every other node by a very small distance L such that the average shortest route in the network is very small (216,242).

Replication of network topology. A key aim in network biology is to assess module reproducibility in an independent dataset. A typical approach, which does not require rigorous statistical methods and is computationally inexpensive, involves visual inspection (e.g. correlation heatmaps) or cross-tabulation of genes within a module across datasets (249,250). This type of approach can be applied to various module detection strategies including basic clustering methods. However, it requires modules to be detected in the test dataset as well and does not yield information about network topology (the relationship between genes). To address these limitations, module preservation statistics have been developed that assess the patterns of gene connections within a module in the test data, using network topological metrics (251,252). For example, Ritchie *et al*. have recently developed a tool called NetRep, which uses a permutation-based approach to assessing module preservation (251).

1.6.1.6 Differential co-expression network analysis

Differential co-expression analysis (DCA) extends on the idea of the gene co-expression network, by aiming to identify groups of genes with altered dependencies across two classes or conditions, i.e. a group of genes or modules highly co-expressed in one condition but not other (253). DCA is a useful complement to the conventional analysis for identifying differential gene expression, as it can identify genes that differ mildly in their expression yet have a strong opposing effect on downstream genes between conditions (254). Several computations methods have been developed for differential co-expression analysis such as CoXpress (255), DICER (256), DiffCoEx (253) and others (257,258). DiffCoEX and MODA use WGCNA to identify gene networks.

1.6.1.7 Functional enrichment analysis

One of the primary goals of gene expression analysis is to ascertain the functional properties of gene sets obtained from differential expression analysis, clustering or network analysis. Functional enrichment analysis is a popular statistical method to look for gene enrichment in a list of DE genes or gene clusters annotated for particular biological pathways or processes (259,260). Several databases such as Gene Ontology (GO) (261), Kyoto Encyclopedia of Genes and Genomes (KEGG) (262), and Biocarta Pathways (263), can be utilised for functional analysis. Over the years a number of tools like DAVID (264,265), GeneCodis (266,267), GeneTrail (268), and GOrilla (269) have integrated information from other databases (e.g. GO, KEGG, BioCarta, disease and protein databases) in order to make the enrichment analysis more comprehensive. For most of these tools, GO database is widely used to identify statistically overrepresented GO terms in a given gene set. In addition, methods that prioritise gene sets based on their association with a given condition have been developed (270). Among those, Gene Set Enrichment Analysis GSEA (271) is the most popular, which performs differential analysis to determine whether a predefined group of genes show significant expression difference between two groups.

1.6.2 Using transcriptome profiling to assess the immune system

Blood transcriptomic profiling has been widely used to capture the overall immune and inflammatory state of the body, as reflected in circulating leukocyte activity. Changes in

global gene expression profiles have been used to identify immune response signatures for a number of diseases, including autoimmune (272,273), cardiovascular (274,275) and infectious diseases (276–279); to assess response to vaccine and drug therapy (280–284); to distinguish between types of infections (285–287); and to differentiate between infection states (279,288). In addition, a few studies have extended blood transcriptomics to population-level studies, showing that heterogeneity in immune responses is due to environment and genetics (289–293). Furthermore, network analysis has been applied to blood transcriptomics to identify networks of highly co-expressed genes that underlie the immune response. These networks have been shown to replicate in cross-study comparisons (249,294). For instance, gene networks exhibiting unique transcriptional signatures have been associated with specific diseases (272) or different types of vaccines (295). Others have identified immune-related gene networks in healthy populations (249,294,296). Moreover, gene networks have also been constructed in B- and T-cells to gain insight into the biological processes involved with these immune cells (297,298). Additionally, differential co-expression analysis has been able to capture pathways that were perturbed or differentially co-expressed in diseased individuals (299).

1.6.3 The genetic architecture of gene expression levels

Initial studies in monozygotic twins, siblings, and family pedigrees provided evidence for heritability of gene expression (300,301). Since then, an increasing number of studies have followed to directly map genetic variants that influence gene expression, both across different tissues and in different populations (301–311).

Genomic regions that contain variable DNA sequences such as single nucleotide polymorphisms (SNPs) and copy number variants (CNVs) have been shown to affect the expression level of one or many genes. These are named expression quantitative trait loci (eQTLs). EQTLs have been identified to regulate the expression of genes that are either within close proximity (< 1 Mb) to the eQTL (*cis*-eQTL), or located distant (> 5 Mb) from the eQTL (*trans*-eQTL). Of the two types, *cis*-eQTLs are predominant and are enriched near transcription start sites, where they have a comparatively larger expression effect on adjacent genes (292). In an eQTL (SNP) analysis, the association between each variant-allele dosage is tested with gene expression abundance (312).

The localisation of most disease-associated loci to intronic regions suggests that they have regulatory roles; these loci may influence immune-related disease phenotype by altering gene expression levels, which might then be correlated with various functional or malfunctioning states of the immune system (164,313,314). Several studies within human populations have demonstrated that genetic variants associated with complex traits alter expression levels of nearby genes (300–308). Hence, dissecting the genetic architecture underlying gene expression changes through eQTL mapping provides a means to identify key genetic drivers of immunological processes and associated diseases.

Earlier eQTL (SNP) mapping studies used human biopsy samples from multiple tissue types – including liver (315), adipose (308), brain (316,317), and others (318) – to report tissue-specific regulation of gene expression. These and other studies have also reported cell-type specific eQTLs in blood (308,317,319). Some of the largest population-based eQTL mapping studies have been performed in whole blood (310,320–322). More recent population-based eQTL studies, which have focused on immune cell types suspected to be involved in some diseases, have also demonstrated eQTLs exerting cell-type specific regulation (289–292,318,323,324). A resource that has emerged recently is the tissue-specific eQTL database from the Genotype-Tissue Expression (GTEx) portal (311). As of now, this database contains eQTLs detected in 43 tissues from 175 postmortem donors (311), which can be queried by other eQTL studies.

Epigenetic modifications of the DNA, which are a heritable change in the chromatin structure rather than the actual DNA sequence, is another mechanism that regulates gene expression. Epigenetic regulation of gene activity, during development, differentiation or in response to environmental cues, occurs by altering the accessibility of the transcription machinery and other DNA binding proteins to specific DNA regions (325–327). Several epigenetic mechanisms including DNA methylation, posttranslational modifications (PTMs) of histone proteins, small and non-coding RNA's, and remodelling of chromatin structure play a part in regulating gene expression (328).

1.7 Genetic architecture of cytokines levels

Early genetic studies identified numerous SNPs and a handful of microsatellite polymorphisms located mainly within regulatory regions of cytokine genes. Most of these studies showed that these variants influenced cytokine gene expression *in vitro,* and were clinically associated with a number of diseases (329,330). Hence, it was hypothesised that these cytokine polymorphisms may also regulate cytokine levels as well. Support for this hypothesis came from an *in vitro* study by Pravica *et al*., where they showed that a SNP within the first intron of *IFNG* gene was correlated with its production (331). In 2005, Craen *et al.* demonstrated evidence for heritability of cytokine levels by comparing variation among monozygotic twins, dizygotic twins, and siblings (332). They found that more than half of the variance in the five cytokines they assessed was due to genetics (332). Since then, a few population-based studies have shown that cytokine polymorphisms are associated with cytokine levels in diseased individuals. For instance, gene polymorphisms in *IFNG, IL-12B, TNF, IL-17A, IL-10,* and *TGFB1* influenced the differential production of these cytokines in tuberculosis patients undergoing treatment (333). Several population-based studies have also linked cytokine gene polymorphisms to diseases such as type 2 diabetes (334), *Helicobacter pylori* infection (335), coronary heart disease (336), cancer (337), and rheumatoid arthritis (338). This suggests that these polymorphisms may also directly contribute towards the development of these diseases. So far, all studies discussed above were limited to polymorphisms located within and proximate to the cytokine genes themselves. EQTL studies have further linked disease-associated SNPs with the expression of cytokine genes and their receptors (339). In addition, several population-based genome-wide eQTL studies have also identified both *cis*- and *trans*- eQTLs for cytokines genes (310,320,322).

A few GWA studies have identified multiple loci associated with cytokine levels (340–343). However these GWASs have only focussed on either one or just a couple of cytokines. Recently, Ahola-Olli *et al*. performed one of the largest GWAS studies for circulating concentrations of 41 cytokines profiled in more than 8,000 individuals. The authors identified 27 loci to be associated with one or more cytokines, which also harboured eQTLs for cytokine genes (344).

1.8 Metabolomics

Metabolites are end-products of a cellular process, and their global measurements can give a close-up picture of the physiological or pathological state of an individual at a specific point in time (345–347). This image can be reflective of endogenous or exogenous (environmental) influences. Additionally, the highly metabolically active microbial community in the human gut can influence the host metabolome (348).

Metabolomics utilises high throughput analytical technology to identify and quantify small molecules such as amino acids, lipids, lipoproteins, carbohydrates, and fatty acids, collectively referred to as metabolites, in biological samples (349). These biological samples mainly include urine, tissues, serum or plasma (350).

1.8.1 Metabolite profiling

The two main technologies for metabolite profiling are nuclear magnetic resonance (NMR) spectroscopy and mass spectroscopy (MS). NMR is by far the most widely-used since it is non-destructive (the same sample can be reused for different analyses); requires minimum sample preparation, and allows lipoprotein subfraction measurements (351). In NMR, the nuclei (neutrons and protons) within a sample are exposed to a magnetic field and excited by a frequency pulse, whereby the motion of magnetic moments in these nuclei results in an NMR spectra for that sample (352). The area under each distinct peak representing a particular metabolite corresponds to its concentration (352,353). MS can quantify thousands of proteins for even low concentration samples. The ionised molecules within a sample are separated according their mass-to-charge ratio and then passed through a detector to measure their abundance (354), which are also displayed as a spectrum. Unlike NMR, MS is far more sensitive to low concentrations (at a micromolar range), and MS can measure up to thousands of different metabolites (355). MS also usually requires an additional separation step before detection by chromatography such as gas chromatography (GC) or high-performance liquid chromatography (356).

1.8.2 Metabolomics to assess the immune system

Metabolomics has gained a lot of interest in medical research, mainly because metabolites reflect the end products of gene activity, transcription, and protein metabolism, and small changes in either of these processes can substantially alter the metabolic fingerprints within a biological system. Metabolomics has been used to identify biomarkers in diseases including type 2 diabetes (357), gestational diabetes (358), and tuberculosis (359). Additionally, metabolite signatures have also been used to predict cardiovascular and metabolic disease risk (360–362), all-cause mortality (363), and death in septic patients (364); assess treatment response (365); differentiate between tumour subtypes (366); and profile host metabolic response to Hepatitis C virus infection (367). Altered metabolite profiles have been associated with obesity status (368), body weight change (369), menopause status (370), age (370), hormonal contraceptive use (371), and insulin resistance (372), which can all contribute to cardiometabolic disease risk. Recently, the trajectory has shifted towards understanding the interplay between metabolic reprogramming and immune function, termed "immunometabolism". The metabolic requirements of immune cells vary according to cell type, stage and function, which support their survival and proliferation (373). It is becoming evident that the metabolic machinery operating within immune cells is important in regulating their cell fate, function, and ultimately shaping an immune response (374). Recently, studies have begun exploring immunometabolism in human blood by integrating matched transcriptomic and metabolomic data (294,296,355,369,375–378). Networks of correlated genes involved in immune function have been found to be associated with blood lipid and serum metabolite levels (296,355,375,378). Immunometabolism offers an additional dimension to understanding the central role of the immune system in health and disease.

1.9 Research objectives

The overall functionality of the immune system relies on the coordinated interactions between sub-systems, and substantial interplay and regulation at systems level. A systems-wide analysis is required to understand the mechanistic underpinnings of the interactions among multiple levels contributing towards the net behaviour of the

immune system, and how it works across other biological systems. High-throughput profiling technologies have allowed measurements of various immunological parameters at the genome-wide level, which captures multi-level information. The application of integrative approaches to these large-scale data is necessary to delineate the complexity and gain insights that are not possible otherwise.

The overall objective of this thesis was to apply integrative bioinformatics methods to multi-omics data obtained from humans and immune cells to understand immune function and regulation. Here, four different aspects of the immune system were explored. First, this thesis globally explored the immune system and its relationship with metabolism in human blood. Next, it focused on the regulation of the immune system by exploring the genetic architecture underlying circulating cytokines in human populations. Then, a specific component of the immune system, immunological memory, was studied with emphasis on TRM cells. Finally, the role of one particular signalling molecule, TGF-β, in influencing tissue residency of TRM cells was assessed.

The specific objectives of this thesis were:

1. To create a robust interaction map of circulating metabolites, immune gene networks, and their genetic regulation in a population-based study.

2. To identify and characterise genetic variants regulating a network of circulating cytokines in a population-based study.

3. Characterise gene networks underlying the transcriptional signature regulating the development and establishment of tissue resident memory T-cells.

4. To perform RNA-seq analysis to explore the role of TGF-β, an extrinsic tissue-derived factor, in influencing the transcriptional signature of TRM cell.

Chapter 2

An interaction map of circulating metabolites, immune gene networks and their genetic regulation

2.1 Introduction

The focus of this chapter was to explore the relationship between two fundamental biological systems at a molecular level, immune function and metabolism, in natural human populations. The aims of this chapter have been addressed in an article published in *Genome Biology* (378), which contains materials presented in this chapter (see Preface). A comprehensive catalogue of diverse metabolite interactions across a spectrum of immune-related processes and their genetic regulation may provide insights into how metabolism is linked to pathogen sensing and immune response.

Investigating the interplay between the immune system and metabolism, coined as immunometabolism, is an emerging area of research. Until recently, these two systems were regarded as separate processes that occur within an organism. Here the role of the immune system was to protect the host against external (e.g. microbes and viruses) and internal threats. While, metabolism was regarded a set to biochemical processes that facilitated the energy requirements for cellular process, which also included cells involved in immune function. However, in the recent years, through immunometabolic studies, it has become increasingly clear that the relationship between metabolism and immunity is more than just energy supply. Metabolic preprograming within immune cells plays a key role in shaping an immune response (373,374) From a health perspective, interest in this area grew with the realisation that both low-grade inflammation and metabolic dysfunction jointly contribute towards metabolic disorders

such as obesity, type 2 diabetes, and cardiovascular diseases (379). Despite considerable progress in the field, a clear understanding of key interactions of immunometabolism in population-based studies is lacking. To identify these key interactions, an actual map of the immune-metabolite interactions needs to be created. Understanding the complex interactions between metabolism and immune function will provide insight into the potential pathogenic mechanisms underlying cardiometabolic diseases and offer ways to manipulate metabolism, which can help boost or suppress immunity. In addition, investigating the genetic basis of inter-individual difference in immune function and how this might remodel the immune-metabolic crosstalk, may explain the differential disease susceptibility in individuals. Hence, assessing immunometabolism against a genetic background will provide an additional dimension to our knowledge of the immune system and its role in health and disease.

2.1.1 Role of immunometabolism in cardiometabolic diseases

Immunometabolism has a key role in both type 2 diabetes (T2D) and atherosclerosis. In T2D, it is generally appreciated that immune overactivation in adipose tissue is a key driver (380,381). Studies have shown that macrophage infiltration and subsequent overexpression of proinflammatory cytokines such as TNF in adipose tissues is associated with insulin resistance (380–382). Moreover, evidence for metabolic inflammation has also been shown in other tissues where, in blood, elevated glucose and free fatty acid levels potentiate IL-1B mediated destruction of pancreatic ß-cells and subsequent T2D progression (383–385). Lipid-induced inflammatory response mechanisms have also been implicated in atherosclerosis and myocardial infarction (386). For example, oxidised phospholipids in an atherogenic lesion lead to a new macrophage phenotype, which promotes inflammasome activation (387) and proinflammatory cytokine secretion (388) .

2.1.2 Immunometabolism in population-based studies

Despite its role in pathogenesis, few large-scale human studies have assessed the systems-level interactions between the immune system and metabolites by systematically integrating matched transcriptomic and metabolomics data. Early

studies have utilised previously collected blood samples from population-based cohorts (294,296,355,369,375). Blood is an easily accessible tissue, which serves as a channel through which cells of both the innate and adaptive arms of the immune system perform their function in proximity to diverse circulating metabolites, making it an ideal tissue to study their interactions. As discussed in **Chapter 1**, blood transcriptomic profiling has been widely used to capture the overall status of the immune system.

Inouye *et al*. were the first to show this link, where they identified an immune related (mast cell/basophil activity) gene co-expression network constructed from whole blood transcriptome, the lipid leukocyte (LL) module, to be associated blood lipid and serum metabolite levels (296,375). Consistent findings were also reported by Wahl *et al*., where authors correlated the LL module with both, a metabolite network consisting of mainly VLDL subclass of lipoprotein and triglycerides, and change in body weight (369). In another study, Bartel *et al*. constructed an integrated correlation network combining blood transcriptome and serum metabolites and were also able to capture pathway level cross-talk between metabolic pathways and immune processes (355). Recently, Ritchie *et al*. showed that a gene co-expression network enriched for neutrophil function was associated with GlycA, a biomarker predictive of cardiovascular disease and all-cause mortality (294). The findings of these above studies suggest that an intimate link exists between immune response pathways and circulating metabolites.

2.1.3 Existing gap in understanding the immune-metabolite interactions in population-based studies

However, the studies discussed above had modest sample sizes (several hundred) and did not fully explore the interplay between the diverse range of immune processes and metabolites. Furthermore, while these integrative studies have investigated the underlying biology of the immunometabolic interactions, eQTL mapping of immune networks promises to provide insight into how genetic variation may further affect these interactions as well as relate to diseases. Despite substantial progress of these early studies, a robust systems-level map of immunometabolic relationships and their genetic

regulation in a natural human population is still incomplete. A detailed map may further guide in inferring which these connections get rewired in diseases.

2.2 Research objectives

The central aim of this chapter was to create a robust integrated map of immunometabolic relationships and their genetic regulation in human blood. The specific objectives of this chapter have been addressed in an article published in *Genome Biology* (378).

The specific objectives of this research chapter were:

1. To perform gene co-expression network analysis for network discovery and cross-cohort topological replication to identify robust gene modules enriched for immune-related functions.

2. To perform association analysis to identify metabolites that show significant associations with each of the immune-related gene modules.

3. To perform genome-wide scans to identify QTLs, both *cis* and *trans*, influencing the overall expression of immune-related modules.

4. To investigate the time-varying affects on metabolite and genetic associations of immune co-expression networks over a 7-year follow-up period.

2.3 Methods

2.3.1 Study populations

An overview of the study populations, molecular data, and study design is given in **Figure 2.1**. This study used data from two population-based cohorts, the Dietary, Lifestyle, and Genetic determinant of Obesity and Metabolic syndrome (DILGOM; N=518) and the Cardiovascular Risk in Young Finns Study, (YFS; N=1,650), which have been described in detail elsewhere (296,389). All subjects enrolled in these studies gave written informed consent.

The DILGOM study is a cross-sectional population-based survey conducted in 2007, which randomly recruited 5,325 unrelated individuals aged between 25–74 years of age from the Helsinki region of Finland, 630 of whom underwent at least one of the genotyping, transcriptomics or metabolomics profiling considered here. Ethics approval was given by the Coordinating Ethical Committee of the Helsinki and Uusimaa Hospital District. In 2014, a follow-up study was conducted, for which 1,273 individuals from the original study re-participated. Samples collected in 2007 and 2014 are referred to as DILGOM07 and DILGOM14, respectively.

The YFS is a longitudinal prospective cohort study that started in 1980, with follow-up studies carried out every three years, to monitor cardiovascular disease risk factors in children and adolescents from the five major regions of Finland (Helsinki, Kuopio, Turku, Oulu, and Tampere). In the baseline study a total of 3,596 children and adolescents in age groups 3, 6, 9, 12, 15, and 18 years participated, who were randomly selected from the national public register, details of which are described in (389). In this current study, data collected from the 2011 follow-up study (participants aged 34, 37, 40, 43, 46, and 49 years) were analysed. Ethics approval for the study research protocols was given by the Joint Commission on Ethics of Turku University and Turku University Central Hospital.

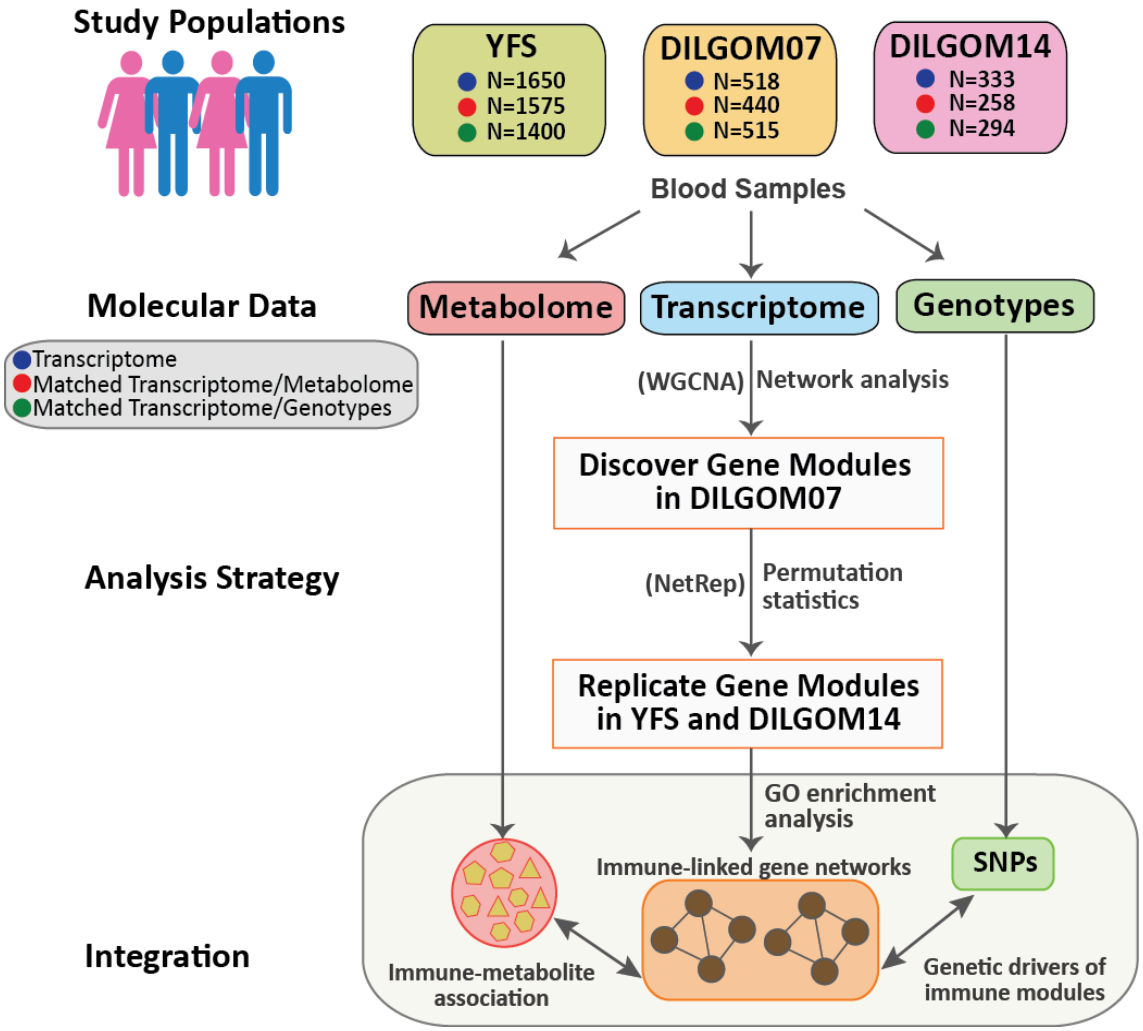


Figure 2.1: Overview of the study design.

For DILGOM, samples collected in 2007 and 2014 are referred to as DILGOM07 and DILGOM14, respectively. GO refers to Gene Ontology. Gene co-expression networks were constructed using the weighted gene co-expression network analysis (WGCNA) tool. Network topologies were assessed using the NetRep tool.

2.3.2 Sample collection

Venous blood was collected following an overnight fast in all three studies. Samples were centrifuged, the resulting plasma and serum samples were aliquoted into separate tubes and stored at −70°C for analyses. Protocols for the blood sampling, physiological measurements, and clinical survey questions were similar across the YFS and DILGOM studies, and are described extensively in (296,390).

2.3.3 Genotyping and imputation

Whole blood genomic DNA obtained from both cohorts was genotyped using the Illumina 610-Quad SNP array for DILGOM07 (N=555) (296) and a custom generated 670K Illumina BeadChip array for YFS (N=2,443) (391). The 670K array shares 562,643 SNPS with the 610-quad array. The 670K array removes poorly performing SNPs from the 610-quad array and improves copy number variation coverage (391). Genotype calling was performed with the Illuminus clustering algorithm (392). Quality control was performed as previously described in (296) and (391) for DILGOM and YFS, respectively. Genotypes were imputed to the 1000 Genomes Phase 1 version 3 reference panel using IMPUTE2 in both DILGOM and YFS (393). Poorly imputed SNPs based on low call-rate (< 0.90 for DILGOM, < 0.95 for YFS), low-information score (< 0.4), minor allele frequency < 1%, and deviation from Hardy-Weinberg equilibrium ($P < 5 \times 10^{-6}$) were then removed. A total of 7,263,701 SNPs in DILGOM and 6,721,082 in YFS passed quality control, with 6,485,973 common between the two. A total of N=518 samples in DILGOM and N=2,443 samples in YFS individuals passed quality control filters.

2.3.4 Metabolomics profiling

Metabolite concentrations for DILGOM07 (N=4,816), DILGOM14 (N=1,273), and YFS (N=2,046) were quantified from serum samples utilizing a high-throughput ^{1}H-NMR metabolomics platform (349,351). Details of the experimental protocol including sample preparation, NMR spectroscopy and metabolite identification has been previously described in (296,349). A total of 158 metabolite measures were assessed, of which 148 were directly measured and 10 were derived (**Table 2.1**). The 148 measures

include the constituents of 14 lipoprotein subclasses (98 measurements total), sizes of 3 lipoprotein particle, 2 apolipoproteins, 8 fatty acids, 8 glycerides and phospholipids, 9 cholesterols, 9 amino acids, 1 inflammatory marker, and 10 small molecules (involved in glycolysis, critic acid cycle and urea cycle). The lipoprotein subclasses are classified according to size (**Table 2.1**). Measurements with very low concentration, set as zero by the NMR pipeline, were set to the minimum value of that particular metabolite. Measurements rejected by automatic quality control or with detected irregularities were treated as missing. Undefined derived ratios arising from measurements with very low concentration (i.e. zero) were also treated as missing. Measurements were log2 transformed to approximate a normal distribution.

C-reactive protein (CRP), an inflammatory marker, was quantified from serum using a high sensitivity latex turbidimetric immunoassay kit (CRP-UL assay, Wako Chemicals, Neuss, Germany) and an automated analyser (Olympus AU400) in DILGOM07 (N=5000), DILGOM14 (N=1308), and YFS (N=2046). CRP levels were log2 transformed.

2.3.5 Gene expression, processing and normalisation

Transcriptome-wide gene expression levels were quantified by microarrays from peripheral whole blood using similar protocols in all three cohorts, and have been previously described for DILGOM07 (296) and YFS (394). Stabilised total RNA was obtained from whole blood using a PAXgene Blood RNA System and the protocols recommended by the manufacturer. In DILGOM07, RNA integrity and quantity was evaluated using an Agilent 2100 Bioanalyzer. In YFS, RNA integrity and quantity were evaluated spectrophotometrically using an Eppendorf BioPhotomer and the RNA isolation process was validated using an Agilent RNA 6000 Nano Chip Kit. RNA was hybridised to Illumina HT-12 version 3 BeadChip arrays in DILGOM07 and to Illumina HT-12 version 4 BeadChip arrays in DILGOM14 and YFS.

Table 2.1: List of 159 NMR based metabolites analysed in this study.

Metabolite	Description	Units
Amino acids		
ALA	Alanine	mmol/L
GLN	Glutamine	mmol/L
GLY	Glycine	mmol/L
HIS	Histidine	mmol/L
ILE	Isoleucine	mmol/L
LEU	Leucine	mmol/L
PHE	Phenylalanine	mmol/L
TYR	Tyrosine	mmol/L
VAL	Valine	mmol/L
Small molecules and energy metabolism related metabolites		
ACACE	Acetoacetate	mmol/L
ACE	Acetate	mmol/L
ALB	Albumin	signal area
BOHBUT	3-hydroxybutyrate	mmol/L
CIT	Citrate	mmol/L
GLC	Glucose	mmol/L
LAC	Lactate	mmol/L
PYR	Pyruvate	mmol/L
GLOL	Glycerol	mmol/L
Fatty acids and fatty acid ratios (relative to total fatty acids)		
TOT_FA	Total fatty acids	mmol/L
UNSAT*	Estimated degree of unsaturation	mmol/L
DHA	22:6, docosahexaenoic acid	mmol/L
LA	18:2, linoleic acid	mmol/L
FAW3	Omega-3 fatty acids	mmol/L
FAW6	Omega-6 fatty acids	mmol/L
PUFA	Polyunsaturated fatty acids	mmol/L
MUFA	Monounsaturated fatty acids; 16:1, 18:1	mmol/L
SFA	Saturated fatty acids	mmol/L
DHA/FA*	Ratio of 22:6 docosahexaenoic acid to total fatty acids	%
LA/FA*	Ratio of 18:2 linoleic acid to total fatty acids	%
FAW3/FA*	Ratio of omega-3 fatty acids to total fatty acids	%
FAW6/FA*	Ratio of omega-6 fatty acids to total fatty acids	%
PUFA/FA*	Ratio of polyunsaturated fatty acids to total fatty acids	%
MUFA/FA*	Ratio of monounsaturated fatty acids to total fatty acids	%
SFA/FA*	Ratio of saturated fatty acids to total fatty acids	%
Cholesterol		
SERUM_C	Serum total cholesterol	mmol/L
EST_C	Esterified cholesterol	mmol/L
FREE_C	Free cholesterol	mmol/L
REMNANT_C	Serum total cholesterol (Non-HDL, non-LDL cholesterol)	mmol/L
HDL_C	Total cholesterol HDL	mmol/L
LDL_C	Total cholesterol LDL	mmol/L
VLDL_C	Total cholesterol VLDL	mmol/L
HDL2_C	Total cholesterol in HDL2	mmol/L
HDL3_C	Total cholesterol in HDL3	mmol/L

Glycerides and phospholipids		
SERUM_TG	Serum total triglycerides (mmol/l)	mmol/L
HDL_TG	Triglycerides in HDL (mmol/l)	mmol/L
LDL_TG	Triglycerides in LDL (mmol/l)	mmol/L
VLDL_TG	Triglycerides in VLDL (mmol/l)	mmol/L
TOT_PG	Total phosphoglycerides (mmol/l)	mmol/L
TG/PG*	Ratio of triglycerides to phosphoglycerides	ratio
PC	Phosphatidylcholine and other cholines (mmol/l)	mmol/L
SM	Sphingomyelins (mmol/l)	mmol/L
TOT_CHO	Total cholines (mmol/l)	mmol/L
Apolipoproteins		
APOA1	Apolipoprotein A-I	g/L
APOB	Apolipoprotein B	g/L
APOB/APOA1*	Ratio of apolipoprotein B to apolipoprotein A-I	ratio
Lipoprotein particle size		
VLDL_D	Mean diameter for VLDL particles	nm
LDL_D	Mean diameter for LDL particles	nm
HDL_D	Mean diameter for HDL particles	nm
Lipoprotein subclasses and their constituents		
Chylomicrons and extremely large VLDL particles (average particle diameter at least 75.0 nm)		
XXL_VLDL_P	Concentration of chylomicrons and extremely large VLDL particles	(mol/L)
XXL_VLDL_L	Total lipids in chylomicrons and extremely large VLDL	mmol/L
XXL_VLDL_PL	Phospholipids in chylomicrons and extremely large VLDL	mmol/L
XXL_VLDL_C	Total cholesterol in chylomicrons and extremely large VLDL	mmol/L
XXL_VLDL_CE	Cholesterol esters in chylomicrons and extremely large VLDL	mmol/L
XXL_VLDL_FC	Free cholesterol in chylomicrons and extremely large VLDL	mmol/L
XXL_VLDL_TG	Triglycerides in chylomicrons and extremely large VLDL	mmol/L
Very large VLDL particles (average particle of 64.0 nm)		
XL_VLDL_P	Concentration of very large VLDL particles	(mol/L)
XL_VLDL_L	Total lipids in very large VLDL	mmol/L
XL_VLDL_PL	Phospholipids in very large VLDL	mmol/L
XL_VLDL_C	Total cholesterol in very large VLDL	mmol/L
XL_VLDL_CE	Cholesterol esters in very large VLDL	mmol/L
XL_VLDL_FC	Free cholesterol in very large VLDL	mmol/L
XL_VLDL_TG	Triglycerides in very large VLDL	mmol/L
Large VLDL particles (average particle diameter of 53.6 nm)		
L_VLDL_P	Concentration of large VLDL particles	(mol/L)
L_VLDL_L	Total lipids in large VLDL	mmol/L
L_VLDL_PL	Phospholipids in large VLDL	mmol/L
L_VLDL_C	Total cholesterol in large VLDL	mmol/L
L_VLDL_CE	Cholesterol esters in large VLDL	mmol/L
L_VLDL_FC	Free cholesterol in large VLDL	mmol/L
L_VLDL_TG	Triglycerides in large VLDL	mmol/L
Medium VLDL particles (average particle diameter of 44.5 nm)		
M_VLDL_P	Concentration of medium VLDL particles	(mol/L)
M_VLDL_L	Total lipids in medium VLDL	mmol/L
M_VLDL_PL	Phospholipids in medium VLDL	mmol/L
M_VLDL_C	Total cholesterol in medium VLDL	mmol/L

M_VLDL_CE	Cholesterol esters in medium VLDL	mmol/L
M_VLDL_FC	Free cholesterol in medium VLDL	mmol/L
M_VLDL_TG	Triglycerides in medium VLDL	mmol/L
Small VLDL particles (average particle diameter of 36.8 nm)		
S_VLDL_P	Concentration of small VLDL particles	(mol/L)
S_VLDL_L	Total lipids in small VLDL	mmol/L
S_VLDL_PL	Phospholipids in small VLDL	mmol/L
S_VLDL_C	Total cholesterol in small VLDL	mmol/L
S_VLDL_CE	Cholesterol esters in small VLDL	mmol/L
S_VLDL_FC	Free cholesterol in small VLDL	mmol/L
S_VLDL_TG	Triglycerides in small VLDL	mmol/L
Very small VLDL particles (average particle diameter of 31.3 nm)		
XS_VLDL_P	Concentration of very small VLDL particles	mol/L
XS_VLDL_L	Total lipids in very small VLDL	mmol/L
XS_VLDL_PL	Phospholipids in very small VLDL	mmol/L
XS_VLDL_C	Total cholesterol in very small VLDL	mmol/L
XS_VLDL_CE	Cholesterol esters in very small VLDL	mmol/L
XS_VLDL_FC	Free cholesterol in very small VLDL	mmol/L
XS_VLDL_TG	Triglycerides in very small VLDL	mmol/L
Intermediate density lipoprotein (IDL) particles (average particle diameter of 28.6 nm)		
IDL_P	Concentration of IDL particles	mol/L
IDL_L	Total lipids in IDL	mmol/L
IDL_PL	Phospholipids in IDL	mmol/L
IDL_C	Total cholesterol in IDL	mmol/L
IDL_CE	Cholesterol esters in IDL	mmol/L
IDL_FC	Free cholesterol in IDL	mmol/L
IDL_TG	Triglycerides in IDL	mmol/L
Large LDL particles (average particle diameter of 25.5 nm)		
L_LDL_P	Concentration of large LDL particles	mol/L
L_LDL_L	Total lipids in large LDL	mmol/L
L_LDL_PL	Phospholipids in large LDL	mmol/L
L_LDL_C	Total cholesterol in large LDL	mmol/L
L_LDL_CE	Cholesterol esters in large LDL	mmol/L
L_LDL_FC	Free cholesterol in large LDL	mmol/L
L_LDL_TG	Triglycerides in large LDL	mmol/L
Medium LDL particles (average particle diameter of 23.0 nm)		
M_LDL_P	Concentration of medium LDL particles	mol/L
M_LDL_L	Total lipids in medium LDL	mmol/L
M_LDL_PL	Phospholipids in medium LDL	mmol/L
M_LDL_C	Total cholesterol in medium LDL	mmol/L
M_LDL_CE	Cholesterol esters in medium LDL	mmol/L
M_LDL_FC	Free cholesterol in medium LDL	mmol/L
M_LDL_TG	Triglycerides in medium LDL	mmol/L
Very large HDL particles (average particle diameter of 14.3 nm)		
XL_HDL_P	Concentration of very large HDL particles (mol/l)	mol/L
XL_HDL_L	Total lipids in very large HDL (mmol/l)	mmol/L
XL_HDL_PL	Phospholipids in very large HDL	mmol/L
XL_HDL_C	Total cholesterol in very large HDL	mmol/L
XL_HDL_CE	Cholesterol esters in very large HDL	mmol/L

XL_HDL_FC	Free cholesterol in very large HDL	mmol/L
XL_HDL_TG	Triglycerides in very large HDL	mmol/L
Large HDL particles (average particle diameter of 12.1 nm)		
L_HDL_P	Concentration of large HDL particles	mol/L
L_HDL_L	Total lipids in large HDL	mmol/L
L_HDL_PL	Phospholipids in large HDL	mmol/L
L_HDL_C	Total cholesterol in large HDL	mmol/L
L_HDL_CE	Cholesterol esters in large HDL	mmol/L
L_HDL_FC	Free cholesterol in large HDL	mmol/L
L_HDL_TG	Triglycerides in large HDL	mmol/L
Medium HDL particles (average particle diameter of 10.9 nm)		
M_HDL_P	Concentration of medium HDL particles	mol/L
M_HDL_L	Total lipids in medium HDL	mmol/L
M_HDL_PL	Phospholipids in medium HDL	mmol/L
M_HDL_C	Total cholesterol in medium HDL	mmol/L
M_HDL_CE	Cholesterol esters in medium HDL	mmol/L
M_HDL_FC	Free cholesterol in medium HDL (mmol/l)	mmol/L
M_HDL_TG	Triglycerides in medium HDL (mmol/l)	mmol/L
Small HDL particles (average particle diameter of 8.7 nm)		
S_HDL_P	Concentration of small HDL particles	mol/L
S_HDL_L	Total lipids in small HDL	mmol/L
S_HDL_PL	Phospholipids in small HDL	mmol/L
S_HDL_C	Total cholesterol in small HDL	mmol/L
S_HDL_CE	Cholesterol esters in small HDL	mmol/L
S_HDL_FC	Free cholesterol in small HDL	mmol/L
S_HDL_TG	Triglycerides in small HDL	mmol/L
Inflammation		
GlycA	Glycoprotein acetyls, mainly alpha-1-acid glycoprotein	mmol/L

* Derived metabolites (N=10)

For DILGOM07, data was preprocessed as described in Inouye *et al.* (296). Briefly, for each array the background corrected probes were subjected to quantile normalisation at the strip-level. Technical replicates were combined by bead count weighted average and replicates with Pearson correlation coefficient < 0.94 or Spearman's rank correlation coefficient < 0.60 were removed. Expression values for each probe were then log2 transformed. For YFS, background corrected probes were subjected to quantile normalisation followed by log2 transformation. For DILGOM14, probes matching to the erythrocyte globin components (N=4) and those that hybridized to multiple locations spanning more than 10Kb (N=507) were removed. Probes with average bead intensity of 0 were treated as missing. The average bead intensity was then log2 transformed and quantile normalised. A total of 35,425 (for DILGOM07), 36,640 (for DILGOM14) and 37,115 (for YFS) probes passed quality control. In order to preserve information on alternative exon usage, which is captured by multiple probes targeting different exons for a given gene, signals from multiple probes were not summarised. All downstream analyses were done at probe level.

2.3.6 Gene coexpression network analysis and replication

Gene co-expression network modules were identified in DILGOM07 (N=518 individuals with gene expression data) as previously described (294) using WGCNA version 1.47 (239,395) on all probes passing quality control. Briefly, probe co-expression was calculated as the Spearman correlation coefficient between each pair of probes after adjusting for the effects of age and sex. A linear model was fit for each probe on age and sex, and the resulting residuals were taken as the adjusted probe expression. The weighted interaction network was calculated as the element-wise absolute co-expression exponentiated to the power 5. This power was selected through the scale-free topology criterion (239), which acts as a penalization procedure to enhance differentiation of signal from noise. Probes were subsequently clustered hierarchically (average linkage method) by topological overlap dissimilarity (239) and modules were detected through dynamic tree cut of the resulting dendrogram with default parameters and a minimum module size of 10 probes (241). Similar modules were merged together in an iterative process in which modules whose eigengenes clustered together below a height of 0.2 were joined. Module eigengenes, representative

summary expression profiles, were calculated as the first eigenvector from a principal components analysis of each module's expression data.

Module reproducibility and longitudinal stability were assessed in YFS (N=1,650 with gene expression data) and DILGOM14 (N=333 with gene expression data) respectively using the NetRep R package version 0.30.1 (251). Briefly, a permutation test (20,000 permutations) of seven module preservation statistics was performed for each module in YFS and DILGOM14 separately. These statistics test the distinguishability and similarity of network features (density and connectivity) for each module in a second dataset (252). Modules were considered reproducible where permutation *P*-values for all seven statistics were < 0.001 (Bonferroni correcting for 40 modules) in YFS, and modules were considered longitudinally stable where *P*-values were < 0.001 for all seven statistics in DILGOM14. Probe co-expression in YFS was calculated as the Spearman correlation coefficient between age and sex adjusted expression levels and the weighted interaction network was calculated as the element-wise absolute co-expression exponentiated to the power 4 as previously described (294). Probe co-expression in DILGOM14 was calculated as the Spearman correlation coefficient between each pair of probes, and the weighted interaction network defined as the element-wise absolute co-expression exponentiated to the power 5.

To filter out genes spuriously clustered into each module by WGCNA we performed a two-sided permutation test on module membership (Pearson correlation between probe expression and the module eigengene) for each reproducible module in DILGOM07 and YFS. Here, the null hypothesis was, for each module, that its probes did not truly coexpress with the module. The null distribution of module membership for each module was empirically generated by calculating the membership between all non-module genes and the module's eigengene. *P*-values for each probe were then calculated using the following permutation test *P*-value estimator (396):

$$p = \frac{b+1}{\mathrm{v}+1} - \int_0^{0.5/v_t+1} F(b;\mathrm{v},v_t)dv_t$$

Where b is taken as the number of non-module genes with a membership smaller or greater than the test gene's module membership, whichever number is smaller. v, the number of permutations calculated, and v_t, the total number of possible permutations,

are both the number of non-module genes. The resulting *P*-value was multiplied by 2 because the test was two-sided. To adjust for multiple testing, false discovery rate (FDR) correction was applied to the *P*-values separately for each module using the Benjamini and Hochberg method (213). We rejected the null hypothesis at FDR adjusted *P*-value < 0.05 in both DILGOM07 and YFS, deriving a subset of core probes for each module.

2.3.7 Functional annotation of immune-related gene modules

Immune modules were identified through over-representation analysis of Gene Ontology (GO) terms in the core gene set for each of the 20 reproducible modules using the web based tool GOrilla (397) with default parameters (performed March 2016). GOrilla was run on two unranked gene lists where core module genes were given as the target list and the background list was given as the 25,233 human RefSeq genes corresponding to any probe(s) passing quality control in both DILGOM07 and YFS. A hypergeometric test was calculated to test whether each module was significantly enriched for genes annotated for each GO term in the "Biological Process" ontology. A GO term was considered significantly over-represented in a module where it's FDR corrected *P*-value was < 0.05. FDR correction was applied in each module separately. Significant GO terms for each module were further summarised into a subset of representative GO terms with REVIGO (398) using the RELSIM semantic similarity measure and a similarity cut-off value $C = 0.5$ on genes from *Homo sapiens*. A module was considered to be immune-linked where the representative GO term list contained the parent GO term GO:0002376 (immune system process) and/or GO:0002682 (regulation of immune system processes).

To delineate the function of the genes in GIMA and GIMB, I further queried a more specialised database for innate immunity research, InnateDB (http://www.innatedb.com) (399). GO enrichment analysis for innate immunity was carried out using the Ensembl IDs of core genes present in GIMA and GIMB as an input in the InnateDB "gene ontology over-representation analysis (ORA)" tool. Gene ontology ORA analysis was performed using a hypergeometric test. The resulting enrichment *P*-values were adjusted using Benjamini-Hochberg (213) FDR correction,

and significance level was established at FDR < 0.05.

2.3.8 Statistical analyses

Reproducible module–metabolite associations were identified through linear regression of each immune module eigengene on each of the 159 metabolites in both DILGOM07 and YFS. Prior to analysis, metabolite data was first subsetted to individuals with matching gene expression profiles, followed by removal of subjects on cholesterol lowering drugs, for YFS (N=62) and DILGOM07 (N=74). Pregnant women in YFS (N=10) and DILGOM (N=2) were further removed from the analysis. A total of 440 individuals in DILGOM07 and 1,575 individuals in YFS had matched gene expression and metabolite data, excluding pregnant women and those individuals taking lipid-lowering medication. Models were adjusted for age, sex, and use of combined oral contraceptive pills. Module eigengenes and metabolite levels were scaled to standard deviation units. To maximize statistical power, a meta-analysis was performed on the DILGOM07 and YFS associations using the fixed-effects inverse variance method implemented in the "meta" R package downloaded from CRAN (https://cran.r-project.org/web/packages/meta/index.html). The meta-*P*-values for the 159 metabolite associations within each module were FDR corrected. An association was considered significant at FDR adjusted *P*-value < 6.25×10^{-3} (0.05/8 modules). This Bonferroni adjusted threshold was chosen to further adjust for the multiple modules being tested. To assess the potential confounding effects of blood cell type abundance on metabolite-module association, the model was rerun in YFS adjusting for leukocyte (for CCLM, VRM, BCM, NM, LLM, GIMA, GIMB) and platelet (for PM) counts available for this cohort. The beta values and *P*-values generated with and without adjusting for cell count were then compared. Additionally, to assess the possible effect of cell counts on expression profiles, cell counts were associated with module eigengenes.

Module–metabolite associations were tested for longitudinal stability in DILGOM14 using a linear regression model of each immune module eigengene on each of the 159 metabolites. A total of, 216 individuals in DILGOM had matched gene expression and metabolite data in both 2007 and 2014, after removing pregnant women and individuals on lipid lowering medication at either time points (N=70). Models were adjusted for age

and sex. Information on use of oral contraceptives was not available for this cohort. It is worth noting that > 60% of women were more than 50 years old, hence we would expect that rates of contraceptive use would be low and therefore not be a significant confounder. Module eigengenes and metabolite levels were scaled to standard deviation units. An association was considered longitudinally stable where the association was significant (FDR adjusted *P*-value < 6.25×10^{-03}) in both DILGOM14 and DILGOM07. For sensitively analysis, the model in DILGOM07 was run without adjusting oral contraceptive use and this did not affect the significant immune-metabolite associations maintained over the two time-points.

Module quantitative trait loci (mQTLs) were identified through genome-wide association scans with each immune module eigengene using the PLINK2 version 1.90 software (https://www.cog-genomics.org/plink2) (400) in DILGOM07 and YFS. A total of 518 individuals had matched gene expression and genotype data in DILGOM07 and 1400 individuals had matched gene expression and genotype data in YFS. Associations were tested using a linear regression model of each eigengene on the minor allele dosage (additive model) of each SNP. Models were adjusted for age, sex, and the first 10 genetic principal components (PCs). Genetic PCs were generated from a linkage-disequilibrium (LD) pruned set of approximately 200,000 SNPs using flashpca (168). *P*-values for each association in DILGOM07 and YFS were combined in a meta-analysis using the METAL software (401), which implements a sample size weighted Z-score method. A SNP was considered an mQTL if meta-analysis *P*-value (meta-*P*-value) was < 5×10^{-8}. Blood cell count data available for YFS was utilised to test the robustness of module associations with mQTLs, where the same model was run with and without adjusting for leukocyte and platelet cell counts.

Significant mQTLs were subsequently tested as expression quantitative trait loci (eQTLs) for genes within their respective modules using Matrix eQTL in both DILGOM07 and YFS (312). Of note, genome-wide *cis* and *trans* eQTLs analysis were not performed, but rather the *cis* and *trans* effects of significant mQTLs on individual genes within a respective module were tested, where *cis* was defined as an mQTL within 1Mb of a given probe and *trans* as greater than 5Mb from a given probe or on a different chromosome. Associations were tested using a linear regression model of probe expression on minor allele dosage (additive model) of the mQTL. Models were

adjusted for age, sex, and the first 10 genetic PCs. For *trans*-eQTL associations *P*-values in DILGOM07 and YFS were combined in a meta-analysis using the weighted Z-score method and considered significant where the meta-*P*-value < $5x10^{-8}$. For *cis*-eQTL associations where the meta-*P*-value < $5x10^{-8}$, permutation tests were further performed to test if the association was robust. For the permutation test, gene expression sample labels were shuffled 10,000 times to compute an empirical *P*-value. The permuted model *P*-values and nominal *P*-value were combined across DILGOM and YFS07 in a meta-analysis using the weighted Z-score method when computing the permutation test *P*-value. A mQTL was considered a *cis*-eQTL where the typically used permutation test *P*-value < 0.05.

2.4 Results and Discussion

2.4.1 Summary of cohorts and data

In this study genome-wide genotype, whole blood transcriptomic and serum metabolomics data from two population-based cohorts were analysed (**Figure 2.1**). Detailed description for each cohort, with regards to the number of individual with relevant omics data before and after filtering, is provided in **Table 2.2**.

Table 2.2: Covariate and data information for each cohort.

Characteristics	YFS	DILGOM07	DILGOM14
Collection year	2011	2007	2014
Covariate information			
Age range (years)	34-49	25-74	32-38
Pregnant women (N)	10	2	0
Individuals on lipid lowering drugs (N)	62	74	65
Women on oral contraceptives (N)	92	33	N/A
Data available for individuals in each cohort			
Metabolome (N)	2,046	4,816	1,273
Transcriptome (N)	1,650	518	333
Genotype (N)	2,443	518	518
C-reactive protein (N)	2,046	5,000	1,308
Number of Individuals (N) profiled with matched data after filtering			
Matched metabolome & transcriptome Total N (Male/female)	1,575 (709/866)	440 (191/249)	258 (155/168)
Matched genotype & transcriptome Total N (Male/female)	1400 (635/765)	515 (239/276)	294 (158/136)

N refers to the total number of individuals. N/A refers to data not available.

DILGOM and YFS genotyping were performed using Illumina Human 610 and 670 arrays, respectively, with subsequent genotype imputation, performed using IMPUTE2 (393) and the 1000 Genomes Phase I version 3 reference panel. For both cohorts, whole blood transcriptome profiling was performed using Illumina HT-12 arrays, and serum metabolomics profiling was carried out using the same ^{1}H-NMR platform (349). Individuals on lipid-lowering medication and pregnant women were excluded from the metabolome analyses. Of the 158 metabolites analysed, 148 were directly measured, and 10 derived (**Table 2.1**). After filtering, individuals with matched data in each cohort (**Table 2.2**) were utilised in subsequent association analyses discussed below.

2.4.2 Inference of robust immune gene co-expression networks in whole blood

In this study, first, networks of tightly coexpressing genes were identified in DILGOM07. From the 35,422 probes subjected to network analysis, a total of 40 modules of coexpressed genes were identified. Then, using NetRep (251), the preservation of the network topology for each of these 40 modules was tested in YFS. Of the 40 DILGOM07 modules, 20 were strongly preserved in YFS, which ranged in size from 14 – 4452 probes. A module was considered strongly preserved if the *P*-value was < 0.001 for all seven preservation statistics (Bonferroni correction for 40 modules) (**Table 2.3**) Next, for each of the 20 replicated modules, the core gene probes were defined, those which are most tightly coexpressed and thus robust to clustering parameters, using a permutation test of module membership. As expected, larger modules (> 1,000 probes) retained a smaller proportion of core gene probes (< 10% of the initial number of probes) as compared to the smaller, more tightly coexpressed modules (>80%) (Table A.1 in Appendix A).

It is also worth noting that 50% of the modules did not replicate between cohorts, possibly due to a number of factors. Firstly, microarray datasets contain systematic (or technical) noise that is introduced during sampling and data generation process. Adequately removing systematic noise from the data during the pre-processing and normalisation steps is challenging and may affect the statistical inference of gene co-expression networks (402). As a result, a proportion of the networks discovered will be

false positives, especially the larger more weakly connected ones, and fail to replicate in an independent dataset. Similarly, technical variation in the replication dataset can hinder the replication of the true networks discovered. Secondly, the study cohorts analysed have a number of differences, for example, a noticeable difference in age distribution. Moreover, cohort-specific environmental factors can and do have an influence on the measured transcript levels, which could have an impact on gene-gene relationships (403). For example, there were a couple of modules (Modules 20, 30, 31, and 36) that failed to replicate in YFS, but not in DILGOM14. Thirdly, a stringent filtering criteria (P-value < 0.001 for all seven preservation statistics) was used to identify modules that replicated across cohorts. All of the 21 modules except two, which failed to replicate, met the *P*-value threshold for at least three preservation statistics.

2.4.3 Identification and characterization of immune-related gene networks

To identify modules of putative immune function, analysis of GO terms "immune system processes" (GO: 0002376) and/or "regulation of immune system processes" (GO:0002682) were performed for the core genes of each replicated modules. Each immune module's gene content and putative biological function is summarised in **Table 2.4**.

Six out of the 20 modules were enriched for at least one of these two terms, of which two have been previously identified. This included a platelet module (PM) that substantially overlaps with a previously reported module for platelet aggregation activity (310) and the neutrophil module (NM) (294). In addition, I also identified another well characterised gene coexpression module, lipid-leukocyte module (LLM) (296), which has been related to mast cell and basophil function. Of note, this module was not significantly enriched for any GO terms owing to small module size. Since it is well appreciated that apart from their classical role in haemostasis and blood clotting, platelets also play a role in inflammation and immune response (47,404), I also characterized PM in subsequent analyses. Hence, in total, eight immune-related module were characterised for subsequent analyses.

Table 2.3: The seven module preservation statistics of gene networks (discovered in DILGOM07) in YFS.

Module	#Probes DILGOM	#Probes YFS	Mean Adj	PVE	Corr. Coexp	Corr. kIM	Corr. MM	Mean Coexp	Mean MM	Rep.
0	8,990	6,418	1	9.2×10^{-2}	5×10^{-5}	6.1×10^{-1}	5×10^{-5}	5×10^{-5}	5×10^{-5}	NO
1	8,680	6,766	3.7×10^{-3}	5×10^{-5}	5×10^{-5}	5×10^{-5}	1	5×10^{-5}	1	NO
2	5,403	4,452	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	YES
3	3,258	2,374	1	5×10^{-5}	5×10^{-5}	5×10^{-5}	9.8×10^{-1}	5×10^{-5}	3.3×10^{-1}	NO
4	1,775	1,666	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	YES
5	1,734	1,722	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	YES
GIMA	1,019	1,009	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	YES
7	604	574	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	YES
8	598	411	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	1.2×10^{-1}	5×10^{-5}	5.4×10^{-2}	NO
9	545	411	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	YES
GIMB	339	335	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	YES
11	265	188	9.8×10^{-1}	5.3×10^{-2}	5×10^{-5}	3.5×10^{-1}	1×10^{-3}	5×10^{-5}	5×10^{-5}	NO
12	255	255	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	YES
13	239	239	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	YES
14	225	225	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	YES
15	208	194	5.5×10^{-4}	3.8×10^{-2}	5×10^{-5}	1.3×10^{-2}	7×10^{-1}	5×10^{-5}	2.4×10^{-2}	NO
CCLM	179	177	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	YES
PM	138	138	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	YES
VRM	112	111	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	YES
19	84	80	2.4×10^{-3}	4.8×10^{-3}	5×10^{-5}	9×10^{-4}	1	5×10^{-5}	1	NO
20	80	79	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	1	5×10^{-5}	1	NO
21	77	57	4.5×10^{-4}	1.1×10^{-2}	5×10^{-5}	4.7×10^{-2}	5×10^{-5}	5×10^{-5}	5×10^{-5}	NO
BCM	67	67	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	YES
23	64	61	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	YES
24	63	37	9×10^{-1}	4.6×10^{-1}	9.7×10^{-1}	2.6×10^{-2}	9.8×10^{-1}	1.4×10^{-1}	3.4×10^{-2}	NO
25	43	42	5×10^{-5}	5×10^{-5}	5×10^{-5}	2.3×10^{-2}	1.1×10^{-3}	5×10^{-5}	5×10^{-5}	NO
26	40	40	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	YES
27	39	39	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	YES
28	34	29	2.5×10^{-2}	1.7×10^{-3}	1×10^{-4}	4.2×10^{-3}	6.3×10^{-3}	5×10^{-5}	5×10^{-5}	NO
NM	31	31	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	YES
30	31	31	5×10^{-5}	5×10^{-5}	5×10^{-5}	1.4×10^{-3}	1.6×10^{-3}	5×10^{-5}	5×10^{-5}	NO
31	31	31	5×10^{-5}	5×10^{-5}	5×10^{-5}	1×10^{-4}	7.8×10^{-3}	5×10^{-5}	5×10^{-5}	NO
32	30	30	5×10^{-5}	5×10^{-5}	5×10^{-5}	1.3×10^{-3}	5×10^{-5}	5×10^{-5}	5×10^{-5}	NO
33	28	28	5×10^{-5}	5×10^{-5}	5×10^{-5}	1×10^{-4}	5×10^{-5}	5×10^{-5}	5×10^{-5}	YES
34	25	25	5×10^{-5}	5×10^{-5}	9.2×10^{-3}	5.4×10^{-1}	4.5×10^{-1}	5×10^{-5}	5×10^{-5}	NO
35	20	20	5×10^{-5}	5×10^{-5}	5×10^{-5}	1.1×10^{-2}	5×10^{-5}	5×10^{-5}	5×10^{-5}	NO
36	19	18	5×10^{-5}	5×10^{-5}	5×10^{-5}	1.1×10^{-2}	9.1×10^{-3}	5×10^{-5}	5×10^{-5}	NO
37	18	18	9.3×10^{-1}	9.1×10^{-1}	3.1×10^{-1}	9.4×10^{-1}	6.7×10^{-1}	6.7×10^{-1}	8.1×10^{-1}	NO
LLM	15	14	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	YES
39	10	10	3.5×10^{-4}	5×10^{-5}	4×10^{-4}	2.6×10^{-2}	7.3×10^{-3}	5×10^{-5}	5×10^{-5}	NO
40	10	10	1×10^{-4}	5×10^{-5}	1.6×10^{-3}	5.3×10^{-1}	2.1×10^{-1}	5×10^{-5}	5×10^{-5}	NO

The seven statistics are as follows: (1) Mean Adjacency (**Mean Adj**), assesses how densely the genes are connected in a module across the datasets; (2) Proportion of variance explained by the module eigengenes (**PVE**); (3) Correlation of the module co-expression across the two datasets (**Corr.Coexp**); (4) Correlation of connectivity (**Corr.kIM**) assesses whether the most highly connected genes in a module are the same across the datasets; (5) Correlation of module membership (**Corr.MM**), which is the correlation between each module gene and the module eigengene, assesses whether the contribution of each gene to the summary expression of a module is same across the datasets; (6) mean co-expression (**Mean Coexp**); and (7) mean module membership (**Mean MM**) assesses whether the signs of the correlation are in the same direction in the two datasets. Module 0 is the background module. The immune-related modules are: GIMA; GIMB; CCLM; PM; VRM; BCM; NM; LLM. Rep. – refers to module replication.

Table 2.4: Immune module gene content and putative biological function based on GO terms (top three shown) and literature.

Module	Size	GO terms	Literature-based immune related function of genes
Cytotoxic cell-like module (CCLM)	130 (115)	Immune system process Defence response Immune response	Cytotoxic effectors (*GZMA, GZMB*, *GZMM*, *CTSW, PRF1* (405))*;* surface receptors (*IL2RB, SLAMF6, CD8A, CD8B, CD2, CD247, KLRD1*, *KLRG1* (405–407))*;* T and NK cell differentiation (*ID2* and *EOMES* (70,408)), activation (*ZAP70* and *CBLB* (409,410)), and recruitment (*CX3CR1*, *CCL5*, *CCL4L2* (411)).
Viral response module (VRM	95 (88)	Response to virus Type I interferon signalling pathway Response to biotic stimulus	Type I interferon-induced antiviral activity (*IFITM1*, *IFIT1*, *IFIT2*, *IFIT3*, *IFIT5*, *IFI44*, *IFI44L*, *IFI6*, *MX1*, *ISG15*, *ISG20*, *HERC5* (412,413)); viral RNA degradation (*OAS1*, *OAS2*, *OAS3*, *OASL*, *DDX60* (414)); type 1 interferon-signalling pathway (*IRF9*, *STAT1*, *STAT2* (415,416)).
B cell activity module (BCM)	54 (49)	Immune system process Immune response B cell activation	B cell surface markers (*CD79A*, *CD79B*, *CD22* (417,418)); B cell activation (*BANK1*, *BTLA*, *CD40*, *TNFRSF13B*, *TNFRSF13C* (419)), development (*POU2AF1*, *BCL11A*, *RASGRP3* (420)), migration (*CXCR5*, *CCR6* (420,421)), and their regulation (*CD83*, *FCER2*, *FCRL5* (422)); antigen presentation (*HLA-DOA, HLA-DOB* (423)).
*Platelet module (PM)	114 (106)	Coagulation Blood coagulation Cell activation	Platelet receptor signalling, activation, and coagulation (*GP6*, *GP9*, *ITGA2B*, *ITGB3*, *ITGB5*, *MGLL*, *MPL*, *MMRN1*, *PTK2*, *VCL*, *THBS1, F13A1*, *VWF*, (424)); regulating platelet activity (*SEPT5*, *TSPAN9* (425,426)).

*Neutrophil module (NM)	26 (26)	Killing of cells of other organism Cell killing Response to fungus	Anti -microbial, -fungal, and -viral activity (*DEFA1, DEFA1B, DEFA3, DEFA4, ELANE, BPI, RNASE2, RNASE3* (427–430)); neutrophil mediated activity (*AZU1, LCN2, MPO, CEACAM6, CEACAM8*, *OLFM4* (430,431)) and its regulation (*LCN2, CAMP*, *OLR1* (432–434))
*Lipid-leukocyte module (LLM)	13 (13)	**Mast cell and basophil function	Mast cell and basophil related immune response and allergic inflammation (*FCER1A, HDC, GATA2, SLC45A3, CPA3, MS4A3* (296,435,436))
General immune module A (GIMA)	509 (482)	Immune system process Defence response Regulation of response to stimulus	These modules contain genes involved in a broad range of immune processes and their regulation such as signalling; cell death; defence response to stress, inflammation, and external stimuli; leukocyte activation, migration, and adhesion.
General immune module B (GIMB)	74 (69)	Immune response-activating signal transduction Positive regulation of immune response Activation of immune response	

* Modules previously reported to have immune related function. ** LLM module was not significantly enriched for any GO term. Size refers to the number of core genes in each module and the subset of these core genes with GO term annotations are listed in parenthesis. Functions were assigned to each of these modules based on GO enrichments and literature-based searches for genes in the modules.

The eight modules encoded diverse immune functions, including cytotoxic, viral response, B cell, platelet, neutrophil, mast cell/basophil, and general immune-related functions.

To further delineate the function of the genes present in GIMA and GIMB, I investigated the enrichment of these genes within innate immunity genes manually annotated in the innateDB (399) database. It was seen that genes in these two modules were enriched in innate immune-related pathways such as recognition and response to bacterial lipoproteins, phagocytosis, and signalling pathways triggered during innate immune response (**Table 2.5)**.

2.4.4 Immune module association analysis for eQTLs and metabolite levels

For each gene module, I performed a genome-wide scan to identify module QTLs (mQTLs) that regulate expression. In DILGOM07 and YFS, the module eigengene was regressed on each SNP, and then mQTL test statistics were combined in a meta-analysis. Significant mQTLs were further examined at individual gene expression levels. A genome-wide significance level (*P*-value < $5x10^{-8}$) was used to identify mQTLs (**Figure 2.2; Table 2.6**). Immune-metabolite associations for all the modules have been summarised in **Figure 2.3**.

Given the exploratory nature of this study, the GWAS significant threshold ($P < 5 \times 10^{-8}$) was chosen for mQTL detection. Since the GWAS threshold is considered to be highly stringent, further accounting for the 8 modules tested (that would raise the cutoff to $P < 6.25 \times 10^{-9}$) would lead to an overly conservative threshold and potential false negative associations. It is pertinent to note here that five out of the nine mQTLs detected, which might of biological importance, did not achieve the 6.25×10^{-9} significance, necessitating further investigation and reconfirmation from replication and/or validation studies. The same *P*-value < $5x10^{-8}$ threshold applied in mQTL analysis was also used to identify significant *trans* effects on individual gene expression because genes within each module are highly correlated with their respective module eigengenes.

Table 2.5: Top InnateDB functional annotations for genes in GIMA and GIMB.

Module	GO ID ~ innate immunity GO annotation terms	P-value	Count	Genes
GIMA	GO:0045087 ~ innate immune response	4.65×10^{-17}	86	1384
	GO:0007165 ~ signal transduction	6.27×10^{-10}	70	1368
	GO:0032496 ~ response to lipopolysaccharide	1.46×10^{-7}	17	154
	GO:0090382 ~ phagosome maturation	2.39×10^{-7}	9	39
	GO:0006954 ~ inflammatory response	1.89×10^{-6}	23	315
	GO:0019221 ~ cytokine-mediated signaling pathway	2.09×10^{-6}	20	249
	GO:0038096 ~ Fc-gamma receptor signaling pathway involved in phagocytosis	3.50×10^{-6}	11	82
	GO:0050900 ~ leukocyte migration	3.62×10^{-6}	13	116
	GO:0031663 ~ lipopolysaccharide-mediated signaling pathway	3.90×10^{-6}	7	29
	GO:0007596 ~ blood coagulation	4.64×10^{-6}	29	483
GIMB	GO:0045087 ~ innate immune response	9.26×10^{-10}	22	1384
	GO:0071726 ~ cellular response to diacyl bacterial lipopeptide	1.23×10^{-5}	2	2
	GO:0006928 ~ cellular component movement	2.88×10^{-5}	5	101
	GO:0060715 ~ syncytiotrophoblast cell differentiation involved in labyrinthine layer development	3.68×10^{-5}	2	3
	GO:0097194 ~ execution phase of apoptosis	7.10×10^{-5}	3	23
	GO:0034097 ~ response to cytokine	7.24×10^{-5}	4	63
	GO:0048713 ~ regulation of oligodendrocyte differentiation	7.34×10^{-5}	2	4
	GO:0038096 ~ Fc-gamma receptor signaling pathway involved in phagocytosis	2.02×10^{-4}	4	82
	GO:0007165 ~ signal transduction	2.53×10^{-4}	14	1368
	GO:0034142 ~ toll-like receptor 4 signaling pathway	3.70×10^{-4}	4	96

GO – refers to Gene Ontology. The GO terms listed are significant at FDR < 0.05.
Counts – the number of genes in the modules (GIMA or GIMB) that were classified to a particular GO annotation in InnateDB database. Genes – the number of module genes in InnateDB for a particular GO annotation.

2.4.4.2 Effect of blood cell counts on immune module associations with mQTLs and metabolites

Leukocyte and platelet counts were available for YFS and were used to test the robustness of module associations with mQTLs and metabolites. Six modules showed statistically significant association with platelet or leukocyte counts (*P*-value < 0.05) (**Table 2.7**), however adjustment for leukocyte counts did not affect mQTL nor metabolite-module associations, with the exception of the PM and CCLM discussed below. Since we did not have cell counts available for DILGOM07, all the immune-metabolite associations discussed below, unless otherwise noted, have not been adjusted for cell counts.

2.4.4.3 Cytotoxic cell-like module (CCLM) associations with mQTLs and metabolites

CCLM was not significantly associated with any mQTL at genome-wide significance, however it was associated with 24 metabolites, mainly consisting of fatty acids, intermediate density lipoproteins, and CR (**Figure 2.3**; Table A.2 in Appendix A). The top associated metabolite was docosahexaenoic acid (DHA) (meta-*P*-value = 5.34×10^{-08}). The role of CRP in augmenting cytotoxic responses has been reported, which includes the ability of CRP to bind to NK cells and also influence their activity (60), enhance cytotoxic response of NK cell against tumour cells (61), and sensitize endothelial cells to cytotoxic T-cell mediated destruction (62).

However, when the CCLM-metabolite associations were adjusted for leukocyte counts, four existing associations (CRP, creatinine, ratio of polyunsaturated fatty acids to total fatty acids, and VLDL particle size) were no longer significantly associated (Table A.3 in Appendix A). I also observed a gain of 38 additional significant associations, mainly the LDL and VLDL subclass of lipoproteins, when leukocyte levels were accounted for (Table A.3 in Appendix A).

2.4.4.4 Viral response module (VRM) associations with mQTLs and metabolites

mQTLs for the VRM

Three genome-wide significant mQTLs were identified for the VRM (**Figure 2.2; Table 2.6**). The strongest mQTL, rs182710579 (meta-*P*-value = 9.22 x 10^{-09}), lies within a relatively unstudied nearly 900bp lncRNA RP11-608O21.1 region (**Figure 2.4A**). Rs182710579 was a *trans* eQTL for 3 genes in the VRM (Table A.4 in Appendix A). The strongest association was seen with *CCL2* (meta-*P*-value = 6.78 x 10^{-12}), a pro-inflammatory chemokine involved in leukocyte recruitment during infection and elevated levels have also been reported during viral infections (63,64).

The next strongest mQTL, rs151234502, resides within intron 4 of *ZNF212* (**Figure 2.4B**). *ZNF212* encodes a zinc finger (ZNF) protein and is part of a *ZNF* gene cluster on chromosome 7q36.1 that contains a conserved Kruppel-associated box (KRAB) domain (65). KRAB-ZNF proteins are usually involved in transcription repression (66) and have been noted to show high expression in immune cells (65). Rs151234502 regulated the expression of 12 probes (corresponding to 11 unique genes) present in the VRM in *trans* (Table A.4 in Appendix A) and the strongest association was detected with *OAS2* (meta-*P*-value = 8.98 x 10^{-10}), an interferon-induced gene that encodes an enzyme responsible for promoting RNase L-mediated cleavage of viral and cellular RNA (67).

The final mQTL for the VRM, rs147742798, was an intergenic lead SNP located between *SHANK2* and *DHCR7* at 11q13.4 (**Figure 2.4C**). Rs147742798 was a *trans* eQTL for 2 genes in the VRM (Table A.4 in Appendix A) and was most strongly associated with *BST2* (meta-*P*-value = 6.10 x 10^{-09}). *BST2,* an interferon-induced gene, encodes a transmembrane protein with antiviral function through inhibition of the egress of mature virions from infected cells by tethering them to the cell surface (68). For the three mQTLs, none were in *cis* to any genes in the VRM.

Metabolites associated with the VRM

VRM was associated with eight metabolites, including amino acids (alanine, phenylalanine), fatty acids (omega-6 fatty acids, polyunsaturated fatty acids, saturated

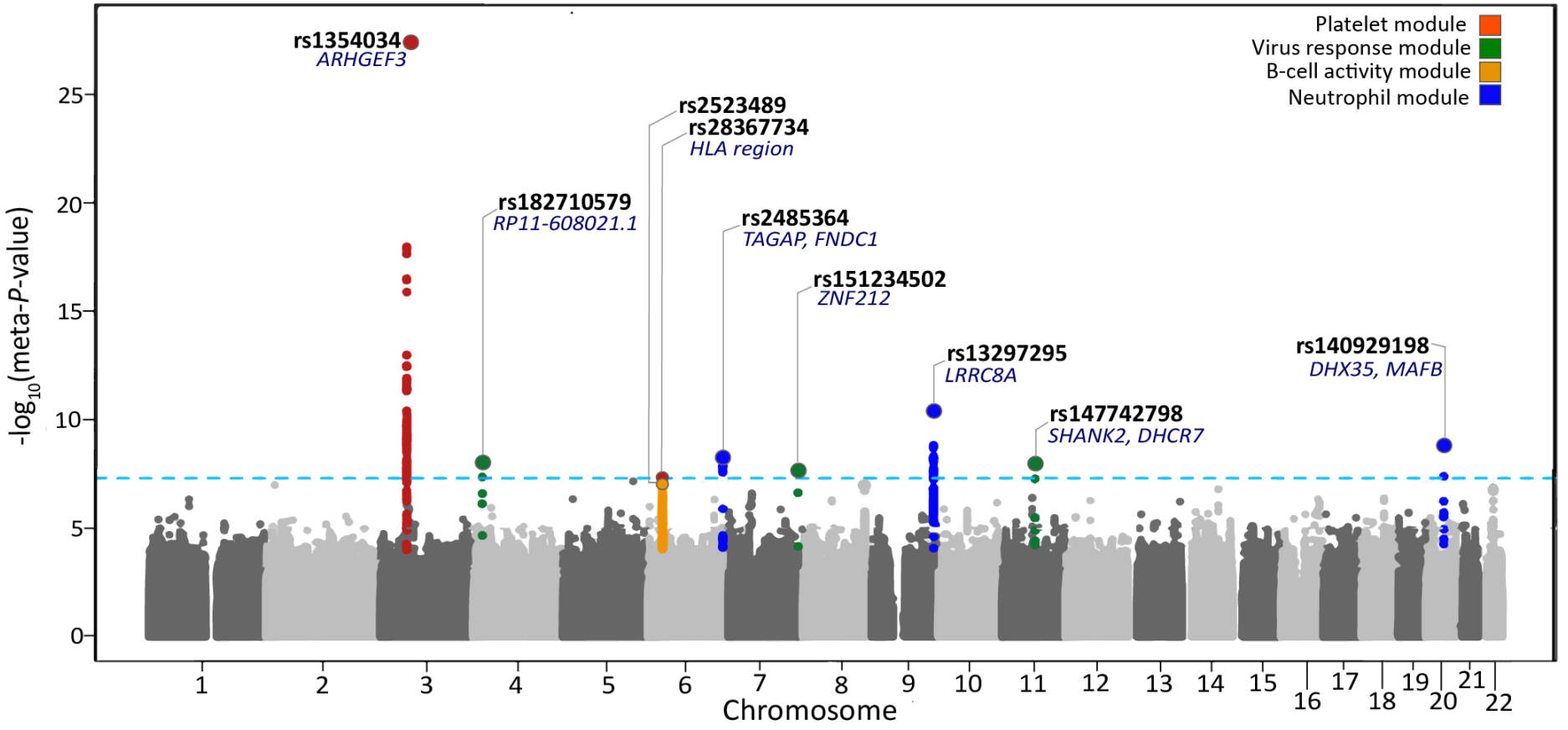


Figure 2.2: Manhattan plot of meta-analysed *P*-values from the DILGOM/YFS module QTL analysis.

The *y*-axis shows the $\log_{10}$(meta-*P*) values plotted against all the SNPs tested (*x*-axis). The lead SNP and its closest genes are shown. The associated loci for each module are highlighted with a separate colour designated for each module. The sky-blue line represents genome-wide (meta *P*-value < 5 x10^{-8}) threshold.

Table 2.6: QTLs for immune gene modules. Modules: VRM (viral response module), BCM (B cell activity module), PM (platelet module), NM (neutrophil module).

Module	Top SNP	CHR	Hg19 Pos. (Mb)	Allele (minor/major)	MAF (Avg)	*P*-value DILGOM07 (effect size)	*P*-value YFS (effect size)	Meta-*P*-value
VRM	rs182710579	4	19768086	G/T	0.012	2.01×10^{-04} (0.05)	8.10×10^{-06} (0.02)	9.23×10^{-09}
	rs151234502	7	148950168	T/C	0.012	2.59×10^{-01} (0.01)	5.31×10^{-09} (0.03)	2.46×10^{-08}
	rs147742798	11	70947761	T/C	0.016	1.51×10^{-03} (0.04)	1.66×10^{-06} (0.02)	9.43×10^{-09}
BCM	rs2523489	6	31348878	T/C	0.186	1.42×10^{-1} (0.005)	5.29×10^{-08} (0.006)	6.27×10^{-08}
PM	rs1354034	3	56849749	T/C	0.284	7.11×10^{-14} (-0.02)	1.51×10^{-16} (-0.008)	7.35×10^{-28}
	rs28367734	6	3128657	A/G	0.108	5.40×10^{-04} (0.02)	2.02×10^{-05} (0.006)	5.44×10^{-08}
NM	rs2485364	6	159512260	C/T	0.466	1.78×10^{-03} (0.009)	6.05×10^{-07} (0.004)	3.93×10^{-09}
	rs13297295	9	131659724	C/T	0.085	4.26×10^{-02} (0.009)	8.39×10^{-11} (0.01)	3.93×10^{-11}
	rs140929198	20	38555870	A/G	0.031	2.98×10^{-02} (0.03)	8.47×10^{-09} (0.01)	1.41×10^{-09}

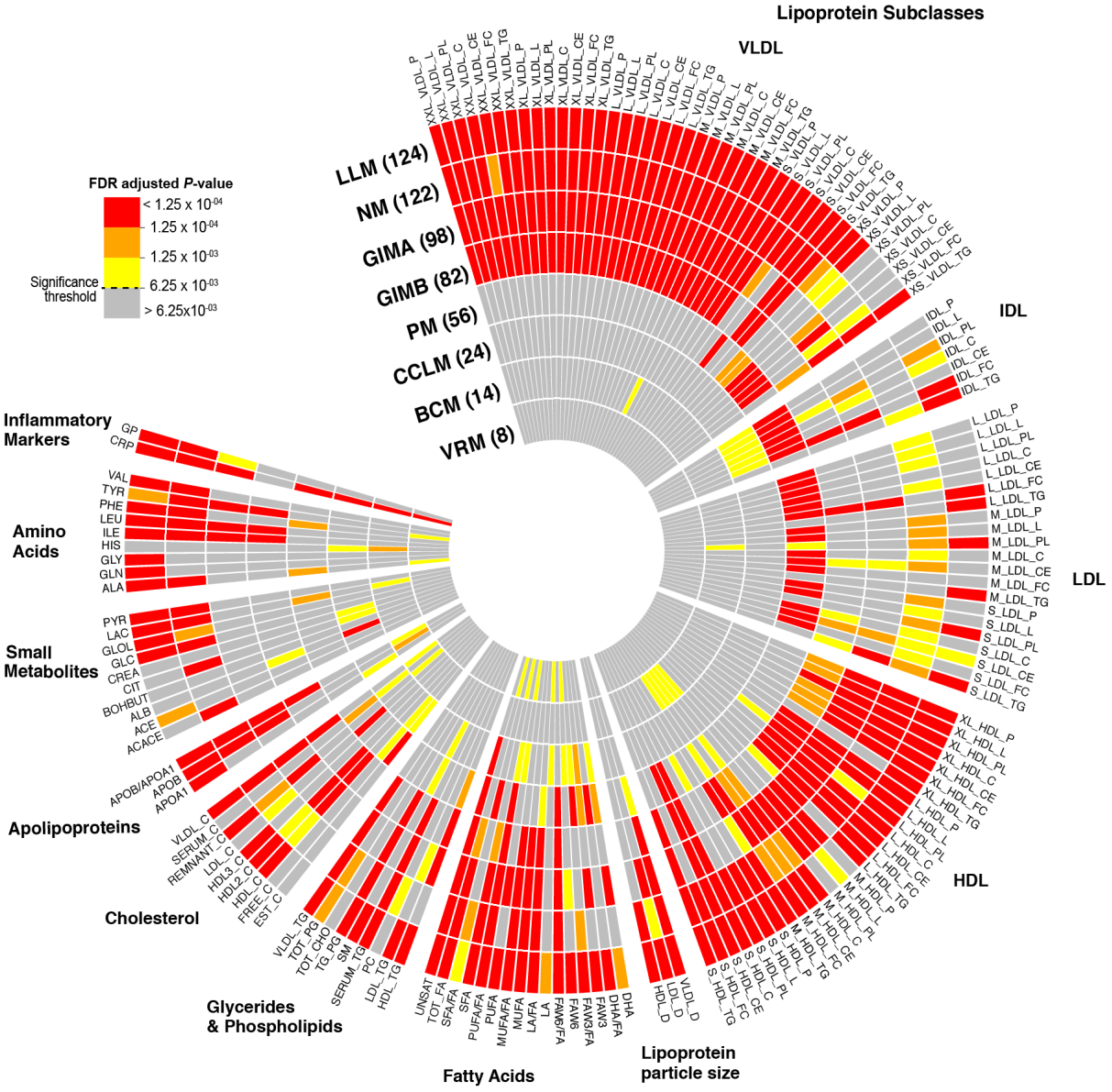


Figure 2.3: Metabolite associations with immune gene modules.

Circular heatmap of associations between individual metabolites and the module eigengene of each module (coloured by FDR-adjusted *P*-values). Concentric circles represent modules, with numbers in parentheses denoting total number of metabolites associated with that module at FDR-adjusted *P*-value < 6.25 x 10^{-03}. Each segment of the circle represents a metabolite labelled on the outside of the heatmap. NM (neutrophil module), LLM (lipid leukocyte module), GIMA, and GIMB (General immune modules A and B), PM (platelet module), CCLM (cytotoxic cell-like module), BCM (B cell activity module), and VRM (viral response module). See **Table 2.1** for full metabolite descriptions.

Table 2.7: Association between immune-related modules and blood cell counts (leukocyte and platelet counts) in YFS.

Module	Cell counts	Beta estimates	Standard error	*P*-value
CCLM	Leukocytes	-0.17	0.03	1.57×10^{-11}
VRM	Leukocytes	-0.04	0.03	8.77×10^{-02}
BCM	Leukocytes	-0.06	0.03	2.55×10^{-02}
NM	Leukocytes	0.24	0.02	6.23×10^{-22}
LLM	Leukocytes	-0.13	0.03	2.11×10^{-07}
GIMA	Leukocytes	0.22	0.03	5.79×10^{-17}
GIMB	Leukocytes	0.11	0.03	2.78×10^{-05}
PM	Platelets	0.29	0.02	8.23×10^{-30}

Results from the linear regression of immune module summary expression profiles on leukocyte and platelet counts. The regression module was adjusted for age and sex.

fatty acids, and total fatty acids), and cholesterol esters in medium VLDL (**Figure 2.3**; Table A.5 in Appendix A). Consistent with its putative role in viral response, VRM was strongly associated with CRP (meta *P*-value = 2.38 x 10^{-10}). Viral infection has been shown to trigger a mild acute phase response, which causes a moderate increase in CRP levels (437–439). Amino acids, phenylalanine and alanine, were also associated with this module. There is increasing evidence supporting the association between viral infection, and altered amino acid metabolism (440–444). Phenylalanine is an essential amino acid shown to be involved in modulating immune response processes (445,446) and elevated levels in circulation have been seen in patients infected with dengue virus (443).

2.4.4.5 B-cell activity module (BCM) associations with mQTLs and metabolites

mQTLs for the BCM

While no mQTLs for the BCM exceeded genome-wide significance, the plausible MHC class I locus showed some evidence in YFS (**Figure 2.2; Table 2.6**). The intergenic index SNP at this locus, rs2523489 (meta-P-value = 6.27 x 10-08), is located between HLA-B/C and MICA (**Figure 2.5**). The HLA class I region is known to be associated with a number of autoimmune diseases. This SNP is within ~1.8Kb and in strong LD with a variant (rs1521; r2 = 0.9) associated with Graves disease (69) an autoimmune disease characterised by hyperthyroidism. The role of B cells in the development of autoimmune diseases is well recognised (70) including Graves disease where their numbers are shown to be elevated (71,72). Rs2523489 was associated in trans with CD79B (meta-P-value = 1.16 x 10-09), a gene present in the BCM, which encodes for the CD79B subunit of the B cell receptor complex that binds to antigens (73). CD79B forms part of the CD79A/B heterodimer that is essential for B-cell receptor functioning (417,447,448) and altered signalling through this receptor has been suggested to contribute to B-cell induced autoimmunity (449). Particularly, antibodies against CD79b have been shown to suppress autoimmune diseases (450–452), This implies that signals emanating from CD79, which may t be influenced by variants at the MHC class I locus, might be crucial in the development of B-cell-mediated autoimmunity.

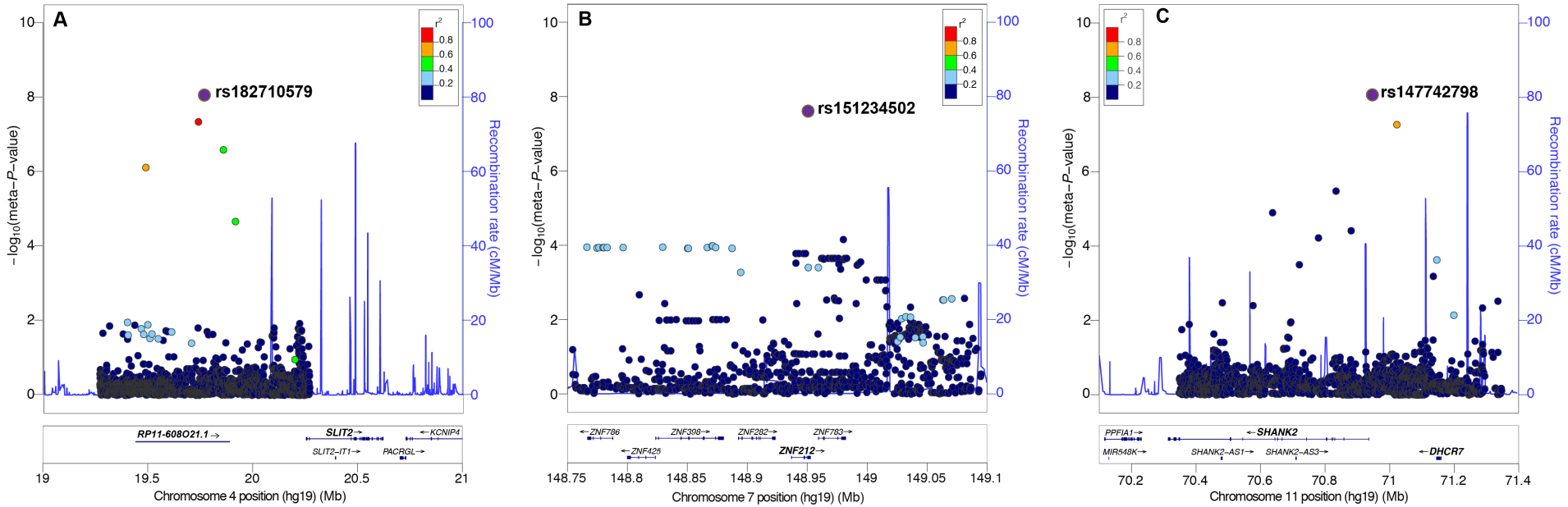


Figure 2.4: Regional plots of the mQTLs associated with the viral response module (VRM) at the (A) 4p15.31, (B) 7q36.1, and (C) 11q13.4 regions.

For each plot, the circles represent the -$\log_{10}$ meta-analysed *P*-values (y-axis) of SNPs plotted against their chromosomal position (x-axis). The lead mQTL (rsID) in each plot is denoted by a purple circle, and its pairwise LD (r^2) strength with other SNPs in the region, estimated from the "1000 genomes Mar 2012 EUR" population, is indicated by colour. The blue lines indicate the recombination rates. The plots were generated using the LocusZoom online tool (http://locuszoom.sph.umich.edu/locuszoom/).

Metabolites associated with the BCM

The BCM was associated with 14 metabolites including CRP, histidine, lactate, apolipoproteins, and mainly medium HDL subclass of lipoproteins. (**Figure 2.3**; Table A.6 in Appendix A). The strongest association was seen with CRP (meta-*P*-value = 2.65 x 10^{-08}). Histidine, the second most strongly associated metabolite, is catabolised to histamine by histidine decarboxylase (a component of LL module). The relationship between B cells and histamine is a central part of the allergic reaction where IgE released by B cells blankets mast cells, causing them to release histamine.

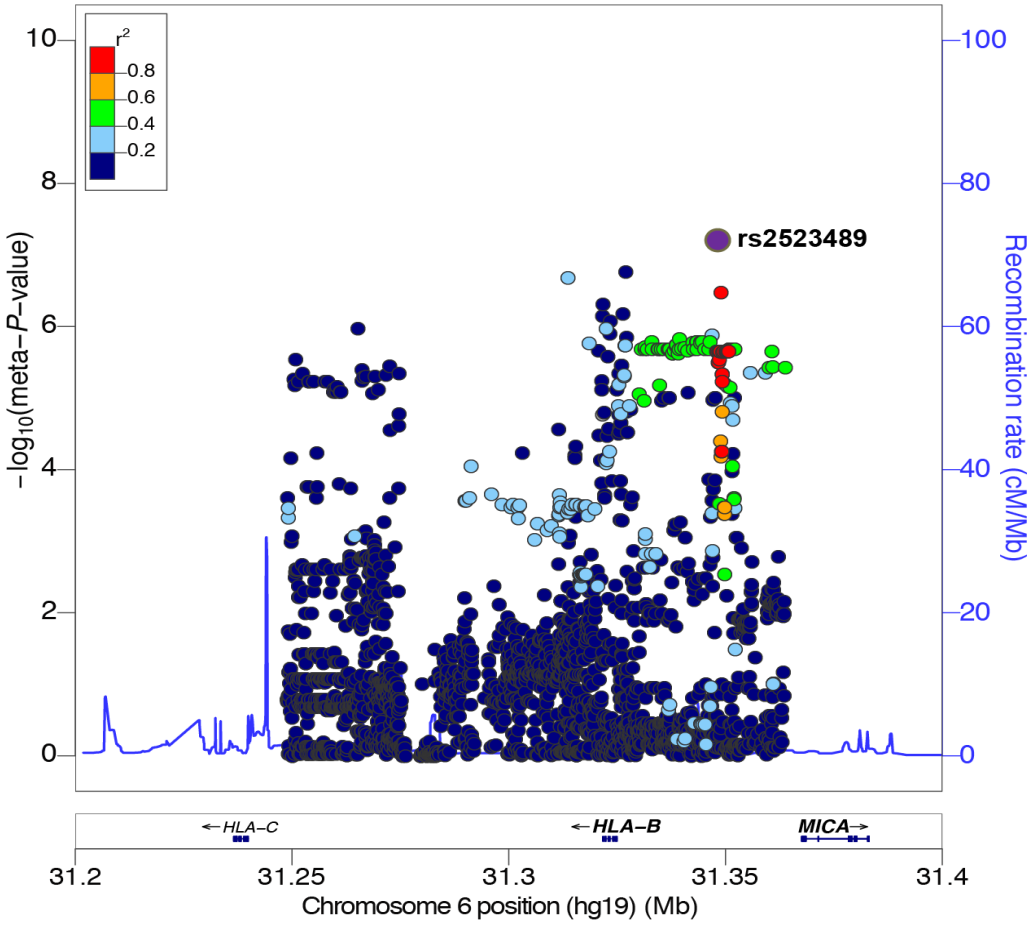


Figure 2.5: Regional plots of the mQTLs associated with the B-cell activity module (BCM) at the 6p21.33 (HLA) region.

The circle represents the -log10 meta-analysed *P*-values (y-axis) of SNPs plotted against their chromosomal position (x-axis). mQTL (rsID) is denoted by a purple circle, and its pairwise LD (r2) strength with other SNPs in the region, estimated from the "1000 genomes Mar 2012 EUR" population, is indicated by colour. The blue lines indicate the recombination rates. The plots were generated using the LocusZoom online tool (http://locuszoom.sph.umich.edu/locuszoom/).

2.4.4.6 Platelet Module (PM) associations with mQTLs and metabolites

mQTLs for PM

Two genome-wide significant mQTLs were identified for the PM (**Figure 2.2; Table 2.6**). The strongest mQTL for the PM, as well as strongest of any gene module, was an intronic SNP (rs1354034; meta-*P*-value = 7.35 x 10^{-28}) located in the *ARHGEF3* gene at 3p14.3 (**Figure 2.6A**). *ARHGEF3* encodes Rho guanine nucleotide exchange factor 3, which mediates the activation of Rho GTPases by catalysing its conversion from an inactive GDP-bound to an active GTP–bound form. No genes within 1Mb of rs1354034 were present in the PM, indicating that this mQTL functions in *trans*. The *ARHGEF*3 mQTL (rs1354034) exhibited a strong *trans*-regulatory effect and was associated with 61 PM genes (65 unique probes) (**Figure 2.7**; Table A.7 in Appendix A). The top *trans* eQTL was *ITGB3* (meta-*P*-value = 5.09 x 10^{-42}), a gene encoding the β_3 subunit of the heterodimeric integrin receptor (integrin $\alpha_{IIb}\beta_3$). This integrin receptor is most highly expressed on activated platelets and plays a key role in mediating platelet adhesion and aggregation upon binding to fibrinogen and Willebrand factor (453,454). Our data are consistent with previous observations of the diverse *trans* eQTL effects of rs1354034 (310). I was able to replicate 26 of the *trans* associations that were previously identified for rs1354034 in an eQTL analysis on RNA-sequencing based expression profiles obtained from whole blood (57). The same study also identified rs1354034 as a splice-QTL for *TPM4*, a significant eGene in the PM. Additionally, An intergenic SNP, rs2836773 (meta-*P*-value = 5.4 x 10^{-08}), at the HLA locus was also identified as a lead mQTL (**Figure 2.6B**). No HLA genes were present in this module.

ARHGEF3 itself is of intense interest to platelet biology. It has previously been shown that silencing of *ARHGEF3* in zebrafish prevents thrombocyte formation (135). To test whether *ARHGEF3* expression had an effect on PM genes, we regressed out *ARHGEF3* levels and re-ran the eQTL analysis. Adjusting for *ARHGEF3* did not attenuate the *trans*-associations of rs1354034, suggesting either independence of downstream function for *ARHGEF3* and rs1354034 or post-transcriptional modification of ARHGEF3. Previous GWAS studies have shown rs1354034 is associated with platelet count and mean platelet volume (135), however, perhaps due to power, we found no significant relationship between platelet counts and rs1354034 in YFS. While platelet counts were positively associated with the PM (β = 0.29; *P*-value = 8.23 x 10^{-30}) **(Table**

2.7), the association between rs1354034 and the PM was still highly significant when conditioning on platelet counts (β = -0.33; *P*-value = 1.40 x 10^{-17}).

Metabolites associated with the PM

PM displayed diverse metabolic interactions and was associated 55 metabolites, largely comprising of lipoprotein subclasses and fatty acids, as well as CRP (**Figure 2.3**; Table A.8 in Appendix A). Cholesterol esters in small HDL particles were most strongly associated with the PM (meta-*P*-value = 9.45 x 10^{-20}). HDL has been shown to exhibit antithrombotic properties by modulating platelet activation, aggregation and coagulation pathway (455). On the other hand, pro-atherogenic lipoproteins effects on platelets have been recognised as an important driver in the development of atherosclerosis. For example, LDL has been shown to influence platelet activity either by enhancing platelet responsiveness to aggregating stimuli or inducing aggregation (456,457). Moreover, LDL specific binding sites on platelets have also been reported (458,459).

CRP was also strongly associated with the PM (meta-*P*-value = 4.12 x 10^{-08}). Several studies have shown the link between CRP and platelet activity, for example, infusion of recombinant human CRP in humans led to the activation of coagulation pathway (460). In addition, CRP has also been shown to promote the adhering of platelets to monocytes (461) and endothelial cells (462), a consequence of platelet activation.

As noted above, the PM was associated with platelet counts, and adjustment of PM-metabolite associations for platelet counts in the YFS resulted in attenuation of approximately half of the weakest metabolite associations; however, the strongest were maintained (Table A.9 in Appendix A). Association with VLDL particle size and three others were gained following the adjustment (Table A.9 in Appendix A).

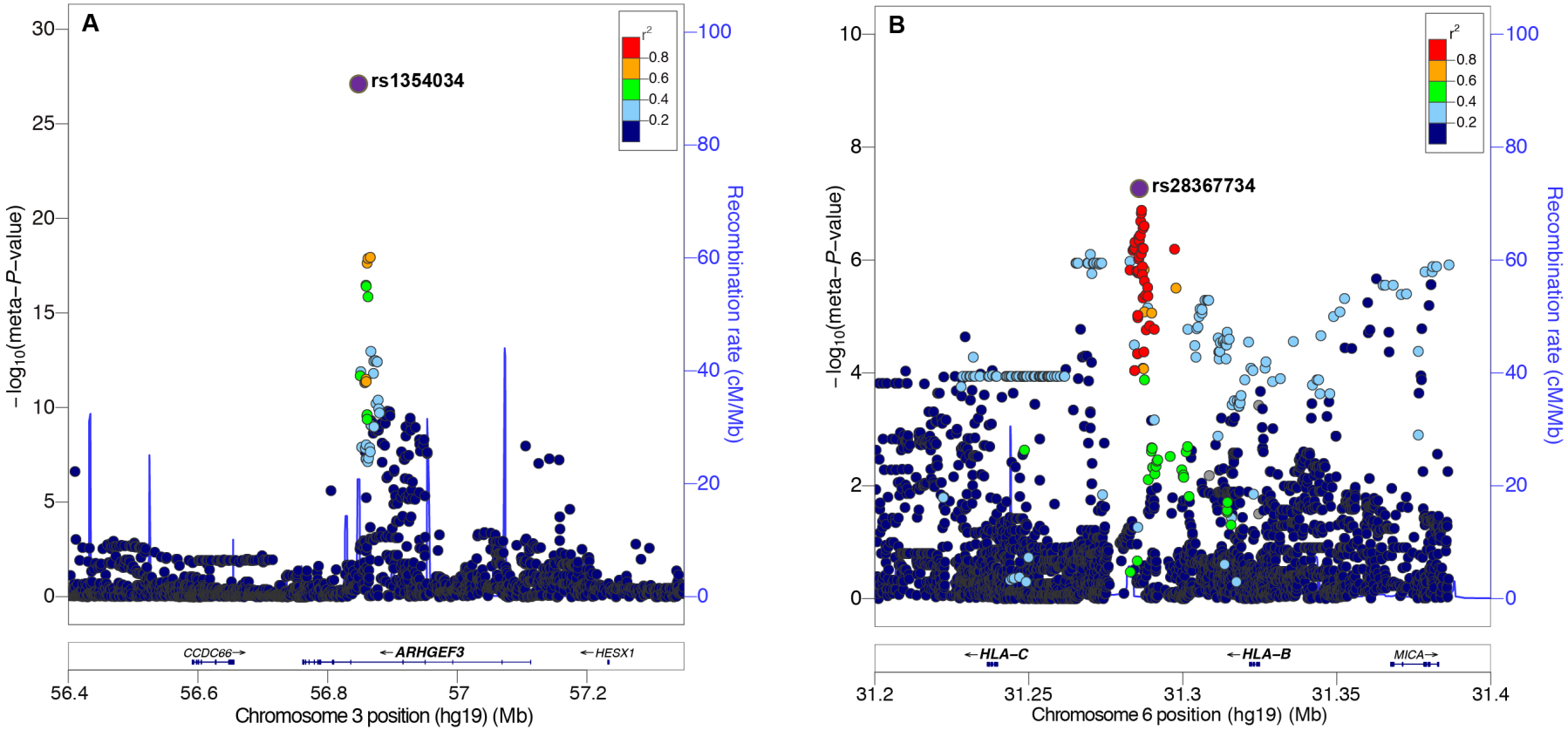


Figure 2.6: Regional plots of the mQTLs associated with the platelet activity module (PM) at regions (A) 3p14.3 and (B) 6p21.33.

For each plot, the circles represent the -$\log_{10}$ meta-analysed *P*-values (y-axis) of SNPs plotted against their chromosomal position (x-axis). The mQTL (rsID) in each plot is denoted by a purple circle, and its pairwise LD (r^2) strength with other SNPs in the region, estimated from the "1000 genomes Mar 2012 EUR" population, is indicated by colour. The blue lines indicate the recombination rates. The plots were generated using the LocusZoom online tool (http://locuszoom.sph.umich.edu/locuszoom/).

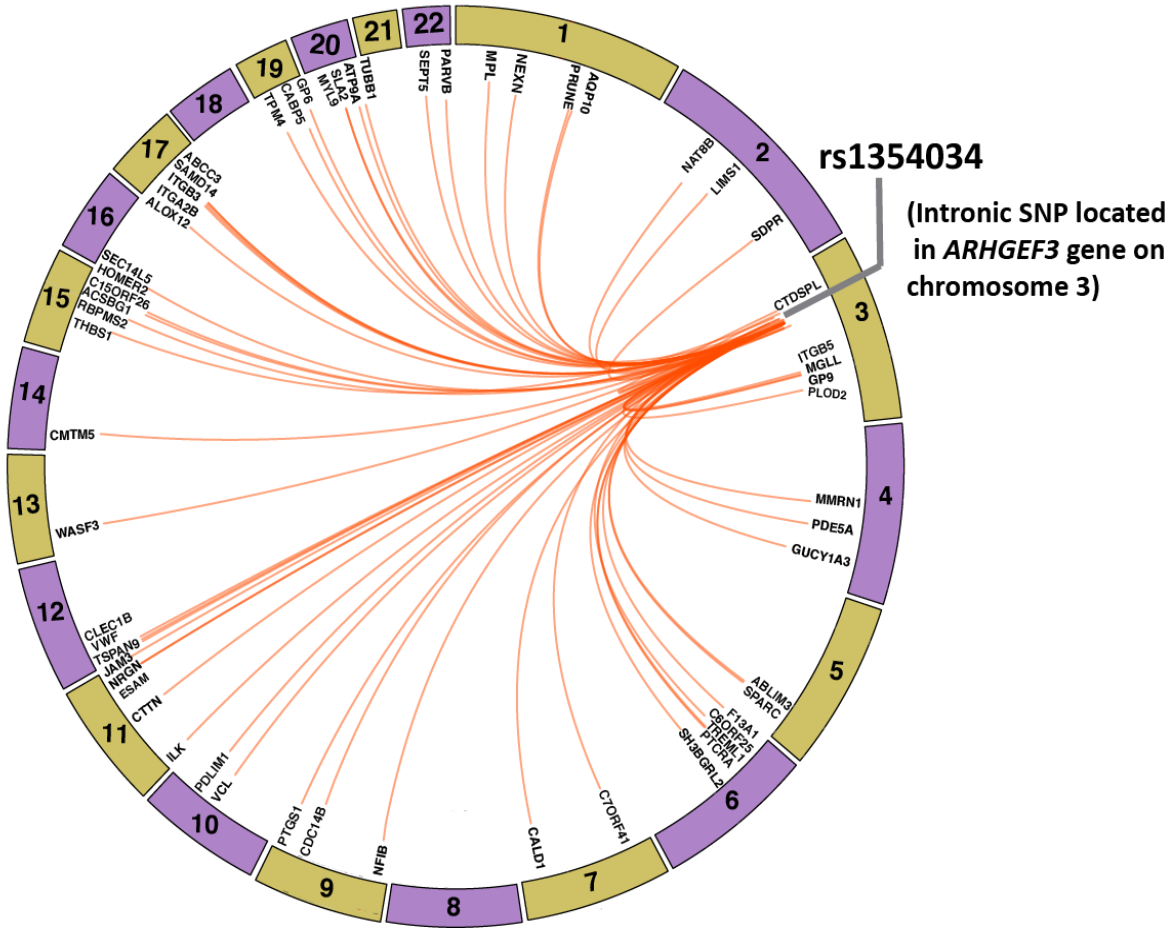


Figure 2.7: rs1354034 is a strong *trans* regulator of genes in the platelet module.

The circular plot shows the *trans* eQTL associations (meta-*P*-value < 5 x10^{-8}) between the lead module QTL (rs1354034) and the genes in the platelet module. The ring presents the genome arranged by the autosomal chromosomes, 1 to 22, showing the location of genes in the platelet module. Lines are pointing from the lead SNP (labelled outside the ring) to the respectively associated gene.

2.4.4.7 Neutrophil Module (NM) associations with mQTLs and metabolites

mQTLs for NM

Three loci were identified as mQTLs for NM (**Figure 2.2; Table 2.6**). The top mQTL was intronic to *LRRC8A* at 9q34.11 (rs13297295; meta-*P*-value = 3.93 x 10^{-11}, **Figure 2.8A)**. *LRRC8A* encodes for a trans-membrane protein shown to play a role in B- and T-cell development and T-cell function (463,464). Two additional intergenic mQTLs were located at the *TAGAP* locus at 6q25.3 (rs2485364; meta-*P*-value = 3.93 x 10^{-9}) and at 20q12 (rs140929198; meta-*P*-value = 1.41 x 10^{-9}) (**Figure 2.8B-C**).

At the LRRC8A locus, rs13297295 was a strong *trans* regulator of the NM and was associated with 8 genes (10 unique probes), in particular, the major alpha defensins (*DEFA1-DEFA4*) which formed the core genes of highest centrality in the module (Table A.10 in Appendix A), The strongest *trans*-eQTL was *DEFA1B* (meta-*P*-value = 3.17 x 10^{-14}). Additionally, rs13297295 was also a *cis*-eQTL for another core NM gene, *LCN2* (meta-*P*-value = 3.81 x 10^{-09}; permuted meta-*P*-value = 1 x 10^{-04}) (Table A.10 in Appendix A). *LCN2* is induced in response to TLR activation and acts as an antimicrobial agent by sequestering bacterial siderophores to prevent iron uptake (86–88). *LCN2*'s role in acute phase response also appears to be related to cardiovascular diseases, such as heart failure (89).

At the *TAGAP* locus, rs2485364 was a *trans*-eQTL for 8 genes (10 probes) in the NM and was also a strong driver of *LCN2* (meta-*P*-value = 9.11 x 10^{-17}) (Table A.10 in Appendix A). Consistent with our findings, neutrophils from *LCN2* deficient mice have been shown to exhibit impaired chemotaxis, phagocytic capability, and increased susceptibility to bacterial and yeast infections compared to wild type (90). The *TAGAP* locus has also been linked to autoimmune diseases, variants in moderate LD (r^2 = 0.52) with rs2485364, have been reported to be associated with celiac disease (91), Crohn's disease (91–93) and rheumatoid arthritis (94). With regards to arthritis, *LCN2* knockout mice showed reduced neutrophil migration and developed a more severe form of the disease than their wild-type counterparts (95). Taken together, it can be speculated that a functional role of *TAGAP* variants is the regulation of neutrophil migration through *LCN2*.

At 20q12, rs140929198 was a *trans*-eQTL for 5 genes (7 probes) in the NM and was most strongly associated with *OLR1* (meta-*P*-value = 6.75 x10^{-09}) (Table A.10 in Appendix A), an endothelial cell surface receptor for pro-atherogenic oxidized-LDL (ox-LDL) (96). *OLR1* expression has been reported to play a role in neutrophil migration during sepsis (97).

Metabolites associated with the NM

NM was associated with 121 circulating metabolites (~76% of all metabolites analysed) as well as CRP (**Figure 2.3**; Table A.11 in Appendix A). The strongest is the previously reported association with inflammatory biomarker GlycA (meta-*P*-value = 2.68 x 10^{-25}) (294), however NM's association with various lipoproteins subclasses, particle sizes of lipoproteins, fatty acids, cholesterol, apolipoproteins, glycerides and phospholipids, amino acids, and other small molecules indicates it has a potentially major role in linking neutrophil function to metabolism. Lipoproteins have been reported to have an immunomodulatory function during bacterial infections; for example, lipoproteins such as HDL, VLDL, and LDL have been shown to bind to lipopolysaccharide (LPS), an outer membrane component of gram-negative bacteria, neutralising their activity (98–101). In addition, the associations seen with LDL lipoprotein subclasses provide possible mechanistic insights into the emerging proatherogenic role of neutrophils. Neutrophil derived defensins (DEFA1–DEFA3) can complex with LDL particles enhancing their binding and uptake by endothelial, smooth muscle cells, and fibroblasts (465) . Furthermore, it has been suggested that defensins induce LDL modification upon binding that accelerates its removal from circulation and subsequent deposition in vascular cells (466).

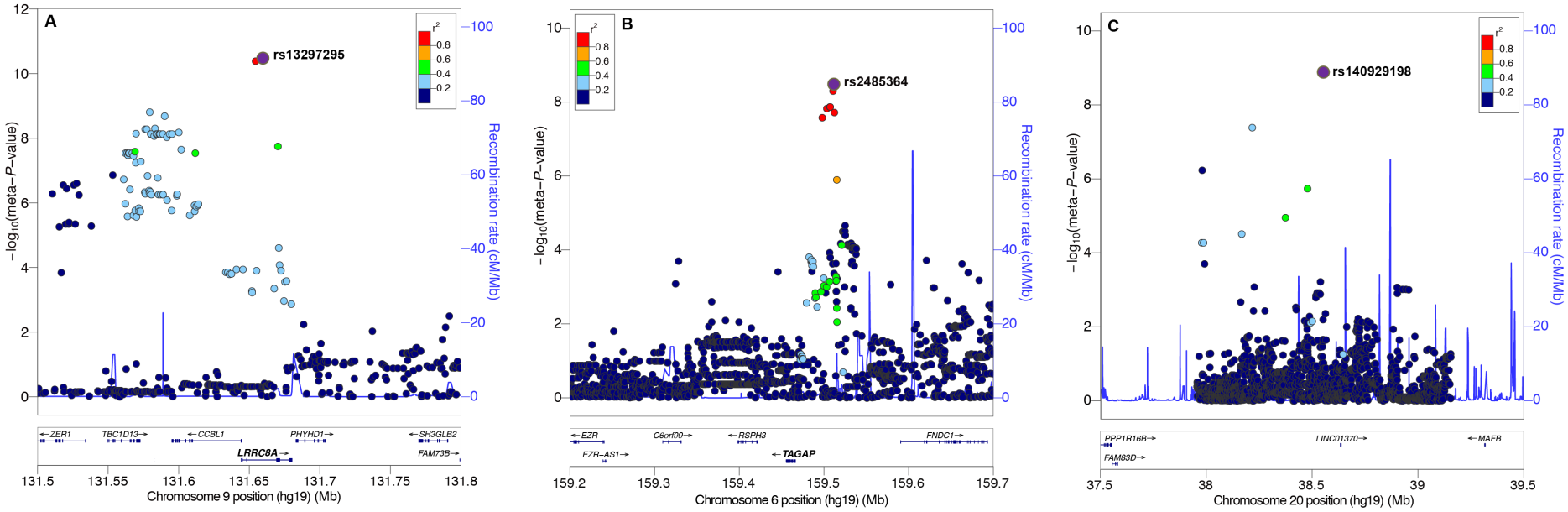


Figure 2.8: Regional plots of the mQTLs associated with the neutrophil module (NM) at the (A) 9q34.11, (B) 6p25, and (C) 20q12 regions.

For each plot, the circles represent the -$\log_{10}$ meta-analysed *P*-values (y-axis) of SNPs plotted against their chromosomal position (x-axis). The lead mQTL (rsID) in each plot is denoted by a purple circle, and its pairwise LD (r^2) strength with other SNPs in the region, estimated from the "1000 genomes Mar 2012 EUR" population, is indicated by colour. The blue lines indicate the recombination rates. The plots were generated using the LocusZoom online tool (http://locuszoom.sph.umich.edu/locuszoom/).

2.4.4.8 Lipid-Leukocyte module (LLM) associations with mQTLs and metabolites

Consistent with previous studies, no mQTLs were detected for LLM. Together with NM, the LLM showed extensive metabolic associations. Overall, 123 metabolites and CRP were associated with LLM, with the strongest being the ratio of triglycerides to phosphoglycerides (meta-*P*-value = 5.16×10^{-138}, **Figure 2.3**; Table A.12 in Appendix A). With the inclusion of the YFS, these findings strongly replicate previous LMM-metabolite associations (375) as well as highlight additional metabolite associations. We also confirm the previous strong negative association between CRP and LLM (meta-*P*-value = 8.16×10^{-20}). The other top associations mainly consisted of the VLDL subclass of lipoproteins.

2.4.4.9 General Immune Module A (GIMA) and General Immune Module B (GIMB associations with mQTLs and metabolites

No mQTLs were associated with GIMA and GIMB. However, these modules were associated with 97 and 82 metabolites, respectively (**Figure 2.3**; Tables A.13 – 14 in Appendix A). Cholesterol esters in small HDL and the mean diameter for VLDL particles exhibited the strongest associations with GIMA (meta-*P*-value = 1.56×10^{-30}) and GIMB (meta-*P*-value = 1.83×10^{-15}), respectively. The GIMA was also associated with CRP (meta-*P*-value = 5.7×10^{-05}). Other metabolites associations with these two modules include mainly the VLDL and HDL subclass of lipoproteins and fatty acids, however, due to their large size and heterogeneous composition, interpretation of metabolic relationships of GIMA and GIMB is limited.

2.4.5 Temporal preservation of immune-linked networks and their interaction with metabolites and mQTLs

Finally, we tested the robustness of each gene network's co-expression and association with metabolites over a 7-year period using the DILGOM07 and DILGOM14 datasets. Between time points, 23 of 40 modules were significantly preserved (all preservation statistics' permutation P-values <0.001), including all 8 immune-related modules (**Table 2.7**). Furthermore, we also observed largely consistent correlation structure in the metabolite profiles between DIGOM07 and DILGOM14 (**Figure 2.9**).

Next, we examined how the interactions between immune-related modules and metabolites changed over the 7-year time period. While power was somewhat limited (N=216 individuals shared between DILGOM07 and DILGOM14), across all modules with significant metabolite associations in DILGOM07, only those for the LLM were maintained over the 7-year period. The LLM was significantly associated with 90 and 79 metabolites in DILGOM07 and DILGOM14 (**Figure 2.10**), respectively, of which 70 metabolite associations overlapped across the two time points (Table A.15 in Appendix A). The stable metabolite associations predominantly included VLDL and HDL subclasses of lipoproteins together with fatty acids and lipids. The direction and effect size of LLM-metabolite associations were largely maintained (**Figure 2.11**). Our findings indicate that the LLM is not only stably coexpressed over time but that its interactions with circulating metabolites are maintained at a level greater than other gene co-expression networks. Across all significant mQTLs, only the *trans*-eQTL effects of rs1354034 on the platelet module appeared to be temporally stable (mQTL *P*-value = 4.87×10^{-07}). However, no metabolite associations were significantly maintained for this module. No other mQTLs reached significance for temporal stability.

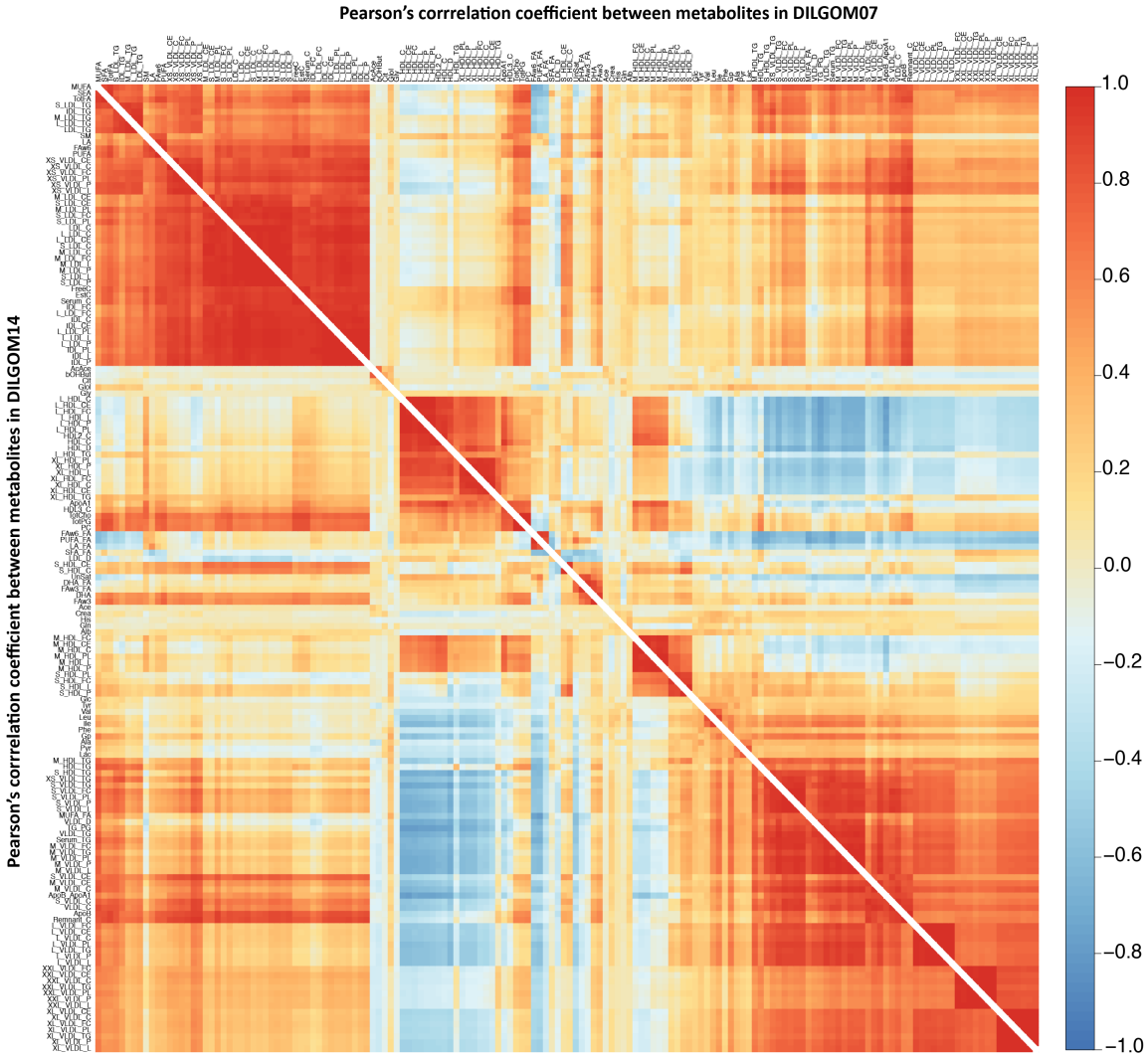


Figure 2.9: Heatmap comparing the correlations between metabolites in DILOM07 with those in DILGOM14.

Comparison of the correlations between the 158 metabolites within DILGOM07 (upper triangle) with those in DILGOM14 (lower triangle). Each square in each triangle represents the Pearson's correlation coefficient calculated between the metabolites within each cohort separately. The correlation matrix in DILGOM07 was hierarchically clustered using distance as, 1-absolute value of the correlations. The ordering of rows and columns in DILGOM2014 (lower triangle) was based on DILGOM07. Red and blue indicates positive and negative correlations, respectively.

Table 2.8: Module preservation statistics of the DILGOM07 immune-related gene co-expression modules in DILGOM14.

Module	Mean Adj	PVE	Corr. Coexp	Corr. kIM	Corr. MM	Mean Coexp	Mean MM	Rep.
GIMA	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	YES
GIMB	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	YES
CCLM	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	YES
PM	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	YES
VRM	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	YES
BCM	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	YES
NM	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	YES
LLM	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	YES

The seven statistics are as follows: (1) Mean Adjacency (**Mean Adj**), assesses how densely the genes are connected in a module across the datasets; (2) Proportion of variance explained by the module eigengenes (**PVE**); (3) Correlation of the module co-expression across the two datasets (**Corr.Coexp**); (4) Correlation of connectivity (**Corr.kIM**) assesses whether the most highly connected genes in a module are the same across the datasets; (5) Correlation of module membership (**Corr.MM**), which is the correlation between each module gene and the module eigengene, assesses whether the contribution of each gene to the summary expression of a module is same across the datasets; (6) mean co-expression (**Mean Coexp**); and (7) mean module membership (**Mean MM**) assesses whether the signs of the correlation are in the same direction in the two datasets. Rep. – refers to module replication.

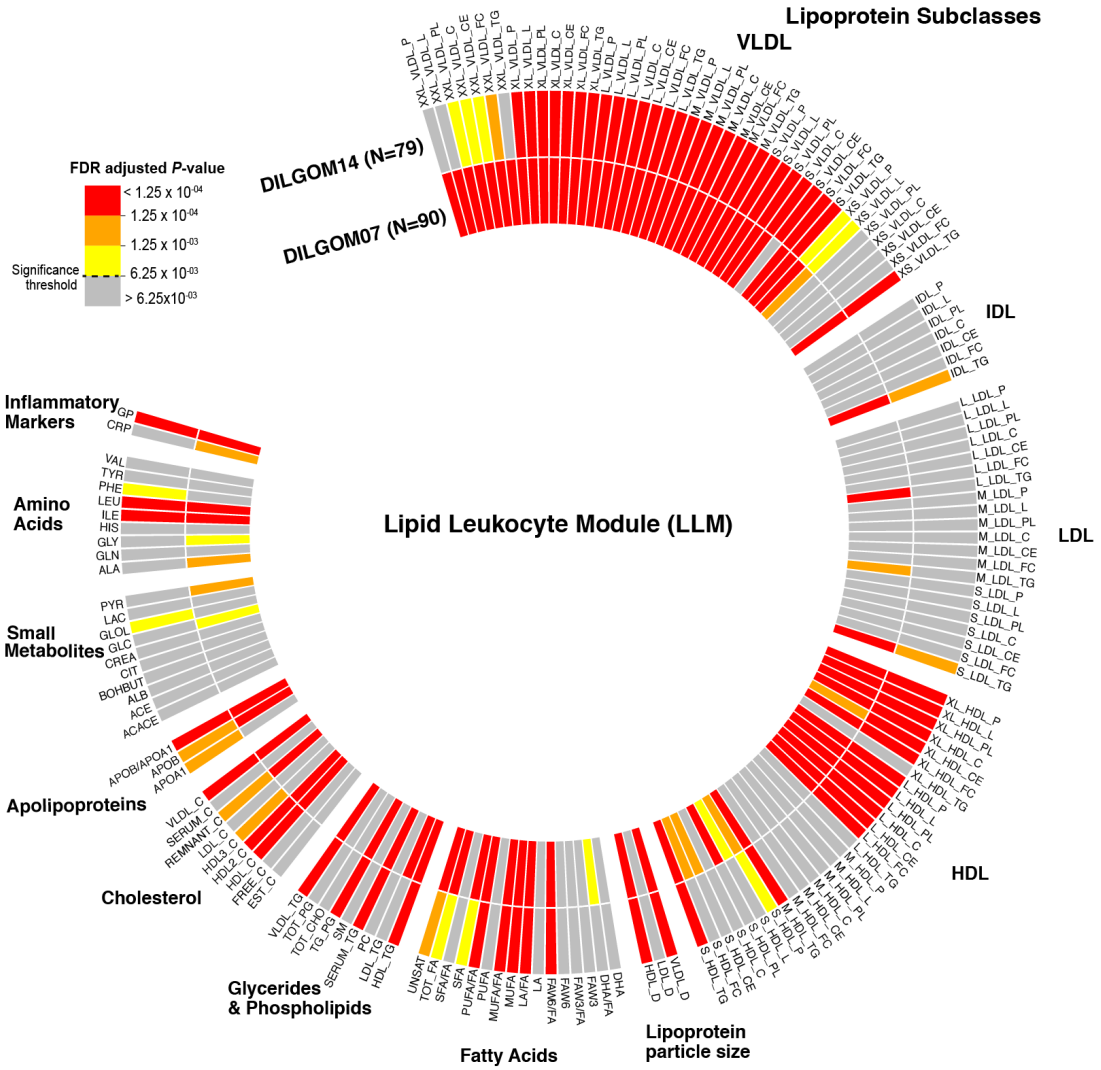


Figure 2.10: Temporally stable metabolite associations with the LLM.

Circular heatmap for association between individual metabolites and the LLM across two time points, DILGOM07 and 7-years later (DILGOM14).

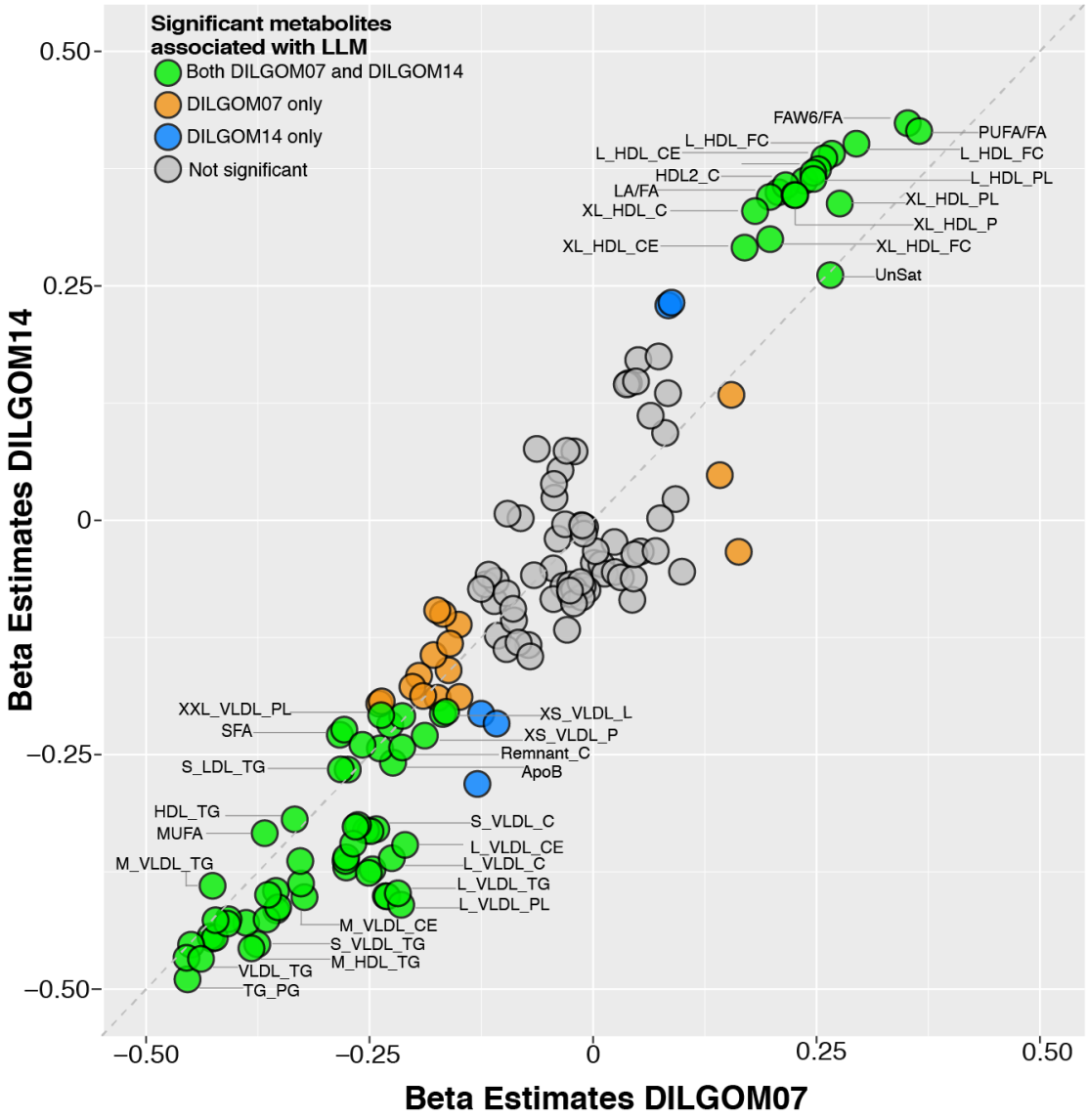


Figure 2.11: Comparison of the effect size estimates of metabolite association with LLM in DILGOM07 and DILGOM14.

The plot shows that the overall association patterns are consistent across the two time-points. Colours denote metabolites that are significantly associated with the LLM in DILGOM07 only (orange), DILOM14 only (blue), and across both time-points (green). The grey dashed line is the x=y line.

2.5 Discussion

This study has utilised over 2,000 individuals to map the immuno-metabolic crosstalk operating in circulation. We have identified and characterised eight robust immune gene modules, their genetic control and interactions with diverse metabolites, including many of clinical significance (e.g. triglycerides, HDL, LDL, branched-chain amino acids). Furthermore, our findings are consistent with and build upon those of previous studies (294,296,310,355,375). In addition to five newly identified gene modules, their mQTLs and metabolite interactions, we have replicated the previously characterised LL module and confirm its association with lipoprotein subclasses, lipids, fatty acids, and amino acids (296,375). Associations between the core genes in the LL module and isoleucine, leucine, and various lipids were also identified independently in the KORA cohort (355). Importantly, we have shown the long-term stability of LL and neutrophil module co-expression and metabolite interactions, and we have greatly expanded the number of known biomarkers associated with the NM from one (GlycA) to 123 (294). Our study has also expanded the widespread *trans* eQTL effects at the *ARHGEF3* locus (310), shows them to be strongly maintained within individuals over time, and further identifies extensive lipoprotein and fatty acid metabolite interactions that may be a consequence of these of these *trans* effects.

Taken together, the findings of this study adds to the growing body of evidence intimately linking the immunoinflammatory response to systemic metabolism, which is consistent with the view that this interplay contributes to myriad complex diseases of the metabolic, cardiovascular, autoimmune and infection aetiologies. Since earlier studies on immunometabolism have mainly been done in the context of disease, using *in-vitro* or non-human design models, focusing on delineating the metabolic configuration in mainly adipocytes or on specific immune mediators, they have provided fragmentary insight on systemic immunometabolism. For example, multiple studies have shown that increased expression of TNF in adipose tissue of both obese rodents and humans plays a critical role in mediating obesity-related insulin resistance (380,467,468). My study addressed this gap by integrating blood transcriptomic and metabolite datasets from two population-based cohorts in a systematic manner to

provide insight into how systemic metabolism integrates with immunological responses. Metabolites of several metabolic pathways were associated with more than one immune-related processes, which is consistent with the notion that complex patterns of interaction exists between inflammatory processes and the underlying metabolic rewiring linked to chronic diseases. Hence, studies exploring immune-targeted therapeutic opportunities for metabolic diseases should not only also focus on single inflammatory mediators, but rather also on immune response pathways. In the case of targeting single inflammatory molecules, a few biologic (therapeutic intervention) studies have also yielded variable results. For instance, studies have found that treatment with anti-TNF drug enhances insulin sensitivity and reduces the risk of diabetes, while others have failed to reach the same conclusions (469,470).

Furthermore, current anti-inflammatory strategies are further evidence for the causative role of inflammatory process in metabolic diseases such as diabetes. Commonly used anti-diabetic drugs such as metformin and thiazolidinedione, as well as exercise has been shown to reduce pro-inflammatory cytokine levels, inflammation, and insulin resistance (471–473). With finer-resolution maps of these interactions, new biomarkers of chronic and acute inflammatory states are likely to emerge. Hence, together with *in vivo* follow-up and interventional studies to modulate metabolite-immune interactions, existing lipid-lowering medications, anti-inflammatory therapies or lifestyle interventions may provide new ways that the immune system itself can be utilised to lessen the burden of cardiometabolic diseases.

Finally, the magnitude of the human immune response indicates a high level of inter-individual variability and 20-40% of this variation is due to genetic diversity (474). As a result, this can lead to varied treatment responsiveness among patients. For example, a recent clinical trial demonstrated that following treatment with an inhibitor of the protein kinase IKKε, only a subgroup of obese diabetic patients responded with a reduced gene expression signature of inflammation and improved insulin sensitivity (475). My study highlights several genetic drivers of immune-related processes and genes, which clinical intervention studies can leverage to characterise sub-populations of individuals who are enriched for specific mQTLs, gene co-expression network levels, or metabolites. The robust integrated map of

immunometabolic relationships and their genetic regulation provided by this study may guide future studies focusing on designing effective therapeutic and preventative approaches for cardiometabolic diseases at the immune-metabolic interface to improve human health. The catalogue of immune-metabolite interactions provided can be explored with targeted experiments to gain insight into underlying disease mechanisms.

Chapter 3

Multivariate genome-wide association analysis identifies eight loci associated with a network of circulating cytokines

3.1 Introduction

The immune system has evolved to provide effective host defence against various threats while maintaining self-tolerance against autoimmunity. Such feature is enabled by cytokines, which are essential regulatory components of the immune system. Their controlled release is important in mediating and regulating an appropriate immune response. The focus of this chapter was to characterise the genetic variants influencing cytokine levels in natural healthy populations. This may provide insight into how inter-individual genetic differences in cytokine levels shape immune responses and subsequently impact disease risks.

3.1.1 Existing gap in understanding the genetic regulation of circulating cytokine levels in population-based studies

Several studies have identified variants in cytokine genes to be associated with circulating cytokine levels (331,333,476). However, these studies have mainly focussed on specific allelic variants located within cytokine gene(s). Others have investigated the effect of certain cytokine gene polymorphisms on disease risk and outcome (329,330).

Importantly, there is a paucity of studies examining the effects of genome-wide variation on cytokine levels in population-based studies. Few population-based GWAS of individual cytokine levels have been performed (340–343,477), however simultaneous assessment of the multiplicity of cytokines is necessary to capture the immune state of the individual. A recent study by Ahola-Olli *et al*. performed a univariate genome-wide scan for loci associated with circulating concentrations of 41 cytokines (344). The study identified 27 loci associated with at least one cytokine, of which 17 were novel (344). Yet, cytokine levels are tightly regulated with the relative levels of both pro and anti-inflammatory cytokines critical to the health of the individual. This tight regulation induces correlations amongst phenotypes, which are rarely considered by genetic association studies.

Simulation studies have previously shown that multivariate analysis of correlated phenotypes can result in increased power to detect genetic associations with small or pleiotropic effects across these phenotypes (478–481). Recent studies have demonstrated increased power empirically, typically using correlations amongst lipids, lipoproteins, and triglyceride levels (482–485). Application of a multivariate test on 4 lipid traits across all combinations of 2, 3, and 4 lipid traits led to the identification of 21% more independent genome-wide significant SNPs compared to the univariate analysis (482). Likewise, simultaneously testing four metabolic traits either in combinations of four or two leads to richer findings over the univariate approach (483). Moreover, complex genotype-phenotype dependencies have been revealed when jointly testing rare variants with lipoprotein traits (484). Of particular relevance to our study, Inouye *et al*. showed that association testing of individual SNPs with networks of highly correlated circulating metabolites increased power to identify additional loci not identified in univariate testing (485). They further identified variants at the top novel multivariate signals as *cis* eQTLs for *SERPINA1* and *AQP9* in multiple tissues. Additionally, they found these genes to be expressed at higher levels in human atherosclerotic plaques.

Hence, exploiting the dependency structure among cytokines jointly through multivariate analysis can provide deeper insight into the shared genetic architecture

between cytokines, giving new perspectives on immune function and disease mechanisms.

3.2 Research objectives

The central aim of this chapter was to leverage the correlation structure between a network of 11 cytokines to perform a multivariate genome-wide scan to identify genetic variants regulating this network in 9,263 healthy individuals from three independent population-based studies.

The specific objectives of this research chapter were:

1. To identify network(s) of correlated cytokines.

2. To identify genetic variants associated with correlated cytokines.

3. To demonstrate that simultaneous analysis of cytokines increases power to detect novel genetic variants.

4. To perform whole blood *cis*- and *trans*-eQTL analyses for lead and tagging GWAS variants.

5. To query the eQTLs identified at the novel loci against the GTex eQTL database to identify if these eQTLs regulate gene expression in a tissue-specific manner.

3.3 Methods

3.3.1 Study populations

Approval for the study protocols for each cohort was obtained from their respective ethics committees and all subjects enrolled in the study gave written informed consent. An overview of the study populations, molecular data, and study design is given in **Figure 3.1**.

The Cardiovascular Risk in Young Finn Study (YFS) cohort used in this analysis was from the year 2007 (YFS07) aged 30, 33, 36, 39, 42 and 45 years. A total of 2,202 individuals had measures available on various physical and clinical variables. In addition, gene expression profiles available for 1,650 individuals from the 2011 follow-up were also analysed. Details of the YFS and specifics on gene expression profiling for this cohort has been described in detail in **Chapter 2**. Ethics were approved by the Joint Commission on Ethics of the Turku University and the Turku University Central Hospital.

The FINRISK cohorts are part of the cross-sectional population-based survey, carried out every 5 years since 1972 to evaluate the risk factors of chronic diseases in the Finnish population (486). Each survey recruits a representative sample of 6,000-8,800 individuals, within the age group of 25-74 years, who are chosen from the national public register. This study utilised samples from the 1997 (FINRISK97) and 2002 (FINRISK02) collections, which recruited individuals from five major regional and metropolitan areas of Finland: the provinces of North Karelia, Northern Savo, and Northern Ostrobothnia and Kainuu; the Turku and Loimaa region of south-western Finland; and the Helsinki and Vantaa metropolitan area. In total, 8,444 (aged 24 – 74 years) and 2,775 (aged 51 – 74 years) individuals were recruited in the FINRISK97 and FINRISK02 studies, respectively. Ethics were approved by the ethical committee of the National Public Health Institute, Finland.

The Dietary, Lifestyle, and Genetic determinant of Obesity and Metabolic syndrome (DILGOM) study, a sub-sample of FINRISK aged between 25 – 75 years, recruited from the Helsinki region of Finland, was conducted in 2007. Gene expression profiles were available for 518 individuals, details of which have been described in **Chapter**

2. Ethics approval was given by the Coordinating Ethical Committee of the Helsinki and Uusimaa Hospital District.

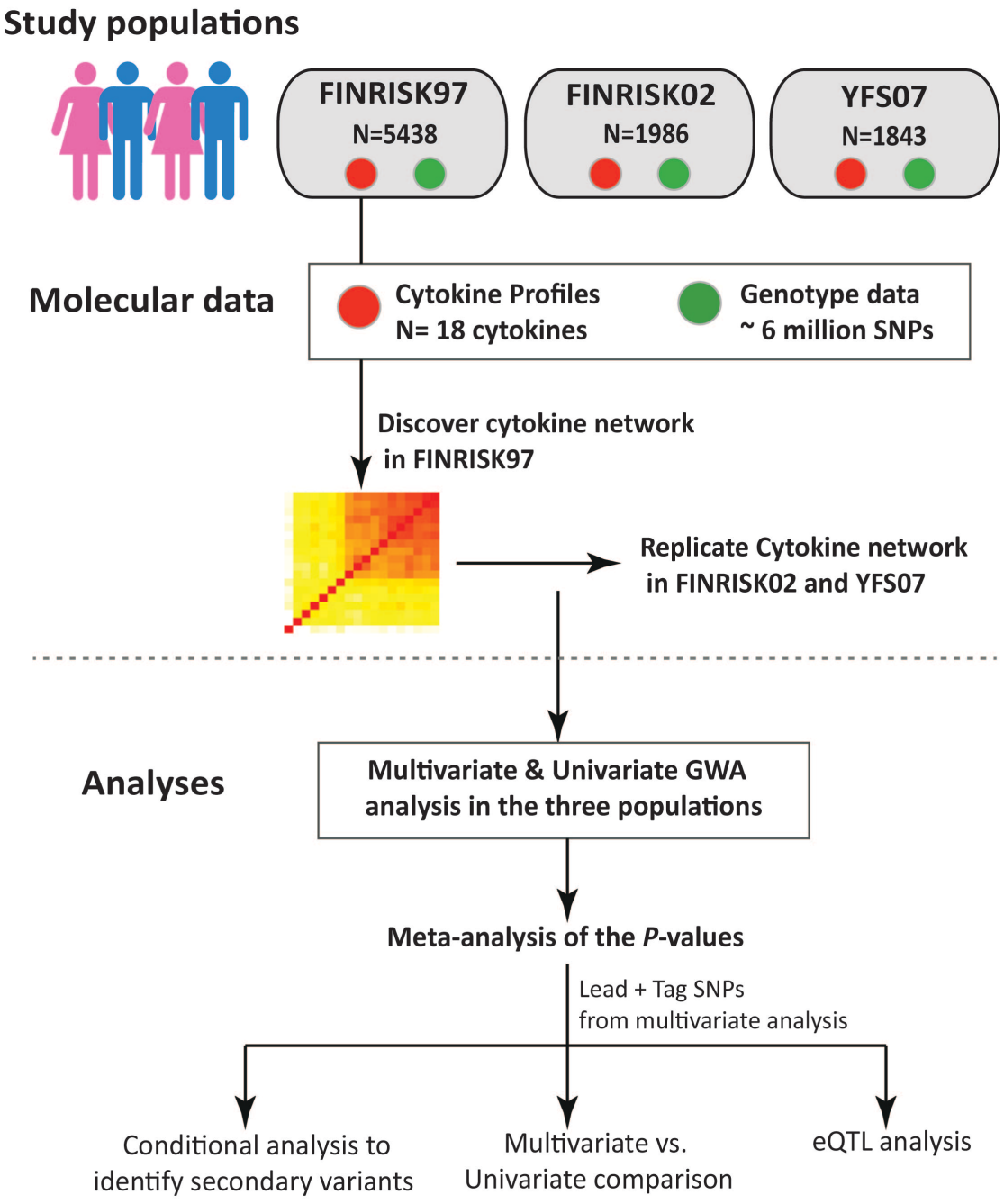


Figure 3.1: Overview of the study populations, design, and the analyses conducted.

3.3.2 Blood sample collection

Blood samples and detailed information on various physical and clinical variables for the YFS and FINRISK cohorts were collected using similar protocols as described previously (389,486). Venous blood was collected following an overnight fast for the YFS cohorts and non-fasting for FINRISK cohorts. Samples were centrifuged, the resulting plasma and serum samples were aliquoted into separate tubes and stored at −70°C for analyses.

3.3.3 Genotype processing and quality control

Genotyping in YFS and the FINIRISK cohorts was performed on whole blood genomic DNA. For YFS07 (N=2,442), a custom 670K Illumina BeadChip array was used for genotyping. For FINRISK97 (N=5798), individuals were genotyped on the Human670-QuadCustom Illumina BeadChip platform. Genotyping in FINRISK02 (N=5988) was performed with the Human670-QuadCustom Illumina BeadChip (N=2447) and the Illumina Human CoreExome BeadChip (N=3541). The Illuminus clustering algorithm was used for genotype calling (392) and quality control was performed using the Sanger genotyping quality control (QC) pipeline. This included removing SNPs and samples with > 5% genotype missingness followed by the removal of samples with gender discrepancies. Genotypes were then imputed with IMPUTE2 (393) using the 1000 Genomes Phase 1 version 3 of the reference panel followed by removal of SNPs with call rate < 95%, imputation "info" score < 0.4, minor allele frequency < 1%, and Hardy-Weinberg equilibrium $P < 5 \times 10^{-6}$. Overlapping SNPs were merged in PLINK (400) where multiple genotyping platforms were used. A total of 6,664,959, 7,370,592 and 6,639,681 genotyped and imputed SNPs passed quality control in YFS, FINRISK97 and FINRISK02, respectively. Cryptic relatedness was assessed using identity by descent (IBD) estimates and in cases where the pi-hat relatedness > 0.1, one of the two individuals was randomly removed (N = 44 for YFS, N=291 for FINRISK97, and N=39 for FINRISK02). Genetic PCs were obtained through principle component analysis (PCA) using FlashPCA (168) on ~60,000 LD pruned SNPs.

3.3.4 Measurement of cytokines, chemokines and growth factors (referred to as cytokines)

Cytokine concentrations were measured in serum (YFS07), EDTA plasma (FINRISK97), and heparin plasma (FINRISK02) using multiplex fluorescent bead based immunoassays (Bio-Rad). A total of 48 cytokines were measured in YFS07 (N=2,200) and FINRSK02 (N=2,775) using two complementary array systems: the Bio-Plex ProTM Human Cytokine 27-plex assay and Bio-Plex ProTM Human Cytokine 21-plex assay. For FINRISK97, 19 cytokines were assayed on the Human Cytokine 21-plex assay system. All assays were performed in accordance with the manufacturer's instructions except that the amount of beads, detection antibodies, and streptavidin-phycoerythrin conjugate were used at half of their recommended concentration. Fluorescence intensity values determined using the Bio-Rad's Bio-Plex 200 array reader were converted to concentrations from the standard curve generated by the Bio-Plex^{TM} Manager 6.0 software. For each cytokine, a standard curve was derived by fitting a five-parameter logistic regression model to the curve obtained from standards provided by the manufacturer. Cytokines with concentrations at the lower and upper asymptotes of the sigmoidal standard curve were set to the concentration corresponding to the fluorescent intensity 2% above or below the respective asymptotes.

3.3.5 Cytokine data filtering, normalisation and clustering

The analysis was limited to 18 cytokines (**Table 3.1**) assayed in all three cohorts used in this study. Although Interleukin 1 receptor, type I (IL-1Ra) was assayed in all three cohorts, it was excluded from the analyses due to inconsistent measurement across the datasets.

Before normalisation, cytokine data was subsetted to individuals with matched genotype data, YFS07 (N=2018), FINRISK97 (N=5728), and FINRISK02 (N=2775). Individuals in YFS07 reporting infection with fever in the two weeks prior to collection were also excluded. To identify extreme outlier samples, PCA was performed on the log2 transformed cytokine values using the "missMDA" R package (487), which first imputes the missing cytokine values using a regularised iterative

PCA algorithm implemented in the "imputePCA" function before performing PCA. 3 and 2 outlier samples were removed from the FINRISK97 and FINRISK02 datasets, respectively. Based on IBD analysis described above N=44 (YFS07), N=291 (FINRISK97), and N=39 (FINRISK02) individuals were further removed. After filtering, a total of N=1,843, N=5,434 and N=1,986 individuals who passed quality control in YFS07, FINRISK97 and FINRISK02, respectively, were utilized for downstream analysis.

Since all the 18 cytokines displayed non-Gaussian distribution, normalisation was necessary. For YFS07, the lower limit of detection (LOD), the lowest concentration of a cytokine that can be measured, was available for each cytokine. Values reported below the LOD are highly unreliable as they could be likely due to background noise signals or instrument error (488). Treating them incorrectly could introduce biases in down stream analysis; hence they were treated as missing. For the FINRISK97 and FINRISK02 datasets, the detection limits were not available, however, it was observed that the leftmost peak of cytokines in these two datasets exhibiting a bimodal distribution pattern comprised primarily of values that were below the LOD. Individuals in the leftmost peak were set to missing. The log2-transformed cytokine values were then normalised to follow Gaussian distribution (mean of 0 and sd of 1) using ranked-based inverse normal transformation function (rntransform) implemented in the GenABEL R package (489). For each study group, residuals for all the cytokines were calculated by regressing the normalised cytokine values on age, sex, BMI, lipid and blood pressure medication, pregnancy status (FINRISK97), and the first 10 genetic PCs using a multiple linear regression model.

Detection of groups of correlated cytokines was done in FINRISK97, the cohort with the largest sample size. Pairwise Pearson correlation coefficients calculated between the residuals of 18 cytokines were hierarchically clustering with 1 minus the absolute correlation coefficient given as the dissimilarity matrix. A group of 11 cytokines, moderate to highly correlated ($r > 0.57$), was identified as the cytokine network to use in the multivariate analysis.

Table 3.1: Cytokine characteristics for the YFS07, FINRISK97 and FINRISK02 cohorts

Cytokine symbol	Cytokine name	YFS07	FINRISK97	FINRISK02
		Median concentration (interquartile range)		
Eotaxin	Eotaxin	115.5 (90.6–148.6)	68.3 (52.7–88.9)	43.9 (27.8–75.3)
FGF-Basic/ bFGF/FGF2	Basic fibroblast growth factor	66.5 (57.1–78.2)	23.8 (13.0–39.9)	31.3 (22.5–44.8)
G-CSF/ CSF3	Granulocyte-colony stimulating factor	136.4 (117.9–157.3)	128.6 (79.7–193.3)	41.62 (31.1–55.9)
HGF	Hepatocyte growth factor	505.9 (405.4–644.6)	324.4 (265.3–403.7)	462.1 (375.9–569.8)
IFN-γ	Interferon-gamma	262.9 (224.4–308.6)	82.4 (49.6–130.0)	47.2 (36.3–65.8)
IL-4	Interleukin 4	11.4 (10.4–12.5)	3.8 (2.7–5.2)	1.0 (0.77–1.42)
IL-6	Interleukin 6	11.7 (10.1–13.6)	10.36 (7.5–14.0)	5.7 (4.23–9.16)
IL-10	Interleukin 10	18.7 (13.2–24.6)	1.9 (0.9–3.3)	5.3 (3.6–7.9)
IL-12p70	Interleukin 12 heterodimer consisting of p35 and p40	66.1 (47–90.5)	19.35 (11.7–31.1)	19.8 (13.6–30.1)
IL-17	Interleukin 17	266.3 (229.1–311.5)	54.6 (27.5–93.1)	34.1 (25.1–50.3)
IL-18	Interleukin 18	65.4 (50–85.8)	245.1 (191.4–313.3)	197.1 (152.8–258.2)
MCP-1/ CCL2	Monocyte chemoattractant protein-1	32.6 (26.6–40.5)	25.1 (20.6–30.7)	84.7 (67.8–103.9)
MIP-1b/ CCL4	Macrophage inflammatory protein-1beta	85.5 (68.8–104.6)	52.1 (41.1–65.7)	63.9 (53.1–80.1)
PDGF-BB	Platelet derived growth factor BB	8,526 (6,886–10,500)	924.4 (470.9–1653.0)	447.3 (313.2–616.9)
SCF	Stem cell factor	90.8 (74.1–109.3)	108.4 (90.7–128.2)	273.1 (224.7–328.1)
SDF-1a/ CXCL12	Stromal cell derived factor - 1alpha	70 (54.5–88.5)	113.4 (74.1–161.1)	81.3 (61.6–101.5)
TRAIL/ TNFSF10	TNF–related apoptosis inducing ligand	133.2 (96.8–172.5)	92.9 (62.8–130.7)	179.4 (141.7–224.3)
VEGF-A	Vascular endothelial growth factor A	69.2 (48.5 – 103.6)	14.7 (6.3–29.7)	36.0 (26.3–49.0)

Values are reported in pg/ml

3.3.6 Statistical Analysis

Univariate association analysis was carried out using linear regression model in PLINK version 1.90 software (https://www.cog-genomics.org/plink2) (400), where the residuals of each cytokine were regressed onto each SNP. *P*-values at each marker across three datasets were combined using the METAL software program (401), which implements a weighted Z-score method.

Multivariate testing (MV) was performed under the canonical correlation (CCA) framework implemented in PLINK (MV-PLINK) (478), which extracts the linear combination of traits most highly correlated with the genotypes at a particular SNP. The test is based on Wilks' Lambda ($\lambda = 1-\rho^2$), where ρ is the canonical correlation coefficient between the SNP and cytokine network. Corresponding *P*-values were computed by transforming Wilks' Lambda to a statistic that approximates an F distribution and the loadings for each cytokine shows their individual contribution towards the multivariate association result (478). The multivariate *P*-values were combined using the weighted Z-score method (490,491) implemented in the "metap" R package. Briefly, the *P*-values for each dataset were transformed into z-scores, weighted by their respective sample sizes and the sum of these weighted z-scores were then divided by the square root of the sum of squares of the sample size for each study. The combined weighted Z-score obtained was back transformed into a one-tailed *P*-value.

To assess the inflation of the test statistics as a result of population structure, quantile-quantile (Q-Q) plots of observed *vs.* expected $-\log_{10}$ *P*-values were generated from the multivariate analysis done on the three datasets separately and meta-analysed. Corresponding genomic inflation factor (λ) was calculated by taking the ratio of the median observed distribution of *P*-values to the expected median.

3.3.7 Gene expression profiling and expression quantitative trait loci (eQTL) analysis

An eQTL meta-analysis was performed for 1,916 peripheral blood samples from two cohorts, DILGOM (N= 515) and YFS (N=1,401). Details of gene expression profiling

and data pre-processing for both the cohorts are described in **Chapter 2.** After pre-processing, a total of 35,425 (for DILGOM) and 37,115 (for YFS) probes were available.

For eQTL analysis, matching SNP and gene expression data was available for a 515 and 1,401 unrelated individuals for DILGOM and YFS, respectively. First, all proxy SNPs in linkage disequilibrium (r2 > 0.5) with each of the 8 multivariate index SNPs and 4 conditional lead SNPs were retrieved. Then, Matrix eQTL (312), an R package, was used to search for eQTLs. Both cis (SNP-Probe distance < 1Mb) and trans (SNP-Probe distance > 5MB) eQTL mapping was conducted using an additive linear model, where the normalised and scaled probe intensities were regressed on each SNP genotype while adjusting for sex, age, and the first 10 genetic PCs as covariates in the model. The test statistics from the cohort-level association analysis was then combined in a meta-analysis using the fixed-effects inverse variance method implemented in the "meta" R package. The combined *P*-values were FDR adjusted for multiple tests using the Benjamini-Hochberg procedure (213). To assess significance, a total of 7,808 *cis* and 10,793,892 *trans* associations were tested, respectively. For both *cis* and *trans* associations, only probe-SNP associations with cohort-level *P*-value < 0.05 and FDR adjusted meta-*P*-value < 0.05 were considered as significant.

3.4 Results

3.4.1 Summary of cohorts and data

Multivariate genome-wide association scans were performed on a total of 9,267 healthy individuals enrolled in three population-based studies, YFS07 (N=1843), FINRISK97 (N=5438), and FINRISK02 (N=1986), which all had genotype data and cytokine profile measurements of 18 cytokines (**Table 3.1**). Characteristics of the study cohorts are summarised in **Table 3.2.** Genotypes for the three datasets were imputed with IMPUTE2 (393) using the 1000 Genomes Phase 1 version 3 of the reference panel. After quality control, a total of 6,022,229 imputed and genotyped SNPs common across the datasets were available. Cytokine levels were measured in serum and plasma using Bio-Plex ProTM Human Cytokine 27-plex and 21-plex assays, then subsequently normalised and adjusted for covariates including age, sex, BMI, pregnancy status, blood pressure treatment, lipid treatment, and population substructure (see Methods). Whole blood gene expression profiles from the Illumina HT-12 array were available for YFS11 and DILGOM07. An overview of the study workflow is shown in (**Figure 3.1**).

Table 3.2: Characteristics of the study population.

Characteristics	FINRISK97	FINRISK02	YFS07
Collection year	1997	2002	2007
Individuals with matched cytokine & genotype data	5438	1986	1843
Male, n (%)	2637 (48.5)	991(49.9)	841 (45.6)
Mean age in years (and range)	47.6 (24-74)	60.3(51-74)	37.7 (30-45)
BMI: mean ± SD kg/m^2	26.6 ± 4.6	28.1 ± 4.5	25.9 ± 4.6
Individuals on lipid lowering drugs	174	284	40
Individuals on blood pressure treatment drugs	698	512	127

3.4.2 Identification of the cytokine network

To identify groups of correlated cytokines for multivariate association analysis, I utilised normalised cytokine residuals for the FINRISK97 cohort, the largest dataset. Hierarchical clustering was applied to the pair-wise Pearson correlation coefficients calculated between the residual values of the 18 cytokines. I identified a cytokine network containing 11 cytokines, with median cytokine-cytokine Pearson correlation of 0.75 (range r=0.57–0.9; **Figure 3.2**). While cytokine-cytokine correlations overall were lower in YFS07 and FINRISK02, the cytokine network was still distinct with moderate median intra-module correlations (r=0.42 and r=0.46, respectively; **Figure 3.3**).

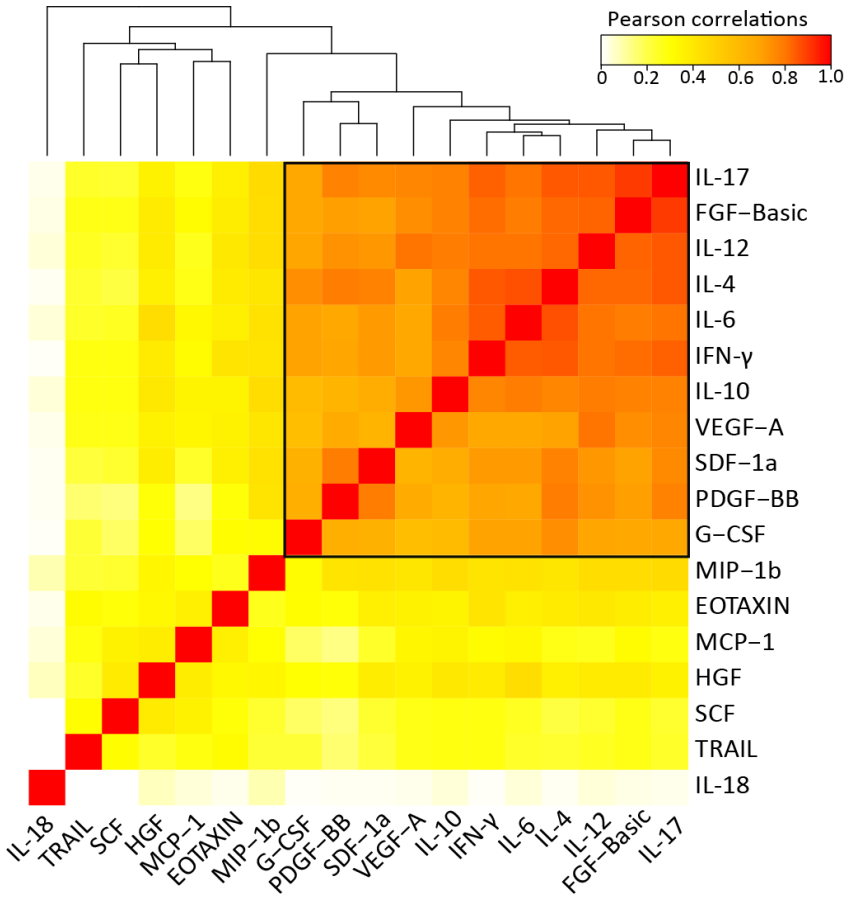


Figure 3.2: Correlation heatmap of the 18 cytokines in the FINRISK97 cohort.

Each cell presents the pair-wise Pearson's correlation coefficient between the normalised cytokine residuals. The cytokines are ordered by hierarchical clustering, using 1 minus the absolute value of the correlations as the distance matrix. The colour scale denotes the strength of the correlations, where red is a high positive correlation. The group of 11 tightly correlated cytokines (black box) was used for multivariate analysis

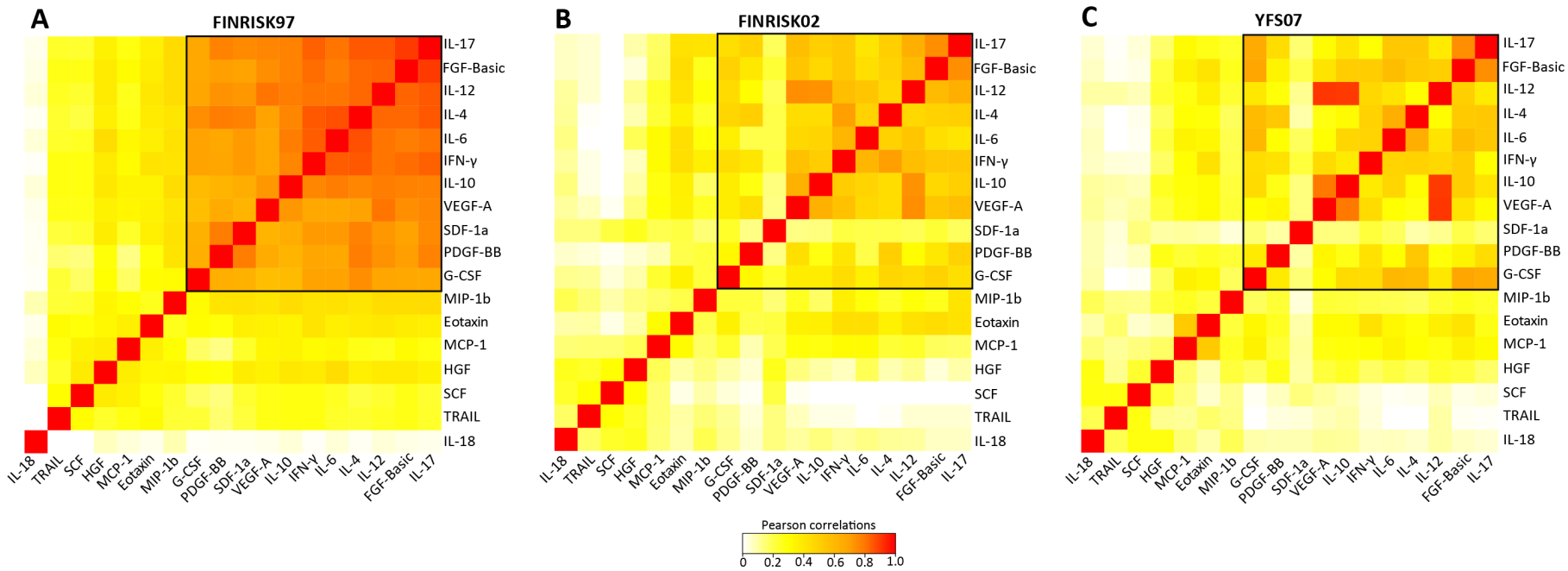


Figure 3.3: Comparison of cytokine-cytokine correlation in FINRISK07, FINRISK02, and YFS07.

The heatmaps show the correlations between the normalised cytokines residuals in the discovery dataset, **(A)** FINRISK97, and the replication datasets, **(B)** FINRISK02 and **(C)** YFS07. Each square represents the Pearson's correlation coefficient between the cytokines. The black box shows the correlation patterns among the 11 correlated cytokines (discovered using the FINRISK97) across the three datasets. The correlation matrix in FINRISK07 was hierarchically clustered using distance as 1 minus the absolute value of the correlations. The ordering of rows and columns in FINRISK02 and YFS07 was defined by the ordering in FINRISK07. The strength of the correlations is indicated by the colour on the scale.

The composition of the cytokine network included both anti-inflammatory (IL-10, IL-4, IL-6) and pro-inflammatory (IL-12, IFN-γ, IL-17) cytokines as well as growth factors (FGF-basic, PDGF-BB, VEGF-A, G-CSF) and a chemokine (SDF-1a) involved in promoting leukocyte extravasation and wound healing (118,492,493). These cytokines were positively correlated with each other suggesting counter-regulatory mechanisms exist between the pathways that release pro-inflammatory and anti-inflammatory cytokines.

3.4.3 Multivariate genome-wide association analysis for cytokine loci

A multivariate GWAS was performed on the cytokine network in each cohort separately, and then cohort-level results were combined using meta-analysis (see Methods)). Since one hypothesis test (corresponding to the cytokine network) was performed for each SNP, a genome-wide significance threshold of $P < 5 \times 10^{-8}$ was used. Minimal inflation was observed for the cohort-level and meta-analysis test statistics with lambda (λ) inflation ranging between 1.00-1.02 (**Figure 3.4**). The meta-analysis identified 8 distinct genomic loci (562 SNPs in total) reaching genome-wide significance for association with the cytokine network (**Figure 3.5; Table 3.3**).

The strongest association was seen with rs7767396 (meta-*P*-value = 6.93×10^{-306}), a SNP located 172kb downstream of vascular endothelial growth factor A (*VEGFA*) gene on chromosome 6p21.1 (**Figure 3.6A; Table 3.3**). The *VEGFA* locus was previously identified in univariate analyses as associated with cytokine levels including VEGF-A, IL-7, IL-10, IL-12, and IL-13 (343,494). Our multivariate GWAS detected other loci previously associated with levels of at least one cytokine present in our cytokine network (343,494,495). This includes loci at *SERPINE2* (rs6722871; meta-*P*-value = 1.19×10^{-59}), *ZFPM2* (rs6993770; meta-*P*-value = 4.73×10^{-08}), *VLDLR* (rs7030781; meta-*P*-value = 3.78×10^{-13}), and *PCSK6* (rs11639051; meta-*P*-value = 1.93×10^{-58}) (**Figure 3.6B-E; Table 3.3**). The F5 locus was also associated with the cytokine network (rs9332599; $P_{meta} = 7.17 \times 10^{-12}$) (**Figure 3.6F; Table 3.3**) and has been previously associated with cytokines, stem cell factor (SCF) and stem cell growth factor beta (SCGF-b) (494), which are not part of the cytokine network identified in this chapter.

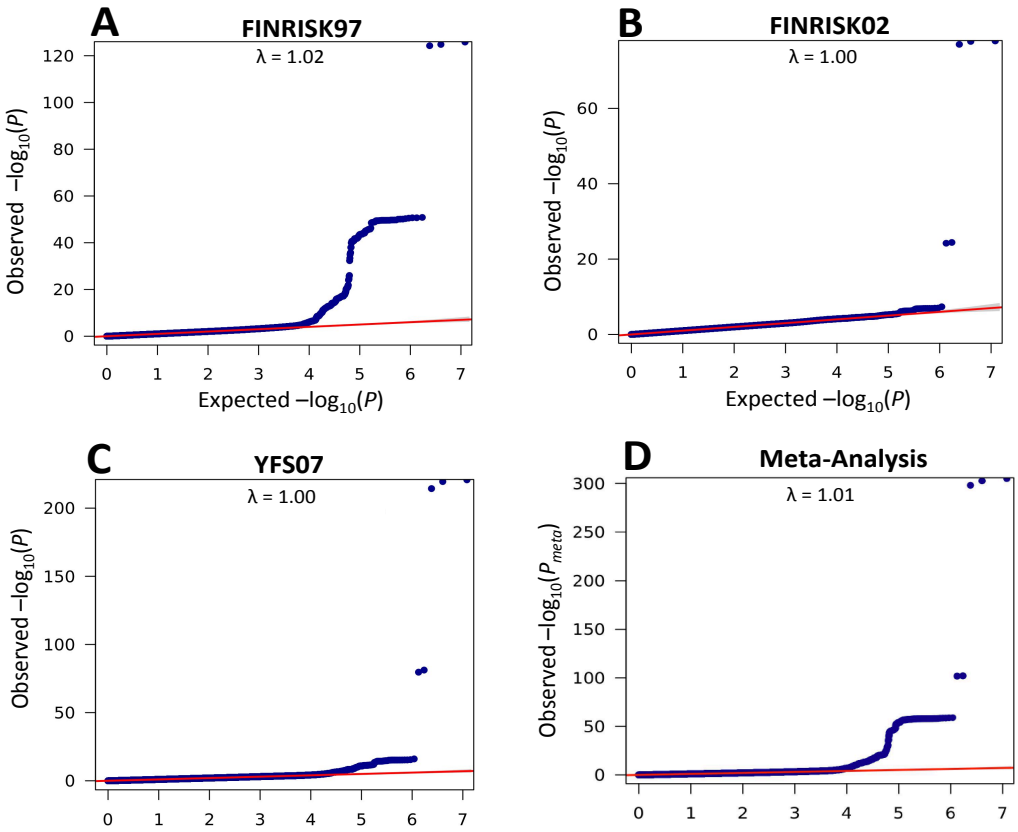


Figure 3.4 Quantile-quantile (Q-Q) plots resulting from the multivariate GWAS in the three cohorts and meta-analysis.

Q-Q plots of observed (*y*-axis) *vs.* expected *P* values (*x*-axis) for each SNP from the multivariate genome-wide association in **(A)** FINRISK97, **(B)** FINRISK02, **(C)** YFS07, and **(D)** Meta-analysis. The diagonal red line (*y*=*x*) indicates null hypothesis of no association. The inflation factor (λ) was between 1.0-1.2 suggesting that inflation from population substructure or other confounders was appropriately adjusted for.

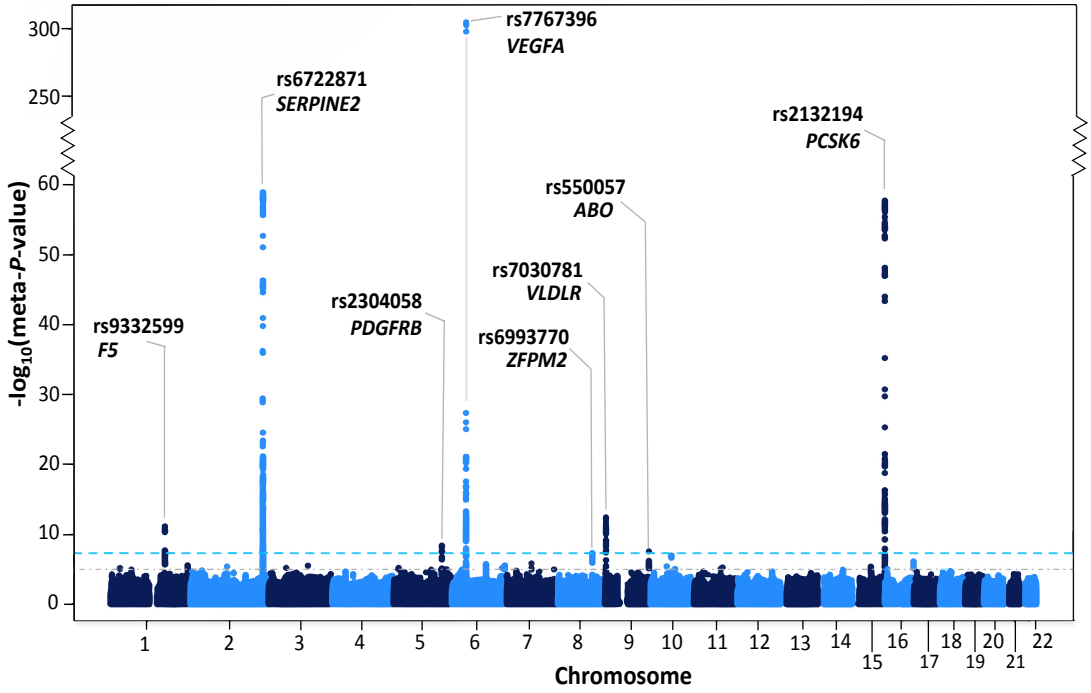


Figure 3.5: Manhattan plot for meta-analysis results from the multivariate genome-wide association analysis of the cytokine network.

The statistical strength of association ($-\log_{10}$ meta-*P*-value) is plotted against all the SNPs ordered by chromosomal position. The sky-blue and grey horizontal dashed lines represent the thresholds for genome-wide (meta-*P*-value < 5×10^{-08}) and suggestive significance (meta *P*-value < 1×10^{-05}), respectively. The lead SNP (lowest meta-*P*-value) at each locus and the nearby genes are shown.

Table 3.3: Meta-analysed results from the multivariate and univariate GWA analysis of the cytokine network and individual cytokines in the cytokine network, respectively.

Locus	Locus Region	Top SNP	Average MAF	Top Multivariate Meta-*P*-value	Univariate Meta-*P*-value (Top Cytokine)	Detection
F5	1q24.2	rs9332599	0.294	7.17×10^{-12}	9.21×10^{-03} (SDF1a)	Multivariate
SERPINE2	2q36.1	rs6722871	0.311	1.19×10^{-59}	3.55×10^{-18} (PDGF-BB)	Both
PDGFRB	5q32	rs2304058	0.379	4.06×10^{-09}	1.52×10^{-05} (IL4)	Multivariate
VEGFA	6p21.1	rs7767396	0.471	6.93×10^{-306}	3.10×10^{-201} (VEGF-A)	Both
ZFPM2	8q23.1	rs6993770	0.221	4.73×10^{-08}	1.01×10^{-07} (IL12p70)	Multivariate
ABO	9q34.2	rs550057	0.306	2.75×10^{-08}	4.9×10^{-03} (IL4)	Multivariate
VLDLR	9p24.2	rs7030781	0.413	3.78×10^{-13}	6.78×10^{-14} (VEGF-A)	Both
PCSK6	15q26.3	rs11639051	0.255	1.93×10^{-58}	1.19×10^{-26} (PDGF-BB)	Both
JMJD1C	10q21.3	rs9787438	0.374	*1.30×10^{-07}	*8.96×10^{-12} (VEGFA)	Univariate

The table shows the meta-analysis *P*-values for the top SNP (lowest *P*-value) at each locus associated with the cytokine network in the multivariate analysis at genome-wide significance threshold (5×10^{-08}). The corresponding lowest meta-*P*-value for the same top SNP in the univariate analysis with any single cytokine present in the cytokine network, given in brackets beside the meta-*P*-value, was also reported. *Instance where the top SNP at a locus crossed only the univariate significance threshold ($P < 4.55 \times 10^{-09}$), then the corresponding meta-*P*-value for that SNP in the multivariate was also given. The univariate significance threshold was calculated from a Bonferroni correction of 5×10^{-8}/ 11 cytokines tested.

In addition to these, multivariate GWAS identified a further two loci not shown to be associated with cytokine levels in our current univariate testing of the 11 cytokines or previous GWAS studies. These included rs2304058 (meta-*P*-value = 4.06 x 10^{-09}), which is located within the tenth intron of the platelet-derived growth factor receptor-beta (*PDGFRB*) gene on chromosome 5q32, and rs550057 (meta-*P*-value = 2.75 x 10^{-08}), situated within the first intron of the *ABO* gene on chromosome 9q34.2 (**Figure 3.6G-H; Table 3.3**).

3.4.4 Conditional analysis revealed multiple independent signals

To find independent signals at each locus associated with the cytokine network, I performed a stepwise conditional multivariate meta-analysis on the lead SNPs at each of the 8 loci (see Methods)). Three loci exhibited evidence of multiple independent variants associated with the cytokine network (**Table 3.4**), including two intergenic SNPs at the *SERPINE2* (rs55864163; conditional-meta-*P*-value = 9.03 x 10^{-29}) and *VEGFA* (rs4714729; conditional-P_{meta} = 7.49 x 10^{-10}) locus; a SNP located on PCSK6 (rs6598475, conditional-P_{meta} = 2.63 x 10^{-17}). A third SNP was identified at the *SERPINE2* locus in the second round conditional analyses (rs112215592, conditional-P_{meta} = 2.10 x 10^{-12}). I repeated the conditional analyses in the univariate model with the aforementioned multivariate SNPs and found a secondary signal at the *VEGFA* locus associated with VEGF-A levels (rs4714729; conditional-P_{meta} = 8.8 x 10^{-13}) after one step of the conditional test (**Table 3.4**).

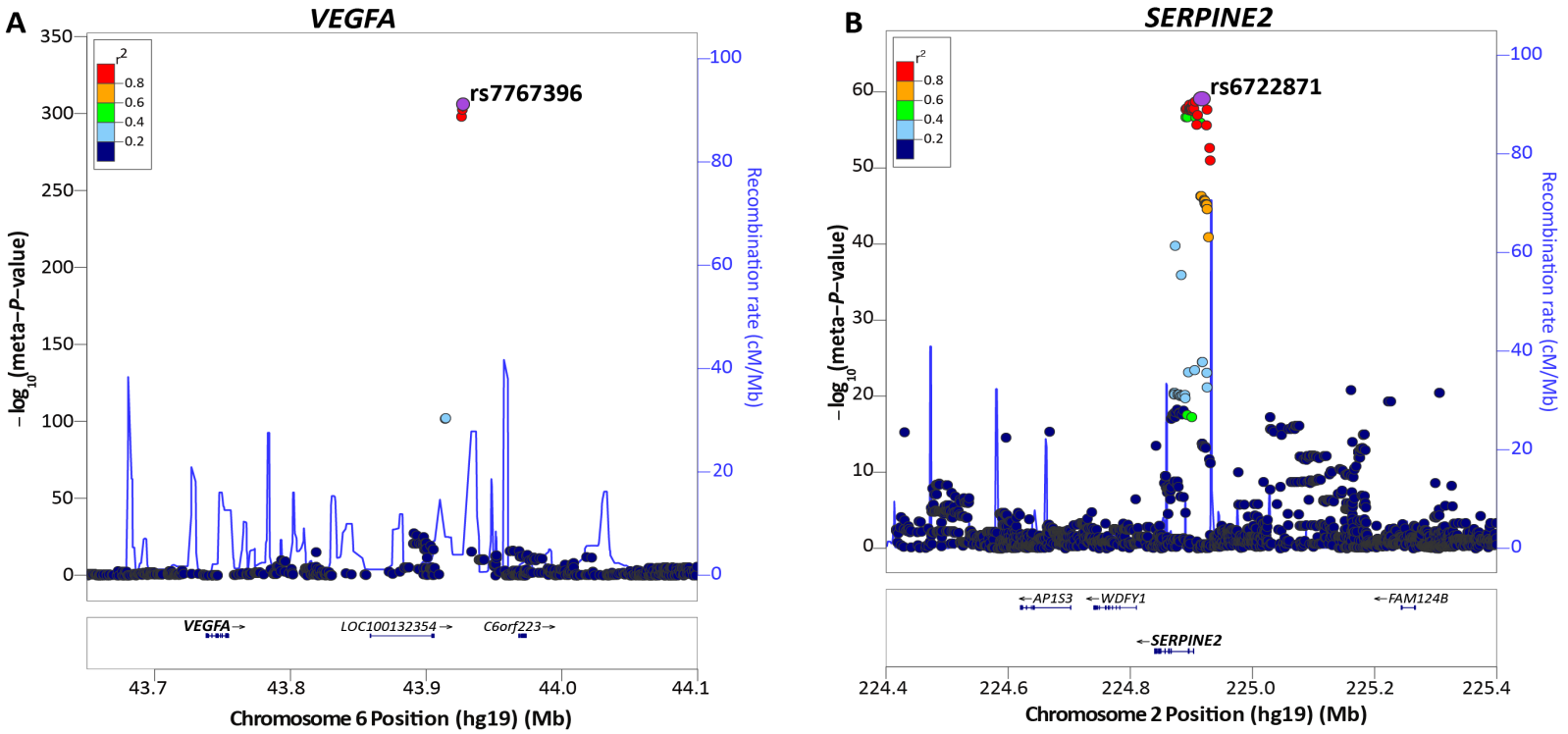
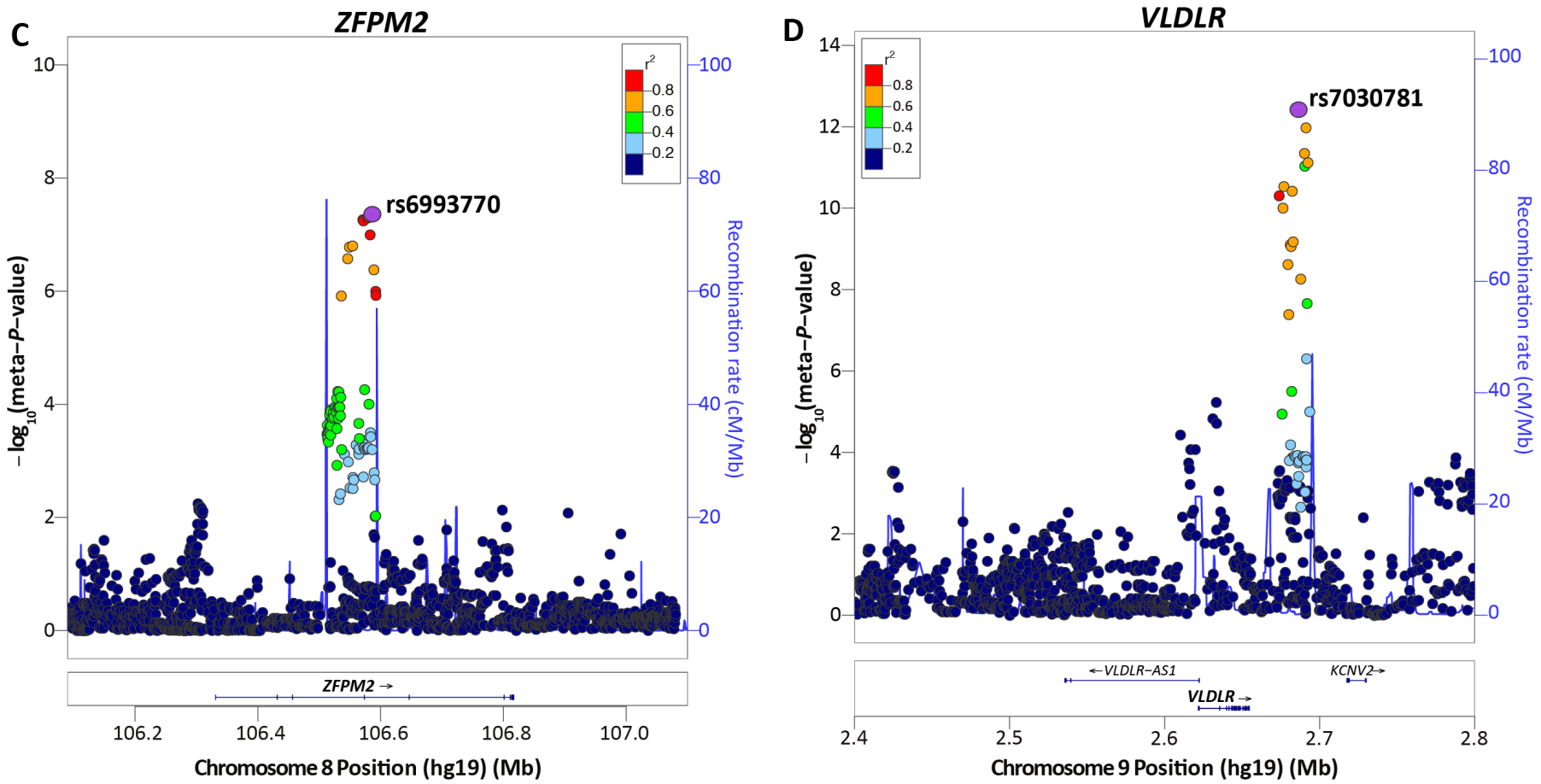
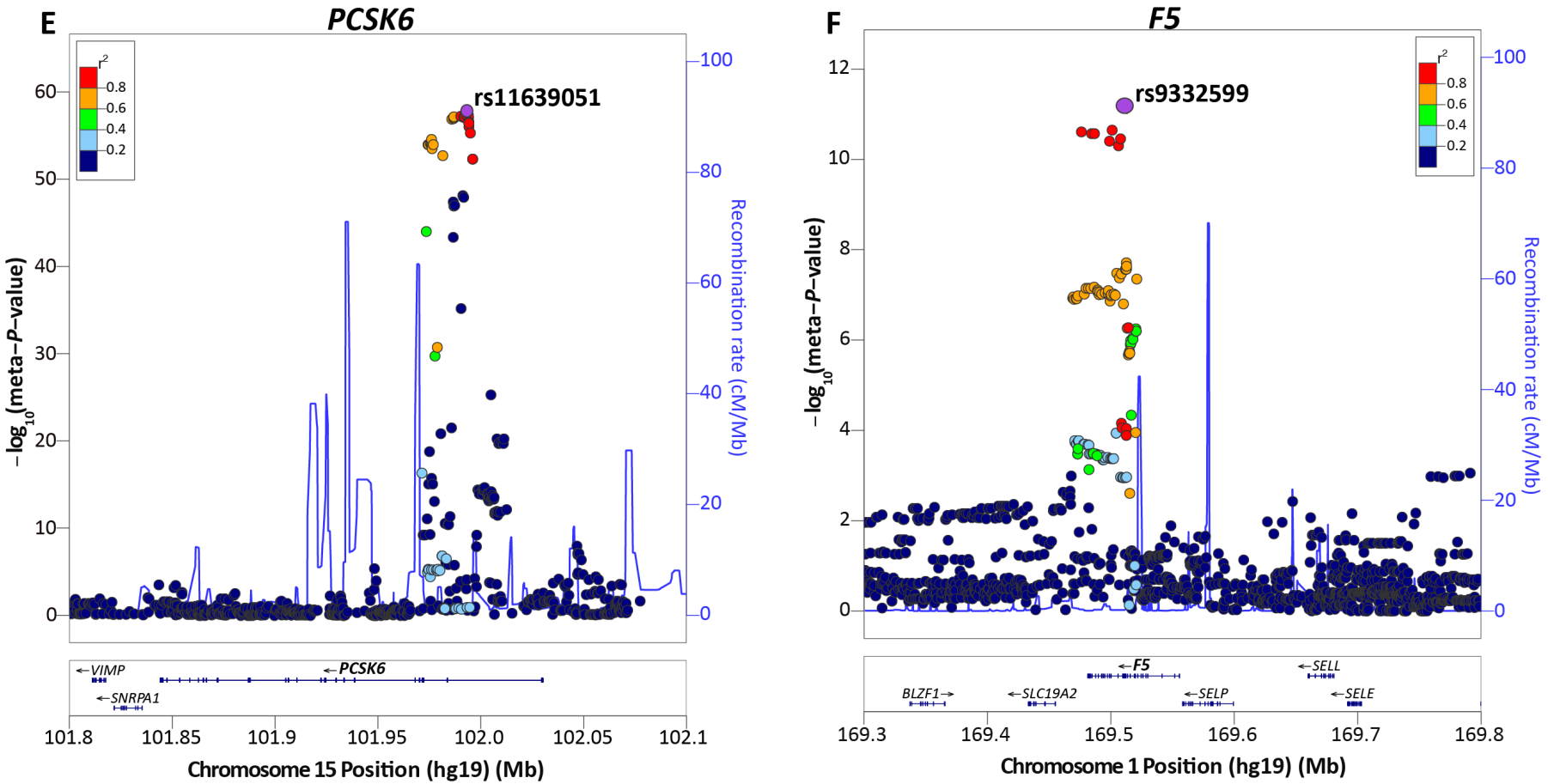


Figure 3.6: Regional association plots for each of the 8 loci associated with the cytokine network from the meta-analysed multivariate GWA analysis

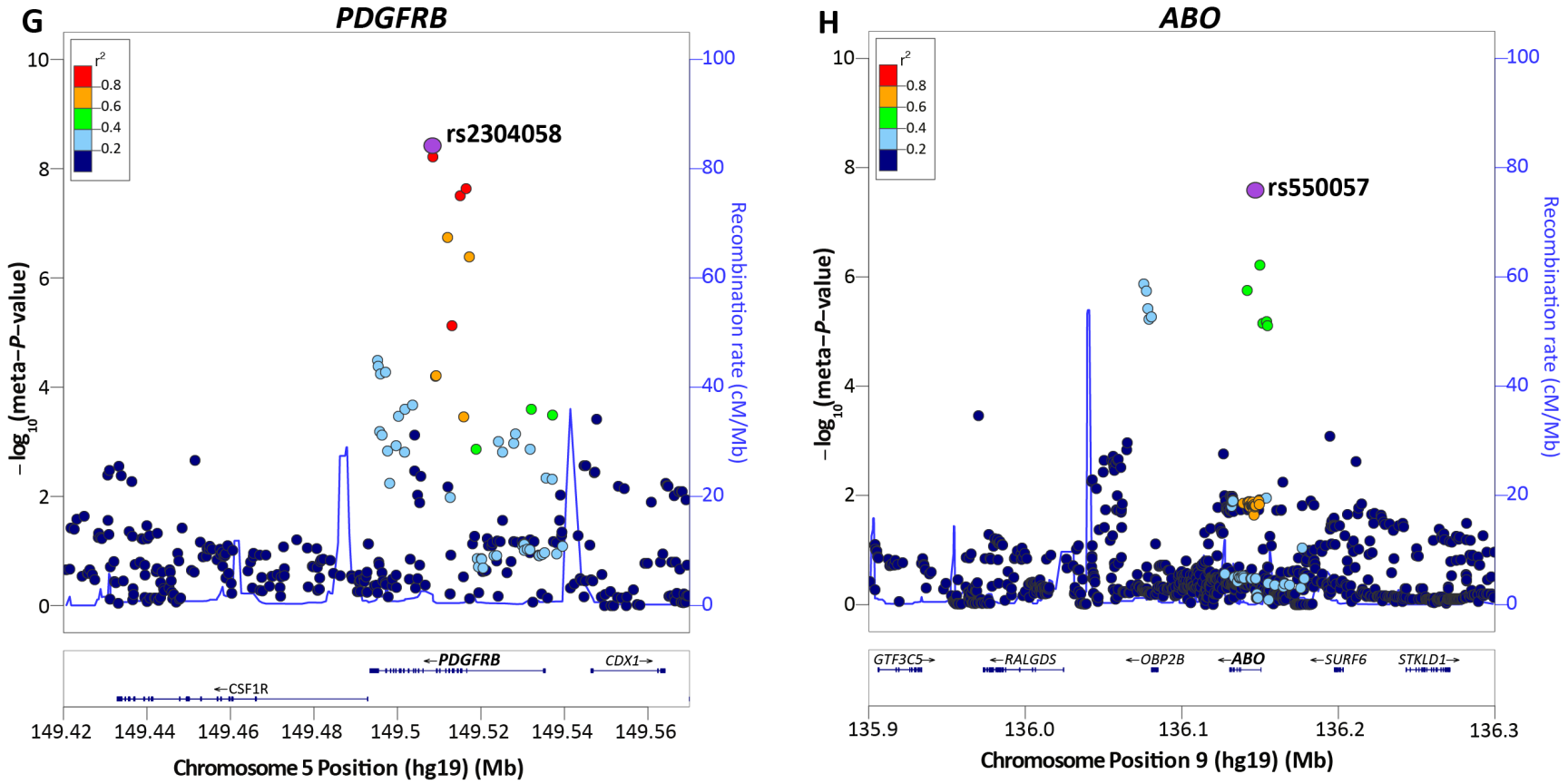
(A) VEGFA locus, rs7767396 is an intergenic SNP located 172.83kb downstream of vascular endothelial growth factor A (*VEGFA*) gene on chromosome 6p21.1. **(B)** SERPINE2 locus, rs6722871 lies 10.9kb upstream of *SERPINE2* on chromosome 2q36.1. For each plot, the circles represent the -$\log_{10}$ meta-analysed *P*-values (y-axis) of SNPs plotted against their chromosomal position (x-axis). The lead SNP in each plot is denoted by a purple circle, and its pairwise LD (r^2) strength with other SNPs in the region, estimated from the "1000 genomes Mar 2012 EUR" population, is indicated by colour. The blue lines indicate the recombination rates. The plots were generated using the LocusZoom online tool (http://locuszoom.sph.umich.edu/locuszoom/).



(C) ZFPM2 locus, rs6993770 lies within intron 4 of the zinc finger protein multitype 2 (*ZFPM2*) gene on chromosome 8q23.1. **(D)** VLDLR locus, rs7030781 is situated ~31.8kb away from the very low-density lipoprotein receptor (*VLDLR*) gene on chromosome 9p24.2.



(E) PCSK6 locus, rs11639051 is located in the second intron of *PCSK6* (proprotein convertase subtilisin/kexin type 6) on chromosome 15q26.3. **(F)** F5 locus, rs9332599 is located within intron twelve of factor V (*F5*) gene on chromosome 1q24.2.



(E) PCSK6 locus, rs11639051 is located in the second intron of *PCSK6* (proprotein convertase subtilisin/kexin type 6) on chromosome 15q26.3. **(F)** F5 locus, rs9332599 is located within intron twelve of factor V (*F5*) gene on chromosome 1q24.2.

Table 3.4: Results from the conditional (regional) multivariate and univariate GWA analysis of the cytokine network.

Locus	Locus region	Cond. SNP	Lead SNP after Cond.	A1/ A2	Cond. meta-*P*-value Multivariate	Cond. meta- *P*-value Univariate (Cytokine)	*r^2
Round 1							
SERPINE2	2q36.1	rs6722871	rs55864163	A/G	9.03×10^{-29}	1.25×10^{-7} (PDGF-BB)	0.010
VEGFA	6p21.1	rs7767396	rs4714729	T/C	7.49×10^{-10}	8.84×10^{-13} (VEGF-A)	0.002
PCSK6	15q26.3	rs11639051	rs6598475	G/T	4.08×10^{-17}	7.364×10^{-7} (PDGF-BB)	0.100
Round 2							
SERPINE2	2q36.1	rs6722871, rs55864163	rs112215592	G/A	4.56×10^{-16}	1.819×10^{-6} (PDGF-BB)	0.008

*Linkage Disequilibrium (LD; r^2) was calculated between the lead SNP associated with the cytokine network and secondary lead SNP. LD was calculated in FINRISK97. Cond. – refers to conditional. A1/A2 refer to the minor/major allele in the Finnish population.

3.4.5 Comparison of multivariate and univariate meta-analyses

To directly compare the power of multivariate to univariate GWAS on the cytokine data, we first performed univariate analysis in each dataset by regressing each of the 11 cytokines in the CM individually on each SNP, then combined the summary statistics in a meta-analysis (see Methods)). Since 11 hypothesis tests were performed for each SNP, genome-wide significance was formally set at $P < 4.55 \times 10^{-9}$, however, we also compared to the standard $P < 5 \times 10^{-8}$ threshold. To compare we used the smallest univariate meta-analysis *P*-value at a given locus. Overall, the multivariate analysis yielded more significant *P*-values in the meta-analysis (**Table 3.3**) while also detecting 4 loci not identified in the univariate analysis at either genome-wide significance threshold. One locus that was significant in univariate analysis, 10q21 (rs9787438, meta $P=9 \times 10^{-12}$, top cytokine *VEGFA*), dropped slightly below genome-wide significance in the multivariate analysis (**Table 3.3**).

3.4.6 Loci associated with the cytokine network harbour eQTLs

To characterise the regulatory effects of the multivariate loci, *cis*-and *trans*-eQTL meta-analysis was performed in 1,916 whole blood samples from two studies, DILGOM07 and YFS11. I tested 270 SNPs across the 8 significant loci, which were either multivariate index SNPs (primary or secondary) or their proxies in linkage disequilibrium ($r^2 > 0.5$) (see Methods). The eQTL analysis was done separately for DILGOM07 and YFS11, and the association test statistics obtained from the additive linear model were then combined in a meta-analysis using the fixed-effects inverse variance method. Only those SNP-probe associations were considered as statistically significant where both the FDR corrected meta-analysed *P*-values (meta-*P*-value$_{cis\text{-}eQTL}$), and cohort-level *P*-values were < 0.05. Given that eQTL studies are generally underpowered, a permissive significance threshold was chosen to identify weaker, but potentially relevant *cis* associations. The *cis*-eQTLs identified were further tested for replication to reduce false positive findings.

Of the 8 loci associated with the cytokine network, 7 harboured *cis*-eQTLs, which, in total, influenced the expression levels of 9 genes (**Table 3.5**). The most significant association was observed between the tagging SNP rs920251 and the expression level

of *SERPINE2* gene (meta-*P*-value$_{cis\text{-eQTL}}$ = 2.35 x 10^{-152}), which encodes for a serine protease inhibitor (496). Rs920251 is in high LD (r^2 = 0.76) with the index SNP rs6722871 and is also strongly associated with the cytokine network (GWAS meta-*P*-value = 1.35 x 10^{-58}). Rs3766103, a tagging SNP in the *F5* region, was significantly correlated with the expression level of *F5* (meta-*P*-value$_{cis\text{-eQTL}}$ = 1.05 x 10^{-5}) and *XCL1* (meta-*P*-value$_{cis\text{-eQTL}}$ = 1.85 x 10^{-4}). SNPs in the *VEGFA* region were associated with *ABCC10* (rs12205248**;** meta-*P*-value$_{cis\text{-eQTL}}$ = 7.41 x 10^{-3}) and *CAPN11* (rs9472179**;** meta-*P*-value$_{cis\text{-eQTL}}$ = 9.61 x 10^{-3}) expression. At the *PCSK6* and *VLDLR* loci, eQTLS were identified for the *PCSK6* (rs1552948; meta-*P*-value$_{cis\text{-eQTL}}$ = 2.70 x 10^{-7}) and *VLDLR* (rs10125071; meta-*P*-value$_{cis\text{-eQTL}}$ = 3.90 x 10^{-3}) genes, respectively.

Additionally, eQTLs were also identified at both novel multivariate loci. At the *PDGFRB* region, r2240780, which is located in the 6th intron of the *PDGFRB* gene, was observed to influence the expression of *CSF1R* (meta-*P*-value$_{cis\text{-eQTL}}$ = 3.35 x 10^{-3}). Rs2240780 was in strong LD (r^2 = 0.75) with its lead GWAS variant rs2304058 and was suggestively associated with the cytokine network (GWAS meta-*P*-value = 7.35 x 10^{-6}). At the *ABO* locus, rs532436, which lies within the first intron of the *ABO* gene, was associated with the expression level of *SURF6* (meta-*P*-value$_{cis\text{-eQTL}}$ = 4.44 x 10^{-5}). Rs532436 was in moderate LD (r^2 = 0.75) with the lead GWAS SNP and showed suggestive evidence of association with the cytokine network (GWAS meta-*P*-value = 6.04 x 10^{-7}).

No significant *trans*-eQTLs were detected at meta-analysed *P*-value threshold of < 5.12 × 10^{-7}, a threshold previously used by Westra *et al*. (320) for detecting *trans* associations. When a relaxed threshold of 1 x 10^{-5} (suggestive association) was applied, *trans*-eQTLs for 22 SNPs across 6 loci were obtained (**Table 3.6**). The strongest *trans* associations were seen between 2 variants (rs3816018 and rs2304058) at 5q32 and the expression of levels of *LUC7L* (meta-*P*-value$_{trans\text{-eQTL}}$ = 2.26 x 10^{-6} and meta-*P*-value$_{trans\text{-eQTL}}$ = 2.30 x 10^{-6}), a gene that encodes for an RNA binding protein (497). Multiple SNPs located at the *VEGFA* locus were correlated in *trans* with the expression levels of *OAS2* gene, which mediates an anti-viral innate immune response (414).

Table 3.5: Meta-analysed results of cytokine network SNPs (lead and tag SNPs) representing significant (FDR < 0.05) *cis*-eQTLs in whole blood.

SNP	Gene (Probe ID)	Chr	A1/A2	DILGOM07			YFS11			Meta analysed *P*-values *cis*-eQTL	FDR adj. meta analysed *P*-values *cis*-eQTL	*Meta-*P*-value GWAS
				P-value	Beta	SE	*P*-value	Beta	SE			
rs3766103	*XCL1*	1	C/T	2.58×10^{-4}	0.24	0.065	1.54×10^{-2}	0.094	0.039	7.46×10^{-5}	1.85×10^{-4}	3.63×10^{-4}
rs3766103	*F5*	1	C/T	4.26×10^{-2}	0.13	0.065	3.05×10^{-5}	0.156	0.037	3.46×10^{-06}	1.05×10^{-5}	3.63×10^{-4}
rs920251	*SERPINE2*	2	A/G	9.93×10^{-37}	0.76	0.055	5.16×10^{-97}	0.806	0.036	9.10×10^{-155}	2.35×10^{-152}	1.35×10^{-58}
rs2240780	*CSF1R*	5	A/G	3.87×10^{-2}	-0.13	0.063	1.73×10^{-2}	-0.092	0.039	1.84×10^{-3}	3.35×10^{-3}	7.35×10^{-6}
rs12205248	*ABCC10*	6	C/T	3.73×10^{-2}	0.13	0.063	3.87×10^{-2}	0.079	0.038	4.34×10^{-3}	7.41×10^{-3}	8.00×10^{-299}
rs9472179	*CAPN11*	6	A/G	4.71×10^{-2}	-0.13	0.063	4.34×10^{-2}	-0.076	0.038	5.81×10^{-3}	9.61×10^{-3}	5.68×10^{-11}
rs10125071	*VLDLR*	9	C/T	1.71×10^{-2}	0.15	0.063	3.29×10^{-2}	0.082	0.038	2.16×10^{-3}	3.90×10^{-3}	2.42×10^{-9}
rs532436	*SURF6*	9	A/G	5.40×10^{-4}	0.29	0.084	2.89×10^{-3}	0.139	0.047	1.67×10^{-5}	4.44×10^{-5}	6.04×10^{-7}
rs1552948	*PCSK6*	15	T/C	8.80×10^{-4}	0.23	0.068	1.66×10^{-5}	0.183	0.042	5.36×10^{-8}	2.7×10^{-7}	4.27×10^{-58}

A1/A2 refer to the minor/major allele in the Finnish population.

Variants at the *F5* locus were associated with the expression of *KCND2* and *NCOR2*. *Trans*-eQTLs at the *SERPINE2* and *PCSK6* loci were observed to influence the expression of *GPSM1* and *CACNA2D4*, respectively.

3.4.7 Linking *cis*-eQTLs identified at the 2 novel loci, *PDGFRB* and *ABO*, with publicly available results

First, I assessed the consistency of the *cis*-eQTLs identified at the novel loci with those previously identified in the Westra *et al*. (320) meta-analysis of eQTL studies of 5,311 whole blood samples from European populations. I was able to replicate the rs2240780-*CSF1R* association seen with a much larger sample size in Westra *et al*. Of note, Westra *et al.* (320) also identified the lead GWAS SNP rs2304058 associated with the cytokine network in my analysis as an eSNP for *CSF1R* (P-value-$_{\text{Westra-study}} = 2.90 \times 10^{-17}$) an association most likely underpowered in my analysis. The *cis*-eQTLs identified at the ABO locus did not replicate, but several proxy SNPs, in high LD ($r^2 > 0.7$) with these eQTLs SNPs, were *cis*-eQTLs in Westra *et al*. These proxy SNPs (rs651007, rs579459, rs649129, and rs495828) were excluded from the meta-analysis because they were either absent in one dataset (DILGOM or YFS) or did not meet the cohort level P-value < 0.05 threshold in both datasets. Hence, there is evidence of the ABO locus replicating in Westra *et al*.

Next, I investigated whether the blood *cis*-eQTLs identified at these two novel loci regulate gene expression in a tissue-specific manner by querying them against the eQTL results from 43 tissues available in the GTEx (Genotype-Tissue Expression) portal (311). Rs2240780 was identified as a *cis*-eQTL for *PDGFRB* in Epstein-Barr virus (EBV)-transformed lymphocytes. The *cis*-eQTL, rs532436, at the *ABO* locus exhibited tissue-specific expression across 17 independent tissues and was associated in *cis* with *ABO, RP11-430N14.4* or *SURF1* expression in at least one of these tissues (**Table 3.7**).

Table 3.6: Meta-analysed results of cytokine network SNPs (lead and tag SNPs) representing significant (FDR < 0.05) *trans*-eQTLs in whole blood.

SNP	SNP Chr	Gene (Probe ID)	Gene Chr	A1/ A2	DILGOM07			YFS11			Meta analysed *P*-values *trans*-eQTL
					P-value	Beta	SE	*P*-value	Beta	SE	
rs9332665	1	*KCND2* (ILMN_1748755)	7	G/T	5.30×10^{-4}	-0.24	0.070	1.52×10^{-3}	-0.13	0.042	6.01×10^{-6}
rs3820060	1	*KCND2* (ILMN_1748755)	7	G/T	5.46×10^{-4}	-0.24	0.070	1.58×10^{-3}	-0.13	0.042	6.42×10^{-6}
rs4656185	1	*KCND2* (ILMN_1748755)	7	A/G	8.08×10^{-4}	-0.23	0.069	1.54×10^{-3}	-0.13	0.042	8.12×10^{-6}
rs9287092	1	*NCOR2* (ILMN_2340052)	12	A/C	2.50×10^{-2}	-0.18	0.080	1.23×10^{-4}	-0.17	0.045	8.26×10^{-6}
rs1557570	1	*KCND2* (ILMN_1748755)	7	T/G	5.26×10^{-4}	-0.24	0.070	2.19×10^{-4}	-0.13	0.042	9.33×10^{-6}
rs181196325	2	*GPSM1* (ILMN_1667064)	9	T/C	1.87×10^{-3}	0.63	0.203	4.48×10^{-4}	0.35	0.101	5.56×10^{-6}
rs144898125	2	*GPSM1* (ILMN_1667064)	9	G/C	1.87×10^{-3}	0.63	0.203	4.57×10^{-4}	0.35	0.101	5.68×10^{-6}
rs147862316	2	*GPSM1* (ILMN_1667064)	9	T/C	1.87×10^{-3}	0.63	0.203	4.57×10^{-4}	0.35	0.101	5.68×10^{-6}
rs3816018	5	*LUC7L* (ILMN_1667064)	16	C/T	3.54×10^{-5}	-0.26	0.062	3.05×10^{-3}	-0.11	0.039	2.26×10^{-6}
rs2304058	5	*LUC7L* (ILMN_1667064)	16	C/G	3.54×10^{-5}	-0.26	0.062	3.09×10^{-4}	-0.11	0.039	2.30×10^{-6}
rs11748255	5	*HS.539385* (ILMN_1667064)	12	G/A	4.55×10^{-3}	-0.17	0.060	6.19×10^{-4}	-0.13	0.038	9.90×10^{-6}
rs12214523	6	*OAS2* (ILMN_1736729)	12	C/T	4.46×10^{-2}	-0.12	0.061	6.58×10^{-5}	-0.15	0.037	7.94×10^{-6}
rs6936047	6	*OAS2* (ILMN_1736729)	12	A/G	4.46×10^{-2}	-0.12	0.061	6.58×10^{-5}	-0.15	0.037	7.94×10^{-6}

rs9462951	6	*OAS2* (ILMN_1736729)	12	C/T	4.46×10^{-2}	-0.12	0.061	6.60×10^{-5}	-0.15	0.037	7.97×10^{-6}
rs9472184	6	*OAS2* (ILMN_1736729)	12	A/G	4.46×10^{-2}	-0.12	0.061	6.60×10^{-5}	-0.15	0.037	7.97×10^{-6}
rs3929925	6	OAS2 (ILMN_1736729)	12	A/G	4.45×10^{-2}	-0.12	0.061	7.27×10^{-5}	-0.15	0.037	8.74×10^{-6}
rs3929926	6	*OAS2* (ILMN_1736729)	12	A/G	4.46×10^{-2}	-0.12	0.061	7.31×10^{-5}	-0.15	0.037	8.79×10^{-6}
rs3929927	6	OAS2 (ILMN_1736729)	12	A/C	4.46×10^{-2}	-0.12	0.061	7.31×10^{-5}	-0.15	0.037	8.79×10^{-6}
rs4714722	6	*OAS2* (ILMN_1736729)	12	T/C	4.46×10^{-2}	-0.12	0.061	7.34×10^{-5}	-0.15	0.037	8.82×10^{-6}
rs550057	9	HS.542481 (ILMN_1904400)	17	T/C	1.47×10^{-2}	0.18	0.072	1.03×10^{-4}	0.16	0.041	4.27×10^{-6}
rs7178458	15	*CACNA2D4* (ILMN_2404493)	12	T/C	6.88×10^{-3}	0.19	0.068	3.94×10^{-4}	0.14	0.041	8.98×10^{-6}
rs7172696	15	*CACNA2D4* (ILMN_2404493)	12	A/G	6.88×10^{-3}	0.19	0.068	4.06×10^{-4}	0.14	0.041	9.27×10^{-6}

A1/A2 refer to the minor/major allele in the Finnish population. SE refers to standard error.

Table 3.7: *Cis*-eQTLs identified at the 2 novel loci, *PDGFRB* and *ABO*, exhibit tissue specific regulation in GTex tissues.

SNP	Chr	Gene	*P*-value	Tissue	Effect Size
rs2240780	5	*PDGFRB*	1.80×10^{-5}	Cells - EBV-transformed lymphocytes	0.50
rs532436	9	*RP11-430N14.4*	1.80×10^{-16}	Muscle – Skeletal	0.7
rs532436	9	*RP11-430N14.4*	2.60×10^{-16}	Whole blood	-0.75
rs532436	9	*ABO*	3.40×10^{-10}	Whole Blood	-0.62
rs532436	9	*RP11-430N14.4*	3.30×10^{-15}	Adipose – Subcutaneous	0.64
rs532436	9	*RP11-430N14.4*	1.20×10^{-13}	Adipose – Visceral (Omentum)	0.70
rs532436	9	*RP11-430N14.4*	5.50×10^{-13}	Artery – Tibial	0.42
rs532436	9	*RP11-430N14.4*	1.10×10^{-12}	Esophagus – Muscularis	0.67
rs532436	9	*RP11-430N14.4*	9.20×10^{-11}	Breast – Mammary Tissue	0.72
rs532436	9	*RP11-430N14.4*	1.00×10^{-10}	Heart – Left Ventricle	0.54
rs532436	9	*RP11-430N14.4*	9.60×10^{-10}	Nerve – Tibial	0.55
rs532436	9	*RP11-430N14.4*	2.80×10^{-9}	Esophagus – Mucosa	0.46
rs532436	9	*RP11-430N14.4*	3.80×10^{-8}	Adrenal Gland	0.76
rs532436	9	*RP11-430N14.4*	5.90×10^{-8}	Pituitary	0.72
rs532436	9	*RP11-430N14.4*	2.80×10^{-7}	Colon – Transverse	0.27
rs532436	9	*RP11-430N14.4*	9.40×10^{-7}	Heart – Atrial Appendage	0.67
rs532436	9	*RP11-430N14.4*	2.30×10^{-6}	Lung	0.28
rs532436	9	*RP11-430N14.4*	4.50×10^{-6}	Thyroid	0.32
rs532436	9	*SURF1*	5.30×10^{-6}	Heart – Atrial Appendage	-0.50
rs532436	9	*ABO*	6.40×10^{-6}	Adrenal Gland	0.62
rs532436	9	*SURF1*	2.70×10^{-5}	Skin – Sun Exposed (Lower leg)	-0.32

3.5 Discussion

Characterising the genetic architecture of cytokine concentration control in circulation in a general population may provide in-depth insight into how inter-individual differences in immune function may contribute to differential disease susceptibility. In this study, we first identified a network of 11 correlated cytokines, which were consistently correlated across the three populations analysed. The cytokines within this network most likely participate in a broad array of immune responses occurring in circulation, rather than in a specific immune pathway. These cytokines have been shown to be involved in the classical TH_1 (IL-12, IFN-γ), TH_2 (IL-4, IL-6, and IL-10), TH_{17} (IL-6, IL-17, and G-CSF) and Treg (IL-10) responses (492,493). It also includes pro-angiogenic growth factor (VEGF-A, FGF-basic and PDGF-BB), which promote angiogenesis during wound repair process (118). This suggests that the immune system in apparently healthy individuals is tailored to effectively counteract different classes of pathogens simultaneously through the release of a plethora of polarising cytokines, which activate various lineage-specific effector responses. A recent study has shown that cytokines induced by bacteria or fungi cluster in a pathogen-specific manner (498). All the 11 cytokines in the network, which exert either anti-inflammatory or pro-inflammatory effects, were positively correlated. This indicates counter-regulatory mechanisms exist between these two broad groups of cytokines that control their release and function.

On the other hand, the 11 cytokines from the various TH subsets have been previously shown to cross regulate each other's development. For example, IL-10 produced by Th2 cells supresses IL-12 secretion and subsequently inhibits Th1 polarization (499). Meanwhile, IFN-γ–produced by Th1 cells inhibits Th2 development (500). Despite evidence that some of the cytokines in the network can negatively influence each other's level, in my study all the cytokines were positively correlated. However, the inhibitory effects of these cytokines have been mainly supported by *in vitro* stimulation studies. There are few studies that have assessed the relationship between cytokine levels in population based studies (344,498). A similar correlation pattern between the 11 cytokines was also observed in a recent study utilising similar data (344). There is increasing evidence through *in vitro* and *in vivo* studies that TH

subsets display context dependent functional plasticity, and are able to convert into other subsets (501). For example, Th17 subsets are capable of producing functional Th1-like progeny (502). This implies that multiple effector functions can co-occur in one TH subset, which also acquires additional cytokine producing capacities. This may partially account for the positive correlation seen between the cytokine, but warrants further investigation. Of note, baseline cytokine levels were analysed in this study. Homeostatic balance of pro-inflammatory and anti-inflammatory cytokine baseline levels is necessary to maintain a healthy state, which may explain to some extent the positive correlations observed. This homeostatic balance tends to shift during infection, whereby studies assessing cytokine production after pathogen challenge have reported different correlation patterns between cytokines mentioned in this study (498,503). For example, following fungal challenge, IL-10 was observed to be negatively correlated IL-17 and IFN-γ, and positively correlated with IL-6 (498).

Secondly, in the meta-analysis of multivariate GWAS of the cytokine network in 9,267 individuals, I identified 12 independent variants located across 8 chromosomal loci. Findings of this study further confirm and extend previously identified genetic signals associated with circulating cytokine levels. Of these 8 loci, 6 were consistent with previous studies (343,494,495). Furthermore, I empirically showed that modelling correlated cytokines in a multivariate fashion increases statistical power to detect associations compared to the univariate test. This lead to the detection of two novel signals located on chromosomes 5 (*PDGFRB* locus) and 9 (*ABO* locus) not previously reported to be associated with cytokine levels. Of note, these two novel loci were also not detected in a recent univariate GWAS of 44 cytokines, which utilised similar data from the same populations used in my analysis (494). Moreover, similar power gain has also been shown in a previous comparison, where multivariate GWAS of metabolite networks outperformed the univariate approach, leading to the identification of novel loci (485). The findings of this study and those of others support the emerging evidence of genetic pleiotropy, whereby a genetic locus influences multiple traits (504). Also, one locus, which was identified in the univariate analysis, did not achieve significance in the multivariate analysis. It has been shown through simulation studies that multi-trait QTL analysis may not always provide a power boost, depending on the QTL effects and the residual correlation

between traits (478,505,506). Instances where genetic variants influence correlations between multi-traits, then jointly modelling of these traits using multivariate approaches can increase power to detect additional loci (506).

Thirdly, *cis*-eQTL analysis revealed that 6 out of the 8 multivariate GWAS loci, which include the two novel loci, harbour eQTLs. The strongest *cis*-association was observed at the *SERPINE2* locus between the tag SNP rs920251 and *SEPRINE2* expression level. Rs920251 is located within intron 1 of *SERPINE2*. This gene encodes a serine protease inhibitor, and its role has been implicated in coagulation, fibrinolysis and remodelling of tissue (496). Moreover, *SERPINE2* is also referred to as a cytokine-inducible gene, whereby its expression has been previously shown to be up-regulated by pro-inflammatory cytokines including IL-1 and TNF as well anti-inflammatory cytokines such as TGF-β (507–510). This suggests that rs920251 most likely influences the expression levels of *SERPINE2* in blood by regulating the concentrations of circulating cytokines. On the other hand, studies have also demonstrated that SERPINE2 antagonises the pro-inflammatory effects of IL-1 on the production of matrix metalloproteinases, proteolytic enzymes known to be linked with the pathogenesis of chronic obstructive pulmonary disease (COPD) (507,511). In addition, rs920251 has been previously reported to be associated with COPD in a case-control study, linking *SERPINE2* with COPD susceptibility (512). The mechanistic role of *SERPINE2* in the development of COPD remains unclear. The association observed between variants at the *SERPINE2* locus, its expression, and the cytokine network in my analysis provides a molecular link between the inflammatory processes and structural remodelling of the airway in COPD. The rs920251-*SERPINE2 cis* association has also been previously identified in B cells (289).

At the novel locus 5q32, a *cis*-eQTL was identified for *CSF1R*, a consistent association previously observed in another blood eQTL study (320). *CSF1R* encodes for a transmembrane tyrosine kinase receptor, which is activated by its cytokine ligands CSF-1 and IL-34 (513). It is largely expressed in cells of the myeloid lineage, particularly macrophages, and plays a role in their development (513). Signalling via the CSF1R receptor has been shown to enhance cytokine production (514). In humans, *CSF1R* and *PDGFRB* have similar gene organisation and are in very close to each other on chromosome 5 (~500bp apart), indicating that these two genes arose

from a duplication event (515). This raises the hypothesis that variants on *PDGFRB* influences *CSF1R* level, which in turn regulates myeloid cells numbers and subsequently cytokines levels.

At the other novel locus, 9q34.2, rs532436 influenced the expression of *SURF6* levels. The *SURF6* gene encodes for a nucleolar protein, which is required for ribosome biogenesis in the nucleolus, and is organised within the Surfeit gene cluster (*SURF1-SURF4*) of unrelated housekeeping genes (516). Rs532436 has been previously associated with LDL level, haemoglobin concentration and haematocrit (517,518). This SNP lies within an enhancer box (E-box) motif that binds to the transcription factor USF (upstream stimulatory factor) (519). In addition, rs532436 is identified by the GTEx project (311) to regulate the expression of genes *RP11-430N14.4, ABO*, and *SURF1* in a number of tissues. However, this SNP has an opposite effect on expression of these genes in different tissues. This further supports the notion that eQTLs regulate gene expression in a tissue-specific manner and may influence the expression levels of different genes across tissues or cell types (318,520).

Although the correlation patterns between cytokines were similar across the three populations, the correlations were somewhat weaker in YFS07 and FINRISK02. This may be linked to the small sample sizes in these two populations. Another reason could be the differences in the cytokine measurements (serum *vs.* plasma), their stability, and experimental assays. Comparison studies have shown that inter-variability among multiplex cytokine assays can make cross-population comparisons difficult (114,521,522). In particular, the choice of the collection tube, storage duration and temperature, and freeze-thaw cycles greatly impacts cytokine levels and stability (523).

Cytokines are important intermediate immunological phenotypes and characterising the genetic architecture of cytokines levels can provide insight into inflammatory pathways underlying the link between genetic variants and disease susceptibility. Given the previously established role of cytokines in orchestrating an inflammatory process in COPD (524), and the consistent association between SERPINE2 locus and susceptibility to COPD (512), which may subsequently lead to airway remodelling

and COPD risk. This to the best of my knowledge is the first study that links *SERPINE2* locus with cytokine levels and *SERPINE2* expression. Moreover, the *ABO* locus described above has been previously associated with venous thromboembolism (525). There is evidence that inflammatory cytokine levels may also play a role in venous thromboembolism (526). Findings in this study also provide support that the mechanism underlying the association between the *ABO* locus and venous thromboembolism may in part be due to genetic regulation of cytokine levels via variants at this locus.

Moreover, the loci identified also overlaps with prominent drug targets. The variants at the *VEGFA* locus are proximal to the *VEGFA* gene that encodes for a drug target for angiogenesis (527). This suggests a likely genetic contribution to the inter-individual variation seen in response to drugs targeting this gene (528). Also, the *VEGFA* locus and the VEGF cascade has been linked to ulcerative colitis risk, implicating that drugs targeting VEGFA can be potentially used to treat ulcerative colitis and other inflammatory diseases of the bowel (494,529).

In summary, in this chapter, a total of 8 loci contributing to the genetic regulation of a network of 11 cytokines were identified. This included 2 novel loci previously undetected loci for cytokines in GWA studies. The novel loci harboured eQTLs, which were previously identified as tissue-specific eQTLs. However, it is pertinent to note that the correlations observed between the 11 cytokines, the genetic variants and eQTLs identified, and the subsequent interpretation the results were based on baseline cytokine levels. Thus, these findings might not directly relate to cytokine response following immunological challenge or infection. Recent studies investigating genetic variants influencing cytokine production have shown that stimulation by different pathogens induces pathogen-specific correlation patterns between pro-inflammatory and anti-inflammatory cytokines (498,503). These studies have also identified different sets of genetic variants affecting cytokine response to pathogens (498,503).

The findings from this study and other similar ones suggest that exploiting cytokines in GWA studies can provide insights into how genetic variation can regulate upstream inflammatory processes, which may confer susceptibility to immune-related diseases.

Characterising genetic variants influencing cytokine levels in population-based studies, which may lead to differential immune responses, is necessary to understand mechanisms underlying autoimmune and infectious diseases, and for the development of effective vaccines and therapeutics.

Chapter 4

Differential network analysis identifies a transcriptional network involved in tissue resident memory T-cell development

4.1 Introduction

This chapter focuses on a particular aspect of the immune system, which is immunological memory, with specific emphasis on the tissue resident memory $CD8^+$ T (TRM) cells. Characterisation of transcriptional networks in TRM cells is of importance to understand the mechanisms regulating their homing and maintenance at tissue sites.

Memory is regarded as a unique trait of the adaptive immune system, which provides long-term protection against reinfections and is the major focus for rational vaccine design. The majority of vaccines developed so far rely on circulating responses. However, infectious pathogens commonly cross barrier sites such as skin, gut and lung, and cause localised infections. Hence, it is crucial to generate memory T-cells at these sites for effective site-specific immunity. Since TRM cells provide frontline defence at barrier sites, understanding the transcriptional mechanisms underlying their generation and maintenance is important. Few studies have identified a number of transcriptional regulators that influence the memory gene signature in T-cells (64,530) but, to date, specific lineage-defining genes have not been linked directly to long-term memory T-cell formation. This suggests that transcriptional programs regulating commitment to long-term memory operate as part of a network rather than a few

candidate master regulators. Hence, this has underscored the need to focus on transcriptional networks that regulate the establishment of TRM cells at tissue sites for robust protective immunity.

The application of integrative tools and network-based approaches to gene expression microarray datasets facilitates the representation of functional dependencies between genes as networks. The recent use of gene expression profiling to assess transcriptional programs in TRM cells has made publically available transcriptomic data, which can be leveraged for network analysis.

4.1.1 Tissue resident memory $CD8^+$ T (TRM) cells

For a long time, it had been accepted that the memory $CD8^+$ T-cells comprised of two major subsets: the central memory T (TCM) cells, which circulate through secondary lymphoid sites and the effector memory T (TEM) cells trafficking through blood, spleen and peripheral tissues (531). However, over the last decade accumulating evidence has supported the existence of a third subset of memory T-cells, which reside in peripheral tissues and are incapable of re-entering circulation (95,100,532). These TRM cells have been identified in a number of barrier tissues including the skin, brain, lung, gut, liver, salivary glands, and female reproductive tract where they have been shown to offer superior protection against local infection compared to their circulating memory counterparts (94–102). In addition, TRM cells exhibit a unique transcriptional profile that distinguishes them from TEM and TCM cells (88,103).

4.1.1.1 Evidence for the existence of TRM cells

Several studies performed in both mice and humans have provided considerable evidence for the existence of TRM cells in peripheral tissues, which have been shown to be disconnected from circulation and independently sustained from their circulating (TEM and TCM) counterparts (533,534). Transplantation experiments by Gebhardt *et al*. showed that memory population residing in the dorsal root ganglia (DRG) of herpes simplex virus (HSV) infected mice did not recirculate when transplanted under the kidney capsule of naive (not infected) recipient mice (95). Likewise, they also demonstrated that the memory population present on the skin after HSV infection did

not migrate and persisted for weeks when transplanted into naive recipients (95). In a later study by the same group, the idea that TRM cells are distinct and independent of the circulating subsets was further reinforced using transfer experiments (535). They showed that upon transfer, HSV-specific T-cells from male mice migrated and survived for about 10 weeks in the skin of female mice in the absence of their non-viable circulating counterparts, which were rejected by the female immune system (535). TRM cells have also been observed to persist for several months within the brain parenchyma following an acute vesicular stomatitis virus (VSV) infection in mice (100). Also, *in situ* intracranial labelling with carboxyfluorescein succinimidyl ester (CFSE) revealed that these brain TRM cells were locally confined to the infection site and represented a self-sustaining population, distinct from their circulating subsets that required constant replenishment from the circulation (100). These brain TRM cells further lost the ability to survive once removed from their tissue niche (100). Similar persisting TRM population has also been found in murine intestinal epithelium (536) and salivary gland upon lymphocytic choriomeningitis virus (LCMV) infection (101) and other tissues including skin, lungs, salivary glands (104). For example, treatment of LCMV-immune mice with fingolimod (FTY; a drug that increases migration of recirculating T-cells into lymph node) did not reduce the TRM cell numbers in intestinal epithelium even after 30 days of treatment but led to a drastic reduction of circulating memory T-cells in blood (536). These findings confirmed that TRM cells were indeed long-term resident cells lacking recirculating abilities (536). Moreover, the resident nature of TRM cells has been further confirmed through experiments utilising parabiotic mice (two mice surgically joined together to share their circulatory system), in which TRM cells failed to equilibrate across the skin of these two mice (96).

Likewise, evidence for TRM cells in human tissues also exists. Early studies utilising xenotransplantation experiments showed that pathogenic resident cells present on human skin grafts from psoriatic patients were able to persist for several weeks, locally proliferate, and were sufficient to give rise to psoriasis lesions following transplantation into immuno-compromised (lacking both type I and type II interferon receptors) mice (537–539). TRM-like cells have also been identified in human genital herpes lesions, where a subset of HSV-specific T-cells was observed to infiltrate to the local site of re-infection and persist for months in the skin following lesion

resolution (540,541). The existence of these resident cells at tissue site further correlated with increased viral clearance (542). Influenza-specific TRM cells have also been identified in human lungs (543,544). TRM cells expressing the characteristic resident surface markers have also been isolated from the human gut (90) and skin (545). Additionally, skin TRM cells have also been linked to fixed drug eruption, an allergic response leading to skin lesions at the same site each time after ingesting a particular drug (546,547). Histological staining has identified IFN-γ producing resident $CD8^+$ T-cells at sites of healed lesions and recurrence of these lesions at the same location following challenge with the inducing drug, further supporting the causative role TRM cells (546,547). The most compelling evidence that human TRM cells persist in tissues and are non-recirculating came from studies utilising cutaneous T-cell lymphoma patients (548,549). Treatment of these individuals with a low dose of alemtuzumab, an antibody that destroys CD52-bearing T-cells in blood, leads to the depletion of all circulating T-cells without affecting those residing in the skin (548,549).

4.1.1.2 TRM cells provide superior protection in peripheral tissues

There is now accumulating evidence that TRM cells persist for long term in peripheral tissues forming a frontline defence. In particular skin TRM cells, which exhibit superior protection against local reinfection relative to their circulating memory subsets (95,96,100,550). Using microscopy and *in vivo* experiments, Ariotti *et al.* showed that antigen-specific TRM cells persisting after a herpes infection acquire a dendritic morphology and are able to rapidly recognise antigen-expressing cells by continuously patrolling the epidermis (550). This suggests that the local patrolling by skin TRM cells provides first line of defence against reinfections. Studies employing parabiotic mice revealed that only the mice with skin TRM cells were able to effectively clear a VACV reinfection, whereas mice with only TCM and no TRM cells could not, suggesting that TRM cells functionally confer superior local protection (96). In another study, Gebhardt *et al*. observed that TRM cells lodged in the skin and vagina were able to effectively clear HSV upon reinfection (551). In the lungs, vaccine-generated TRM cell populations (552) provided effective shielding against influenza and lower respiratory infections (553). Others have demonstrated

that TRM cells are capable of rapidly responding to infections by clearing them before the arrival of circulating memory T-cells (96,552).

4.1.1.3 Molecular mechanisms defining TRM cells generation and maintenance

TRM cells have been best characterised by their incapability to recirculate through blood once they are lodged within a tissue following local infection (87,95,96,554). However, the molecular mechanisms underlying their differentiation, migration, and retention in peripheral tissues have just begun to emerge.

Generation from TRM cell precursors. The exact lineage-committed precursor that gives rise to TRM cells is still not clear, but they have been shown to generate from the same $KLRG1^{Low}$ effector-like subset that gives rise to TCM cells as well. For example, two separate studies in mice showed that effector cells with low KLRG1 expression were able to infiltrate into the skin and gut, and later develop into TRM cells (98,103). The localisation of these effector precursors in the skin was driven by high amounts of CXCR3 chemokine expression (103). Consistent with these findings, DNA sequencing of TCR gene isolated from skin TRM cells and lymph node TCM cells generated after skin immunisation revealed that they bear identical TCR motif (CDR3), suggesting that both TRM and TCM cells arise from common naive precursor cells (555).

TRM Cell surface markers. The identification of TRM cells in many non-lymphoid tissues, including skin, gut, lungs, brain, female reproductive tract, salivary glands, and thymus have been based on the expression of two key the surface molecules, CD69 (Cluster of Differentiation 69) and CD103 (95,100,532,556–558). In addition, a subset of TRM cells completely devoid of CD103 expression has been found in the liver (102), intestine (559), and secondary lymphoid organs (560), suggesting that CD103 expression might be tissue-specific. A recent study showed that human spleen and tonsils were populated with two distinct groups of TRM cells ($CD69^{+}CD103^{+}$ and $CD69^{+}CD103^{-}$), which were also anatomically separate within these tissues (561). While on the other hand, it was observed that a considerable fraction of TRM cells in the salivary glands, pancreas and female reproductive tract of mice lacked both CD69 and CD103 expression (562). These studies provide evidence that the TRM cell pool

exhibits phenotypic heterogeneity and the expression of these phenotypic markers might be dependent on tissue type and compartment.

CD103 (encoded by the *Itgae* gene) is the alpha subunit of the $\alpha_E\beta_7$ integrin receptor found on the TRM cells, which promotes their adhesion to skin and gut by interacting with E-cadherin ligands constitutively expressed in the epithelial layers of these tissues (95,101,103,556,558,563,564). Interestingly, E-cadherin expression has also been noted in TRM cells obtained from a range of tissues including skin, gut, lung, and brain (88,101,103). CD103 has been shown to play a functional role in mediating the homing and retention of TRM cells within peripheral tissues (89,100,103,556,565,566). There is also evidence that this receptor might also be involved in directing tissue localisation of TRM cells (567,568). TRM cells defective of CD103 expression were able to infiltrate into the gut, brain, and skin, but failed to persist long-term (98,100,551,556,567,569). Moreover, TRM cell precursors in the skin and gut up-regulated CD103 once they were lodgement into these tissues (103,559). TGF-β, a tissue derived cytokine is essential for the development of TRM cells in the skin, gut and lungs by the inducing CD103 expression (98,103,556,569,570). The role of TGF-β in inducing the residency-related transcriptional profile has explored through RNA-sequencing analysis in **Chapter 5**.

CD69 is encoded by the *Cd69* gene, which is localised within the cluster of Natural Killer receptors, and is a transmembrane glycoprotein (571,572). CD69 is well known as an activation marker that is expressed very early on stimulated T-cells, but its role in mediating tissue residency has been identified lately (573,574). Similar to CD103, CD69 expression is also induced on TRM cells soon after they reach their residency site, which further enhances their retention by antagonising tissue egress signals for TRM cells (103,105). CD69 forms a complex with the migration receptor, sphingosine-1-phosphate receptor-1 (S1PR1), and mediates its internalisation into the cytoplasm for degradation (575). Consequently, the surface expression of S1PR1 is inhibited hampering the receptors ability to chemotactically respond to the exit signals from sphingosine-1-phosphate (S1P), thereby ensuring tissue residency (103,574). T-cells deficient in CD69 expressed functional S1PR1 and as a result were incapable of tissue lodgement (103,105,573). Together with *Itgae*, *S1pr1* gene is also part of the core transcriptional signature defined in epithelial TRM cells from the gut, lung, and

skin (103). Moreover, CD69 inhibited S1P1-induced chemotactic migration of T-cells as soon as they entered the lung, however, their long-term persistence does not depend upon such inhibition (576). This suggests that TRM maintenance might be CD69-exclusive. In addition, the loss of expression of both S1PR1 and its transcriptional activator KLF2 (kruppel-like factor 2) in TRM cells is correlated with their residency at peripheral sites (105).

Other signals regulating TRM cell development and survival. Evidence exists for additional factors that are required for the acquisition of residency. The expression of a number of chemokine receptors on TRM precursor cells, in particular, CXCR3, are required for their homing in skin and lung (103,577,578). Furthermore, the local production of cytokines including IL-15, IL-33, TGF-β, and TNF produced at tissue environments influences the development and establishment of TRM cells at these sites. For instance, the induction of CD103 via TGF-β signalling plays a key role in the formation of TRM cells in the skin, gut and lung (98,103,570,577,579,580). TGF-β, IL-33, and TNF have been shown to synergistically up-regulate CD103 expression and induce tissue resident phenotype (556), as well as reduce KLF2 expression levels in $CD8^+$ T-cells (105). Also, long-term persistence of TRM cells in the skin and lung require IL-15 dependent signal (103,579,581). Increased levels of the pro-survival molecule BCL2 (B-cell lymphoma 2) in murine brain $CD103^+$ TRM cells, implicated its involvement in facilitating TRM cell survival and prolonged maintenance (100). Similarly, the role of antiviral molecule IFITM3 (Interferon-induced transmembrane protein 3) in protecting lung TRM cells from viral infections and promoting their survival has been demonstrated. Elevated expression of *Ifitm3* has also been reported in brain TRM cells, indicating that it might be a key survival gene required for long-term maintenance of TRM cells (558).

Transcriptional signature of TRM cells. Comparative analysis of microarray-based expression profiles of murine $CD103^+$ TRMs isolated from different peripheral sites (skin, gut, lung, and brain) with their circulating counterparts revealed that TRM cells from various tissues exhibit a distinctive transcriptional signature, which demonstrates considerable similarity to one another (88,103,582). This further led to the establishment of a shared core residency signature comprising of up-regulated genes known be involved in adhesion (*Itgae*, *Itga1*, and *Cdh1*) and tissue homing

(*Rgs1*, *Rgs2*), and down-regulated genes involved in tissue exit (e.g. *S1pr1*) (103). Similarly, TRM cells from murine liver, and human lung and skin also displayed a unique transcriptional profile (583).

Several transcriptional factors have been identified as differentially expressed in TRM cells. KLF2 is a key driver controlling the movement of T-cells. The loss of expression of both *Klf2* and its downstream target gene *S1pr1*, noted in almost all TRM cells, facilitates tissues retention of TRM cells (105). As discussed in **Chapter 1**, the reciprocal interplay between T-Bet and EOMES play a central role in effector and memory cell-fate decisions during $CD8^+$ T-cell differentiation (584). *T-bet* and *Eomes* are both down-regulated in TRM cells in the skin, gut and lungs (88,103). Their down-regulation has been shown to be necessary for TRM cell generation, whereby TRM cells failed to develop in the skin and lung when expression of either *Eomes* or *T-bet* was forced in maturing T-cells (579). Moreover, TGF-β signalling dependent generation of TRM cells is reinforced by positive feedback mechanism. TGF-β inhibits the expression of EOMES and T-bet, which are negative regulators of its receptor, and the suppression of these two transcription factors further augments TGF-β activity promoting TRM cell differentiation (577,579). Unlike the transcription factors discussed so far, whose down-regulation is associated with the TRM phenotype, the up-regulation of two other central players (BLIMP1 and HOBIT) is essential for TRM cells (582). The combined effect of transcription factors BLIMP1 and HOBIT (homolog of Blimp1 in T cells; also referred to as ZFP683 or LOC100503878) is not only essential for the establishment of TRM cells in the skin, gut, liver and kidney, but also in other resident populations such as resident natural killer T (NKT) cells form the liver (582). This highlights that BLIMP1 and HOBIT transcriptionally programs a residency-affiliated signature universal to an assortment of tissue resident populations (582). Additionally, analyses of ChIP Sequencing data further revealed that these two transcriptional promote tissue residency by directly binding to and down-modulating tissue egress genes such as *Tcf7*, *Klf2*, *S1pr1*, and *Ccr7* (582). High expression levels of other transcription factors such as *Litaf*, *Nr4a1*, *Ahr*, has also been recorded in TRM cells, and have been shown to be involved in the generation and persistence of TRM cells (103,585–587).

Transcriptional signature extends beyond TRM cells. The Majority of studies describing tissue residency so far have largely focused on $CD8^+$ T cells. However, evidence for antigen-specific tissue resident cells from several lymphocyte lineages also exists. $CD4^+$ memory T cells expressing CD69 and no or low levels of CD103 have been isolated from various non-lymphoid sites in both human and mice (588–590). In addition, regulatory T cell (Treg) subsets showcasing specialised function and phenotype are also present in the adipose tissue, lung, liver, and skin (591,592). Tissue residency also spans innate lymphocyte lineages such as natural killer (NK) and NKT cell (593). Few recent studies examining expression profiles have just started to reveal that the transcriptional requirements for tissue residency in various lymphocyte lineages are shared to some degree (545,582). Li *et al*. demonstrated that the transcriptomes of Tregs and $CD4^+$ memory T-cells found in human skin were enriched for genes previously identified to be part of the common core transcriptional program associated with tissue residency in murine skin, gut and lung TRM cells (545). In another study, Mackay *et al*. also provided consistent evidence and further highlighted that sharing of residency signature is not only confined to adaptive lymphoid resident cells found at epithelial sites but also spans both innate lineage and non-epithelial tissues (582). The authors performed RNA-seq based comparative transcriptional analyses of murine tissue resident lymphocytes, which consisted of innate (NK and NKT cells) and adaptive (TRM cells) subsets from various peripheral sites, leading to the identification of a universal transcriptional signature common to all these cell types (582). This shared universal signature of 30 genes are mainly involved in chemokine receptor signalling (*Xcl1*, *Cxcr6*, and *Cxcr4*), regulating tissue exit (*S1pr1*, *Klf2*), and the establishment of tissue residency (*Osgin1*, *Hobit*, *Tcf7*, *Arhgef18*, *Fam65b*, *S1pr4*, *S1pr5*).

4.1.2 Existing gap in the understanding of tissue residency

Results from genome-wide transcriptome analyses of TRM cells isolated from various tissues have made it clear that residency in TRM cells is transcriptionally programmed. However, understanding how this program is fine-tuned to facilitate the development and maintenance of TRM cells in tissues is still in its early stages. Nearly all studies transcriptionally characterising TRM cells so far have assessed

changes in the mean expression level of individual genes between TRM cells and their circulating counterparts, resulting in a list of differentially expressed genes (88,103,582). The most differentially expressed genes, based on either fold change or statistical significance, were then prioritised as candidate genes associated with the tissue residency. The change in expression of these top candidate genes, which might not be causally linked to the residency phenotype, could be a consequence of mild expression changes of an upstream gene. Hence, differential expression analysis does not take into account the correlation structure that exists amongst differentially expressed genes. Network-based approaches are being increasingly applied to transcriptomic data to delineate the gene-gene interactions, and have also been used in defining networks of coexpressed genes involved in immune cell differentiation and lineage fates (594,595). Recently, differential network analysis tools, which assess the change in pairwise correlations between genes across conditions, have been developed to supplement differential expression analysis (252,253,255). To identify groups of residency-related genes exhibiting altered dependencies across resident and circulating states; a suitable approach would be to construct sub-networks within the boundaries of differentially expressed genes between resident and circulating groups (596,597). The identification of such differentially coexpressed sub-networks will provide deeper understanding into regulatory programs and pathways that might be essential for tissue residency. Additionally, potential drivers of residency can be inferred from the identification genes most central (hub genes) to the network.

Moreover, comparative analyses of transcriptional profiles have also revealed that some degree of shared residency-related transcriptional identity exists between tissue resident cells within and across lineages. However, gaining insight into how widely this shared residency signature is maintained across tissue resident cells present in haematopoietic lineage requires a large compendium of expression profiles. The availability of expression profiles from an array of immune cell types through large consortia such as the ImmGen project (598,599) makes it possible to directly assess the enrichment of the residency-related genes sets.

4.2 Research objectives

The central aim of this chapter was to use network-based approach to characterise gene network(s) underlying the transcriptional signature regulating the development and establishment of tissue resident memory T cells.

In this chapter, two previously published microarray datasets (88,103) generated from murine TRM cells isolated from skin, gut, lung, and brain were integrated for network analysis.

The specific objectives of this research chapter were:

1. To perform differential co-expression analysis on genes differentially expressed between TRM and circulating $CD8^+$ T-cells and identify residency-related gene network(s) in TRM cells.

2. To infer potential drivers of the residency-related gene network(s).

3. To assess whether the residency-related gene signature identified in TRM cells share transcriptional similarity in different populations of tissue resident immune cells.

4.3 Methods

4.3.1 Gene expression data

An overview of data analysis workflow employed in this study is given in **Figure 4.1**. The microarray gene expression profiles of murine TRM and circulating memory CD8+ T-cells analysed in this chapter have been previously published in (103) and (88). All the microarrays were generated on the Affymetrix Mouse Gene 1.0 ST array platform. The raw gene expression data (Affymetrix CEL files) were downloaded from the Gene Expression Omnibus (GEO) database (GEO accession numbers; GSE47045 (103) and GSE39152 (88)). A total of 25 samples, which included both the TRM and circulating memory T-cells, were analysed. The resident samples (N=14) were obtained from TRM cells isolated from skin (at day 30 post-infection (p.i.) with HSV), gut (at day 60 p.i. with LCMV), lung (at day 30 p.i. with influenza virus), and brain (at day 20 p.i. with VSV-OVA) as described in (88,103). The circulating samples were obtained from spleen, which consisted of $CD103^{-}CD8^{+}$ T-cells (isolated at day 20 p.i. with VSV-OVA), and $CD8^{+}$ TCM and TEM cells (isolated at day 30 p.i. with HSV) as described in (88,103). There were three biological replicates for each tissue type, except brain TRM cells and their circulating splenic ($CD103^{-}CD8^{+}$ T-cells) counterparts, which had five replicates each. The resident and circulating samples are hereafter referred to as "resident" and "circulating" groups, respectively.

4.3.2 Microarray data processing and normalisation

Affymetrix CEL files containing the raw microarray intensities of expression for 34,760 probes across all the 25 samples were processed with the Bioconductor "affy" package (184,600) in R. First, each CEL file was background corrected for non-specific hybridisation effects using the robust multichip average (RMA) algorithm. Next, probe-specific correction was performed using the "pmonly" method (184,600). Finally, the probes in each probeset, perfect match (PM) and its corresponding mismatch (MM) probe, were summarised into a single expression value for each probe using the median polish method (184,600).

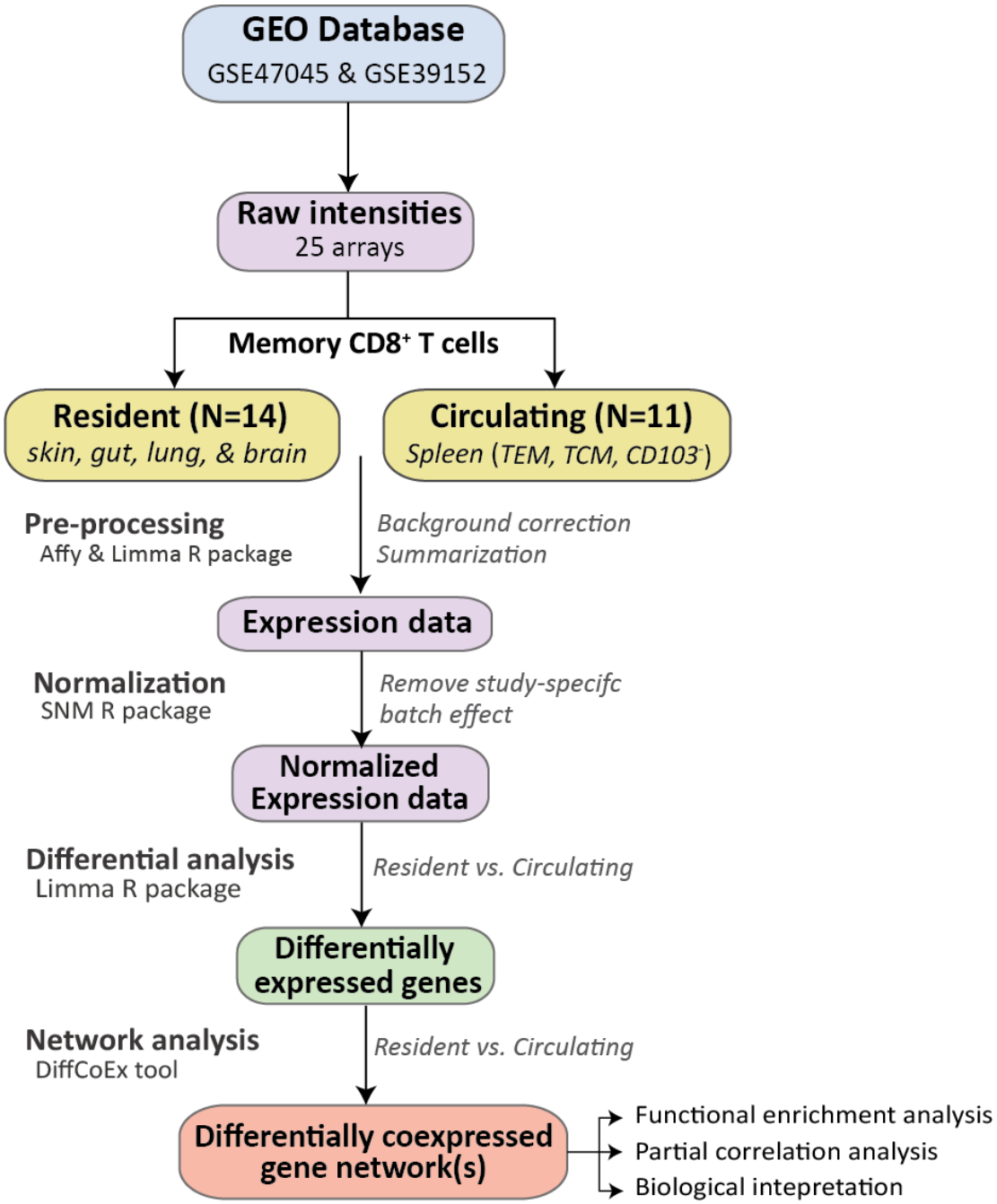


Figure 4.1: Overview of data pre-processing and analysis workflow

Study related batch effects are common when data from different microarray studies are combined. Batch effects were removed from the log2-transformed expression values using the SNM normalisation method implemented in the "snm" R package (195),. SNM jointly fits a study-specific model to all probes and samples using two types of defined variables: those that are of interest to the biological outcome (biological variables), and those that are not (adjustment variables). Here, the resident and circulating phenotypes were modelled as biological variables, and the expression data was adjusted for the effects of study-specific batches (adjustment variable). All downstream analyses were performed with SNM normalised expression values.

4.3.3 Global analysis of the transcriptome

For exploratory analyses of the global expression profiles, lowly expressed probes were excluded to retain only those (N=10,428) with a variance of expression within the 70th percentile. Principal component analysis (PCA) was performed using the "prcomp" function with default settings in R. The samples were hierarchically clustered, with Wards clustering algorithm based on Euclidean distances as a similarity measure, using the "dendextend" R package (601).

4.3.4 Differential gene expression analysis between resident and circulating groups

For differential gene expression analysis, only probes with a gene symbol annotation were considered. The analysis was performed on 24,534 annotated probes using the Bioconductor "limma" R package (602). Briefly, a linear model was fitted to each gene using limma's "lmFit" function. Then, for each gene, the significance of the differential expression between resident *vs.* circulating groups was assessed using the empirical Bayes moderated *t*-statistics computed with limma's "eBayes" function (202). An FDR adjusted *P*-value significance threshold of 0.05 was implemented to identify differentially expressed (DE) genes. The DE genes were used for network analysis as described below.

4.3.5 Differential gene co-expression network analysis

To identify gene network(s) that were differentially coexpressed between the resident *vs.* circulation groups, the DiffCoEx method described in Tesson *et al*. (253) was employed. DiffCoEx is built on a widely used framework for constructing weighted gene co-expression networks, known as weighted gene co-expression network analysis (WGCNA) (239,395). Expression profiles of 2,197 unique genes DE between the resident and circulating groups were used as an input for network analysis. Prior to the analysis, multiple DE probes (> 1) corresponding to the same gene were collapsed to a single representative, using the "CollapseRow" function with the "Max-Mean" method in WGCNA. The Max-Mean method chooses the probe with the maximum mean expression across all samples.

Differential co-expression analysis with DiffCoEx involved four steps. First, an adjacency matrix $C^{[K]}$ was computed, with each group *K* (resident or circulating), by calculating the pair-wise Pearson correlation *c* between all the DE gene pairs (*i,j*).

$$\mathrm{C^{[K]}}: \mathrm{c_{ij}^{[K]}} = \mathrm{cor(gene_i, gene_j)}$$

Next, an adjacency difference matrix *D* was obtained by raising the absolute value of the difference between the signed squared correlation coefficient of the resident and circulating groups to a soft-thresholding power (β=10).

$$D: d_{ij} = \left(\sqrt{1/2 | sign\,(c_{ij}^{[res]}) * (c_{ij}^{[res]})^2 - sign\,(c_{ij}^{[cir]}) * (c_{ij}^{[cir]})^2 |} \right)^{\beta}$$

The cut off value for β was determined based on the scale-free topology criterion (239) using the "PickSoftThreshold" function in WGCNA. Then, the topological overlap matrix (TOM) was calculated using the adjacency matrix followed by hierarchical clustering of genes using 1-TOM as a dissimilarity measure. Finally, module identification was performed using the "Dynamic Hybrid" tree cut algorithm implemented in WGCNA (241) with the following parameters: deep split = 3 and minimum cluster size of 15. DiffCoEx assessed the statistical significance of module-wise co-expression changes between the resident and circulating groups using permutations. Briefly, 1000 permutations of sample labels between the resident and circulating groups were performed. For each permutation, the absolute mean of the module-wise correlation changes (dispersion) between the groups was calculated. This was used to generate a null distribution for each module. The *P*-value for each

module was calculated as the number of permutations with dispersion values greater than or equal to the original value. Connectivity, the degree of interaction between genes, was calculated by summing the weights of the overall edges of a gene. The median co-expression of each module in the resident and circulating groups was calculated as the median absolute value of the correlations between genes within a module. A module with a higher median co-expression in the resident group compared to the circulating group was defined as a RESIDENT module.

4.3.6 Partial correlation analysis in the resident group to infer potential network drivers of the RESIDENT module

To infer potential network drivers of the RESIDENT module, partial correlation analysis was carried out in the resident group using the "ppcor" package in R. Partial correlation can be used to infer a direct correlation relationship between genes *a* and *b* by removing the linear effect of gene *c* (603,604). As a result, the disruption of the extent of co-expression between genes within the network can be used to infer the regulatory role gene *c*. The partial correlation coefficient ($r_{ab,c}$) was calculated for all DE gene pairs (a and b) in the resident group by iteratively conditioning on each gene (gene c) present in the RESIDENT module. For each of the partial correlation matrix obtained, the mean of the difference between the absolute value of the correlation coefficient between genes a and b (r_{ab}) with and without adjusting for gene *c* was calculated, to identify which gene *c* had drastically perturbed the co-expression when its effect was removed.

$$Mean\ (|r_{ab}| - |r_{ab,c}|)$$

A high positive mean value indicates that gene *c* is a potential key hub gene, which greatly affects the co-expression amongst genes present in the resident and other modules

4.4 Results

4.4.1 Overview of the study samples and analyses

In this chapter, I utilised two previously published microarray datasets (88,103) of expression intensities generated from murine TRM cells isolated from skin, gut, lung, and brain, and their circulating memory counterparts from the spleen (**Figure 4.1**). A total of the 25 samples were analysed, which included 14 TRM cell samples from various tissues classified as "resident", and 11 samples from the spleen (TEM and TCM) classified as "circulating". All samples were profiled on the Affymetrix Mouse Gene 1.0 ST arrays, and raw data was then pre-processed using the RMA algorithm (184). The expression profiles from both studies were combined and batch-corrected with the SNM normalisation method (195). DE genes between the resident and circulating groups were identified using the limma statistical analysis package (602). These 2,197 DE genes (significant at an FDR < 0.05) were then utilised for constructing differentially coexpressed gene networks between the resident and circulating groups using the DiffCoEx tool (253). Next, to infer key module regulators, partial correlation analysis was performed to statistically knockdown each gene iteratively in the RESIDENT module, a module identified as highly coexpressed in the resident group. Finally, the 2,197 genes DE in TRM cells were used as TRM-specific transcriptional signature to assess their enrichment across a subset mouse immune cell profiles in the ImmGen dataset (598,599).

4.4.2 Global analysis of the transcriptome in resident and circulating memory T cells reveals differences between their expression profiles

Comparison of the overall distributions of probe expression between samples from the two different microarray studies showed an uneven distribution across studies, which clearly indicates study-specific batch effects (**Figure 4.2A**). After SNM normalisation, the median expression values were very similar across all samples (**Figure 4.2B**), suggesting that the normalisation procedure used performed well and the normalised data was suitable for downstream analyses.

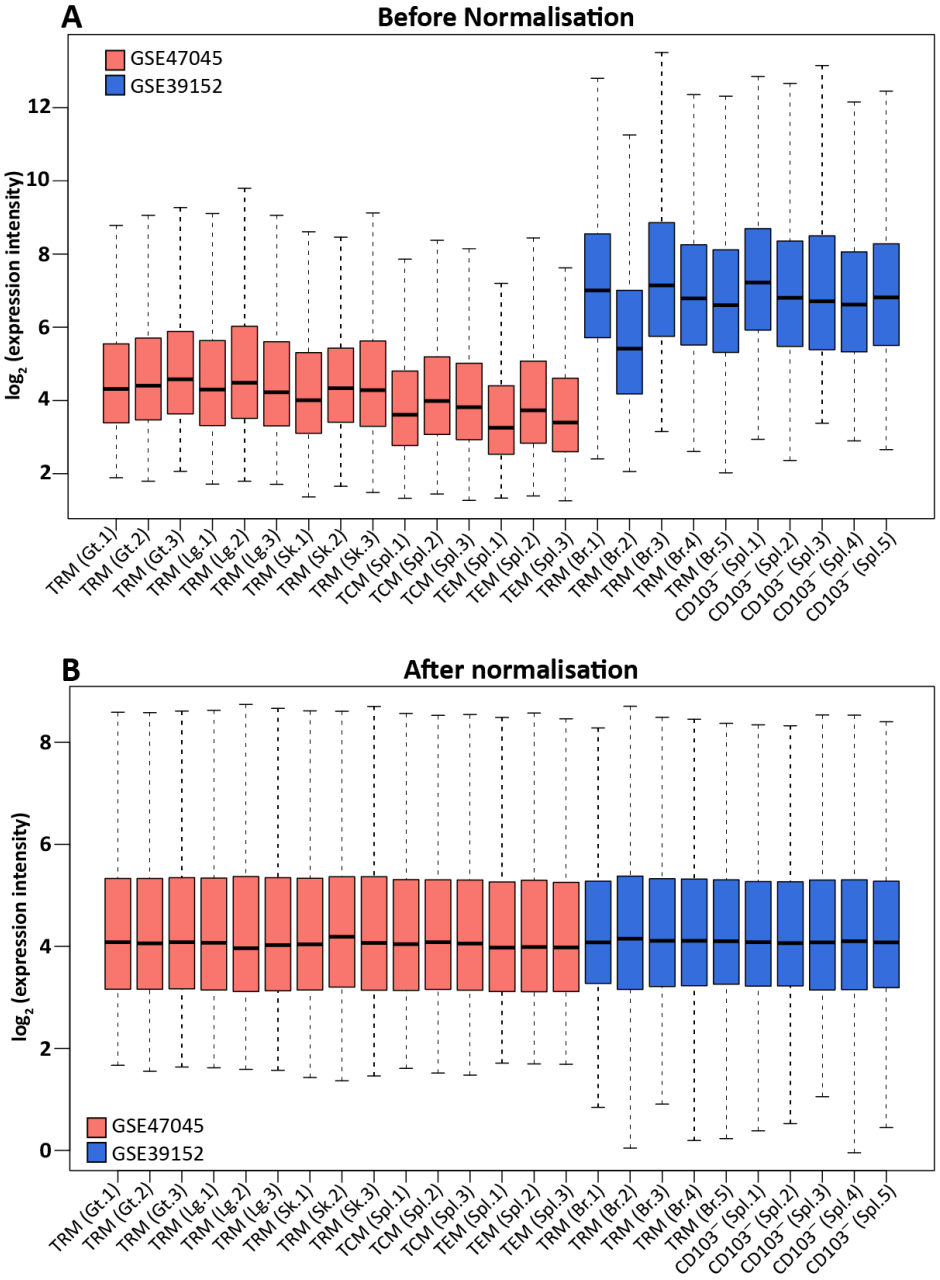


Figure 4.2: Boxplots of log2-transformed expression values of 34,760 probes across all the 25 resident and circulating samples (A) before and (B) after normalisation.

Raw expression intensities were obtained from two independent microarray studies, GSE47045 (red boxplots) and GSE39152 (blue boxplots). For each boxplot, the boxes represent the interquartile range of expression (25% –75%), and the median expression value is denoted by the horizontal black line within the box. The whiskers show the maximum and minimum expression values. The resident (TRM) samples were obtained from the gut (Gt), lungs (Lg), skin (Sk), and brain (Br). The circulating samples were obtained from the spleen (Spl: TCM, TEM and $CD103^{-}$ $CD8^{+}$ T-cells). Numbers at the end of the sample labels indicate biological replicates

An exploratory analysis was performed on 10,428 probes, after removing genes with low expression variance across the 25 samples. PCA revealed that the first five principle components (PCs) captured 47% of the variation in the expression data (**Figure 4.3A**). Grouping within the resident and circulating samples indicate that the TRM cells exhibit a distinct transcriptional signature compared to their circulating counterparts (**Figure 4.3B**). The resident and circulating samples separated along the PC1 axis, which accounted for 15% of the total variance (**Figure 4.3B**). Hierarchical clustering of the samples based on their similarity in gene expression profiles showed clear separation of the resident samples from the circulating, which further highlighted the differences in gene expression profiles between these two groups (**Figure 4.3C**).

4.4.3 DE genes in resident *vs.* circulating memory T-cells

To identify genes DE between the resident and circulating groups, differential analysis was performed on 24,534 annotated probes with the LIMMA statistical package (602). A total of 2,197 unique genes were significantly DE (FDR < 0.05) between the resident and circulating groups. Of those 1,551 were up-regulated and 646 were down-regulated in the resident group. It was observed that there were more up-regulated genes (more than 65%) in both the total set of DE genes and the list of top 50 most DE genes when ranked by statistical significance as shown in the heatmap (**Figure 4.4**). The top 10 most differentially up-regulated and down-regulated genes are given in **Table 4.1**, and the detailed list of all DE genes is provided in Table B.1 in Appendix B. *Itgae* was the most highly expressed gene in resident group. It encodes for CD103, a well-characterised TRM cell surface marker known to mediate the homing and retention of TRM cells at tissues sites (89,100,103,556,565,566). The most down-regulated gene was *S1pr1,* which encodes for a tissue exit receptor (105).

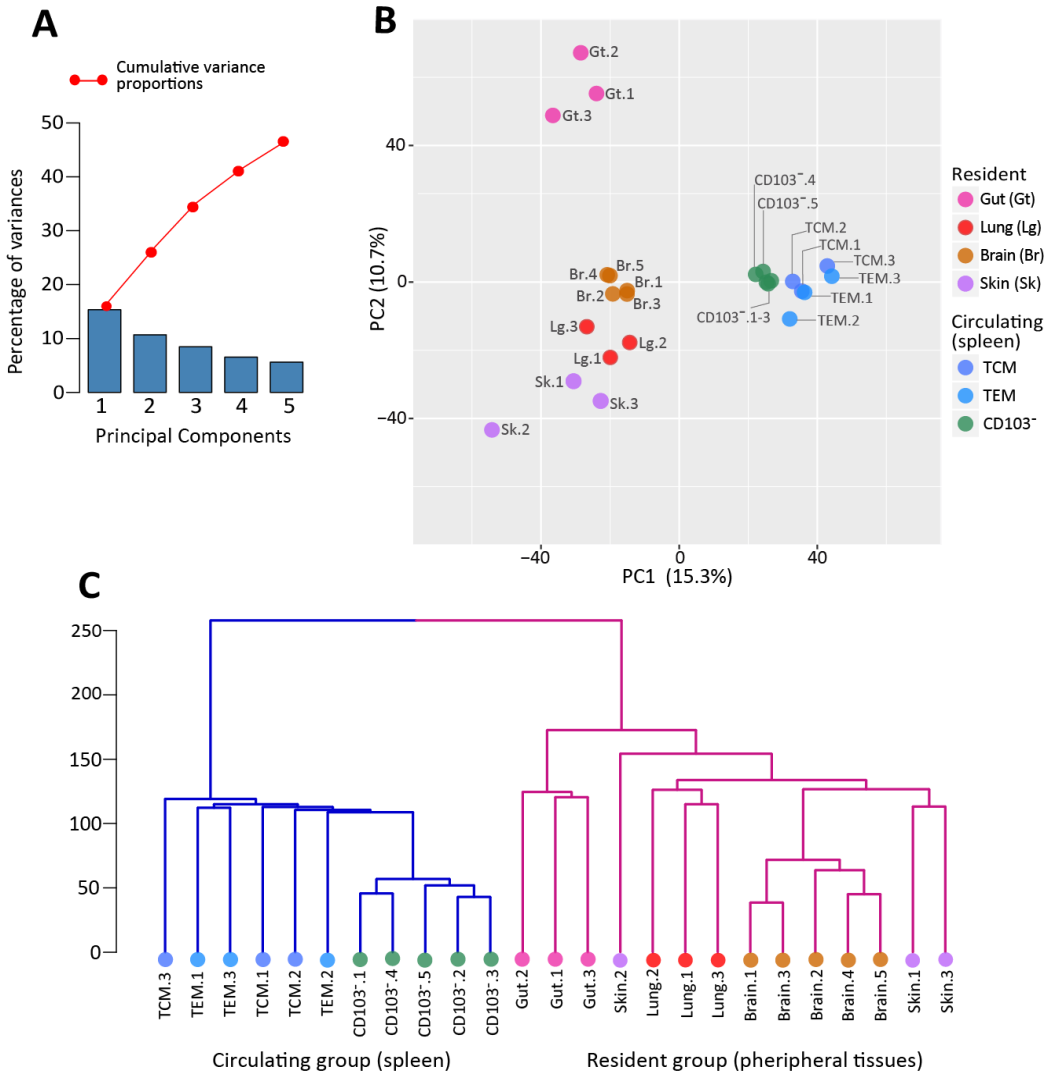


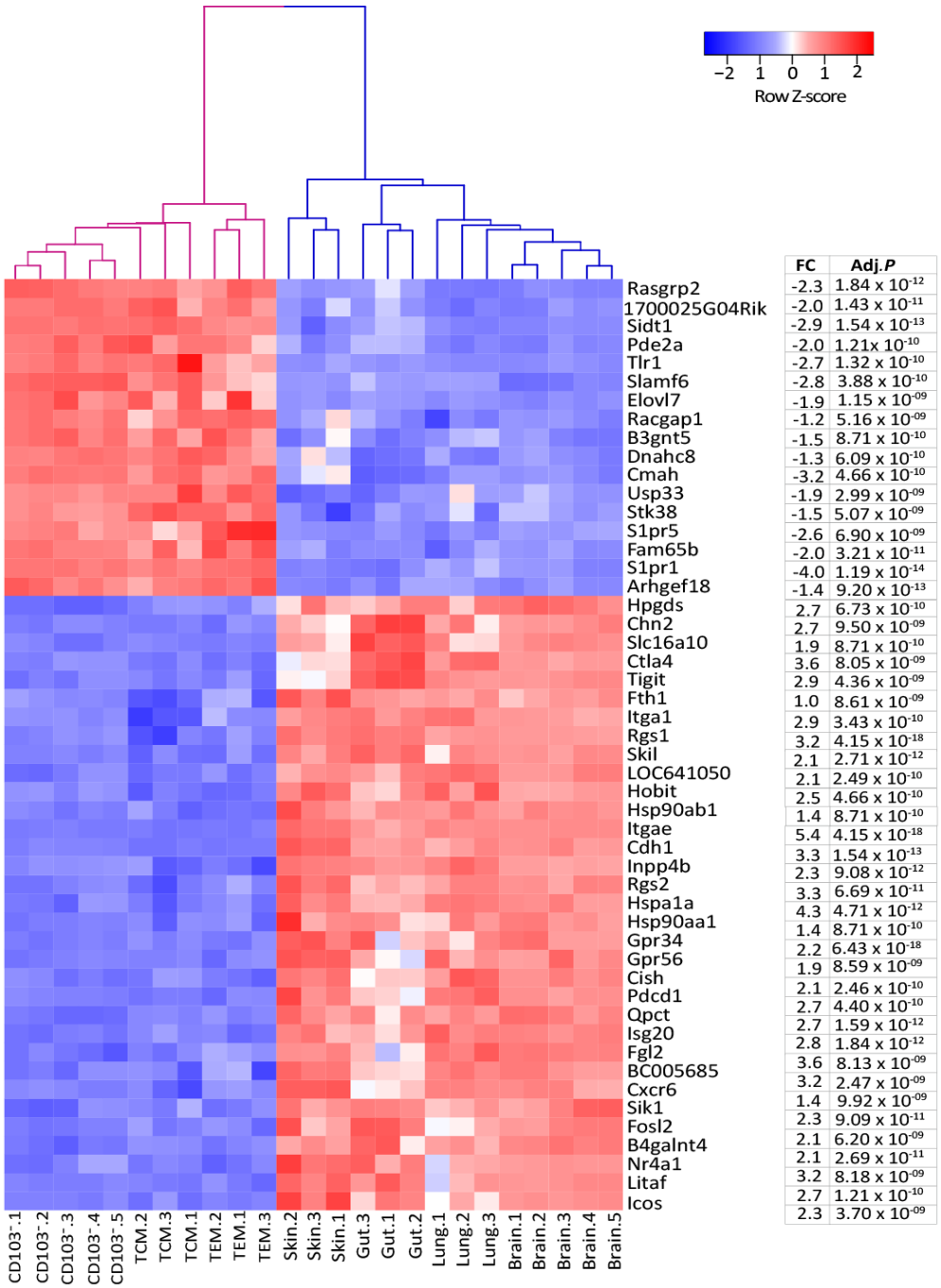
Figure 4.3: Global analysis of the expression profiles obtained across a total of 25 resident (N= 14) and circulating (N=11) samples.

(A) Principal component analysis (PCA) of the transcriptome for the 25 samples. Scree plot showing the amount of variance (bar height) captured by each of the top five principal components (PCs; bars). The cumulative proportion of variance explained by the first five PCs (red line) was 47%. **(B)** The plot of PC1 versus PC2 shows the separation of the samples into resident and circulating clusters along the PC1 axis. The numbers in parenthesis beside the PC labels denote the percentage of variance explained by the respective PCs. Each sample, represented by either three or five biological replicates (1-5), is denoted by dots. **(C)** The dendrogram obtained from the hierarchal cluster analysis (HCA) of the samples based on their transcriptome further confirms the resident *vs.* circulating separation. Clustering was done using the Ward's method with the "Euclidean" distances measure provided as the dissimilarity matrix. The dendrogram branches are coloured according to sample group: resident (blue) and circulating (pink). Dots at the tip of the leaves represent each sample. For both PCA and HCA, the dots representing samples are coloured according to the respective resident or circulating cell type, where biological replicates (indicated by numbers next to each sample label) have the same colour. All analysis was performed on 10,428 probes, log2-transformed SNM normalised.

Table 4.1: The top 10 most significantly up-regulated and down-regulated genes DE between resident *vs.* circulating groups.

Probe ID	Gene symbol	Gene name	FC (log2)	FDR Adj. *P*-value
Up-regulated				
10378286	*Itgae*	Integrin alpha E, epithelial-associated	5.36	4.15×10^{-18}
10575052	*Cdh1*	Cadherin 1	3.34	1.54×10^{-13}
10538892	*LOC641050*	Uncharacterised	2.68	2.28×10^{-13}
10447056	*Qpct*	Glutaminyl-Peptide Cyclotransferase	2.69	1.59×10^{-12}
10554240	*Isg20*	Interferon-stimulated protein	2.83	1.84×10^{-12}
10491300	*Skil*	SKI-like	2.07	2.71×10^{-12}
10451110	*Hsp90ab1*	Heat shock protein 90 alpha (cytosolic), class B member 1	1.09	3.61×10^{-12}
10450369	*Hspa1a*	Heat shock protein 1A	4.29	4.71×10^{-12}
10573082	*Inpp4b*	Inositol polyphosphate-4-phosphatase, type II	2.26	9.08×10^{-12}
10358408	*Rgs1*	Regulator of G-protein signalling 1	3.20	9.27×10^{-12}
Down-regulated				
10501586	*S1pr1*	Sphingosine-1-phosphate receptor 1	−4.04	1.19×10^{-14}
10439583	*Sidt1*	SID1 transmembrane family, member 1	−2.94	1.54×10^{-13}
10569733	*Arhgef18*	Rho/rac guanine nucleotide exchange factor (GEF) 18	−1.43	9.20×10^{-13}
10460968	*Rasgrp2*	RAS, guanyl releasing protein 2	−2.33	1.84×10^{-12}
10358717	*1700025G04-Rik*	RIKEN cDNA 1700025G04 gene	−1.98	1.43×10^{-11}
10404152	*Fam65b*	Family with sequence similarity 65, member B	−1.96	3.21×10^{-11}
10555510	*Pde2a*	Phosphodiesterase 2A, cGMP-stimulated	−2.00	1.21×10^{-10}
10530145	*Tlr1*	Toll-like receptor 1	−2.67	1.32×10^{-10}
10351691	*Slamf6*	SLAM family member 6	−2.76	3.88×10^{-10}
10404132	*Cmah*	Cytidine monophospho-N-acetylneuraminic acid hydroxylase	−3.23	4.66×10^{-10}

FC – refers to fold change. FDR Adj. – refers to FDR adjusted *P*-values. *P*-values were adjusted using the Benjamini-Hochberg false discovery rate (FDR) procedure.



FC	Adj.*P*
-2.3	1.84 x 10^{-12}
-2.0	1.43 x 10^{-11}
-2.9	1.54 x 10^{-13}
-2.0	1.21x 10^{-10}
-2.7	1.32 x 10^{-10}
-2.8	3.88 x 10^{-10}
-1.9	1.15 x 10^{-09}
-1.2	5.16 x 10^{-09}
-1.5	8.71 x 10^{-10}
-1.3	6.09 x 10^{-10}
-3.2	4.66 x 10^{-10}
-1.9	2.99 x 10^{-09}
-1.5	5.07 x 10^{-09}
-2.6	6.90 x 10^{-09}
-2.0	3.21 x 10^{-11}
-4.0	1.19 x 10^{-14}
-1.4	9.20 x 10^{-13}
2.7	6.73 x 10^{-10}
2.7	9.50 x 10^{-09}
1.9	8.71 x 10^{-10}
3.6	8.05 x 10^{-09}
2.9	4.36 x 10^{-09}
1.0	8.61 x 10^{-09}
2.9	3.43 x 10^{-10}
3.2	4.15 x 10^{-18}
2.1	2.71 x 10^{-12}
2.1	2.49 x 10^{-10}
2.5	4.66 x 10^{-10}
1.4	8.71 x 10^{-10}
5.4	4.15 x 10^{-18}
3.3	1.54 x 10^{-13}
2.3	9.08 x 10^{-12}
3.3	6.69 x 10^{-11}
4.3	4.71 x 10^{-12}
1.4	8.71 x 10^{-10}
2.2	6.43 x 10^{-18}
1.9	8.59 x 10^{-09}
2.1	2.46 x 10^{-10}
2.7	4.40 x 10^{-10}
2.7	1.59 x 10^{-12}
2.8	1.84 x 10^{-12}
3.6	8.13 x 10^{-09}
3.2	2.47 x 10^{-09}
1.4	9.92 x 10^{-09}
2.3	9.09 x 10^{-11}
2.1	6.20 x 10^{-09}
2.1	2.69 x 10^{-11}
3.2	8.18 x 10^{-09}
2.7	1.21 x 10^{-10}
2.3	3.70 x 10^{-09}

Figure 4.4: Heatmap from the hierarchical clustering of top 50 most differentially expressed genes between resident *vs.* circulating groups.

The genes were ranked according to statistical significance. The normalised expression values for each gene across the 25 samples were standardised (mean of 0 and standard deviation of 1), such that red denotes increased expression and blue denotes decreased expression. The dendrogram shows the clustering of the samples based on the expression of these 50 genes, and the branches are coloured blue for resident samples and pink for circulating samples. The log2 fold change (FC) and FDR adjusted *P*-values (Adj.*P*) for each gene is given in the table.

4.4.4 Transcription factors (TFs) and cofactors differentially expressed in resident *vs.* circulating groups

Transcription factors and their cofactors are important transcriptional regulators of gene expression and they have been shown to play a fundamental role in programming immune cell fate decisions (70). To identify differentially expressed TFs and cofactors between the resident and circulating groups, the list of DE genes was cross-referenced against known mouse TFs and cofactors downloaded from the AnimalTFDB (605). Among the 36 TFs found to be differentially expressed ($|log2FC| > 1$) between the resident and circulating groups, 27 were up-regulated and 9 were down-regulated (**Table 4.2**). Additionally, 6 transcription cofactors (5 up-regulated and 1 down-regulated) were also identified as differentially expressed (**Table 4.2**).

4.4.5 Functional enrichment analysis of DE genes

To assess function of the DE genes ($FDR < 0.05$, $|log2FC| > 1$), GO (Biological Processes) enrichment analysis was performed using GOrilla (269) on DE genes against the background list of 20,577 genes, which were present on the Mouse Gene 1.0 ST Array and also annotated to a GO term. GO enrichment was performed on three sets of DE genes: up-regulated genes, down-regulated genes, and both sets combined. Significant GO terms ($FDR < 0.05$) were further condensed into non-redundant representative terms by clustering them based on semantic similarity using the REVIGO tool (398). The analysis identified a total of 136 and 15 significantly enriched REVIGO-summarised GO terms among genes up-regulated and down-regulated in the resident group, respectively. The top 10 most over-represented GO terms for DE genes, sorted by their enrichment P-values, are shown in **Figure 4.5**.

Table 4.2: Differentially expressed transcriptional factors (TFs) and cofactors between resident *vs.* circulating groups

Probe ID	Gene symbol	FC (log2)	Adjusted *P*-value	Probe ID	Gene symbol	FC (log2)	FDR Adj. *P*-value
TFs				**TFs continued**			
10437687	*Litaf*	2.66	1.21×10^{-10}	10354111	Aff3	-1.18	1.68×10^{-4}
10520862	*Fosl2*	2.12	6.20×10^{-9}	10391301	Stat3	1.05	3.03×10^{-4}
10427035	*Nr4a1*	3.16	8.18×10^{-9}	10452633	Tgif1	1.55	3.49×10^{-4}
10589994	*Eomes*	-2.03	1.07×10^{-8}	10384725	Rel	1.21	4.92×10^{-4}
10521913	*Rbpj*	2.20	2.05×10^{-8}	10522051	Klf3	-2.95	8.36×10^{-4}
10457205	*Crem*	1.82	2.21×10^{-8}	10363735	Egr2	1.89	9.08×10^{-4}
10580282	*Junb*	1.62	5.00×10^{-8}	10425283	Maff	1.06	1.71×10^{-3}
10482772	*Nr4a2*	3.03	5.00×10^{-8}	10511416	Tox	1.03	2.11×10^{-3}
10514466	*Jun*	1.68	2.91×10^{-7}	10405918	Rsl1	-1.03	4.84×10^{-3}
10409278	*Nfil3*	2.26	4.73×10^{-7}	10468517	Mxi1	1.76	7.97×10^{-3}
10404389	*Irf4*	2.41	7.41×10^{-7}	10397346	Fos	1.56	1.29×10^{-2}
10505911	*Dmrta1*	-1.12	1.22×10^{-6}	10492997	Etv3	1.50	1.62×10^{-2}
10540472	*Bhlhe40*	1.76	1.49×10^{-6}	10368970	Prdm1	1.01	1.78×10^{-2}
10504838	*Nr4a3*	3.18	1.89×10^{-6}	10400006	Ahr	1.24	2.63×10^{-2}
10385776	*Tcf7*	-2.28	3.19×10^{-6}	10560481	Fosb	1.64	4.38×10^{-2}
10463930	*Mxi1*	1.19	3.57×10^{-6}	**Cofactors**			
10545921	*Mxd1*	1.04	3.62×10^{-6}	10491300	*Skil*	2.07	2.71×10^{-12}
10496091	*Lef1*	-1.48	5.82×10^{-6}	10520950	*Pdlim1*	−1.75	3.27×10^{-6}
10454782	*Egr1*	2.08	6.31×10^{-6}	10405994	*Med10*	1.30	1.97×10^{-5}
10572800	*Klf2*	-1.32	2.10×10^{-5}	10406551	*Ssbp2*	1.56	3.35×10^{-5}
10482448	*Zeb2*	-1.61	6.46×10^{-5}	10411126	*Jmy*	1.76	6.47×10^{-4}
10361091	*Atf3*	2.19	9.49×10^{-5}	10531707	*Lin54*	1.54	6.24×10^{-3}

FC – refers to fold change. FDR Adj. – refers to FDR adjusted *P*-values. *P*-values were adjusted using the Benjamini-Hochberg false discovery rate (FDR) procedure.

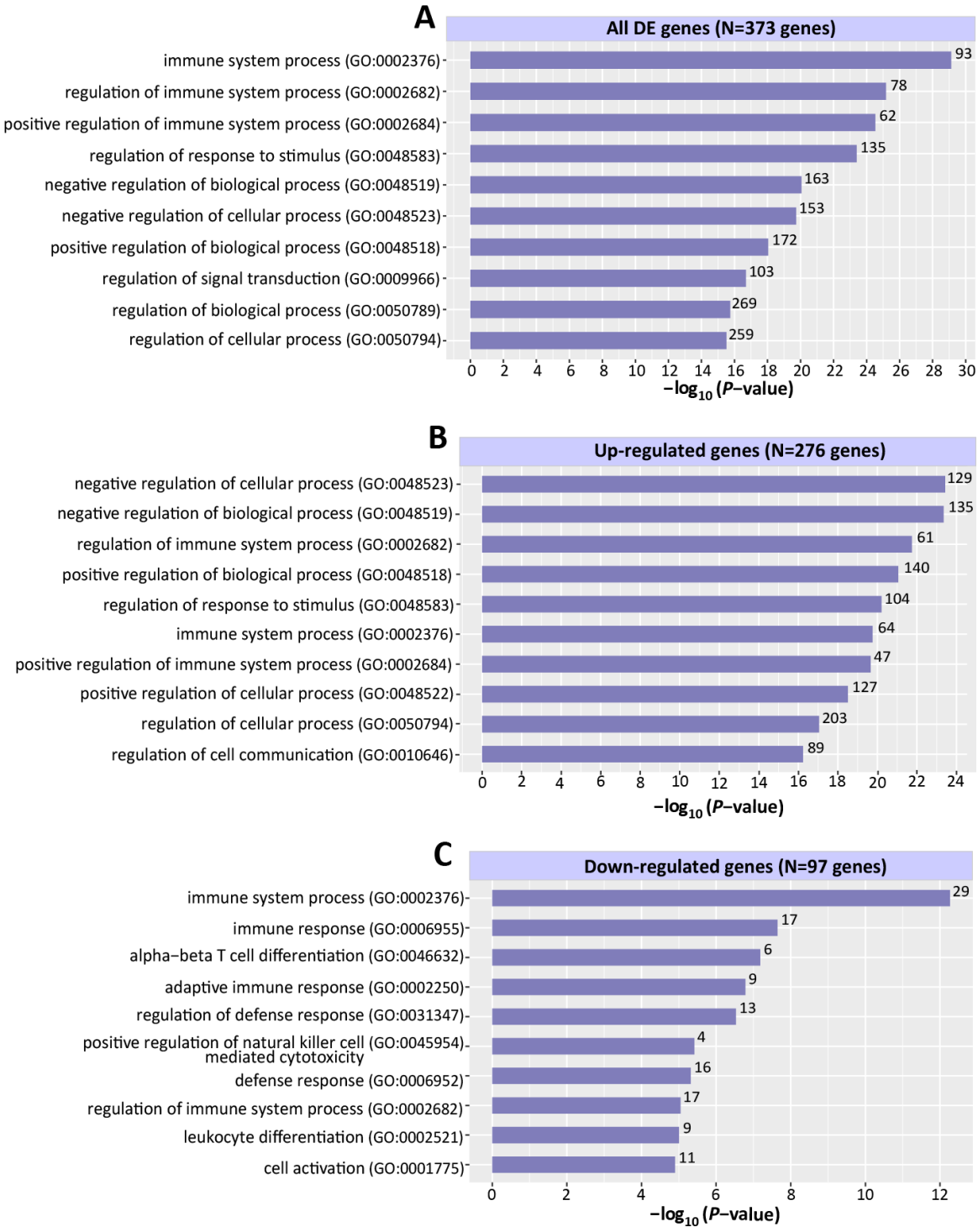


Figure 4.5: Gene Ontology (GO) terms enriched among genes differentially expressed between resident and circulating groups.

Top representative GO (biological processes) terms based on REVIGO output, enriched among **(A)** all the DE genes, **(B)** up-regulated genes, and **(C)** down-regulated genes in the resident group. The enrichment of each GO term is represented by -$\log_{10}$ (*P*-value). The numbers next to each bar denote the total number of up- or down-regulated genes with annotations for a particular GO term. All GO terms listed were significant at FDR < 0.05.

4.4.6 Network analysis identifies a RESIDENT module differentially coexpressed in the resident group

Next, to gain insight into residency-specific gene sets associated with TRM cells, I employed network analysis to identify gene networks with varying co-expression patterns across resident and circulating groups (see Methods). Differential co-expression networks were constructed for 2,197 DE genes (FDR < 0.05) in resident *vs*. circulating groups using the DiffCoEx tool (253). The resulting network was organised into 44 modules, ranging in size from 22 to 124 genes, which were identified as significantly (*P*-value < 0.001) differentially coexpressed between resident and circulating groups. The top 15 most coexpressed modules in the resident group were selected, by ranking their absolute median co-expression, and their differential co-expression patterns were compared to those in the circulating group (**Figure 4.6**). To further identify a resident-related module, the median co-expression values for all differentially coexpressed modules in both groups were plotted against each other. The co-expression of one particular module, the "blue" module (hereafter referred to as the "RESIDENT" module), was higher in the resident group (median correlation = 0.57) compared to the circulating group (median correlation = 0.30) (**Figures 4.6 – 4.7**).

4.4.7 Characterisation of the RESIDENT module

The RESIDENT module contained 88 genes; of which 71 were up-regulated and 17 were down-regulated in the resident group compared to the circulating group. The heatmap of pairwise correlations between these 88 genes in the resident group shows that majority of genes in the RESIDENT module were positively correlated (**Figure 4.8**). Functional enrichment analysis showed that this module was significantly (FDR < 0.05) enrichment for GO biological processes terms mostly related to the regulation of immune processes such as leukocyte differentiation, interleukin-6-production, of interleukin-8 production, and cell-cell adhesion (**Table 4.3**). The top 10 most highly connected hub genes, identified by ranking their connectivity, were *Tnf*, *Tjp1*, *Tigit*, *Gem*, *Dusp1*, *Csf1*, *Styk1*, *Areg*, *Fndc3a*, *Fos*.

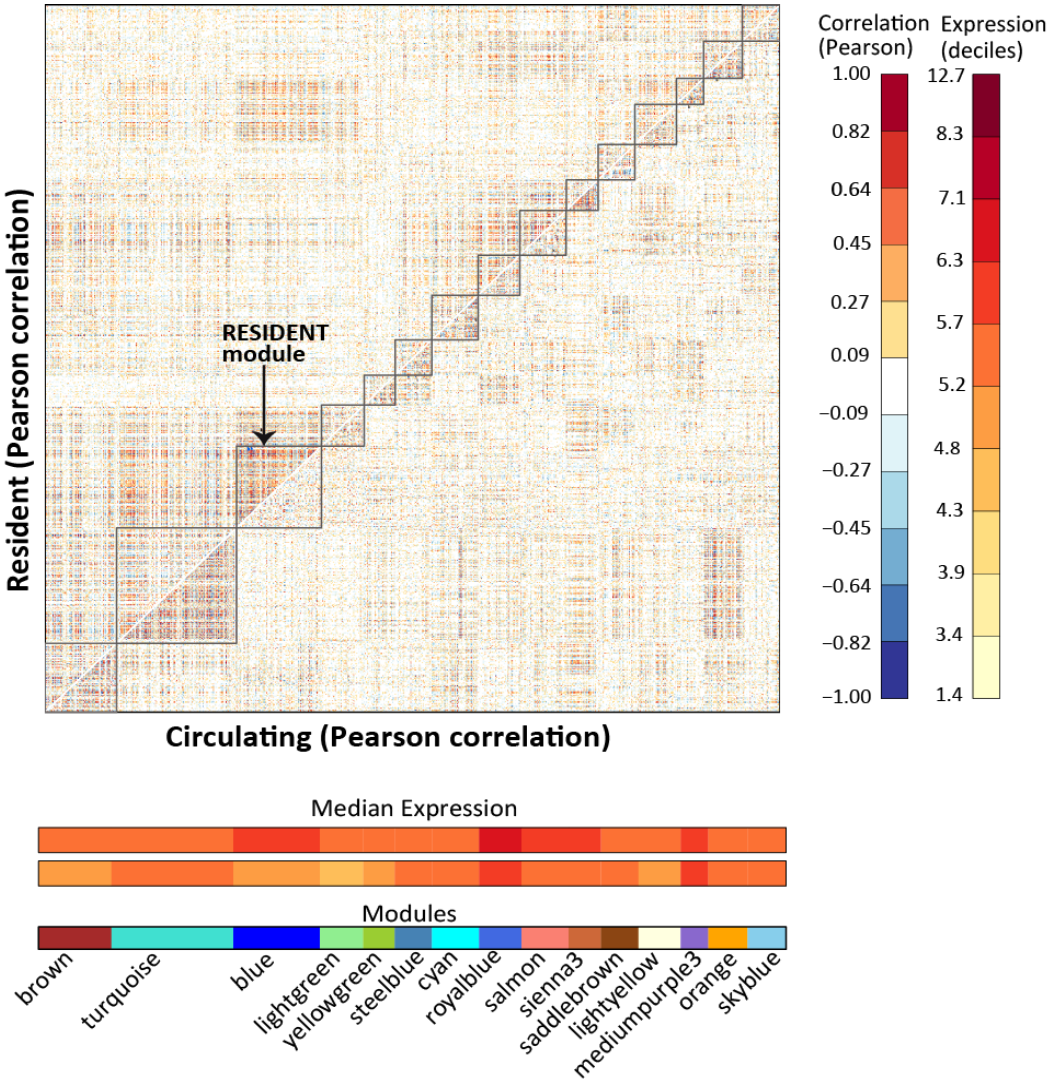


Figure 4.6: Gene modules differentially coexpressed between resident and circulating groups.

Comparative heatmap showing the intra- and inter-module correlation calculated as the pairwise Pearson correlation between genes. Red and blue denote positive and negative correlations, respectively, as depicted by the correlation colour scale on the right-hand side. The upper triangle of the symmetric heatmap shows the pairwise correlation between genes in the resident group. The lower triangle shows the pairwise correlation between the same genes in the circulating group, which follow the same order as in the resident group. The Top 15 coexpressed modules in the resident group were selected (upper triangle), and the changes in their co-expression patterns were compared with those in the circulating group (lower triangle). Modules are highlighted by black boxes on the heatmap with a corresponding colour bar at the bottom. The blue module, which is highly coexpressed in the resident group, is referred to as the RESIDENT module. Bars below the heatmap indicate the median expression of each module in the resident and circulating groups, where light yellow to dark red colour scale represents low to high median expression levels as shown by the median expression (log2-transformed) colour scale shown on the right-hand side.

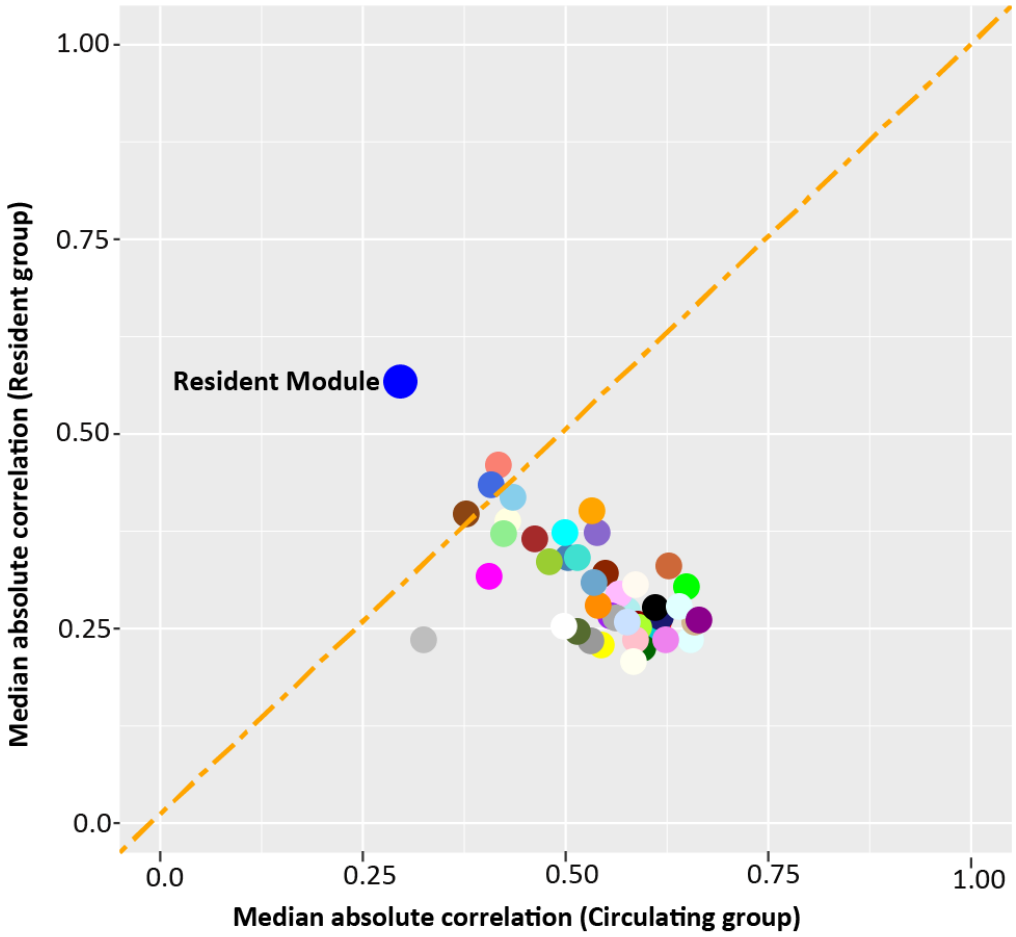


Figure 4.7: Scatter plot comparing the median correlation (absolute values) for each module in the resident (y-axis) and circulating (x-axis) groups.

The median correlation was compared for all the 44 modules detected by DiffCoEx as differentially coexpressed between resident and circulating groups. Each dot represents a module and is coloured according to their respective module colour assignment. The blue module is referred to as the RESIDENT module. The orange dashed line represents the x=y line.

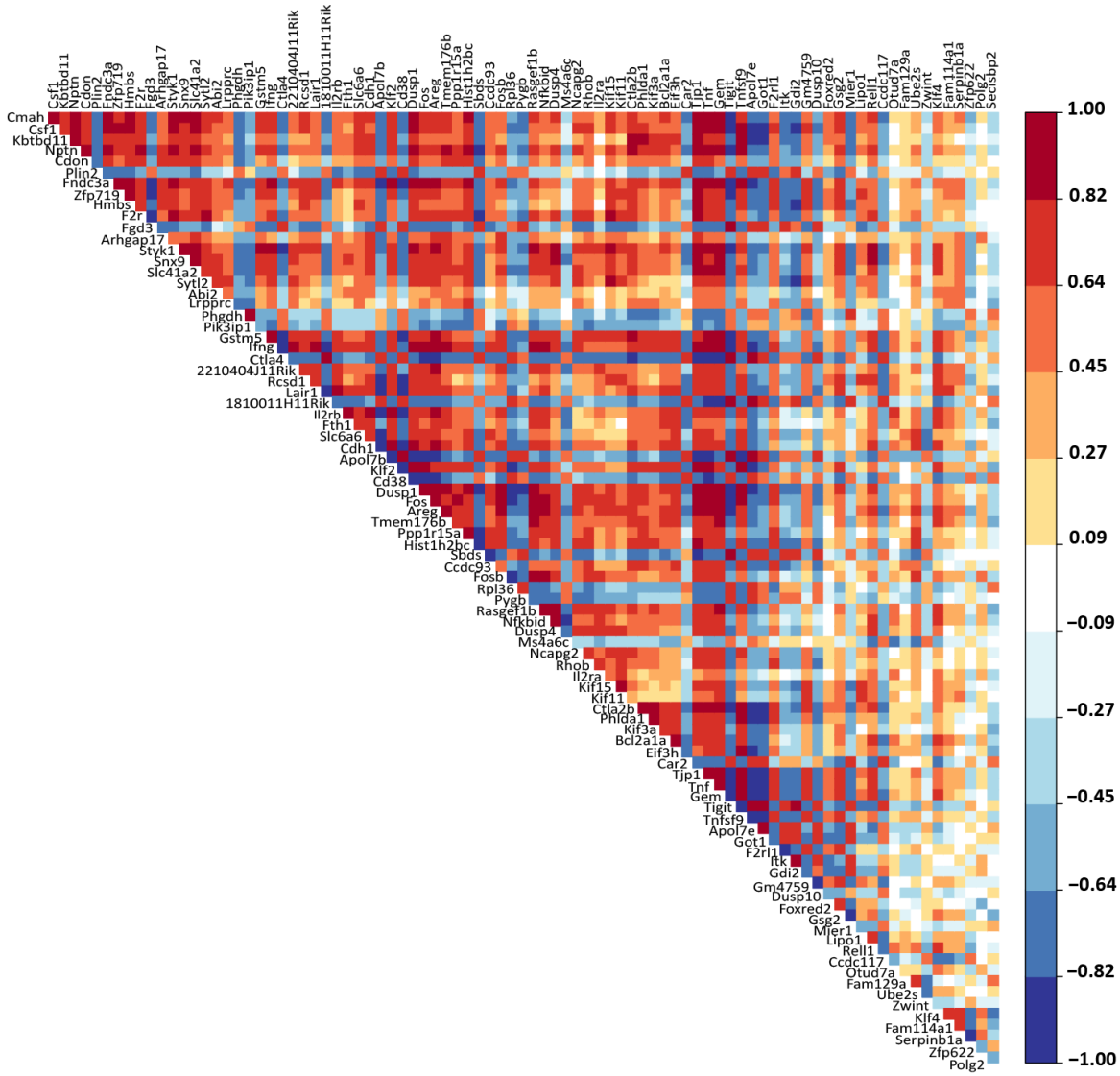


Figure 4.8: Triangular heatmap showing the pairwise correlation coefficients between genes in the RESIDENT module within the resident group.

Each square represents the Pearson correlation coefficient calculated between the genes present in the RESIDENT module within the resident group. The correlation matrix was hierarchically clustered using 1-absolute value of the correlations as the dissimilarity measure.

Table 4.3: Significant GO (biological processes) terms enriched among genes (N=88) present in the RESIDENT module.

GO terms	*P*-value	Genes with GO term annotations
Regulation of leukocyte differentiation (GO:1902105)	1.33×10^{-7}	*Tnfsf9, Ifng, Fos, Tmem176b, Il2ra, Nfkbid, Car2, Tnf, Csf1, Ctla4*
Regulation of interleukin-6 production (GO:0032675)	9.81×10^{-6}	*Tnfsf9, Ifng, F2r, F2rl1, Tnf*
Regulation of cell-cell adhesion (GO:0022407)	2.29×10^{-5}	*Tnfsf9, Klf4, Ifng, Cdh1, Nfkbid, Tnf, Tigit, Ctla4*
Positive regulation of nitric oxide metabolic process (GO:1904407)	2.68×10^{-5}	*Klf4, Klf2, Ifng, Tnf*
Positive regulation of membrane protein ectodomain proteolysis (GO:0051044)	3.20×10^{-5}	*Ifng, Snx9, Tnf*
Positive regulation of calcidiol 1-monooxygenase activity (GO:0060559)	5.29×10^{-5}	*Ifng, Tnf*
Homeostasis of number of cells (GO:0048872)	6.26×10^{-5}	*Bcl2a1a, Klf2, Kif3a, Il2ra, F2r, Csf1*
Regulation of interleukin-8 production (GO:0032677)	7.45×10^{-5}	*Klf4, F2r, F2rl1, Tnf*
Immune system process (GO:0002376)	8.24×10^{-5}	*Bcl2a1a, Hist1h2bc, Ifng, Klf2, Otud7a, Cd38, Tnfsf9, Itk, Sbds, Il2ra, Il2rb, F2rl1, Styk1, Tnf, Ctla4, Csf1*

GO – refers to Gene Ontology. The GO terms listed are significant at FDR < 0.05.

4.4.8 Partial correlation analysis infers TNF as a top potential regulator of the RESIDENT module

To infer key potential regulators of the RESIDENT module, partial correlation analysis was performed on genes in the resident group. The pairwise partial correlation coefficient was calculated between the 2,197 DE genes while iteratively conditioning on the 88 genes present in the RESIDENT module. After each iteration, the co-expression pattern changes for the top 15 aforementioned modules before and after partial correlation analysis was observed. The calculated average of the co-expression differences across all these modules revealed that conditioning on *Tnf* produced the greatest disruption in the co-expression patterns (**Figure 4.9**) *Tnf*, which was the most highly connected gene in the RESIDENT module, greatly affected the coexpresssion within the RESIDENT module, causing a decrease in median module correlation from 0.57 to 0.01 after partial correlation analysis. Additionally, the inter-module co-expression between the RESIDENT and other modules was also lost (**Figure 4.9**).

4.4.9 The underlying transcriptional program in TRM cells extends to other tissue-residing lymphocyte populations

Next, I asked if the transcriptomic fingerprint defining TRM cells shared similarity with other murine tissue resident lymphocyte subpopulations present in the ImmGen dataset (93,94). To do so, comparative analysis of the TRM-associated genes, identified as DE ($|\log2FC| > 1.5$) between resident and circulating groups in section 4.4.3 of this chapter, was performed on a subset of ImmGen immune cells. Expression profiles of ImmGen cell types including splenic $CD8^+$ effector T-cells, $CD8^+$ memory T-cells (brain TRM, TCM, and TEM cells), Tregs (from the adipose, lymph node, and spleen), and NKT cells (from the liver, lung and spleen) were obtained and averaged across replicates. It is worth noting that the brain TRM samples present in the ImmGen data and those analysed in this chapter are the same.

Comparison of expression profiles of 165 TRM-associated genes across the 26 ImmGen samples revealed that the expression profile of adipose Tregs was more similar to brain TRM and CD103− T-cells than its circulating counterparts (**Figure 4.10**). The Pearson correlation coefficients calculated for the TRM-associated genes

across these samples ranged from 0.65 (adipose Tregs *vs.* brain TRM cells) to 0.74 (adipose Tregs *vs.* brain $CD103^-$ T-cells). The resulting dendrogram generated from the hierarchical clustering of the samples based on their similarity in expression profiles further subdivided the samples. I observed a clear separation of the adipose Tregs and brain samples into a separate cluster (**Figure 4.10**; red branches). The circulating effector and memory $CD8^+$ T-cells, and NKT cells were also seen to cluster into two distinct groups (**Figure 4.10**; green and blue branches). Furthermore, I identified genes that had been previously established as part of the core TRM transcriptional signature (103) to be similarly expressed in both adipose Tregs and brain TRM cells. This included up-regulated genes such as *Skil, Vps37b*, *Nr4a1*, *Nr4a2*, *Hspa1a*, *Sik1*, *Ctla4*, *Rgs2, Fgl2*, and *Inpp4b*, and down-regulated genes such as *Cmah*, *S1pr5*, *Sidt1*, *Slamf6, Elovl7,* and *Fgf13*. These results indicate that resident adipose Tregs and brain TRM cells share a notable similarity in expression of genes related to tissue residency.

To further demonstrate that the similarity seen between adipose Tregs and brain TRM cells is related to the residency phenotype as opposed to tissue similarities, I compared the global expression profiles of most highly expressed genes across the 26 ImmGen cell types. Using an arbitrary cut-off of standard deviation < 0.35 (default is usually between 0.2 – 0.5), 7,474 representative set of genes were identified as highly expressed. Based on their overall expression these two resident cells types appear quite different from each other, which most likely reflect tissue-specific and function-specific differences (**Figure 4.11**). Hence, this further confirms that the observed similarity of the residency genes is not merely a reflection of average similarity of expression profiles in the brain and adipocyte.

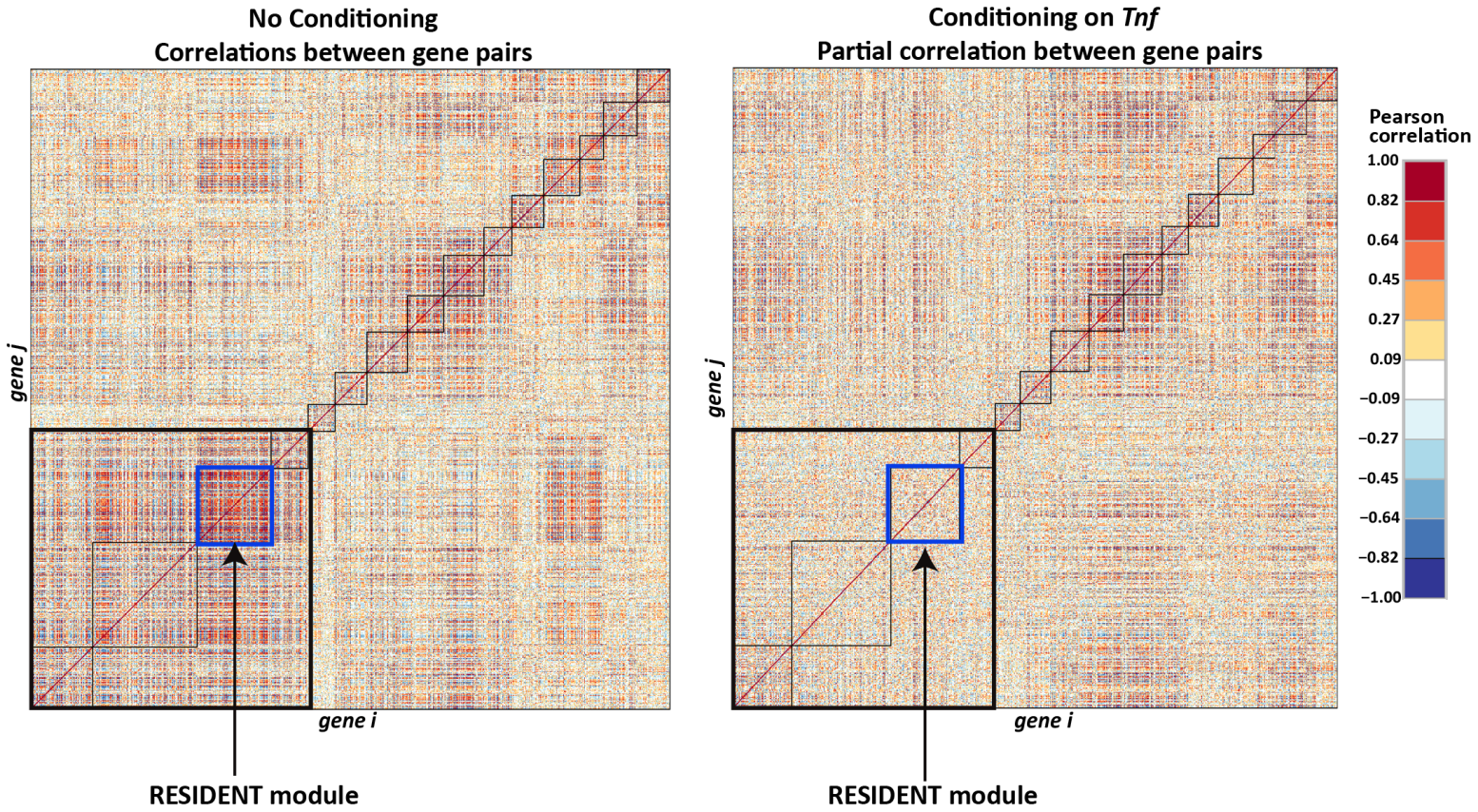


Figure 4.9: Heatmap comparing the co-expression changes between genes in the resident group (A) before and (B) after conditioning on *Tnf* gene.

The heatmap is symmetrical along the diagonal, and the top 15 coexpressed modules in the resident group are shown. The black box indicates the loss of inter-module co-expression between the RESIDENT and other modules after partial correlation analysis.

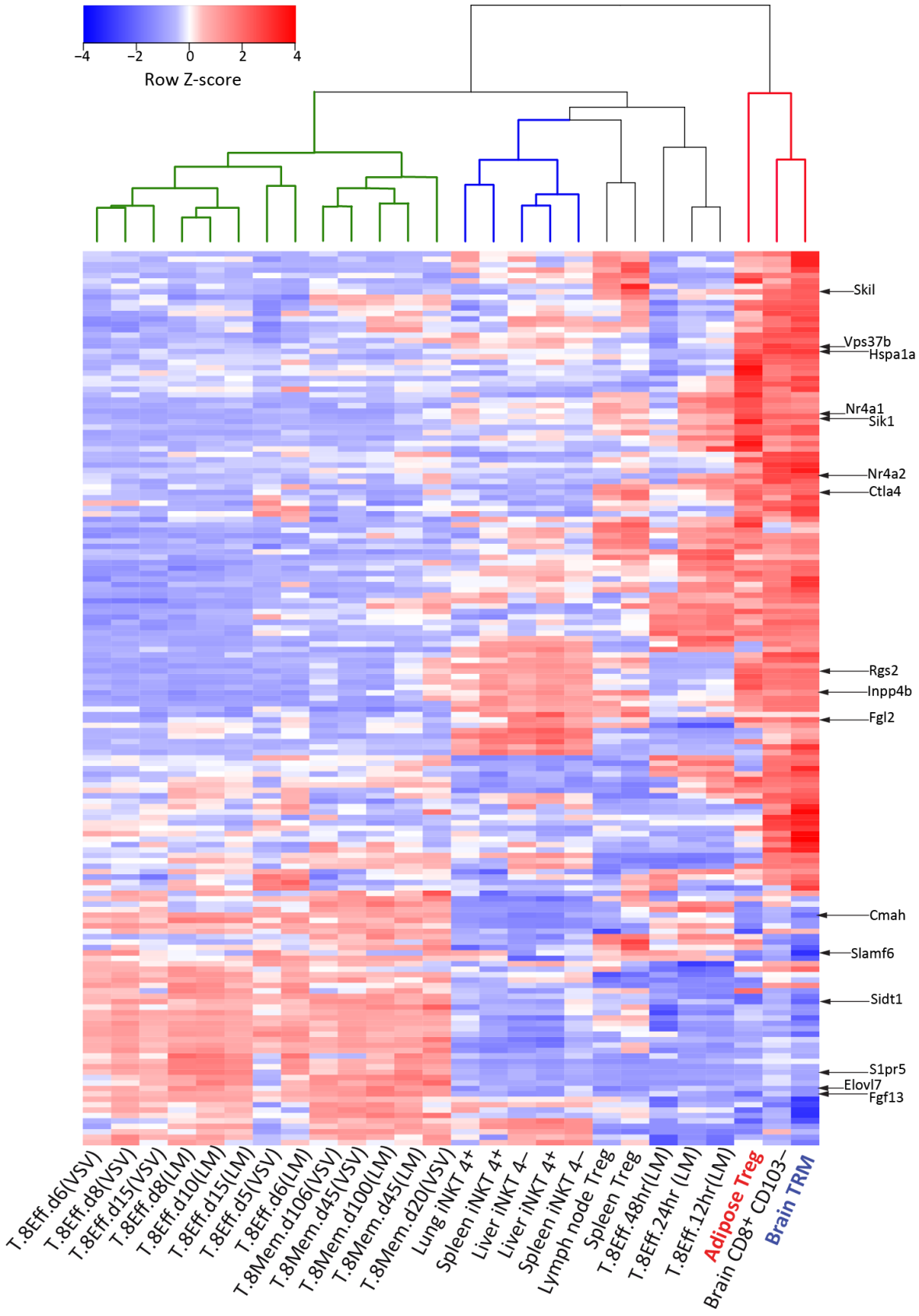


Figure 4.10: Heatmap of 165 TRM-associated genes across 26 lymphocyte populations obtained from the ImmGen data.

Figure 4.10: Heatmap of 165 TRM-associated genes across 26 lymphocyte populations obtained from the ImmGen data.

Each row represents a gene, and each column represents an immune cells type obtained from the ImmGen data. The expression values for each gene across the 26 samples were standardised (mean of 0 and standard deviation of 1), such that red denotes increased expression and blue denotes decreased expression. The dendrogram shows the clustering of the samples based on the expression of these 165 genes. The samples clustered into distinct groups based on their expression similarity, which is denoted by the colour of the dendrogram branches: circulating effector and memory $CD8^+$ T-cells (green); NKT cells (blue); brain TRM cells, brain CD103− T-cells, and adipose Tregs (red). The genes labelled are part of the core transcriptional signature previously defined in TRM cells (103), which show consistent expression across adipose Tregs and brain TRM cells.

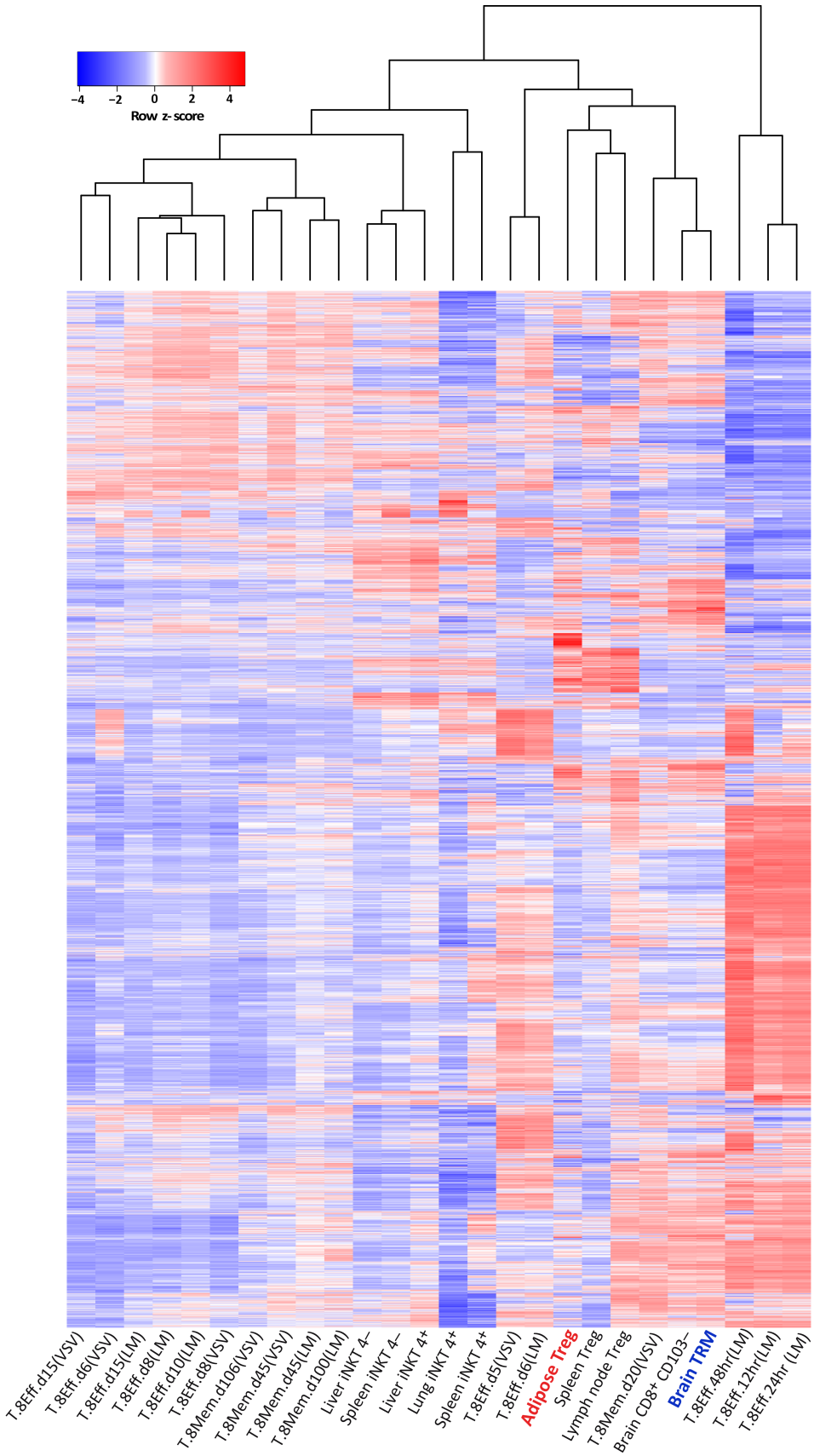


Figure 4.11: Heatmap of global expression profiles of the 26 lymphocyte populations obtained from the ImmGen data.

Figure 4.11: Heatmap of global expression profiles of the 26 lymphocyte populations obtained from the ImmGen data.

The global expression profiles of 7,474 most highly expressed genes across the 26 cells types. Each row represents a gene, and each column represents an immune cells type obtained from the ImmGen data. The expression values for each gene across the 26 samples were standardised (mean of 0 and standard deviation of 1), such that red denotes increased expression and blue denotes decreased expression. The dendrogram shows the clustering of the samples based on the expression of 7,474 genes. The samples clustered into distinct groups based on their expression similarity.

4.5 Discussion

The role of TRM cells in providing localised protective immunity at tissue niches is well documented. Understanding the mechanisms that regulate the homing and long-term maintenance of these cells has implications for TRM-based vaccine design. Although traditional transcriptome-wide analysis of TRM cells from various tissues has been reported, this approach fails to provide insight into regulatory interactions among the DE genes underlying tissue residency.

In this chapter, I took advantage of the growing number of publically available microarray data sets from TRM cells isolated from a range of tissues and circulating splenic TCM and TEM cells to perform a network-based analysis. First, I confirmed findings of previous studies that TRM cells from various tissues cluster together and their transcriptional profiles are more similar to each other than their circulating counterparts (88,103). TRM cells have been shown to exhibit a distinct transcriptional signature (88,103,582). In this chapter, more than 2,000 genes were identified as differentially expressed between the resident and circulating groups. This DE gene list included all the 37 genes previously established as the core TRM signature (103).

Next, differential network analysis was used to identify a network of coexpressed genes with altered connectivities between the resident and circulating groups. This led to the identification of a residency-related sub-network that was highly coexpressed in the resident group. The RESIDENT module comprised of genes that are involved in diverse functions such as transcription (*Fos*, *Fosb*, *Klf2*, and *Klf4*), cytokine/cytokine receptor signalling (*Ifng*, *Tnfsf9*, *Il2ra*, *Il2rb*, and *Tnf*), adhesion (*Tigit*, *Cd38*, *Cdh1*, and *Fndc3a*), and cell development (*Areg*, *Csf1*, and *Styk1*). The GO terms associated with genes present in this module further revealed that interactions between diverse biological processes are required for residency of TRM cells. Moreover, two key genes, *Cdh1* and *Klf2*, previously described to play a crucial role in homing TRM cells within tissues (95,101,103,105,556,558,563,564) were part of the RESIDENT module. Hence, it is likely that the RESIDENT sub-network plays a role in mediating tissue retention of TRM cells.

The RESIDENT module identified through network analysis contains 88 genes that were highly correlated with each other, which poses a challenge to identify putative regulatory genes. Studies have shown that key driver analysis can highlight novel condition-specific (disease) genes (232,606). To further infer key residency-related regulatory genes, partial correlation analysis was performed with genes in this module. Here, *Tnf* was inferred as a potential key driver of the RESIDENT module. As a result of *Tnf* conditioning, the intra-and inter-module disruptions in co-regulation patterns found in the resident group indicates that this gene is not only a potential key regulator of the RESIDENT module, but also facilitates interaction between the RESIDENT and other modules required for tissue residency. *Tnf* has been shown to be expressed by both effector $CD8^+$ T-cells as well as antigen exposed memory subsets (70,607). Similarly, in agreement with these studies, *Tnf* was found to be up-regulated in TRM cells. This finding also provides further support for the role of cytokines in tissue residency, where previous studies have shown that TRM cells produce cytokines upon antigen re-exposure (608,609). Additionally, the role of tissue-derived cytokines including TNF in TRM cell development has also been established (105,556). TNF induces the expression of CD69, a key surface molecule expressed on TRM cells that facilitate TRM cell homing by inhibiting tissue egress signals (105,556). *Tnf* represses *Klf2* expression (105), a gene that is also a part of the RESIDENT module. The down-regulation of both *Klf2* and its downstream target *S1pr1* promotes tissue retention by disabling TRM cells to respond to chemotactic signals required for tissue exit (103,533,574). These studies further provide support for the role of TNF in regulating the tissue homing of TRM cells. The fact that TNF is produced both locally in tissues and by TRM cells suggests that an increased or sustained dose of the TNF signal might be required for the retention of TRM cells within tissues.

There is considerable evidence indicating that a number of immune cell subpopulations tend to be tissue resident and disconnected from circulation (610,611). This raises the question whether these resident cells are transcriptionally related to each other. Finally, based on available expression data from various immune cells, I demonstrated that residency-related transcriptional similarities exist between brain TRM cells and adipose Tregs. Consistent with two recent studies, which also confirm the validity of my analysis, this finding suggests that shared commonality exists

between the transcriptional programs driving residency across various subsets of tissue-resident lymphocytes (545,582). As previously highlighted, genes linked to promoting tissue homing and suppressing tissue exit tend to be similarly expressed across resident populations (103,582). However, many of the signature TRM cell genes were not consistently expressed in adipose Tregs, e.g. *Itgae*, *Klf2, and S1pr1*, reflecting the tissue-specific influence on the transcriptional profile. In agreement with other studies, it was seen that adipose Tregs did not express the key TRM cell surface marker CD103 (*Itgae*) (591,612). Effector and memory-like Tregs, but not resident adipose Tregs, have been shown to express high levels of CD103 (613). This raises the fundamental question about the necessity of CD103 expression for residency. Hence, further supports the idea that tight local transcriptional regulation might be important for the specialised functions of resident immune cells at tissue sites.

A limitation of this study is that the sample size was small. In such case a subset of outlying expression data points can lead to extremely high Pearson correlations, particularly when most of the data points are not correlated (614). This may impact the scale-free topology fit. As a result, there is a possibility of detecting unstable gene networks. However, studies have shown that WGCNA perform well with relatively small samples size (~ $n<20$) and can robustly infer biologically meaningful gene modules (395,615,616). As a precaution to avoid detecting noisy gene networks (false positives) for the reasons discussed above, gene networks in this chapter were inferred from differentially expressed genes.

In summary, this chapter provides a residency-related co-expression sub-network, furthering our understanding into the molecular mechanisms underpinning TRM cell development and persistence in tissues. Additionally, results of this chapter consistently support previous findings that the transcriptional signatures of tissue resident populations are shared to some extent. This shared feature might be an essential requisite for tissue dwelling. Understanding the molecular mechanisms underlying tissue residency and the degree to which they are shared may provide insights for designing vaccines and immunotherapies that will provide rapid and site-specific immunity.

Chapter 5

Comparative transcriptional analysis reveals the role of TGF-β in defining the transcriptional signature in TRM cells

5.1 Introduction

This chapter focuses on the role of transforming growth factor-beta (TGF-β) signalling, a local extrinsic factor present at tissue sites, in influencing the transcriptional program of TRM cells.

The local tissue-derived signals that control the development and persistence of TRM cells at tissue sites is not well understood. The role of cytokines in the differentiation and maintenance of circulating memory T-cell subsets is well documented (70,607). Few studies have established links between local tissue-derived cytokines and tissue residency (98,103,105,556,569,570), but the precise mechanisms by which these cytokines regulate the establishment of TRM cells are lacking. In particular, TGF-β has been shown to regulate few key genes involved in the homing and tissue egress of TRM cells. Recently, through comparative transcriptional analysis, it has become clear that TRM cells isolated from various tissues exhibit overlapping residency-related transcriptional signature that distinguishes them from their circuiting TCM and TEM counterparts (88,103). Even though it is known that TGF-β imprints homing potential on TRM cells to some extent (98,103,556,569,570), its role in shaping the transcriptional signature of TRM cells is still not clear. *In vitro* characterisation of the TGF-βinduced

transcriptional signature in $CD8^+$ T-cells will provide insight into the underlying mechanisms by which TGF-β signalling at tissue sites regulate residency of TRM cells.

5.1.1 TGF-β plays a role in up-regulating CD103 and the acquisition of TRM phenotype

As mentioned previously, several studies have demonstrated that TRM-specific cell surface markers CD103 and CD69 are crucial for TRM cell formation and retention in multiple tissues (95,103,556,558). However, factors regulating their expression remain largely unknown. The tissue-restricted expression of CD103 and CD69 suggests that TRM cells are under the influence of local tissue-derived signals. Studies have implicated the role of local cytokines such as IL-15, IL-33, *TGF-, and TNF* in the acquisition of residency in TRM cells (98,103,105,556,569,570).

In particular, accumulating evidence has shown that TGF-β activity is critical for the development of TRM cells in the skin, gut and lungs via the induction of CD103 (98,103,556,569,570). TGF-β is commonly appreciated as an anti-inflammatory cytokine for $CD8^+$ T-cells and is essential to prevent autoimmunity and maintain immune homoeostasis (617). TGF-β signalling is initiated by the binding of active TGF-β to the extracellular domain of the TGF-β type II receptor (TGF-βRII), which then phosphorylates and activates TGF-β type I receptor (TGF-βRI) followed by the activation of both smad-dependent and -independent signalling pathways (618). It has been known for some time that addition of TGF-β to *in vitro* cultures greatly enhances the expression of CD103 on effector $CD8^+$ T-cells obtained from mice and humans (556,619–622). Moreover, constitutive expression of TGF-β has been observed at epithelial sites, including the skin and small intestine (623–625).

Also*, in vivo* studies have also demonstrated that TGF-β induces the up-regulation of CD103 on T-cells, which then promotes their retention in several peripheral tissues (556,569,626). Using a mouse model of graft-versus-host disease (GVHD), El-Asady *et al.* showed that effector $CD8^+$ T-cells expressing *a* dominant-negative mutant TGF-βRII (dnTGF-βRII) were devoid of CD103 expression in the small intestinal epithelium (626). In a similar finding, Casey *et al.* reported that transgenic mice deficient in TGF-

βRII were defective of CD103 expression on $CD8^+$ T-cells infiltrating the gut following infection with either vesicular stomatitis virus encoding ovalbumin (VSV-Ova) or lymphocytic choriomeningitis virus (LCMV) in separate experiments (556). Their findings further provided support that TGF-β dependent induction of CD103 on these cells was a necessary requisite for their maintenance in the gut (556). Likewise, $CD8^+$ T-cells from dnTGF-βRII mice that migrate to the lungs following influenza virus infection failed to express CD103 (569). Although the findings of these studies provide insight into the role of TGF-β in driving the residency phenotype in $CD8^+$ T-cells, they have focused on T-cells from the effector phase of immune response. Besides, these earlier work have utilised a transgenic mouse model expressing the truncated form of the TGF-βRII (missing the kinase domain), which has been shown to be mildly functional (627). As a result, detectable levels of CD103 expression were noted in cells with this leaky mutation (556,569,626).

More recent studies have assessed the role of TGF-β signalling in retaining memory phase $CD8^+$ T-cells and have employed a model system that is completely deficient in TGF-βRII activity on T-cells (98,103,570). Zhang and Bevan showed that conditional deletion of TGF-βRII on T-cells led to dramatically reduced expression of retention markers CD103, CD69, and integrin β7 in the IEL compartment of the gut following LCMV infection (570). They further showed that absence of TGF-β responsiveness severely impaired the homing and long-term retention of TRM cells (570). Another study utilising the same transgenic model system also reported consistent findings, whereby TRM cells that were incapable of responding to TGF-β could not be maintained in the gut after oral infection with Listeria monocytogenes (98). Similarly, defective generation and maintenance of TRM cells devoid of TGF-βRII were observed in skin (103).

5.1.2 Existing gap in understanding the role of TGF-β in establishing tissue residency in TRM cells

The above studies have not only provided compelling evidence to support the role of TGF-β in the development and lodgement of TRM cells in the peripheral tissues, but

have also implicated its role in TRM cell maintenance. However, since these studies have utilised a knockout system where TRM development was already impaired, there has been no direct assessment of whether or to what extent TGF-β signalling is required for the maintenance of TRM cells in tissues. Also, previous studies investigating the role of TGF-β in TRM cell development have only assessed a few TRM-related genes. Moreover, it has recently been shown that TRM cells isolated from various tissues share a core transcriptional program, suggesting a common molecular machinery underlying their development, maintenance, and possibly function in peripheral tissues (103). Hence, characterising the global transcript of $CD8^+$ T-cells induced *in vitro* by TGF-β and linking this transcriptional profile to the previously identified core TRM signature will give insight into the role of TGF-β signalling in establishing the residency phenotype in TRM cells.

5.2 Research objectives

The central aim of this chapter was to utilise an *in vitro* inducible model system with exogenous TGF-β to understand the role of TGF-β signalling in establishing tissue residency in TRM cells.

The specific objectives of this research chapter were:

1. To use RNA-seq analysis to characterise the global transcriptional profiles of murine $CD8^+$ T-cells stimulated *in vitro* with TGF-β.

2. To compare the overall transcriptional profiles of the TGF-β stimulated T-cells with the known core TRM-specific transcriptional signature previously established in TRM cells isolated from murine lung, skin, and gut.

5.3 Methods

5.3.1 *In vitro* cell culture and RNA extraction

The schematic diagram of the experimental protocol is illustrated in **Figure 5.1**. C57BL/6 (wild-type [WT] B6) and gBT-I female mice, 8 to 15 weeks old, used in this study were bred and maintained under specific pathogen-free conditions in the Department of Microbiology and Immunology, University of Melbourne. The gBT-I mice express a transgenic T-cell receptor (TCR) that recognises the herpes simplex virus type 1 (HSV-1) glycoprotein B (gB) peptide (628). Spleen was harvested from gBT-I and C57BL/6 mice and processed into single cell suspensions by teasing the cells through a mesh. The B6 cells were incubated with gB peptide (100 μg/ml) at 37°C for 15 minutes, washed and stimulated with 2 mg/ml of Lipopolysaccharides (LPS) in 10ml RPMI-1640 liquid medium. 5ml of peptide coated B6 splenocytes were added to gBT-I cells present in each of the two T75 flasks containing half of a spleen cell suspension in 35ml of RPMI-1640 liquid medium. The gBT-I cells in culture were then activated and expanded with 500 U/ml of recombinant human interleukin-2 (rhIL-2), which was added on day 2, 3, and 4. At day 5, cells were seeded at 12 million/5ml RPMI-10 in 4 conditions: untreated cells (Untreated); cells treated with TGF-β only (TGF-β); cells treated with IL-2 only (IL-2); cells treated with both IL-2 and TGF-β (IL-2/TGF-β). Subsequently, the TGF-β treatment groups (TGF-β; IL-2/TGF-β) were then stimulated with 3 ng/ml of TGF-β for 40 hours. The experiment was repeated three times, for three independent biological replicates. A total of 12 samples were prepared.

Total RNA was extracted from each sample by adding 200μl chloroform per 1mL TRIzol directly to cells, vortexing briefly, and incubating at room temperature for 5 minutes. The samples were centrifuged at 12,000g for 20 minutes at 4°C. The colourless upper aqueous layer was transferred to a new tube containing 500μl of Propan-2-ol, kept at room temperature for 10 minutes, and then centrifuged at 12,000g for 20 minutes. The supernatant was removed, RNA pellet was washed with 1ml of 75% ethanol, and the samples were centrifuged at 7,500g for 5 minutes at 4°C.

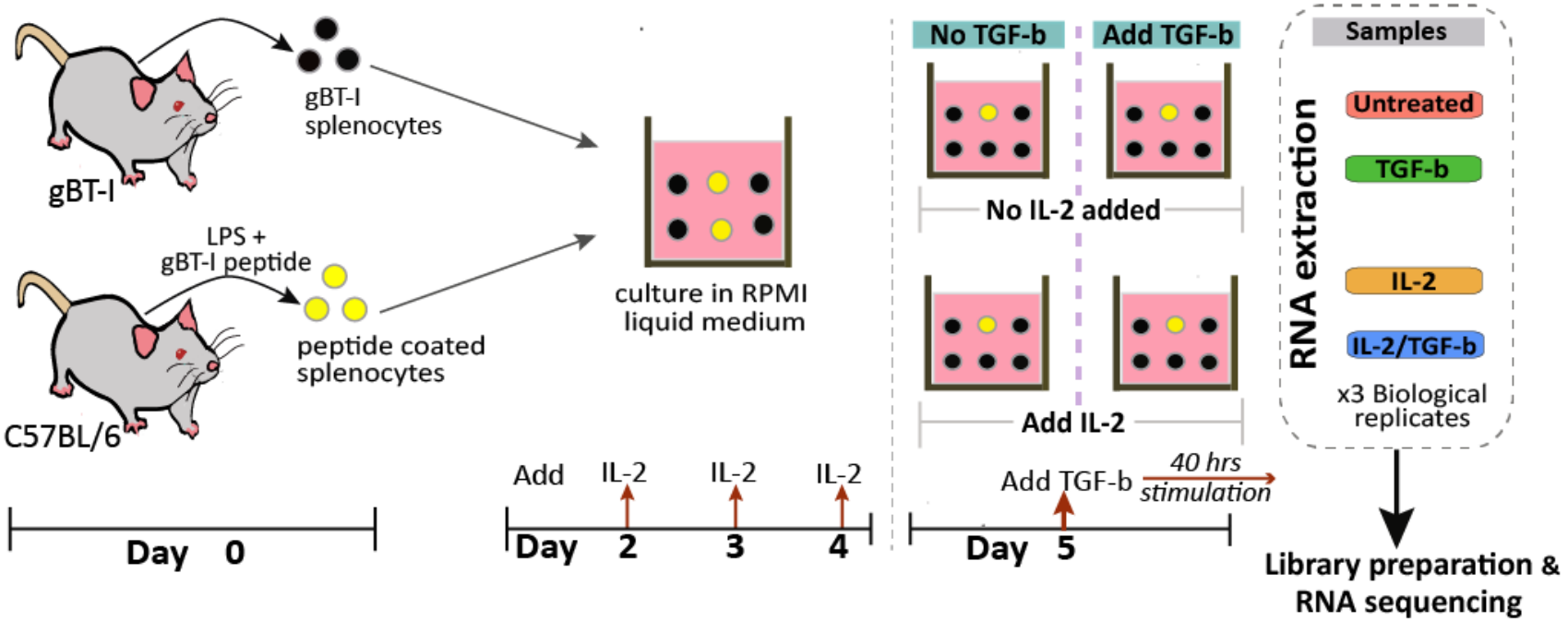


Figure 5.1: Schematic overview of the experimental design.

gBT-I splenocytes from gBT-I mice were harvested and co-cultured with gBT-I peptide coated splenocytes harvested from C57BL/6 mice. Exogenous IL-2 was added at day 2, 3 and 4 to stimulate the gBT-I $CD8^+$ T cells, which recognise the gBT-I peptide and subsequently expand and differentiate into effector cells. On day 5, the cells were stimulated with TGF-β for 40 hours in the presence or absence of IL-2. Four different samples were prepared from the *in vitro* culture: untreated cells (Untreated); cells treated with TGF-β only (TGF-β); cells treated IL-2 only (IL-2); cells treated with both IL-2 and TGF-β (IL-2/TGF-b). Total RNA was extracted from each of the four samples using the combination of TRIzol and phenol/chloroform procedure. The experiment was repeated three times, for 3 independent biological replicates. A total of 12 RNA-seq libraries were prepared using the TruSeq Stranded mRNA sample prep protocol (Illumina) and subsequently sequenced on the Illumina HiSeq 2500 system.

Following the removal of the supernatant, the RNA pellet was air dried no longer than 5 minutes, and then resuspended in 20µl of sodium citrate dissolved in RNase-free water. DNA digestion with DNase-I was carried out with the RNeasy MinElute Cleanup Kit (Qiagen, CA).

5.3.2 DNA Library construction, paired-end (PE) RNA sequencing and data pre-processing

Library preparation and sequencing were both performed by the Australian Genome Research Facility (AGRF; Melbourne, Australia). All the 12 RNA samples were processed with the TruSeq Stranded mRNA sample prep protocol (Illumina) to make cDNA libraries. The resulting normalised and pooled libraries were clustered on the Illumina cBot *cluster amplification system* using the HiSeq PE Cluster Kit v4 reagents followed by sequencing on the Illumina HiSeq 2500 system with the HiSeq SBS Kit v4 reagents. Base calling and quality scoring were done with the standard Illumina pipeline, Real-Time Analysis (RTA) version 1.18.64 software. De-multiplexed raw FastQ files containing 100bp PE reads were generated using Illumina's bcl2fastq version 1.8.4 pipeline.

An average of 22.84 million PE 100bp reads were obtained per sample. The qualities of the raw sequence reads were assessed using FastQC version 0.11.3 (629). Based on the quality reports, adapter and quality trimming was not required.

5.3.3 Read mapping, gene expression estimation and differential expression analysis

The analysis pipeline is summarised in **Figure 5.2**. The reads from each sample were aligned to the mouse (mm10) reference genome, downloaded and indexed from UCSC Genome Browser, using Tophat2 version 2.1.1 (630). Briefly, Tophat2 uses Bowtie2 version 2.2.9 (631) first to map reads that fall entirely within an exon and later tries to map reads that span splice junctions by segmenting them. All mappings were performed with default options, except the mate pair inner distance and standard deviation, which was set to 0 and 65, respectively. The alignment for each biological replicate was

performed independently and only reads that mapped uniquely as pairs were retained for downstream analysis. Assembly and quantification of transcripts were carried out with Cufflinks2 version 2.2.1 (196) using the reference annotation file (GTF format downloaded from the UCSC browser). Gene-level abundance was expressed as fragments per kilobase of exons per million mapped (FPKM) values. Differential analysis was performed using Cuffdiff2 version 2.2.1 (206), which employs a beta negative binomial distribution model and estimates the between-group variance from the FPKM values using Student's *t*-tests. The model takes into account variability between replicates and read mapping ambiguity. Three comparisons were made: (1) TGF-b *vs.* Untreated, (2) IL-2/TGF-b *vs.* Untreated, and (3) IL-2/TGF-b *vs.* IL-2. Genes with Benjamini and Hochberg (213) adjusted *P*-values < 0.05 were considered as significantly differentially expressed. Genes DE with FPKM values higher or lower in the TGF-b-treated groups than those in the -untreated groups were defined as "up-" and "down-regulated" genes, respectively. List of DE (FDR < 0.05) genes obtained in each of these comparisons were: (1) TGF-b *vs.* Untreated (N=103), (2) IL-2/TGF-b *vs.* Untreated (N=184), and (3) IL-2/TGF-b *vs.* IL-2 (N=118).

5.3.4 Assessment of the global quality of the RNA-seq data

An important preliminary step in RNA-seq analysis is to assess how the samples in different treatment groups separate based on their global gene expression profile. Before such analysis, lowly expressed genes (FPKM ≤ 0.3) for any sample were removed. Exploratory analysis of the global FPKM expression values across all the samples was performed on 10,941 genes, log2 transformed. Principal component analysis was performed using the "prcomp" function with default settings in R. Samples were hierarchically clustered using the Wards clustering algorithm. Similarities were calculated using the maximum distance measure. The dendrogram was generated using the "dendextend" R package (601).

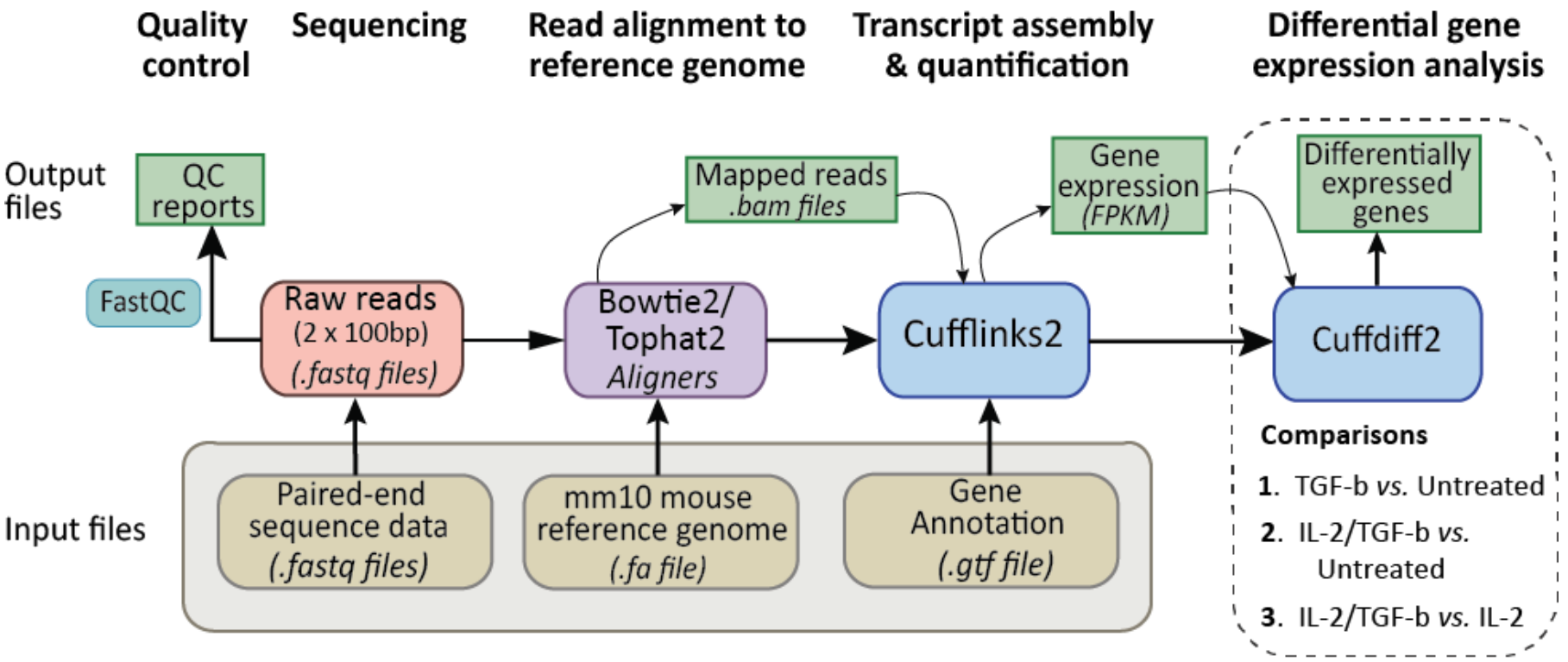


Figure 5.2: Analysis pipeline for RNA-seq data.

Quality control of raw PE reads (FASTQ format) was performed using FastQC. Bowtie2 and Tophat2 were used to align the raw reads to the mm10 version of the mouse reference genome (downloaded from the UCSC browser). Mapped reads from BAM files together with a reference gene annotation file (GTF format) were supplied to Cufflinks2 for transcript assembly and quantification. Differential analysis was performed using Cuffdiff2. Three pairwise comparisons were made: (1) TGF-b *vs.* Untreated; (2) IL-2/TGF-b *vs.* Untreated; (3) IL-2/TGF-b *vs*. IL-2.

5.3.5 Functional enrichment analysis of differentially expressed (DE) genes

I compared the overlap between the lists of DE genes across the three comparisons. The subset of common genes was functionally characterised for enriched GO (biological processes) terms among up- and down-regulated genes in the TGF-b treated groups. As described in **Chapter 2,** GO biological processes enrichment for up- and down-regulated genes was carried out using GOrilla (269), with 23,997 annotated genes in the *Mus musculus* genome (UCSC version mm10) provided as the background set. GO terms reaching significance (FDR < 0.05) were then summarised into representative terms based on semantic similarity using REVIGO (398). Summarisation analysis was performed using the RELSIM semantic similarity measure with a medium similarity cut-off ($C = 0.7$) on genes from *Mus muluscus.* Top ten summarised GO terms, ranked by enrichment *P*-values, were represented on a bar plot.

5.3.6 Comparison with the core TRM transcriptional signature

I next sought to gain insight into the role of TGF-β in influencing the transcriptional signature of TRM cells, by examining the degree of overlap between genes induced by TGF-β with those present in the previously defined TRM core signature. The TRM core signature comprises of 37 genes that were identified as commonly DE in murine TRM cells from skin, gut, lung with respect to their circulating spleen (TEM and TCM cells) counterparts (103). Since the TRM core genes were profiled on microarray-based platform (Affymetrix Mouse Gene 1.0ST arrays), only 35 core TRM signature genes, which were common to both platforms (RNA-seq *vs.* microarray), were considered. These 35 genes were compared to the DE gene list obtained in each of three TGF-b-treated *vs.* TGF-b-untreated comparisons. Only genes expressed in the same direction were considered as overlapping.

Additionally, bootstrapping was used to evaluate the statistical significance of the observed overlap. A total of 10,000 bootstraps were performed. Each time the intersection between k and m number of genes randomly selected (with replacement) was calculated, where k is number of genes in the TRM core (N=35) and m is the

number of genes DE in each of three comparisons. Enrichment *P*-value was calculated as the probability of observing an overlap as extreme as the true overlap.

5.3.7 Gene set enrichment analysis (GSEA) with TRM gene sets

GSEA was further carried out to test the enrichment of TRM associated gene sets against the ranked list of genes DE between TGF-b-treated *vs.* TGF-b-untreated groups using the GSEA version 2.2.3 software downloaded from the Broad Institute website (http://www.broadinstitute.com/gsea/index.jsp) (271). The predefined TRM-related gene sets analysed included genes previously identified as significantly up- or down-regulated (|log2FC| > 1.5) in TRM cells isolated from skin, gut, lung, brain (88,103). The list of genes DE between TCM and naive T-cells was used as negative control gene set. A total of 12 different gene sets were tested, which includes 6 up-regulated and 6 down-regulated gene sets (Table C.1 in Appendix C). I performed enrichment analysis on standardised, log-transformed FPKM values for 10,941 expressed genes (FPKM > 0.3) across the 12 samples. First, GSEA ranked all the genes differentially expressed between the TGF-b-treated *vs.* TGF-b-untreated groups by expression fold change using the 'Signal2Noise' ranking metric, which scales the mean expression within each group by their respective standard deviation. This resulted in a list of genes sorted according to their association with TGF-β treatment, with the most up-regulated genes at the top end and the most down-regulated genes at the lower end. Next, the genes in the predefined gene set were tested for their overrepresentation at the top (or bottom) end of the ranked list. The degree of enrichment was defined by an Enrichment Score (ES). The enrichment *P*-values were computed by running 1,000 permutations of gene sets and an FDR < 0.05 was used for significance threshold.

5.4 Results

5.4.1 Experimental design and analysis of the RNA-seq data

As outlined in the experimental design workflow (**Figure 5.1**), RNA samples were harvested from *in vitro* activated murine gBT-I cells that were stimulated with or without TGF-β in the presence or absence of IL-2. The experiment was repeated with exogenous IL-2 to mitigate any potential negative effects on cell survival *in vitro*. Three biological samples were obtained for each TGF-b-stimulated (TGF-b; IL-2/TGF-b) and TGF-b-unstimulated (Untreated; IL-2) groups. A total of 12 RNA-seq libraries were prepared and sequenced on the Illumina HiSeq2500 platform at depths of 19.8 – 24.6 million 100-bp PE reads per sample (**Table 5.1**). Raw reads were processed using the pipeline in **Figure 5.2**. Transcriptome assembly, gene level quantification and differential expression analysis was performed using the Tophat2/Cufflinks2/Cuffdiff2 pipeline as detailed in Methods.

Table 5.1: Summary of PE reads alignment to the mm10 reference genome

Samples	Treatment	Biological replicate	Total PE reads	Mapped reads (%)
Untreated	No treatment	1	24,596,360	21,097,751 (85.8)
		2	24,418,403	20,821,933 (85.3)
		3	24,437,275	21,222,305 (86.8)
TGF-b	Only TGF-b added	1	23,087,215	19,859,206 (86.0)
		2	20,076,757	17,088,647 (85.1)
		3	22,174,312	18,900,874 (85.2)
IL-2	Only IL-2 added	1	19,761,593	16,874,005 (85.4)
		2	23,396,138	20,324,917 (86.9)
		3	22,543,083	19,241,102 (85.4)
IL2/TGF-b	Both IL-2 and TGF-b added	1	23,135,036	19,897,388 (86.0)
		2	24,157,669	20,717,517 (85.8)
		3	22,290,252	19,125,150 (85.8)

Quality assessment of the raw reads with FastQC tool (629) reported high quality reads with average quality (Phred) score greater than 35 for all the libraries. After mapping the reads to the reference genome using Bowtie2/TopHat2, about 86% were aligned as pairs to the mm10 mouse genome across all libraries (**Table 5.1**).

5.4.2 Global expression profiles are distinct between the TGF-b-treated and TGF-b-untreated groups

The mapped reads generated by Tophat2 were assembled and quantified by Cufflinks2. Gene-level abundance estimates obtained from Cufflinks2, expressed as FPKM values, were used as expression values. Exploratory analysis was performed on 10,941 genes that achieved FPKM > 0.3 across all samples. The distribution of gene expression was observed to be similar across all the 12 samples (**Figure 5.3A**). PCA analysis was carried out to characterise the patterns of covariance in gene expression between the TGF-b-treated and TGF-b-untreated samples. The PCA results showed that the first 5 PCs explained a combined 87.2% of total variation in expression levels of 10,941 genes (**Figure 5.3B**). The TGF-b-treated samples separated from the TGF-b-untreated groups along the PC2 axis (**Figure 5.3C**). Hierarchical cluster analysis of the samples further supported these results (**Figure 5.3D**).

5.4.3 Identification of genes DE between TGF-b-treated and TGF-b-untreated groups

Cuffdiff2 was used to identify genes DE between TGF-b-treated and TGF-b-untreated groups. Differential expression analysis was performed using Cuffdiff2 (196) on three pairwise comparison groups: (1) TGF-b *vs.* Untreated; (2) IL-2/TGF-b *vs.* Untreated; and (3) IL-2/TGF-b *vs.* IL-2. The numbers of significantly DE genes obtained in each comparison are summarised in **Table 5.2**.

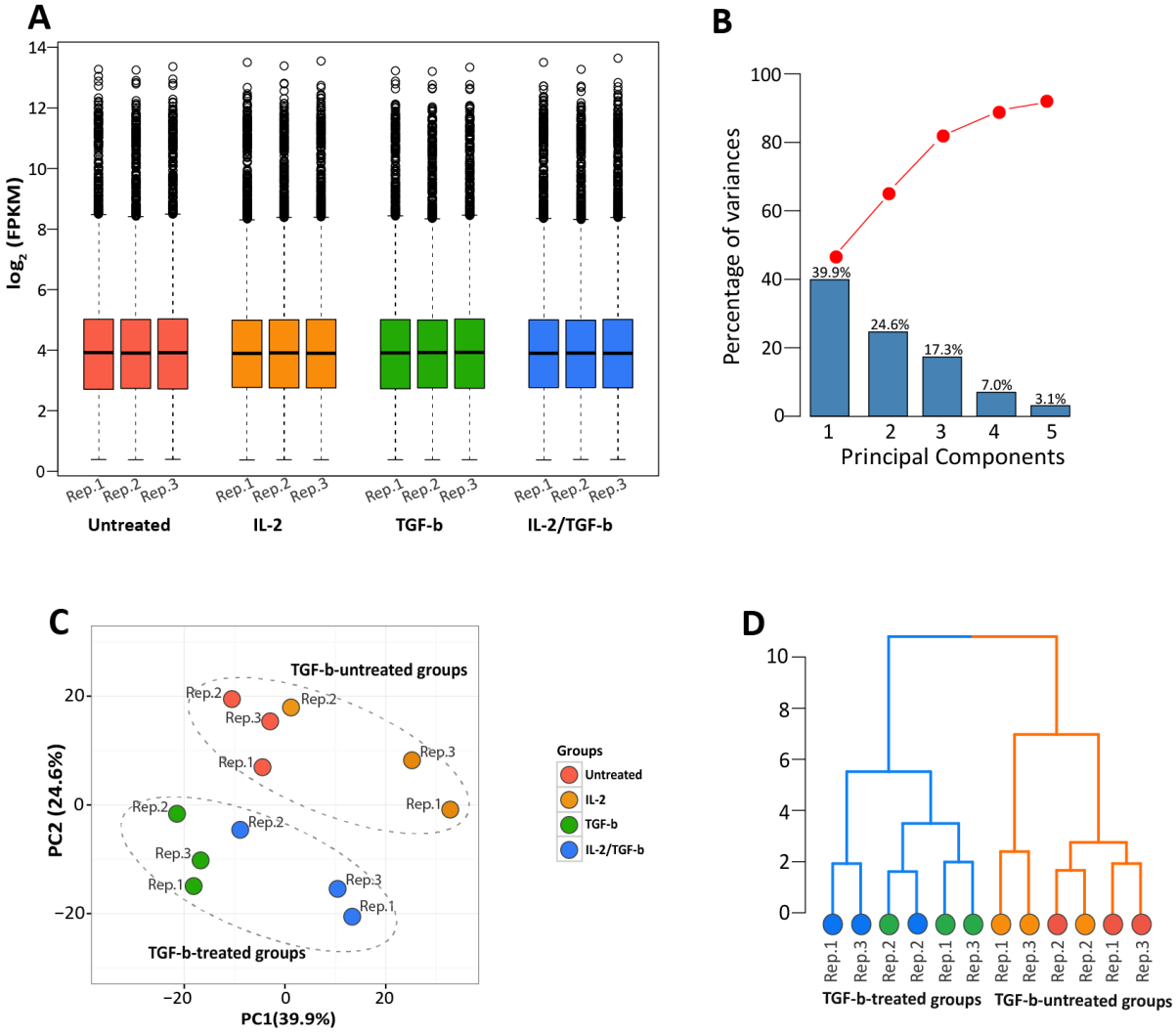


Figure 5.3: Global gene expression analysis of 10,941 expressed genes.

All analyses were performed on log2-transformed FPKM values. There were three biological replicates (Rep.1-3) for each treatment **(A)** Boxplots of log2-transformed FPKM values for each sample. **(B)** Principal component analysis (PCA) performed on the transcriptome across all the 12 samples. The scree plot shows the amount of variance (bar height) captured by each of the top 5 principal components (PCs; bars). The cumulative proportion of variance explained by the first 5 PCs (red line) is 87.2%. **(C)** The first two PCs are plotted against each other. The numbers in parenthesis beside the PC labels denote the percentage of variance explained by the respective PCs. The dots represent biological samples, which are coloured according to the treatment they received. Clusters of TGF-b-treated (TGF-b and IL-2/TGF-B) and TGF-b-untreated (Untreated and IL-2) groups, which separated along PC2, are circled. **(D)** Dendrogram from hierarchical cluster analysis of the 12 samples based on their expression profiles. Clustering was done using the Ward's method with the "maximum" distances measure provided as dissimilarity matrix. Dendrogram branches are coloured by TGF-b treatment: TGF-b-treated groups (blue) and TGF-b-untreated groups (orange). Dots at the tip of the leaves represent biological samples coloured according to TGF-b treatment. Biological replicates (Rep.1-3) for each treatment have the same colour.

Table 5.2: Differentially expressed genes identified by Cuffdiff2

Comparisons	**FDR < 0.05**	**|log2FC| > 1 & FDR < 0.05**	**|log2FC| > 2 & FDR < 0.05**
	Total (up-regulated / down-regulated genes)		
TGF-b *vs.* Untreated	849 (373 / 476)	240 (131 / 109)	57 (41 / 16)
IL2/TGF-b *vs.* Untreated	1261 (839 / 422)	448 (358 / 90)	93 (86 / 7)
IL2/TGF-b *vs.* IL-2	951 (436 / 515)	274 (162 / 112)	60 (52 / 8)

|log2FC| refers to absolute value of log2 fold change (log2FC). FDR refers to false discovery rate.

The list of DE genes was further narrowed down based on fold change: greater than 2 – (|log2FC| > 1) and 4 – fold (|log2F| > 2) **(Table 5.2)**. Across all fold change cut-offs, in all the three comparisons, it was seen that TGF-b treatment resulted in mostly up-regulated genes.

Next, the lists of DE genes obtained from the three comparisons were overlapped to identify those genes that were consistently up- or down-regulated in the TGF-b-treated groups. It was observed that approximately 25% (N= 254) of up-regulated and 18% (N= 162) of down-regulated genes overlapped between the comparisons (**Figure 5.4**).

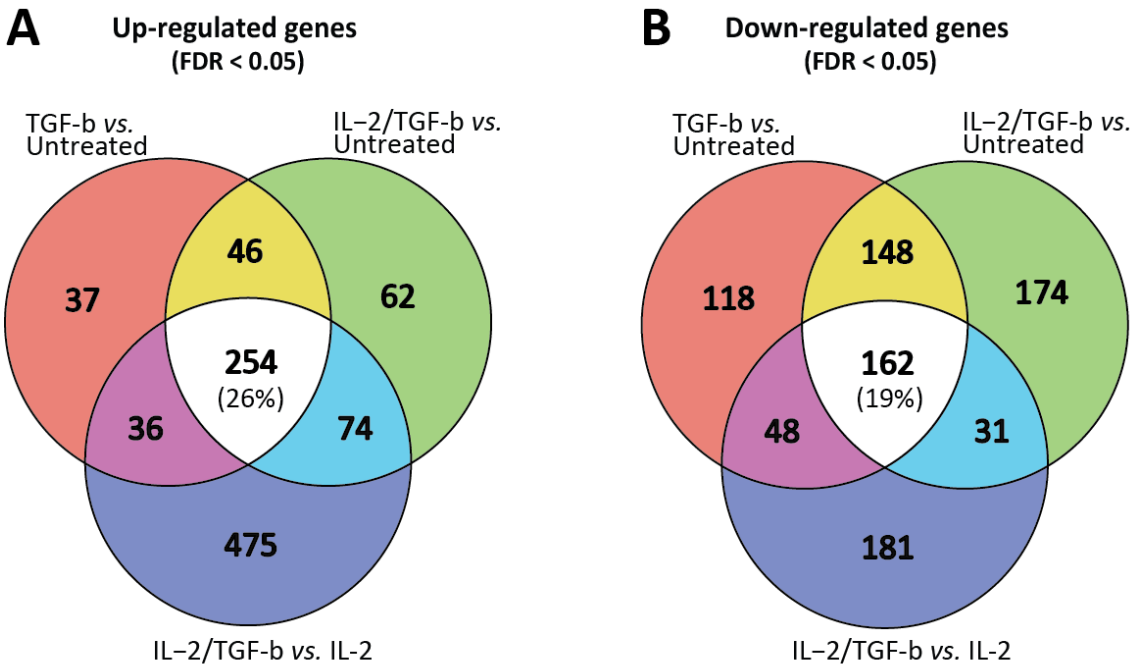


Figure 5.4: Venn diagrams of overlapping up-regulated and down-regulated genes in the TGF-b-treated groups.

(A) Up-regulated and **(B)** down-regulated genes significant at FDR < 0.05. The numbers denoted as percentages in parentheses are the percentage of differentially expressed genes overlapping across all three comparisons.

Also, further assessment of the top 30 most DE genes (based on |log2FC| > 2) in this common subset showed that majority of the genes (90%) in the TGF-b-treated groups were up-regulated (**Figure 5.5**). All the 30 genes were similarly expressed in both the TGF-b-treated groups, suggesting that cells behaved similarly in both settings (without and without IL-2). Extended list of all the common DE genes at FDR < 0.05 can be found in Table C.2 in Appendix C.

5.4.4 Functional analysis of genes DE in TGF-b-treated groups

To further functionally characterise the genes differentially regulated in response to TGF-β stimulation, GO enrichment terms associated with biological processes were assigned to 416 common DE genes among the TGF-b-treated groups. GO enrichment was performed using GOrilla (269) on three sets of DE genes: up-regulated genes, down-regulated genes, and both sets combined. The significant GO terms (FDR < 0.05) were then further summarised into representative terms using REVIGO (398). The top 10 over-represented GO terms among the DE genes in the TGF-b-treated groups, ranked by enriched *P*-values, are shown in (**Figure 5.6**). All the 254 up-regulated genes were associated with at least one GO term. The most enriched biological processes among the up-regulated genes were related to regulation including the regulation of signalling, signal transduction, cell communication and cell movement (**Figure 5.6B**). 160 out of 162 down-regulated genes had GO term annotations. These genes are largely involved in regulation of cell adhesion and response to stimulus (**Figure 5.6C**).

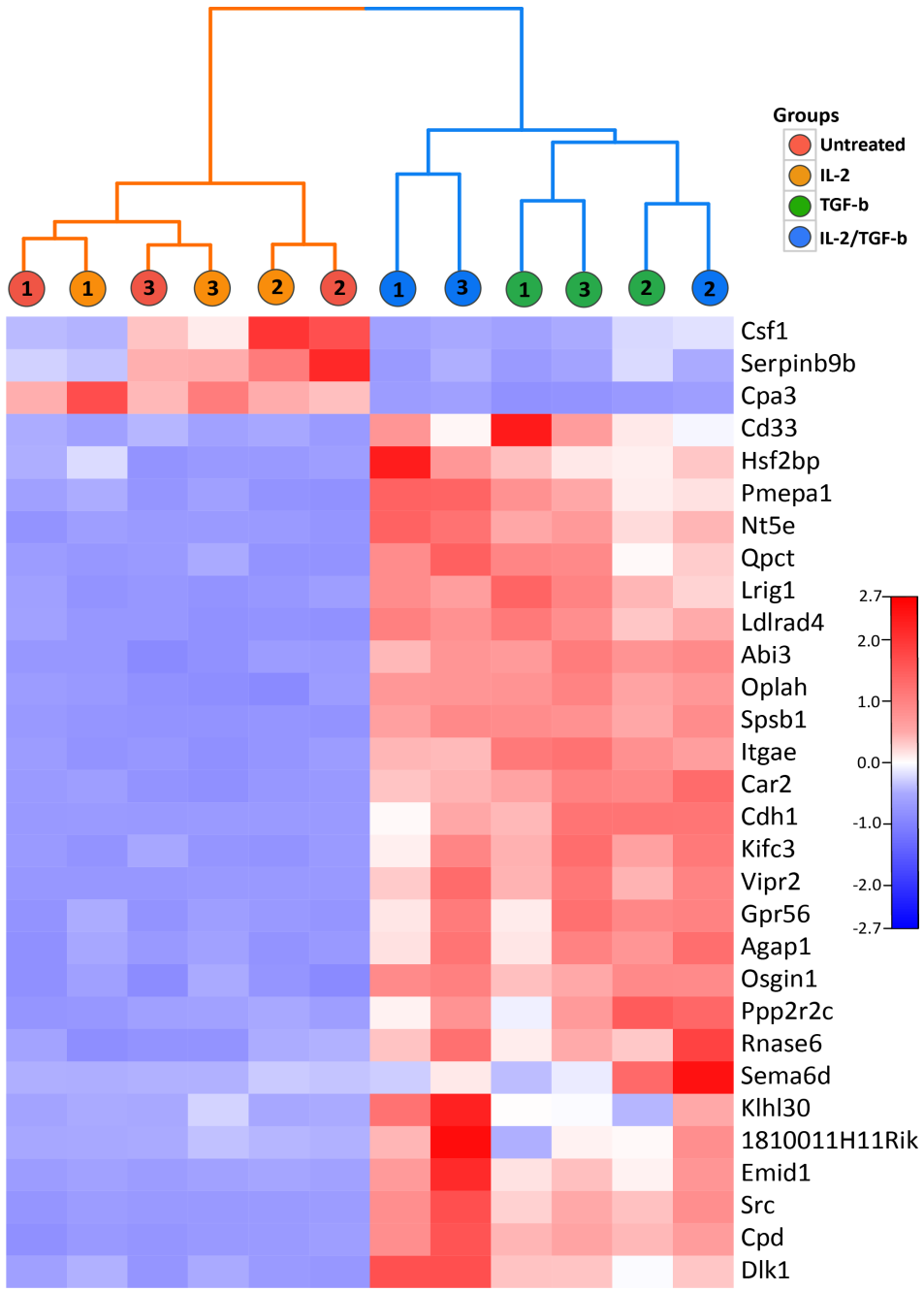


Figure 5.5: Heatmap from the hierarchical clustering of top 30 most differentially expressed genes (FDR < 0.05, |log2FC| > 2) common in all three comparisons.

The FPKM expression values for each gene across the 12 samples are presented after being log2 transformed and scaled (mean of 0 and standard deviation of 1), such that red denotes increased expression and blue denotes decreased expression. The dendrogram shows the clustering of the samples based on the expression of the 30 genes and the branches are coloured blue for TGF-b-treated groups and orange for TGF-b-untreated groups. Circles represent the samples, which are coloured according to the treatment they received, and the numbers inside denote each biological replicate.

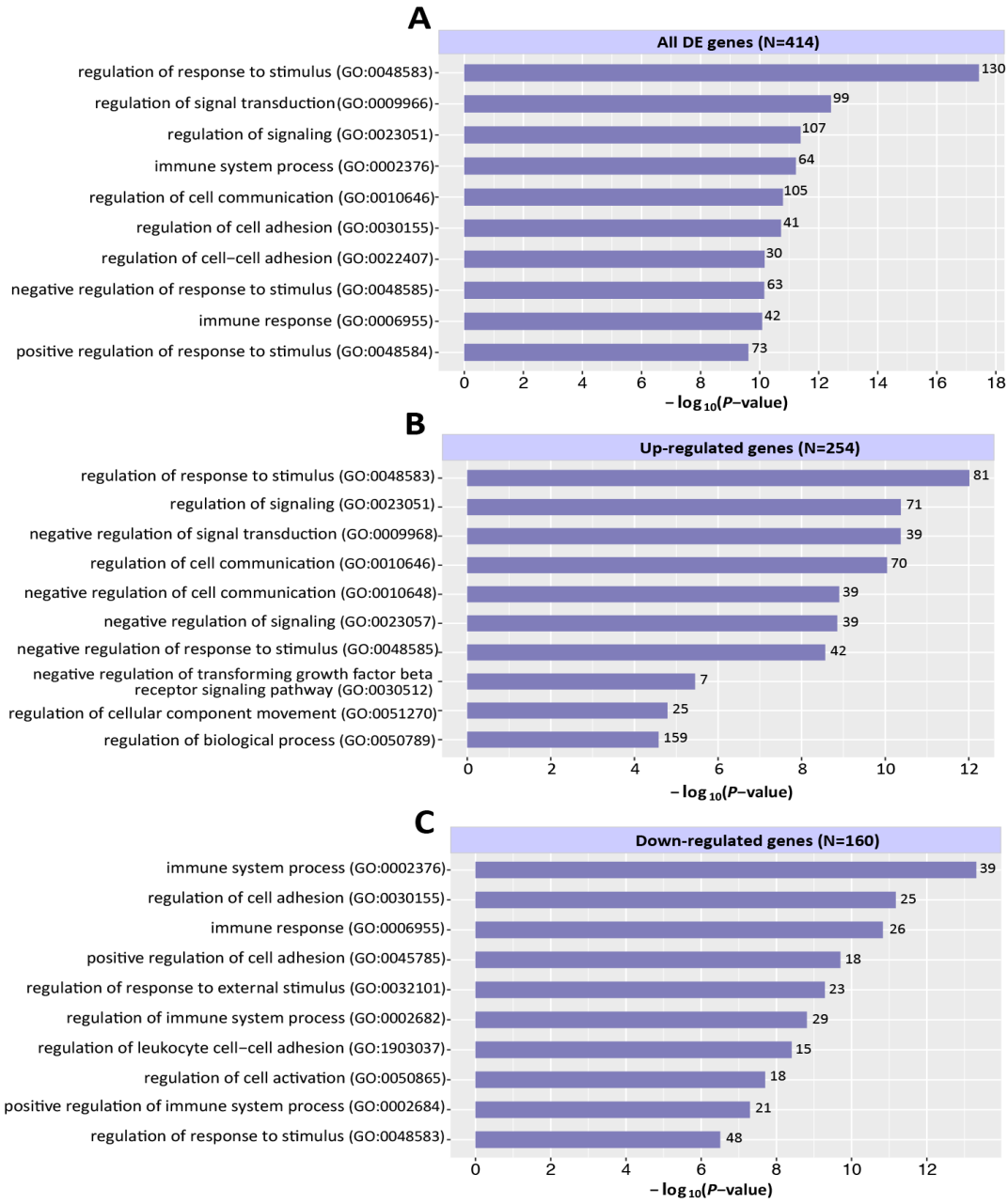


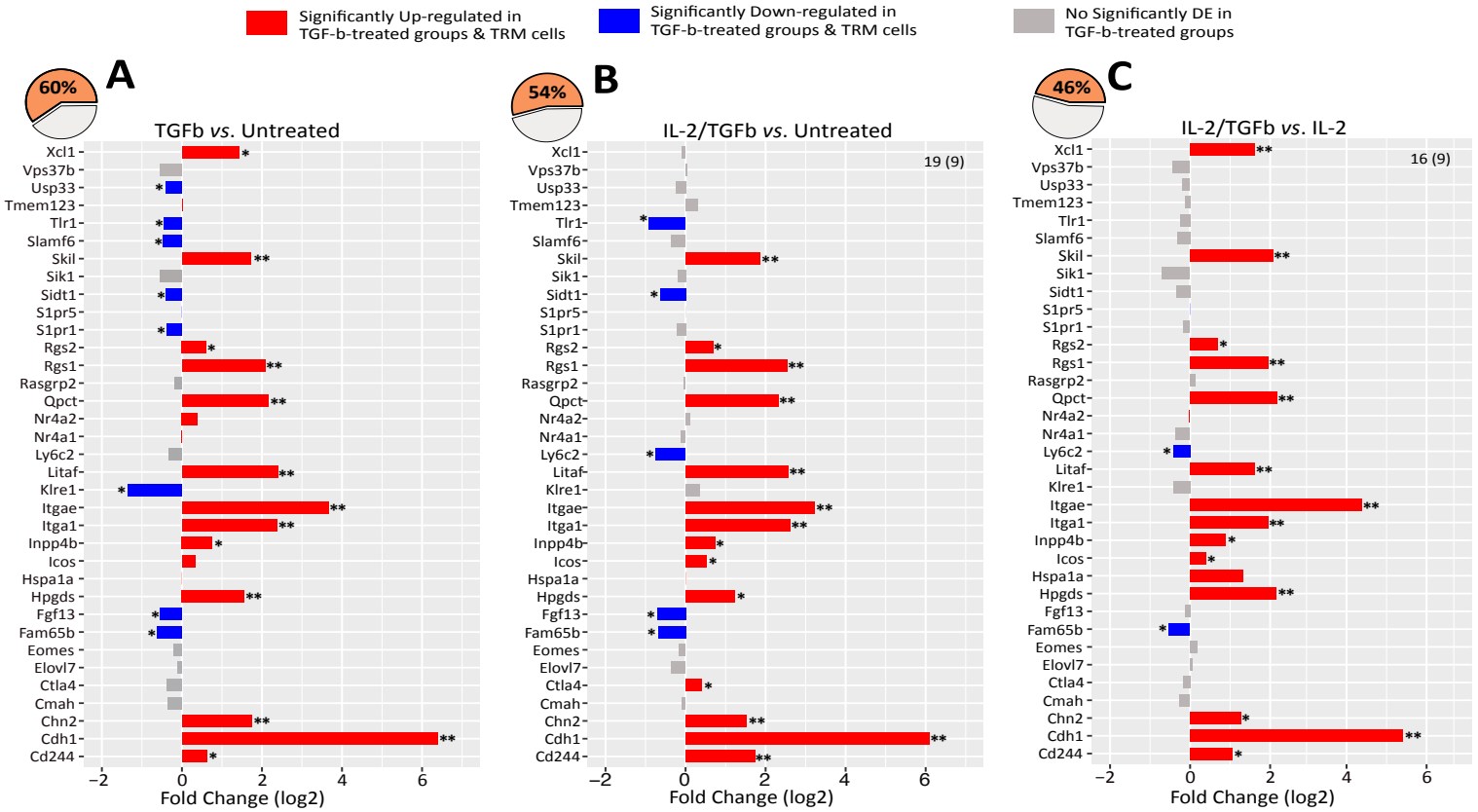
Figure 5.6: GO term enriched among genes DE genes in the TGF-b-treated groups compared to their untreated counterparts.

Top representative GO (biological processes) terms based on REVIGO output, enriched among **(A)** all the DE genes, **(B)** up-regulated genes, and **(C)** down-regulated genes in the TGF-b treated groups. The GO terms (y-axis) were ranked according to their enrichment *P*-values (x-axis). The numbers on top of the bar plots in parenthesis denotes the number of up- and down-regulated genes with GO term annotations. The numbers at the end of each bar represent the actual number of DE genes, up-regulated genes or down-regulated genes that were classified to a particular biological process. All GO terms listed were significant at FDR < 0.05. All of the 254 common up-regulated genes in the TGF-b-treated groups were associated with GO terms. 160 out of 162 common down-regulated genes in the TGF-b groups were annotated with GO terms.

5.4.5 Transcriptional profiles of TGF-b-treated T-cells are significantly enriched for TRM signature genes

I next sought to determine if TGF-b influenced the TRM-related transcriptional profile by comparing the list of DE genes in each of the TGF-b-treated groups with genes previously established as the TRM core signature (103). The TRM core signature analysed here comprised of 35 genes. Most of the genes present in the TRM core signature were consistently expressed in a similar manner in the TGF-b-treated groups (**Figure 5.7A– C**). The overlap in all the three comparisons was confirmed to be statistically significant, *P*-values of < 0.001, using bootstrapping. The majority of these overlapping genes were up-regulated in the TGF-b-treated groups, consistent with the hypothesis that TGF-β induces genes that promote the maintenance of TRM cells in tissues. The TRM core genes that were consistently up-regulated in all the TGF-b-treated groups included *Cdh244, Chd1, Chn2, Hpgds, Inpp4b, Itga1, Itgae*, *Qpct, Rgs1, Rgs2* and *Skil*. While *Fam65b* was the only TRM core gene that was similarly down-regulated across all the TGF-b-treated groups.

GSEA was performed with pre-defined TRM-associated gene sets to further explore the role of TGF-β signalling in regulating residency-related transcriptional profile of TRM cells. Genes previously identified as DE (FDR < 0.05 and |log2FC| > 1.5) in murine TRM cells from skin, gut, lung, and brain in comparison to their circulating counterparts (103,533) were divided into up-regulated and down-regulated gene sets. GSEA confirmed that all the 4 up-regulated and 4 down-regulated gene sets were significantly (FDR < 0.05) enriched (**Figures 5.8 – 5.9**). Genes that were up-regulated in TRM cells had higher expression in the TGF-b-treated group (**Figure 5.8**), whereas down-regulated genes had higher expression in the TGF-b-untreated group (**Figure 5.9**) This suggests that TGF-β plays a role in inducing and repressing the expression of genes up-regulated and down-regulated in TRM cells, respectively. As expected, the TGF-b-treated group showed no enrichment for TEM-related gene sets, which served as a negative control. Since TEM cells are circulating and do not share a common precursor with TRM cells, TGF-β is not expected to influence the transcriptional profile of these cells. Hence, these results strongly support that the transcriptional signature in TRM cells is largely driven by TGF-β signalling.



Bootstrapping enrichment *P*-value calculations

TGF-b vs. Untreated

TRM vs. Cirulating	DE	Not DE	Total
DE	21	14	35
Not DE	797	18,729	19,526
Total	818	18,743	19,561

P-value = 1.16e-20

IL-2/TGF-b vs. Untreated

	DE	Not DE	Total
DE	19	16	35
Not DE	1186	18,340	19,526
Total	1205	18,356	19,561

P-value = 1.38e-14

IL-2/TGF-b vs. IL-2

	DE	Not DE	Total
DE	16	19	35
Not DE	895	18,631	19,526
Total	911	18,650	19,561

P-value = 7.60e-13

Figure 5.7: TGF-b induced transcriptional profiles are enriched for the TRM core signature genes identified in murine TRM cells.

The bar plot shows the log2 fold change in expression (x-axis) of the 35 TRM core signature genes in each of the pairwise comparison of the TGF-b-treated groups *vs.* TGF-b-untreated groups: (A), (B), and (C). The bars are coloured based on their expression in both TRM cells and TGF-b-treated groups. Red represents genes significantly up-regulated, blue represents genes significantly down-regulated, and grey represents genes that were significantly DE in TGF-b group. The asterisks denote genes that are significantly differently expressed in the TGF-b treated groups: *FDR < 0.05, ** FDR < 0.05 and |log2FC| > 1.5. The numbers in the pie chart denote the percentage overlap between the genes in the TRM core and genes DE (FDR < 0.05) in the TGF-b-treated groups. The contingency table below each plot shows the observed overlap between the genes in the TRM core and genes DE (FDR < 0.05) in the TGF-b-treated groups, and genes with no change in expression out of the 19,561 genes common across the RNA-seq and microarray technologies. The enrichment *P*-values were calculated using bootstrapping.

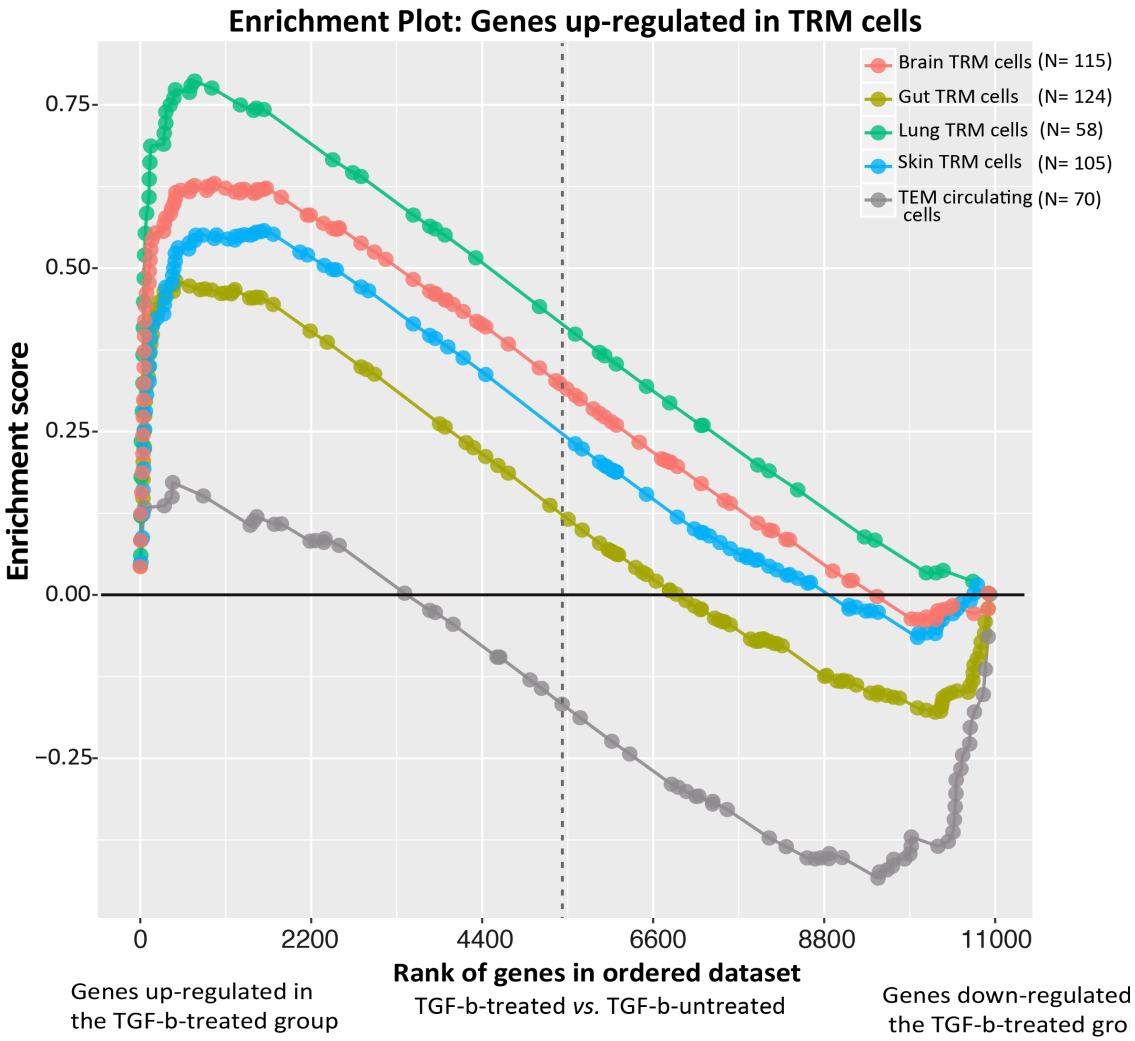


Figure 5.8: Enrichment plot for the 4 TRM-related up-regulated gene sets in the TGF-b-treated group.

The enrichment scores (ES; y-axis) of up-regulated gene sets from skin, gut, lung, or brain TRM cells (88,103) in the ranked list of genes DE between TGF-b-treated-group and TGF-b-untreated group (x-axis). The genes in the rank list are ordered along the x-axis based on fold change, where the most up-regulated genes in the TGF-b-treated group are on the far left and the most down-regulated genes – far right. The dotted vertical grey line represents fold change of zero. The curved lines, coloured by tissue type, show the cumulative enrichment score. The dots denote the positions in the ordered ranked list where the genes in each gene set appear. The TEM gene set served as a negative control for no enrichment.

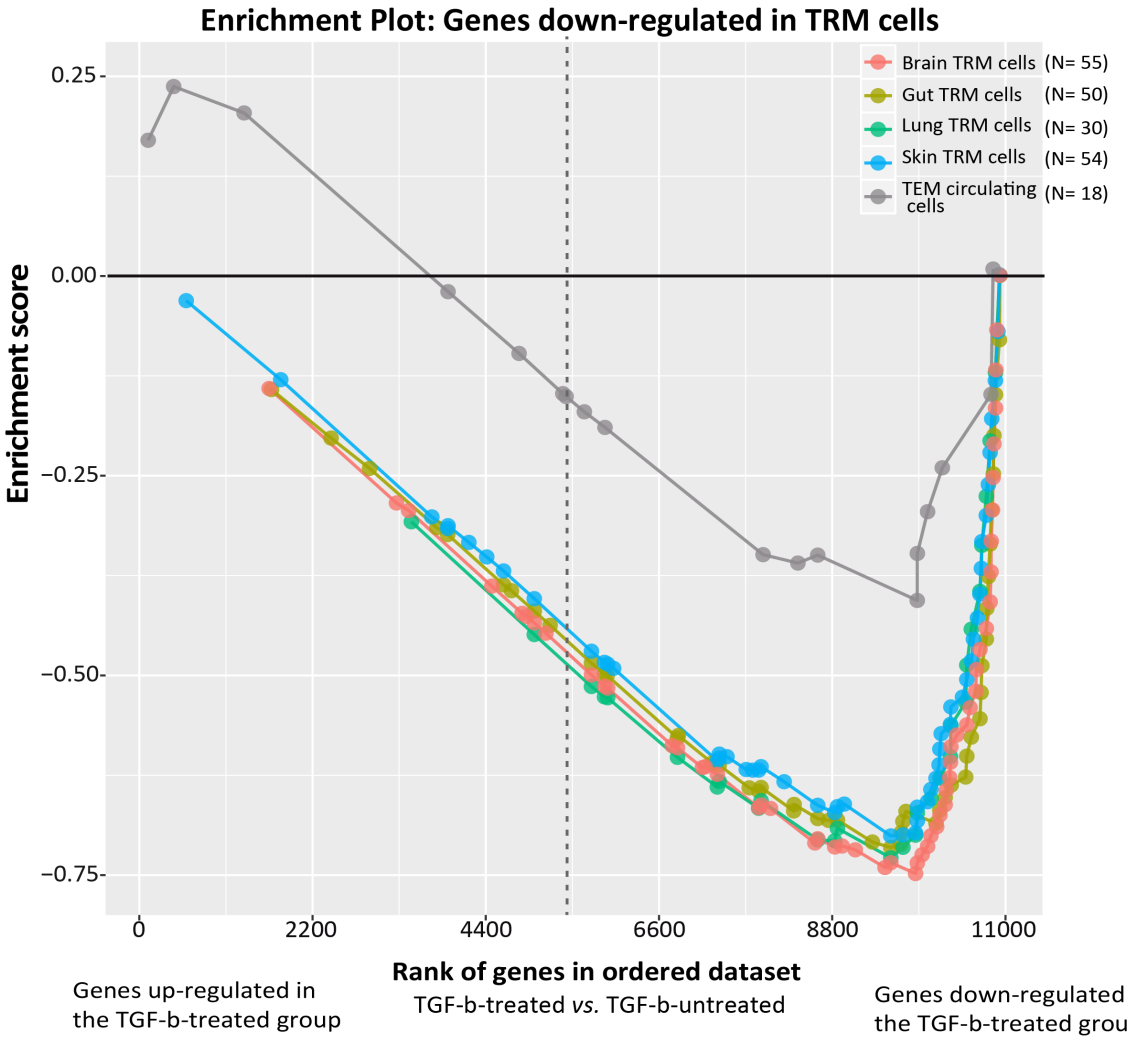


Figure 5.9: Enrichment plot for the 4 TRM-related down-regulated gene sets in the TGF-b-treated group.

The enrichment scores (ES; y-axis) of down-regulated gene sets from skin, gut, lung, or brain TRM cells (88,103) in the ranked list of genes DE between TGF-b-treated-group and TGF-b-untreated group (x-axis). The genes in the rank list are ordered along the x-axis based on fold change, where the most up-regulated genes in the TGF-b-treated group are on the far left and the most-down-regulated genes – far right. The dotted vertical grey line represents fold change of zero. The curved lines, coloured by tissue type, show the cumulative enrichment score. The dots denote the positions in the ordered ranked list where the genes in each gene set appear. The TEM gene set served as a negative control for no enrichment.

5.5 Discussion

Compelling experimental evidence have demonstrated that TGF-β is one of the necessary niche signals required for the differentiation of TRM cells in the skin, gut and lungs (98,103,556,569,570). TRM cells with defective TGF-β receptors are unable to respond to TGF-β signals, and as a consequence do not up-regulate CD103 expression and are incapable of maintaining residency at tissue sites (98,103,556,569,570). However, apart from the role of TGF-β in regulating expression of homing and adhesion receptors in CD8+ T cells, how exactly this cytokine impacts on the establishment and long-term maintenance of TRM cells within tissues remains largely unexplored. Recently, it was revealed that TRM cells exhibit a distinct transcriptional program that distinguishes them from their TEM and TCM circulating counterparts (98,103,556,569,570). Based on the fact that TGF-β is essential for tissue residency in TRM cells, it was hypothesised that the gene signature of TGF-β induced $CD8^+$ T-cells established under *in vitro* conditions may comprise of genes that are part of the transcriptional program previously defined in TRM cells.

To generate a TGF-β specific gene signature, in this study RNA-seq was used to profile the expression of murine $CD8^+$ T-cells stimulated *in vitro* with exogenous TGF-β. The transcriptional profiles from the TGF-b-treated groups were compared with the TGF-b-untreated groups, and it was seen that the two groups differed based on their gene expression signature, revealing that TGF-β stimulation had a widespread effect on gene expression. Differential gene expression analysis further led to the identification of 254 up-regulated and 162 down-regulated genes common across all the three pairwise comparisons made between the TGF-b-treated and TGF-b-untreated groups (TGF-b *vs.* Untreated; IL-2/TGF-b *vs.* Untreated; and IL-2/TGF-b *vs.* IL-2). Functional analysis of these DE genes suggests that changes in gene expression most likely affect a wide range of regulatory processes such as regulation of signalling cell communication, cell movement, cell adhesion, and response to a stimulus. It is likely that these processes might be mediated through TGF-β signalling.

Direct comparison of genes differentially expressed between TGF-b-treated and TGF-b-untreated groups with those previously identified to be part of the core TRM gene signature (103) revealed a significant overlap of nearly 50% of genes, which also exhibited consistent expression pattern. Several core TRM genes that play an essential role in tissue residency were common in TGF-β induced gene signature. *Itgae*, which encodes for CD103, was one of the most significantly up-regulated genes in the TGF-b-treated group, which is consistent with previous findings that TGF-β induces CD103 expression on TRM cells (556,569,626). CD103 binds to its ligand E-cadherin expressed on the epithelial surfaces of the skin and gut; possibly mediates the retention of TRM cells by tethering them within these tissues (567,620). Its has been shown that CD103-deficient mice had lower T-cell numbers in skin, intestinal, and vaginal epithelium (103,567,632) in the memory phase, further suggesting that TGF-β induced expression of CD103 is important for the homing of TRM cells. Additionally, *Chd1* and *Itga1*, which also encode for adhesion molecules, were up-regulated in the TGF-b-treated group. *Cdh1 and Itga1* genes encode for E-cadherin and alpha 1 subunit of integrin receptors, respectively, and have been previously reported to be up-regulated in TRM cells (88,103,582). In Langerhans cells (a subtype of dendritic cells), TGF-β dependent induction of E-cadherin is crucial for their residency and maintenance in the skin (633), implicating a similar requisite in TRM cells. Increased expression of the chemokine *Xcl1* was also noted in the TGF-b-treated group. Similarly, several studies have reported high expression levels of *Xcl1* in TRM cells (88,103,582). This finding and those of others have led to the speculation that TRM cell-derived XCL1 facilitates the recruitment of $CD103^+$ dendritic cells (DCs), which express both the receptor (XCR1) for XCL1 and produce TGF-β. As a consequence, the $CD103^+$ DC-derived TGF-β promotes the persistence of TRM cells in the skin and gut (634,635). Other genes previously found to be highly expressed in TRM cells were also consistently up-regulated in the TGF-b-treated group analysed in this chapter. This includes genes encoding for costimulatory receptors involved in immunomodulation (*Ctla4*, and *Icos),* enzymes *(Inapp4b* and *Qpct)*, and signalling regulators that mediate tissue retention (*Rgs1 and Rgs2*) (88,103,582). Hence, the significant overlap seen between genes involved in TRM cell retention and those in TGF-induced signature suggests that TGF-β dependent tissue homing might be an

important process for the establishment of residency by TRM cells at tissue sites. Moreover, gene set enrichment analysis showed that genes associated with TRM cells from various tissues were significantly enriched in the TGF-b-treated group, which further demonstrated an overlap in the transcriptional profile of TRM cells and TGF-β induced gene signature.

A possible limitation of this study is using *in vitro* stimulated $CD8^+$ T-cells as a surrogate for the differentiation of TRM cells. The development and tissue specific activation of TRM cells will require precise temporal and spatial regulation of gene expression, which is achieved by epigenetic mechanisms such as histone modifications and DNA methylation. The epigenetic state of TRM cells is most likely to be influenced by signals derived from the local tissue microenvironment. Hence, one cannot exclude the possibility that epigenetic changes and/or tissue-specific local cues may have an impact on the TGF-β induced signature of TRM cells at tissue sites *in vivo*. In light of caveats of this experiment, the findings do not negate the important role TGF-β plays in imprinting tissue homing transcriptional profile on TRM cells. The *in vitro* induced cells appear very similar to TRM cells, since more than half of the genes were regulated in a way that is reminiscent of the regulation of genes in TRM versus circulating cells. However, further *in vivo* studies are required to establish if constant "education" by TGF-β is required for TRM cells to acquire long-term maintenance capacity. An approach to address this question is to carry out time-point conditional knockout of the TGF-β receptor or antibody blockade of TGF-β to assess to what extent TGF-β signalling is required for the establishment of tissue residency.

Chapter 6

Conclusions

The immune system has evolved to provide effective host defence against a diverse range of internal and external threats, and its aberrant regulation has been associated with a number of diseases (636). Understanding the mechanisms underlying immune function and its regulation may offer strategies to harness and manipulate the immune system to treat diseases and improve health. The highly complex and multi-level nature of the immune system means that systems-wide analysis is necessary to achieve mechanistic insights. High-throughput "-omic" profiling technologies have yielded large-scale data characterising the immune system at multiple organisational levels. Integrative methods applied to such large-scale data allows one to comprehensively evaluate the immune system and its relationship with other biological systems at a level of detail previously restricted to a single organisational layer. This thesis interrogated various aspects of immune processes in terms of genetics, transcriptional networks, cytokine signalling, and interactions with metabolism using multi-omic analysis. The findings demonstrate the power of bioinformatics approaches in providing fine resolution insights into immune function and its regulation, that would not have been possible with traditional methods.

In the first part (**Chapter 2**), to gain insight into these complex interactions, I integrated blood transcriptomic, metabolomic, and genomic profiles from two Finnish population-based cohorts, including a subset of individuals profiled 7-years apart. Through gene co-expression network analysis, I identified networks enriched for diverse immune functions, which topologically replicated between cohorts. I then performed association analysis of these immune-related modules with circulating metabolites and CRP, showing that each of these modules were significantly

associated with at least one metabolite including lipoprotein subclasses, lipids, fatty acids, amino acids, and CRP. Further, genome-wide scans revealed mQTL with both *cis* and *trans* effects. Finally, I assessed the long-term stability of these interactions, finding that the metabolite associations for a mast cell and basophil-related module and the *trans*-QTL effects of rs1354034 were largely maintained over a seven-year period. Taken together, this study provides a detailed map of natural variation at the immuno-metabolic interface in human blood, which may be used to explain differential disease susceptibility between individuals. Several genetic drivers of immune-related gene networks and genes were identified in this chapter, which strongly supports the notion that individuals differ considerably in the magnitude of their immune responses. Consequently, this may affect the cross talk between immune and metabolic systems. This implies that perturbations in the immune-metabolic interplay might further modulate the magnitude of an immune response or contribute to an altered metabolic state, and hence influence susceptibility to cardiometabolic diseases. The catalogue of metabolite interactions identified here strongly support an intimate relationship between the immune response and systemic metabolism, which is consistent with the view that this interplay contributes to many complex diseases of metabolic, cardiovascular, autoimmune or infectious aetiology. Moreover, these interactions can be explored experimentally to gain insight into immuno-metabolic disease mechanisms, as well as stratify patients into subgroups who are enriched for specific mQTLs, gene co-expression network levels, or metabolites. Future studies can expand on the immunometabolic map by exploring interactions between the immune processes, the microbiota and metabolites (1).

In the second part (**Chapter 3**), I performed multivariate GWAS on a network of 11 correlated cytokines using data from over 9,000 individuals. The findings are consistent with and add upon previous knowledge of genetic variation regulating circulating cytokine levels. This study also demonstrated the power gains of a multivariate approach, which led to the identification of two novel loci. These two loci also harboured whole-blood eQTLs and have previously been shown to exhibit tissue-specific regulation of gene expression across various tissues. Recently, studies have begun to characterise genetic variants influencing cytokine production in human immune cells in response to *ex-vivo* stimulation with bacterial, fungi or viruses (498,503). These studies have identified distinct patterns of correlated cytokines

released in an infection-dependent manner (498,503). However, these studies have associated individual cytokines with SNPs while ignoring the correlation structure among the cytokines. The multivariate-versus-univariate comparison provided in this chapter and other publications (478–481) should motivate future researchers to consider multivariate analysis of cytokine data for association studies so that we can better understand the genetic basis of inter-individual differences in immune function and response.

In the third part (**Chapter 4**), to identify residency-related sub-networks, network analysis was performed on genes that were differentially expressed between resident and circulating murine memory T-cells. This led to the identification of a RESIDENT module, most likely involved in tissue homing, which was highly coexpressed in the resident cells. Key driver analysis further revealed TNF as a potential regulator of the RESIDENT module. Furthermore, comparative transcriptome analysis revealed that the residency-related transcriptional signature of brain TRM cells shared similarities with that of resident adipose Tregs. As shown in this chapter, the application of network analysis to expression profiles from immune cells subtypes is useful to furthering our understanding of lymphocyte biology, as it may lead to the identification of sub-networks underpinning lymphocyte development and function. Leveraging trancriptomic data from large-scale consortia such as ImmGen and public repositories will increase the sample size and immune cell subtypes for performing network analysis. This chapter also shows that key driver analysis can identify potential regulators of coexpressed sub-networks as candidate gene targets that can be further investigated through experimental studies.

Finally (**Chapter 5**), to gain insight into the role of tissue-derived TGF-β in driving the transcriptional program of TRM cells, I performed RNA-seq-based transcriptome analysis of T cells stimulated *in vitro* with TGF-β. Here, I showed that TRM cells are enriched for a TGF-β-driven transcriptional signature. The local instructions provided at tissue sites shape the TRM gene profile, which may ultimately affect their survival, function, and interaction with other immune cell subtypes within tissues (103,105,637). The findings here further support the hypothesis that the cellular environment plays an important role in shaping an immune response (637,638).

Hence, local environmental cues should be taken into consideration when designing vaccine or therapeutic strategies to establish TRM for local immunity.

Integrative and comparative bioinformatics can be applied to characterise the immune system and its role in health and disease. Recently, a 10-year megaproject called the Human Cell Atlas (HCA) has been launched to map out 35 trillion human cells, which also include immune cells from the hematopoietic lineage. The HCA project will employ single-cell genomics and high-throughput measurements of other "omes" to generate large-scale omic datasets. This will allow us to create comprehensive interaction maps of immune cells, which can then be used to infer how immune cell interactions differ between and within individuals, and change over time, during human development and disease.

Our understanding of immune cell differentiation and function has mainly come from studies in mice. Despite similarities in the expression pattern of orthologous genes between human and mice, notable transcription differences have been identified across these two species (639,640). This may pose a challenge when translating research results obtained from mice studies to human, mainly with regards to understanding disease mechanisms and designing therapeutics (641). The availability of genome-wide gene expression profiles across a multi-species compendium of immune cells from the ImmGen and HCA project means that it is now be possible to do comparative analyses of immune cell types using similar methods applied in **Chapters 4** and **5**. Creating a reference chart of species-, lineage- and immune cell-specific transcriptional signatures may guide in translating mice findings to human settings. In addition, as demonstrated in **Chapter 2**, network analysis and replication methods can also be applied to ImmGen and HCA datasets to assess conservation of cross-tissue or cross-species co-expression patterns. Few studies have investigated the conservation of tissue-specific gene modules across and within species (251,642), but doing so may help prioritise tissues and pathways that are pertinent to humans when using mouse models to understand the human immune system. Moreover, **Chapter 4** revealed a tissue residency-associated gene network and showed that its transcriptional signature might be shared across resident immune cells of different lineages in mice. Although such analysis has so far been limited to mouse data, the analyses in **Chapter 4** can be applied to human data from the HCA project.

Integrative bioinformatics may create opportunities in personalised medicine to better treat and predict disease risk. Through this thesis, I have demonstrated examples where integrative and comparative bioinformatics methods can be applied to multidimensional omics data to gain novel insight into immune function and its regulation. Such approaches can not only be applied to large-scale data obtained from population studies, but also to small datasets generated from experimental studies with mouse models and *in vitro* setups. Overall this thesis offers a general framework for future studies to integrate and make the most out of multi-level omics data.

List of References

1. Lippolis JD. Immunological signaling networks: integrating the body's immune response. J Anim Sci. 2008;86(14 Suppl):E53-63.
2. Shaw AC, Joshi S, Greenwood H, Panda A, Lord JM. Aging of the innate immune system. Curr Opin Immunol. 2010;22(4):507–13.
3. Chaplin DD. Overview of the immune response. J Allergy Clin Immunol. 2010;125(2 Suppl 2):S3-23.
4. Weih F, Caamaño J. Regulation of secondary lymphoid organ development by the nuclear factor-kappaB signal transduction pathway. Immunol Rev. 2003;195:91–105.
5. Sugiyama T, Nagasawa T. Bone marrow niches for hematopoietic stem cells and immune cells. Inflamm Allergy Drug Targets. 2012;11(3):201–6.
6. Ribatti D, Crivellato E. Miller's seminal studies on the role of thymus in immunity. Clin Exp Immunol. 2006;144(3):371–5.
7. Tiron A, Vasilescu C. Role of the spleen in immunity. Immunologic consequences of splenectomy. Chir. 2008;103(3):255–63.
8. von der Weid PY, Rainey KJ. Review article: lymphatic system and associated adipose tissue in the development of inflammatory bowel disease. Aliment Pharmacol Ther. 2010;32(6):697–711.
9. Larsson J, Karlsson S. The role of Smad signaling in hematopoiesis. Oncogene. 2005;24(37):5676–92.
10. Lai AY, Kondo M. T and B lymphocyte differentiation from hematopoietic stem cell. Semin Immunol. 2008;20(4):207–12.
11. Kondo M. Lymphoid and myeloid lineage commitment in multipotent hematopoietic progenitors. Immunol Rev. 2010;238(1):37–46.
12. Kondo M, Wagers AJ, Manz MG, Prohaska SS, Scherer DC, Beilhack GF, et al. Biology of hematopoietic stem cells and progenitors: implications for

clinical application. Annu Rev Immunol. 2003;21:759–806.

13. Manz MG, Traver D, Miyamoto T, Weissman IL, Akashi K. Dendritic cell potentials of early lymphoid and myeloid progenitors. Blood. 2001;97(11):3333–41.
14. Nakajima H. Role of transcription factors in differentiation and reprogramming of hematopoietic cells. Keio J Med. 2011;60(2):47–55.
15. Kobayashi-Osaki M, Ohneda O, Suzuki N, Minegishi N, Yokomizo T, Takahashi S, et al. GATA motifs regulate early hematopoietic lineage-specific expression of the Gata2 gene. Mol Cell Biol. 2005;25(16):7005–20.
16. Barreda DR, Belosevic M. Transcriptional regulation of hemopoiesis. Dev Comp Immunol. 2001;25(8–9):763–89.
17. Argiropoulos B, Humphries RK. Hox genes in hematopoiesis and leukemogenesis. Oncogene. 2007;26(47):6766–76.
18. Dorritie KA, McCubrey JA, Johnson DE. STAT transcription factors in hematopoiesis and leukemogenesis: opportunities for therapeutic intervention. Leukemia. 2013;28(2):248–57.
19. Holmes M, Turner J, Fox A, Chisholm O, Crossley M, Chong B. hFOG-2, a novel zinc finger protein, binds the co-repressor mCtBP2 and modulates GATA-mediated activation. J Biol Chem. 1999;274(33):23491–8.
20. Kaushansky K. Lineage-specific hematopoietic growth factors. N Engl J Med. 2006;354(19):2034–45.
21. Akashi K, Kondo M, Weissman IL. Role of interleukin-7 in T-cell development from hematopoietic stem cells. Immunol Rev. 1998;165:13–28.
22. Wong FS, Wen L. Innate and adaptive immune responses are highly interconnected at many levels. Curr Mol Med. 2009;9(1):1–3.
23. Amit I, Garber M, Chevrier N, Leite AP, Donner Y, Eisenhaure T, et al. Unbiased reconstruction of a mammalian transcriptional network mediating pathogen responses. Science. 2009;326(5950):257–63.
24. Warrington R, Watson W, Kim HL, Antonetti FR. An introduction to immunology and immunopathology. Allergy Asthma Clin Immunol. 2011;7(Suppl 1):S1.
25. Fujii S, Liu K, Smith C, Bonito AJ, Steinman RM. The linkage of innate to adaptive immunity via maturing dendritic cells in vivo requires CD40 ligation in addition to antigen presentation and CD80/86 costimulation. J Exp Med.

2004;199(12):1607–18.

26. Mogensen TH. Pathogen recognition and inflammatory signaling in innate immune defenses. Clin Microbiol Rev. 2009;22(2):240–73.
27. Thompson MR, Kaminski JJ, Kurt-Jones EA, Fitzgerald KA. Pattern recognition receptors and the innate immune response to viral infection. Viruses. 2011;3(6):920–40.
28. Kawai T, Akira S. The roles of TLRs, RLRs and NLRs in pathogen recognition. Int Immunol. 2009;21(4):317–37.
29. Arango DG, Descoteaux A. Macrophage cytokines: involvement in immunity and infectious diseases. Front Immunol. 2014;5:491.
30. Silva MT, Correia-Neves M. Neutrophils and macrophages: the main partners of phagocyte cell systems. Front Immunol. 2012;3:174.
31. Colucci-Guyon E, Tinevez JY, Renshaw SA, Herbomel P. Strategies of professional phagocytes in vivo: unlike macrophages, neutrophils engulf only surface-associated microbes. J Cell Sci. 2011;124(Pt 18):3053–9.
32. Cheroutre H, Huang Y. Crosstalk between adaptive and innate immune cells leads to high quality immune protection at the mucosal borders. Adv Exp Med Biol. 2013;785:43–7.
33. Clark GJ, Angel N, Kato M, López JA, MacDonald K, Vuckovic S, et al. The role of dendritic cells in the innate immune system. Microbes Infect. 2000;2(3):257–72.
34. Getz GS. Bridging the innate and adaptive immune systems. J Lipid Res. 2005;46(4):619–22.
35. Moignard V, Macaulay IC, Swiers G, Buettner F, Schütte J, Calero-Nieto FJ, et al. Characterization of transcriptional networks in blood stem and progenitor cells using high-throughput single-cell gene expression analysis. Nat Cell Biol. 2013;15(4):363–72.
36. Zhu J, Paul WE. CD4 T cells: fates, functions, and faults. Blood. 2008;112(5):1557–69.
37. Luckheeram R V, Zhou R, Verma AD, Xia B. $CD4^{+}$T cells: differentiation and functions. Clin Dev Immunol. 2012;2012:925135.
38. Yang CY, Best JA, Knell J, Yang E, Sheridan AD, Jesionek AK, et al. The transcriptional regulators Id2 and Id3 control the formation of distinct memory CD8+ T cell subsets. Nat Immunol. 2011;12(12):1221–9.

39. Nothelfer K, Sansonetti PJ, Phalipon A. Pathogen manipulation of B cells: the best defence is a good offence. Nat Rev Microbiol. 2015;13(13):173–84.
40. Topham NJ, Hewitt EW. Natural killer cell cytotoxicity: how do they pull the trigger? Immunology. 2009;128(1):7–15.
41. Kumar V, Sharma A. Neutrophils: Cinderella of innate immune system. Int Immunopharmacol. 2010;10(11):1325–34.
42. Jacobsen EA, Helmers RA, Lee JJ, Lee NA. The expanding role(s) of eosinophils in health and disease. Blood. 2012;120(19):3882–90.
43. Siracusa MC, Kim, B S Jonathan M. Spergel JM, Artis D. Basophils and allergic inflammation. J Allergy Clin Immunol. 2013;132(4):789–788.
44. Gautier EL, Shay T, Miller J, Greter M, Jakubzick C, Ivanov S, et al. Gene-expression profiles and transcriptional regulatory pathways that underlie the identity and diversity of mouse tissue macrophages. Nat Immunol. 2012;13(11):1118–28.
45. Qu C, Brinck-Jensen NS, Zang M, Chen K. Monocyte-derived dendritic cells: targets as potent antigen-presenting cells for the design of vaccines against infectious diseases. Int J Infect Dis. 2014;19:1–5.
46. Miller JC, Brown BD, Shay T, Gautier EL, Jojic V, Cohain A, et al. Deciphering the transcriptional network of the dendritic cell lineage. Nat Immunol. 2012;13(9):888–99.
47. Hundelshausen P Von, Weber C. Platelets as immune cells bridging inflammation and cardiovascular disease. Circ Res. 2007;100(1):27–40.
48. Germain RN. T-cell development and the CD4-CD8 lineage decision. Nat Rev Immunol. 2002;2(5):309–22.
49. Naito T, Tanaka H, Naoe Y, Taniuchi I. Transcriptional control of T-cell development. Int Immunol. 2011;23(11):661–8.
50. Arens R, Schoenberger SP. Plasticity in programming of effector and memory CD8 T-cell formation. Immunol Rev. 2010;235(1):190–205.
51. Kaech SM, Wherry EJ, Ahmed R. Effector and memory T-cell differentiation: implications for vaccine development. Nat Rev Immunol. 2002;2(4):251–62.
52. Malissen B, Bongrand P. Early T cell activation: integrating biochemical, structural, and biophysical cues. Annu Rev Immunol. 2015;33:539–61.
53. Mellman I, Steinman RM. Dendritic cells: specialized and regulated antigen processing machines. Cell. 2001;106(3):255–8.

54. Pozzi LA, Maciaszek JW, Rock KL. Both dendritic cells and macrophages can stimulate naive CD8 T cells in vivo to proliferate, develop effector function, and differentiate into memory cells. J Immunol. 2005;175(4):2071–81.
55. Mempel TR, Henrickson SE, Von Andrian UH. T-cell priming by dendritic cells in lymph nodes occurs in three distinct phases. Nature. 2004;427(6970):154–9.
56. Sallusto F, Schaerli P, Loetscher P, Schaniel C, Lenig D, Mackay CR, et al. Rapid and coordinated switch in chemokine receptor expression during dendritic cell maturation. Eur J Immunol. 1998;28(9):2760–9.
57. Lim TS, Goh JK, Mortellaro A, Lim CT, Hämmerling GJ, Ricciardi-Castagnoli P. CD80 and CD86 differentially regulate mechanical interactions of T-cells with antigen-presenting dendritic cells and B-cells. PLoS One. 2012;7(9):e45185.
58. Doyle C, Strominger JL. Interaction between CD4 and class II MHC molecules mediates cell adhesion. Nature. 1987;330(6145):256–9.
59. Norment AM, Salter RD, Parham P, Engelhard VH, Littman DR. Cell-cell adhesion mediated by CD8 and MHC class I molecules. Nature. 1988;336(6194):79–81.
60. Chittasupho C, Siahaan TJ, Vines CM, Berkland C. Autoimmune therapies targeting costimulation and emerging trends in multivalent therapeutics. Ther Deliv. 2011;2(7):873–89.
61. Henry CJ, Ornelles DA, Mitchell LM, Brzoza-Lewis KL, Hiltbold EM. IL-12 Produced by Dendritic Cells Augments CD8+ T cell Activation through the Production of the Chemokines CCL1 and CCL17. J Immunol. 2008;181(12):8576–84.
62. Lai YP, Lin CC, Liao WJ, Tang CY, Chen SC. CD4+ T cell-derived IL-2 signals during early priming advances primary CD8+ T cell responses. PLoS One. 2009;4(11):e7766.
63. D'Souza WN, Lefrançois L. Frontline: An in-depth evaluation of the production of IL-2 by antigen-specific CD8 T cells in vivo. Eur J Immunol. 2004 Nov;34(11):2977–85.
64. Best JA, Blair D a, Knell J, Yang E, Mayya V, Doedens A, et al. Transcriptional insights into the CD8(+) T cell response to infection and memory T cell formation. Nat Immunol. 2013;14(4):404–12.

65. Curtsinger JM, Mescher MF. Inflammatory cytokines as a third signal for T cell activation. Curr Opin Immunol. 2010;22(3):333–40.
66. Zhu J, Yamane H, Paul WE. Differentiation of effector CD4 T cell populations. Annu Rev Immunol. 2010;28:445–89.
67. Auderset F, Coutaz M, Tacchini-Cottier F. The role of Notch in the differentiation of $CD4^+$ T helper cells. Curr Top Microbiol Immunol. 2012;360:115–34.
68. Zhu J, Paul WE. Peripheral CD4+ T-cell differentiation regulated by networks of cytokines and transcription factors. Immunol Rev. 2010;238(1):247–62.
69. Zhang Y, Zhang Y, Gu W, Sun B. TH1/TH2 cell differentiation and molecular signals. Adv Exp Med Biol. 2014;841:15–44.
70. Kaech SM, Cui W. Transcriptional control of effector and memory CD8+ T cell differentiation. Nat Rev Immunol. 2012;12(11):749–61.
71. Weaver CT, Elson CO, Fouser L a, Kolls JK. The Th17 pathway and inflammatory diseases of the intestines, lungs, and skin. Annu Rev Pathol. 2013;8:477–512.
72. Tran DQ. TGF-β: the sword, the wand, and the shield of FOXP3(+) regulatory T cells. J Mol Cell Biol. 2012;4(1):29–37.
73. Stäger S, Rafati S. CD8(+) T cells in leishmania infections: friends or foes. Front Immunol. 2012;3:5.
74. Kapsenberg M. Dendritic-cell control of pathogen-driven T-cell polarization. Nat Rev Immunol. 2003;3(12):984–93.
75. Joshi NS, Cui W, Chandele A, Lee HK, Urso DR, Hagman J, et al. Inflammation directs memory precursor and short-lived effector CD8+ T cell fates via the graded expression of T-bet transcription factor. Immunity. 2007;27(2):281–95.
76. Sercan O, Stoycheva D, Hämmerling GJ, Arnold B, Schüler T. IFN-gamma receptor signaling regulates memory CD8+ T cell differentiation. J Immunol. 2010;184(6):2855–62.
77. Smith-Garvin JE, Burns JC, Gohil M, Zou T, Kim JS, Maltzman JS, et al. T-cell receptor signals direct the composition and function of the memory CD8+ T-cell pool. Blood. 2010;116(25):5548–59.
78. Teixeiro E, Daniels MA, Hamilton SE, Schrum AG, Bragado R, Jameson SC, et al. Different T cell receptor signals determine CD8+ memory versus effector

development. Science. 2009;323(5913):502–5.

79. Joshi NS, Cui W, Dominguez CX, Chen JH, Hand TW, Kaech SM. Increased numbers of preexisting memory CD8 T cells and decreased T-bet expression can restrain terminal differentiation of secondary effector and memory CD8 T cells. J Immunol. 2011;187(8):4068–76.

80. Crotty S, Johnston RJ, Schoenberger SP. Effectors and memories: Bcl-6 and Blimp-1 in T and B lymphocyte differentiation. Nat Immunol. 2010;11(2):114–20.

81. Johnston RJ, Poholek AC, DiToro D, Yusuf I, Eto D, Barnett B, et al. Bcl6 and Blimp-1 are reciprocal and antagonistic regulators of T follicular helper cell differentiation. Science. 2009;325(5943):1006–10.

82. Kallies A, Xin A, Belz GT, Nutt SL. Blimp-1 transcription factor is required for the differentiation of effector CD8(+) T cells and memory responses. Immunity. 2009;31(2):283–95.

83. Rutishauser RL, Martins GA, Kalachikov S, Chandele A, Parish IA, Meffre E, et al. Transcriptional repressor Blimp-1 promotes CD8(+) T cell terminal differentiation and represses the acquisition of central memory T cell properties. Immunity. 2009;31(2):296–308.

84. Ichii H, Sakamoto A, Hatano M, Okada S, Toyama H, Taki S, et al. Role for Bcl-6 in the generation and maintenance of memory CD8+ T cells. Nat Immunol. 2002;3(6):558–63.

85. Khan AA, Penny LA, Yuzefpolskiy Y, Sarkar S, Kalia V. MicroRNA-17~92 regulates effector and memory CD8 T-cell fates by modulating proliferation in response to infections. Blood. 2013;121(22):4473–83.

86. Ji Y, Pos Z, Rao M, Klebanoff CA, Yu Z, Sukumar M, et al. Repression of the DNA-binding inhibitor Id3 by Blimp-1 limits the formation of memory CD8+ T cells. Nat Immunol. 2011;12(12):1230–7.

87. Mueller SN, Gebhardt T, Carbone FR, Heath WR. Memory T cell subsets, migration patterns, and tissue residence. Annu Rev Immunol. 2013;31:137–61.

88. Wakim LM, Woodward-Davis A, Liu R, Hu Y, Villadangos J, Smyth G, et al. The molecular signature of tissue resident memory CD8 T cells isolated from the brain. J Immunol. 2012;189(7):3462–71.

89. Shin H, Iwasaki A. Tissue-resident memory T cells. Immunol Rev. 2013;255(1):165–81.

90. Sathaliyawala T, Kubota M, Yudanin N, Turner D, Camp P, Thome JJC, et al. Distribution and compartmentalization of human circulating and tissue-resident memory T cell subsets. Immunity. 2013;38(1):187–97.
91. Ahlers JD, Belyakov IM. Memories that last forever: strategies for optimizing vaccine T-cell memory. Blood. 2010;115(9):1678–89.
92. Marzo AL, Klonowski KD, Le Bon A, Borrow P, Tough DF, Lefrançois L. Initial T cell frequency dictates memory CD8+ T cell lineage commitment. Nat Immunol. 2005;6(8):793–9.
93. Gebhardt T, Mackay LK. Local immunity by tissue-resident CD8(+) memory T cells. Front Immunol. 2012;3(November):340.
94. Ariotti S, Hogenbirk MA, Dijkgraaf FE, Visser LL, Hoekstra ME, Song JY, et al. Skin-resident memory $CD8^+$ T cells trigger a state of tissue-wide pathogen alert. Science. 2014;346(6205):101–5.
95. Gebhardt T, Wakim LM, Eidsmo L, Reading PC, Heath WR, Carbone FR. Memory T cells in nonlymphoid tissue that provide enhanced local immunity during infection with herpes simplex virus. Nat Immunol. 2009;10(5):524–30.
96. Jiang X, Clark RA, Liu L, Wagers AJ, Fuhlbrigge RC, Kupper TS. Skin infection generates non-migratory memory CD8+ T(RM) cells providing global skin immunity. Nature. 2012;483(7388):227–31.
97. Schenkel JM, Fraser KA, Beura LK, Pauken KE, Vezys V, Masopust D. Resident memory CD8 T cells trigger protective innate and adaptive immune responses. Science. 2014;346(6205):98–101.
98. Sheridan B, Pham QM, Lee YT, Cauley L, Puddington L, Lefrançois L. Oral infection drives a distinct population of intestinal resident memory CD8(+) T cells with enhanced protective function. Immunity. 2014;40(5):747–57.
99. Shin H, Iwasaki A. A vaccine strategy that protects against genital herpes by establishing local memory T cells. Nature. 2012;491(7424):463–7.
100. Wakim LM, Woodward-Davis A, Bevan MJ. Memory T cells persisting within the brain after local infection show functional adaptations to their tissue of residence. Proc Natl Acad Sci U S A. 2010;107(42):17872–9.
101. Hofmann M, Pircher H. E-cadherin promotes accumulation of a unique memory CD8 T-cell population in murine salivary glands. Proc Natl Acad Sci U S A. 2011;108(40):16741–6.
102. Fernandez-Ruiz D, Ng WY, Holz LE, Ma JZ, Zaid A, Wong YC, et al. Liver-

Resident Memory CD8+ T Cells Form a Front-Line Defense against Malaria Liver-Stage Infection. Immunity. 2016;45(4):889–902.

103. Mackay LK, Rahimpour A, Ma JZ, Collins N, Stock AT, Hafon ML, et al. The developmental pathway for CD103(+)CD8(+) tissue-resident memory T cells of skin. Nat Immunol. 2013;14(12):1294–301.
104. Mueller SN, Mackay LK. Tissue-resident memory T cells: local specialists in immune defence. Nat Rev Immunol. 2015;16(2):1–11.
105. Skon CN, Lee JY, Anderson KG, Masopust D, Hogquist KA, Jameson SC. Transcriptional downregulation of S1pr1 is required for the establishment of resident memory CD8+ T cells. Nat Immunol. 2013;14(12):1285–93.
106. Ikram N, Hassan K, Tufail S. Cytokines. Int J Pathol. 2004;2(1):47–58.
107. Whiteside TL. Cytokines and Cytokine Measurements in a Clinical Laboratory. Clin Diagn Lab Immunol. 1994;1(3):257–60.
108. Lunney JK. Cytokines orchestrating the immune response General properties of cytokines. RevSciTech. 1998;17(1):84–94.
109. Zhang JM, An J. Cytokines, inflammation, and pain. Int Anesth Clin. 2007;45(2):27–37.
110. Nicola N. Cytokine pleiotropy and redundancy: a view from the receptor. Stem Cells. 1994;12(1):3–12.
111. Bartee E, McFadden G. Cytokine synergy: an underappreciated contributor to innate anti-viral immunity. Cytokine. 2013;63(3):237–40.
112. Mäkelä SM, Strengell M, Pietilä TE, Osterlund P, Julkunen I. Multiple signaling pathways contribute to synergistic TLR ligand-dependent cytokine gene expression in human monocyte-derived macrophages and dendritic cells. J Leukoc Biol. 2009;85(4):664–72.
113. de Oliveira CM, Sakata RK, Issy AM, Gerola LR, Salomão R. Cytokines and pain. Rev Bras Anestesiol. 2011;61(2):255–9.
114. Leng SX, McElhaney JE, Walston JD, Xie D, Fedarko NS, Kuchel GA. ELISA and multiplex technologies for cytokine measurement in inflammation and aging research. J Gerontol A Biol Sci Med Sci. 2008;63(8):879–84.
115. Randall RE, Goodbourn S. Interferons and viruses: an interplay between induction, signalling, antiviral responses and virus countermeasures. J Gen Virol. 2008;89(1):1–47.
116. Akdis M, Burgler S, Crameri R, Eiwegger T, Fujita H, Gomez E, et al.

Interleukins, from 1 to 37, and interferon-γ: receptors, functions, and roles in diseases. J Allergy Clin Immunol. 2011;127(3):701–21.

117. Metcalf D. The colony-stimulating factors and cancer. Cancer Immunol Res. 2013;1(6):351–6.
118. Frantz S, Vincent KA, Feron O, Kelly RA. Innate immunity and angiogenesis. Circ Res. 2005;96(1):15–26.
119. Juhász K, Buzás K, Duda E. Importance of reverse signaling of the TNF superfamily in immune regulation. Expert Rev Clin Immunol. 2013;9(4):335–48.
120. Johnston CJ, Smyth DJ, Dresser DW, Maizels RM. TGF-β in tolerance, development and regulation of immunity. Cell Immunol. 2016;299(14–22).
121. Graves DT, Jiang Y. Chemokines, a family of chemotactic cytokines. Crit Rev Oral Biol Med. 1995;6(2):109–18.
122. Sachdeva N, Asthana D. Cytokine quantitation: technologies and applications. Front Biosci. 2007;1(12):682–95.
123. García-Piñeres A, Hildesheim A, Dodd L, Kemp T, Williams M, Harro C, et al. Cytokine and chemokine profiles following vaccination with human papillomavirus type 16 L1 Virus-like particles. Clin Vaccine Immunol. 2007;14(8):984–9.
124. Lalor MK, Floyd S, Gorak-Stolinska P, Ben-Smith A, Weir RE, Smith SG, et al. BCG vaccination induces different cytokine profiles following infant BCG vaccination in the UK and Malawi. J Infect Dis. 2011;204(7):1075–85.
125. Hur YG, Gorak-Stolinska P, Lalor MK, Mvula H, Floyd S, Raynes J, et al. Factors affecting immunogenicity of BCG in infants, a study in Malawi, The Gambia and the UK. BMC Infect Dis. 2014;14:184.
126. James EA, LaFond RE, Gates TJ, Mai DT, Malhotra U, Kwok WW. Yellow fever vaccination elicits broad functional CD4+ T cell responses that recognize structural and nonstructural proteins. J Virol. 2013;87(23):12794–804.
127. Bozza FA, Salluh JI, Japiassu AM, Soares M, Assis EF, Gomes RN, et al. Cytokine profiles as markers of disease severity in sepsis: a multiplex analysis. Crit Care. 2007;11(2):R49.
128. Davis JM, Knutson KL, Strausbauch MA, Crowson CS, Therneau TM, Wettstein PJ, et al. Analysis of complex biomarkers for human immune-mediated disorders based on cytokine responsiveness of peripheral blood cells.

J Immunol. 2010;184(12):7297–304.

129. Lozupone CA, Li M, Campbell TB, Flores SC, Linderman D, Gebert MJ, et al. Alterations in the gut microbiota associated with HIV-1 infection. Cell Host Microbe. 2013;14(3):329–39.

130. Cala CM, Moseley CE, Steele C, Dowdy SM, Cutter GR, Ness JM, et al. T cell cytokine signatures: Biomarkers in pediatric multiple sclerosis. J Neuroimmunol. 2016;297:1–8.

131. Burska A, Boissinot M, Ponchel F. Cytokines as biomarkers in rheumatoid arthritis. Mediators Inflamm. 2014;2014:1–24.

132. Andrade BB, Hullsiek KH, Boulware DR, Rupert A, French MA, Ruxrungtham K, et al. Biomarkers of inflammation and coagulation are associated with mortality and hepatitis flares in persons coinfected with HIV and hepatitis viruses. J Infect Dis. 2013;207(9):1379–88.

133. Cho J. The heritable immune system. Nat Biotechnol. 2015;33:2105.

134. Lee JY, Lee BS, Shin DJ, Woo PK, Shin YA, Joong KK, et al. A genome-wide association study of a coronary artery disease risk variant. J Hum Genet. 2013;58(3):120–6.

135. Gieger C, Kühnel B, Radhakrishnan A, Cvejic A, Serbanovic-Canic J, Meacham S, et al. New gene functions in megakaryopoiesis and platelet formation. Nature. 2011;480(7376):201–8.

136. Soranzo N, Spector TD, Mangino M, Kühnel B, Rendon A, Teumer A, et al. A genome-wide meta-analysis identifies 22 loci associated with eight hematological parameters in the HaemGen consortium. Nat Genet. 2009;41(11):1182–90.

137. Daly AK. Genome-wide association studies in pharmacogenomics. Nat Rev Genet. 2010;11(4):241–6.

138. The International HapMap Consortium. The International HapMap Project. Nature. 2003 Dec;426(6968):789–96.

139. The International HapMap Consortium. A haplotype map of the human genome. Nature. 2005;437(7063):1299–320.

140. The International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. Nature. 2007;449(7164):851–61.

141. 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature. 2015;526(7571):68–74.

142. 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. Nature. 2010;467(7319):1061–73.
143. 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012;491(7422):56–65.
144. Lamy P, Grove J, Wiuf C. A review of software for microarray genotyping. Hum Genomics. 2011;5(4):304–9.
145. Marchini J, Howie B. Genotype imputation for genome-wide association studies. Nat Rev Genet. 2010;11(7):499–511.
146. Stankiewicz P, Lupski JR. Structural variation in the human genome and its role in disease. Annu Rev Med. 2010;61:437–55.
147. Jorgenson E, Cheng I. Genome-wide association studies and cancer. Hawaii Med J. 2010;69(10):249–51.
148. Stranger BE, Stahl EA, Raj T. Progress and promise of genome-wide association studies for human complex trait genetics. Genetics. 2011;187(2):367–83.
149. Witte JS. Genome-wide association studies and beyond. Annu Rev Public Health. 2010;31:9–20.
150. Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, et al. Complement factor H polymorphism in age-related macular degeneration. Science. 2005;308(5720):385–9.
151. Yamazaki K, McGovern D, Ragoussis J, Paolucci M, Butler H, Jewell D, et al. Single nucleotide polymorphisms in TNFSF15 confer susceptibility to Crohn's disease. Hum Mol Genet. 2005;14(22):3499–506.
152. Sladek R, Rocheleau G, Rung J, Dina C, Shen L, Serre D, et al. A genome-wide association study identifies novel risk loci for type 2 diabetes. Nature. 2007;445(7130):881–5.
153. Duerr RH, Taylor KD, Brant SR, Rioux JD, Silverberg MS, Daly MJ, et al. A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. Science. 2006;314(5804):1461–3.
154. Burton P, Clayton D, Cardon L, Craddock N, Deloukas P, Duncanson A, et al. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 2007;447(7145):661–78.
155. Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic

Acids Res. 2014;42:D1001–6.

156. MacArthur, Bowler E, Cerezo M, Gil L, Hall P, Hastings E, et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res. 2017;45(Database issue):D896–D901.
157. Brant SR, Okou DT, Simpson CL, Cutler DJ, Haritunians T, Bradfield JP, et al. Genome-wide association study identifies African-specific susceptibility loci in African Americans with inflammatory bowel disease. Gastroenterology. 2017;152(1):206–17.
158. Marson A, Housley WJ, Hafler DA. Genetic basis of autoimmunity. J Clin Invest. 2015;125(6):2234–41.
159. Chambers JC, Zhang W, Li Y, Sehmi J, Wass MN, Zabaneh D, et al. Genome-wide association study identifies variants in TMPRSS6 associated with hemoglobin levels. Nat Genet. 2009;41(11):1170–2.
160. Gibson G. Rare and common variants: twenty arguments. Nat Rev Genet. 2011;13(2):135–45.
161. Visscher PM, Brown MA, McCarthy MI, Yang J. Five years of GWAS discovery. Am J Hum Genet. 2012;90(1):7–24.
162. Stranger BE, De Jager PL. Coordinating GWAS results with gene expression in a systems immunologic paradigm in autoimmunity. Curr Opin Immunol. 2012;24(5):544–51.
163. Stringer S, Wray NR, Kahn RS, Derks EM. Underestimated effect sizes in GWAS: fundamental limitations of single SNP analysis for dichotomous phenotypes. PLoS One. 2011;6(11):e27964.
164. Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, et al. Systematic localization of common disease-associated variation in regulatory DNA. Science. 2012;337(6099):1190–5.
165. Price AL, Zaitlen NA, Reich D, Patterson N. New approaches to population stratification in genome-wide association studies. Nat Rev Genet. 2010;11(7):459–63.
166. Astle W, Balding DJ. Population structure and cryptic relatedness in genetic association studies. Stat Sci. 2009;24(4):451–71.
167. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. PLoS Genet. 2006;2(12):e190.
168. Abraham G, Inouye M. Fast principal component analysis of large-scale

genome-wide data. PLoS One. 2014;9(4):e93766.

169. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006;38(8):904–9.
170. Hoffman GE. Correcting for population structure and kinship using the linear mixed model: theory and extensions. PLoS One. 2013;8(10):e75707.
171. de Bakker PI, Ferreira MA, Jia X, Neale BM, Raychaudhuri S, Voight BF. Practical aspects of imputation-driven meta-analysis of genome-wide association studies. Hum Mol Genet. 2008;17(R2):R122-8.
172. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods. 2008;5(7):621–8.
173. Murphy D. Gene expression studies using microarrays: principles, problems, and prospects. Adv Physiol Educ. 2002;26(1–4):256–70.
174. Barnes M, Freudenberg J, Thompson S, Aronow B, Pavlidis P. Experimental comparison and cross-validation of the Affymetrix and Illumina gene expression analysis platforms. Nucleic Acids Res. 2005;33(18):5914–23.
175. van Dijk EL, Jaszczyszyn Y, Thermes C. Library preparation methods for next-generation sequencing: tone down the bias. Exp Cell Res. 2014;322(1):12–20.
176. Liao Y, Smyth GK, Shi W. FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014;30(7):923–30.
177. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, et al. Transcript assembly and abundance estimation from RNA-Seq reveals thousands of new transcripts and switching among isoforms. Nat Biotechnol. 2010;28(5):511–5.
178. Garber M, Grabherr MG, Guttman M, Trapnell C. Computational methods for transcriptome annotation and quantification using RNA-seq. Nat Methods. 2011;8(5):469–77.
179. Guo Y, Sheng Q, Li J, Ye F, Samuels DC, Shyr Y. Large scale comparison of gene expression levels by microarrays and RNAseq using TCGA data. PLoS One. 2013;8(8):e71462.
180. Oshlack A, Robinson MD, Young MD. From RNA-seq reads to differential expression results. Genome Biol. 2010;11(12):220.

181. Macaulay IC, Voet T. Single cell genomics: advances and future perspectives. PLoS Genet. 2014;10(1):e1004126.
182. Lim WK, Wang K, Lefebvre C, Califano A. Comparative analysis of microarray normalization procedures: effects on reverse engineering gene networks. Bioinformatics. 2007;23(13):i282-8.
183. Gagnon-Bartsch JA, Speed TP. Using control genes to correct for unwanted variation in microarray data. Biostatistics. 2012;13(3):539–52.
184. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics. 2003;4(2):249–64.
185. Chen C, Grennan K, Badner J, Zhang D, Gershon E, Jin L, et al. Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods. PLoS One. 2011;6(2):e17238.
186. Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP. Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Res. 2003;31(4):15e–15.
187. Wu Z, Irizarry RA, Gentleman R, Martinez-Murillo F, Spencer F. A model-based background adjustment for oligonucleotide expression arrays. J Am Stat Assoc. 2004;99(468):909–17.
188. Wu J, Gentry Ri. gcrma: Background adjustment using sequence information. R package version 2.46.0. 2016;
189. Li C, Wong WH. Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. Proc Natl Acad Sci U S A. 2001;98(1):31–6.
190. Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, et al. Tackling the widespread and critical impact of batch effects in high-throughput data. Nat Rev Genet. 2010;11(10):733–9.
191. Qin S, Kim J, Arafat D, Gibson G. Effect of normalization on statistical and biological interpretation of gene expression profiles. Front Genet. 2012;3(May):160.
192. Kupfer P, Guthke R, Pohlers D, Huber R, Koczan D, Kinne RW. Batch correction of microarray data substantially improves the identification of genes differentially expressed in rheumatoid arthritis and osteoarthritis. BMC Med Genomics. 2012;5(1):23.

193. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007;8(1):118–27.
194. Leek JT, Storey JD. Capturing heterogeneity in gene expression studies by surrogate variable analysis. PLoS Genet. 2007;3(9):1724–35.
195. Mecham BH, Nelson PS, Storey JD. Supervised normalization of microarrays. Bioinformatics. 2010;26(10):1–5.
196. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc. 2012;7(3):562–78.
197. Zyprych-Walczak, J Szabelska A, Handschuh L, Górczak K, Klamecka K, Figlerowicz M, Siatkowski I. The impact of normalization methods on RNA-seq data analysis. Biomed Res Int. 2015;2015:621690.
198. Sreekumar J, Jose KK. Statistical tests for identification of differentially expressed genes in cDNA microarray experiments. Indian J Biotechnol. 2008;7:423–36.
199. Tarca AL, Romero R, Draghici S. Analysis of microarray experiments of gene expression profiling. Am J Obstet Gynecol. 2006;195(2):373–88.
200. Murie C, Woody O, Lee AY, Nadon R. Comparison of small n statistical tests of differential expression applied to microarrays. BMC Bioinformatics. 2009;10:45.
201. Cui X, Churchill GA. Statistical tests for differential expression in cDNA microarray experiments. Genome Biol. 2003;4(4):210.
202. Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol. 2004;3(1):Article3.
203. Bullard JH, Purdom E, Hansen KD, Dudoit S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. BMC Bioinformatics. 2010;11:94.
204. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26(1):139–40.
205. Robinson MD, Smyth GK. Moderated statistical tests for assessing differences in tag abundance. Bioinformatics. 2007;23(21):2881–7.

206. Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L. Differential analysis of gene regulation at transcript resolution with RNA-seq. Nat Biotechnol. 2013;31(1):46–53.
207. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol. 2010;11(10):R106.
208. Dudoit S, Yang YH, Callow MJ, Speed TP. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Stat Sin. 2002;12:111–39.
209. Dudoit S, Shaffer JP, Boldrick JC. Multiple hypothesis testing in microarray experiments. Stat Sci. 2003;18(1):71–103.
210. Bland J, Altman D. Multiple significance tests: the Bonferroni method. BMJ. 1995;310(6973):170.
211. Holm S. A simple sequentially rejective multiple test procedure. Scand J Stat. 1979;6(2):65–70.
212. Storey JD. A direct approach to false discovery rates. J R Stat Soc. 2002;64:479–498.
213. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B. 1995;57(1):289–300.
214. Gollub J, Sherlock G. Clustering microarray data. Methods Enzymol. 2006;411:194–213.
215. Kerr G, Ruskin HJ, Crane M, Doolan P. Techniques for clustering gene expression data. Comput Biol Med. 2008;38(3):283–93.
216. Ruan J, Dean AK, Zhang W. A general co-expression network-based approach to gene expression analysis: comparison and applications. BMC Syst Biol. 2010;4(1):8.
217. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A. 1998;95(25):14863–8.
218. Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM. Systematic determination of genetic network architecture. Nat Genet. 1999;22(3):281–5.
219. Carr DB, Somogyi R, George M. Templates for looking at gene expression clustering. Stat Comput Stat Graph Newslette. 1997;8(1995):20–9.
220. Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, et al. Interpreting patterns of gene expression with self-organizing maps: methods

and application to hematopoietic differentiation. Proc Natl Acad Sci U S A. 1999;96(6):2907–12.

221. Ringnér M. What is principal component analysis? Nat Biotechnol. 2008;26:303–4.
222. Boutros PC, Okey AB. Unsupervised pattern recognition: An introduction to the whys and wherefores of clustering microarray data. Brief Bioinform. 2005;6(4):331–43.
223. Lee YF, Roe T, Mangham DC, Fisher C, Grimer RJ, Judson I. Gene expression profiling identifies distinct molecular subgroups of leiomyosarcoma with clinical relevance. Br J Cancer. 2016;115(8):1000–7.
224. Jaskowiak PA, Campello RJ, Costa IG. On the selection of appropriate distances for gene expression data clustering. BMC Bioinformatics. 2014;15:S2.
225. Hu X, Park EK, Zhang X. Microarray gene cluster identification and annotation through cluster ensemble and EM-based informative textual summarization. IEEE Trans Inf Technol Biomed. 2009 Sep;13(5):832–40.
226. Kluger Y, Basri R, Chang JT, Gerstein M. Spectral biclustering of microarray data: coclustering genes and conditions. Genome Res. 2003;13(4):703–16.
227. Li L, Guo Y, Wu W, Shi Y, Cheng J, Tao S. A comparison and evaluation of five biclustering algorithms by quantifying goodness of biclusters for gene expression data. BioData Min. 2012;5(1):8.
228. Sheng Q, Moreau Y, De Moor B. Biclustering microarray data by Gibbs sampling. Bioinformatics. 2003;19(Suppl 2):ii196-ii205.
229. Zhu D, Hero AO, Cheng H, Khanna R, Swaroop A. Network constrained clustering for gene microarray data. Bioinformatics. 2005;21(21):4014–20.
230. Reiss DJ, Baliga NS, Bonneau R. Integrated biclustering of heterogeneous genome-wide datasets for the inference of global regulatory networks. BMC Bioinformatics. 2006;7:280.
231. Yosef N, Shalek AK, Gaublomme JT, Jin H, Lee Y, Awasthi A, et al. Dynamic regulatory network controlling TH17 cell differentiation. Nature. 2013;496(7446):461–8.
232. Zhang B, Gaiteri C, Bodea LG, Wang Z, McElwee J, Podtelezhnikov AA, et al. Integrated systems approach identifies genetic nodes and networks in late-onset Alzheimer's disease. Cell. 2019;153(3):707–20.

233. Chasman D, Fotuhi S, Roy S. Network-based approaches for analysis of complex biological systems. Curr Opin Biotechnol. 2016;39:157–66.
234. Butte AJ, Kohane IS. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. Pacific Symp Biocomput. 2000;426:418–29.
235. Quackenbush J. Genomics. Microarrays--guilt by association. Science. 2003;302(5643):240–1.
236. van Dam S, Cordeiro R, Craig T, van Dam J, Wood SH, de Magalhães JP. GeneFriends: an online co-expression analysis tool to identify novel gene targets for aging and complex diseases. BMC Genomics. 2012;13(1):535.
237. Ince RA, Giordano BL, Kayser C, Rousselet GA, Gross J, Schyns PG. A statistical framework for neuroimaging data analysis based on mutual information estimated via a gaussian copula. Hum Brain Mapp. 2017;38(3):1541–73.
238. Horvath S, Dong J. Geometric interpretation of gene coexpression network analysis. PLoS Comput Biol. 2008;4(8):24–6.
239. Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. Stat Appl Genet Mol Biol. 2005;4:e17.
240. Yip AM, Horvath S. Gene network interconnectedness and the generalized topological overlap measure. BMC Bioinformatics. 2007;8:22.
241. Langfelder P, Zhang B, Horvath S. Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. Bioinformatics. 2008;24(5):719–20.
242. van Noort V, Snel B, Huynen MA. The yeast coexpression network has a small-world, scale-free architecture and can be explained by a simple model. EMBO Rep. 2004;5(3):280–4.
243. Jordan IK, Mariño-Ramírez L, Wolf YI, Koonin E V. Conservation and coevolution in the scale-free human gene coexpression network. Mol Biol Evol. 2004;21(11):2058–70.
244. Tsaparas P, Mariño-Ramírez L, Bodenreider O, Koonin E V, Jordan IK. Global similarity and local divergence in human and mouse gene co-expression networks. BMC Evol Biol. 2006;6:70.
245. Carlson MRJ, Zhang B, Fang Z, Mischel PS, Horvath S, Nelson SF. Gene connectivity, function, and sequence conservation: predictions from modular yeast co-expression networks. BMC Genomics. 2006;7:40.

246. Lee HK, Hsu AK, Sajdak J, Qin J, Pavlidis P. Coexpression analysis of human genes across many microarray data sets. Genome Res. 2004;14(6):1085–94.
247. Albert R. Scale-free networks in cell biology. J Cell Sci. 2005;118(Pt 21):4947–57.
248. Chen L, Gable GG, Hu H. Communication and organizational social networks: a simulation model. Comput Math Organ Theory. 2012;19(4):460–79.
249. Preininger M, Arafat D, Kim J, Nath AP, Idaghdour Y, Brigham KL, et al. Blood-informative transcripts define nine common axes of peripheral blood gene expression. PLoS Genet. 2013;9(3):e1003362.
250. Boyle AP, Araya CL, Brdlik C, Cayting P, Cheng C, Cheng Y, et al. Comparative analysis of regulatory information and circuits across distant species. Nature. 2014;512(7515):453–6.
251. Ritchie SC, Watts S, Fearnley LG, Holt KE, Abraham G, Inouye M. A scalable permutation approach reveals replication and preservation patterns of network modules in large datasets. Cell Syst. 2016;3(1):71–82.
252. Langfelder P, Luo R, Oldham MC, Horvath S. Is my network module preserved and reproducible? PLoS Comput Biol. 2011;7(1):e1001057.
253. Tesson BM, Breitling R, Jansen RC. DiffCoEx: a simple and sensitive method to find differentially coexpressed gene modules. BMC Bioinformatics. 2010;11(1):497.
254. Lui TW, Tsui NB, Chan LW, Wong CS, Siu PM, Yung BY. DECODE: an integrated differential co-expression and differential expression analysis of gene expression data. BMC Bioinformatics. 2015;16:182.
255. Watson M. CoXpress: differential co-expression in gene expression data. BMC Bioinformatics. 2006;7:509.
256. Amar D, Safer H, Shamir R. Dissection of regulatory networks that are altered in disease via differential co-expression. PLoS Comput Biol. 2013;9(3):e1002955.
257. Choi Y, Kendziorski C. Statistical methods for gene set co-expression analysis. Bioinformatics. 2009;25(21):2780–6.
258. Fang G, Kuang R, Pandey G, Steinbach M, Myers CL, Kumar V. Subspace differential coexpression analysis: problem definition and a general approach. Pac Symp Biocomput. 2010;1(c):145–56.
259. Tseng GC, Ghosh D, Feingold E. Comprehensive literature review and

statistical considerations for microarray meta-analysis. Nucleic Acids Res. 2012;40(9):3785–99.

260. Slonim DK, Yanai I. Getting started in gene expression microarray analysis. PLoS Comput Biol. 2009;5(10):e1000543.
261. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000;25(1):25–9.
262. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M. KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res. 2012;40(Database issue):D109-14.
263. Nishimura D. A view From the web: BioCarta. Biotech Softw Internet Rep. 2001;2(3):117–20.
264. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2009;4(1):44–57.
265. Dennis G, Sherman B, Hosack D, Yang J, Gao W, Lane HC, et al. DAVID: Database for annotation, visualization, and integrated discovery. Genome Biol. 2003;4(5):P3.
266. Carmona-Saez P, Chagoyen M, Tirado F, Carazo JM, Pascual-Montano A. GENECODIS: a web-based tool for finding significant concurrent annotations in gene lists. Genome Biol. 2007;8(1):R3.
267. Tabas-Madrid D, Nogales-Cadenas R, Pascual-Montano A. GeneCodis3: a non-redundant and modular enrichment analysis tool for functional genomics. Nucleic Acids Res. 2012;40(Web Server issue):W478-83.
268. Keller A, Backes C, Al-Awadhi M, Gerasch A, Küntzer J, Kohlbacher O, et al. GeneTrailExpress: a web-based pipeline for the statistical evaluation of microarray experiments. BMC Bioinformatics. 2008;9:552.
269. Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC Bioinformatics. 2009;10(1):48.
270. Tarca AL, Bhatti G, Romero R. A comparison of gene set analysis methods in terms of sensitivity, prioritization and specificity. PLoS One. 2013;8(11):e79217.
271. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA,

et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005;102(43):15545–50.

272. Chaussabel D, Quinn C, Shen J, Patel P, Glaser C, Stichweh D, et al. A modular analysis framework for blood genomics studies: application to systemic lupus erythematosus. Immunity. 2009;29(1):150–64.
273. Pascual V, Chaussabel D, Banchereau J. A genomic approach to human autoimmune diseases. Immunology. 2010;28:535–71.
274. Lin H, Yin X, Lunetta KL, Dupuis J, McManus DD, Lubitz SA, et al. Whole blood gene expression and atrial fibrillation: The Framingham Heart Study. PLoS One. 2014;9(5):e96794.
275. Cappuzzello C, Napolitano M, Arcelli D, Melillo G, Melchionna R, Di Vito L, et al. Gene expression profiles in peripheral blood mononuclear cells of chronic heart failure patients. Physiol Genomics. 2009;352(16):233–40.
276. Anderson ST, Ph D, Kaforou M, Phil M, Brent AJ, Ph D, et al. Diagnosis of Childhood Tuberculosis and Host RNA Expression in Africa. 2014;370(18):1712–23.
277. Zhang Z-N, Xu J-J, Fu Y-J, Liu J, Jiang Y-J, Cui H-L, et al. Transcriptomic analysis of peripheral blood mononuclear cells in rapid progressors in early HIV infection identifies a signature closely correlated with disease progression. Clin Chem. 2013;59:1175–86.
278. Jacobsen M, Repsilber D, Gutschmidt A, Neher A, Feldmann K, Mollenkopf HJ, et al. Candidate biomarkers for discrimination between infection and disease caused by Mycobacterium tuberculosis. J Mol Med (Berl). 2007;85:613–21.
279. Ockenhouse CF, Hu W -c., Kester KE, Cummings JF, Stewart A, Heppner DG, et al. Common and Divergent Immune Response Signaling Pathways Discovered in Peripheral Blood Mononuclear Cell Gene Expression Patterns in Presymptomatic and Clinically Apparent Malaria. Infect Immun. 2006;74(10):5561–73.
280. Furman D, Jojic V, Kidd B, Shen-Orr S, Price J, Jarrell J, et al. Apoptosis and other immune biomarkers predict influenza vaccine responsiveness. Mol Syst Biol. 2013;9(659):659.
281. Nakaya HI, Wrammert J, Lee EK, Racioppi L, Marie-Kunze S, Haining WN, et

al. Systems biology of seasonal influenza vaccination in humans. Nat Immunol. 2012;12(8):786–95.

282. Monaco A, Marincola FM, Sabatino M, Pos Z, Tornesello ML, Stroncek DF, et al. Molecular immune signatures of HIV-1 vaccines in human PBMCs. FEBS Lett. 2009;583(18):3004–8.

283. Sanayama Y, Ikeda K, Saito Y, Kagami SI, Yamagata M, Furuta S, et al. Prediction of therapeutic responses to tocilizumab in patients with rheumatoid arthritis: Biomarkers identified by analysis of gene expression in peripheral blood mononuclear cells using genome-wide DNA microarray. Arthritis Rheumatol. 2014;66(6):1421–31.

284. Hecker M, Hartmann C, Kandulski O, Paap BK, Koczan D, Thiesen HJ, et al. Interferon-beta therapy in multiple sclerosis: The short-term and long-term effects on the patients' individual gene expression in peripheral blood. Mol Neurobiol. 2013;48:737–56.

285. Hu X, Yu J, Crosby SD, Storch GA. Gene expression profiles in febrile children with defined viral and bacterial infection. PNAS. 2013;110(31):12792–7.

286. Ramilo O, Allman W, Chung W, Mejias A, Ardura M, Glaser C, et al. Gene expression patterns in blood leukocytes discriminate patients with acute infections. 2007;109(5):1–2.

287. Zaas AK, Chen M, Varkey J, Veldman T, Hero AO 3rd, Lucas J, et al. Gene expression signatures diagnose influenza and other symptomatic respiratory viral infections in humans. Cell Host Microbe. 2009;6(3):207–17.

288. Berry MPR, Graham CM, McNab FW, Xu Z, Bloch SAA, Oni T, et al. An interferon-inducible neutrophil-driven blood transcriptional signature in human tuberculosis. Nature. 2010;466(7309):973–7.

289. Fairfax BP, Makino S, Radhakrishnan J, Plant K. Genetics of gene expression in primary immune cells identifies cell type–specific master regulators and roles of HLA allele. Nat Gen. 2012;44(5):502–10.

290. Ferraro A, D'Alise AM, Raj T, Asinovski N, Phillips R, Ergun A, et al. Interindividual variation in human T regulatory cells. Proc Natl Acad Sci. 2014;111(38):E1111–20.

291. Lee MN, Ye C, Villani AC, Raj T, Li W, Eisenhaure TM, et al. Common genetic variants modulate pathogen-sensing responses in human dendritic cells.

Science. 2014;343(6175):1246980.

292. Naranbhai V, Fairfax BP, Makino S, Humburg P, Wong D, Ng E, et al. Genomic modulators of gene expression in human neutrophils. Nat Commun. 2015;6:7545.

293. Barreiro L, Tailleux L, Pai A, Gicquel B, Marioni JC, Gilad Y. Deciphering the genetic architecture of variation in the immune response to Mycobacterium tuberculosis infection. Proc Natl Acad Sci. 2012;109:1204–9.

294. Ritchie SC, Würtz P, Nath AP, Abraham G, Havulinna AS, Fearnley LG, et al. The biomarker glycA is associated with chronic inflammation and predicts long-term risk of severe infection. Cell Syst. 2015;1(4):293–301.

295. Li S, Rouphael N, Duraisingham S, Romero-Steiner S, Presnell S, C D, et al. Molecular signatures of antibody responses derived from a systems biological study of 5 human vaccines. Nat Immunol. 2014;15(2):195–204.

296. Inouye M, Silander K, Hamalainen E, Salomaa V, Harald K, Jousilahti P, et al. An immune response network associated with blood lipid levels. PLoS Genet. 2010;6(9):e1001113.

297. Doering TA, Crawford A, Angelosanto JM, Paley MA, Ziegler CG, Wherry EJ. Network analysis reveals centrally connected genes and pathways involved in CD8+ T cell exhaustion versus memory. Immunity. 2012;37(6):1130–44.

298. Nayak RR, Kearns M, Spielman RS, Cheung VG. Coexpression network based on natural variation in human gene expression reveals gene interactions and functions. Genome Res. 2009;19(11):1953–62.

299. Choi JK, Yu U, Yoo OJ, Kim S. Differential coexpression analysis using microarray data and its application to human cancer. Bioinformatics. 2005;21(24):4348–55.

300. Cheung VG, Conlin LK, Weber TM, Arcaro M, Jen K, Morley M, et al. Natural variation in human gene expression assessed in lymphoblastoid cells. 2003;33(march):33–6.

301. Monks SA, Leonardson A, Zhu H, Cundiff P, Pietrusiak P, Edwards S, et al. Genetic inheritance of gene expression in human cell lines. Am J Hum Genet. 2004;75(6):1094–105.

302. Stranger BE, Nica AC, Forrest MS, Dimas A, Bird CP, Beazley C, et al. Population genomics of human gene expression. Nat Genet. 2007;39(10):1217–24.

303. Stranger BE, Forrest MS, Clark AG, Minichiello MJ, Deutsch S, Lyle R, et al. Genome-wide associations of gene expression variation in humans. PLoS Genet. 2005;1(6):e78.
304. Storey JD, Madeoy J, Strout JL, Wurfel M, Ronald J, Akey JM. Gene-expression variation within and among human populations. Am J Hum Genet. 2007;80(3):502–9.
305. Idaghdour Y, Storey JD, Jadallah SJ, Gibson G. A genome-wide gene expression signature of environmental geography in leukocytes of Moroccan Amazighs. PLoS Genet. 2008;4(4):e1000052.
306. Heap GA, Trynka G, Jansen RC, Bruinenberg M, Swertz M a, Dinesen LC, et al. Complex nature of SNP genotype effects on gene expression in primary human leucocytes. BMC Med Genomics. 2009;2:1.
307. Göring HHH, Curran JE, Johnson MP, Dyer TD, Charlesworth J, Cole S a, et al. Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes. Nat Genet. 2007 Oct;39(10):1208–16.
308. Emilsson V, Thorleifsson G, Zhang B, Leonardson AS, Zink F, Zhu J, et al. Genetics of gene expression and its effect on disease. Nature. 2008 Mar;452(7186):423–8.
309. Fehrmann RSN, Jansen RC, Veldink JH, Westra H-JJ, Arends D, Bonder MJ, et al. Trans-eQTLs reveal that independent genetic variants associated with a complex phenotype converge on intermediate genes, with a major role for the HLA. PLoS Genet. 2011;7(8):e1002197.
310. Battle A, Mostafavi S, Zhu X, Potash JB, Weissman MM, McCormick C, et al. Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. Genome Res. 2014;24(650):14–24.
311. GTEx Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. Science (80-). 2015;348(6235):648–60.
312. Shabalin AA. Matrix eQTL: Ultra fast eQTL analysis via large matrix operations. Bioinformatics. 2012;28(10):1353–8.
313. Kidd B a, Peters L a, Schadt EE, Dudley JT. Unifying immunology with informatics and multiscale biology. Nat Immunol. 2014;15(2):118–27.
314. Nicolae DL, Gamazon E, Zhang W, Duan S, Dolan ME, Cox NJ. Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. PLoS Genet. 2010;6(4):e1000888.

315. Schadt EE, Molony C, Chudin E, Hao K, Yang X, Lum PY, et al. Mapping the genetic architecture of gene expression in human liver. PLoS Biol. 2008;6(5):e107.
316. Webster JA, Gibbs JR, Clarke J, Ray M, Zhang W, Holmans P, et al. Genetic control of human brain transcript expression in Alzheimer disease. Am J Hum Genet. 2009;84(4):445–58.
317. Heinzen EL, Ge D, Cronin KD, Maia JM, Shianna K V, Gabriel WN, et al. Tissue-specific genetic control of splicing: implications for the study of complex traits. PLoS Biol. 2008;6(12):e1.
318. Dimas AS, Deutsch S, Stranger BE, Montgomery SB, Borel C, Attar-Cohen H, et al. Common regulatory variation impacts gene expression in a cell type-dependent manner. Science. 2009;325(5945):1246–50.
319. Powell JE, Henders AK, Mcrae AF, Wright MJ, Martin NG, Dermitzakis ET, et al. Genetic control of gene expression in whole blood and lymphoblastoid cell lines is largely independent. Genome Res. 2012;22(3):456–66.
320. Westra HJ, Peters MJ, Esko T, Yaghootkar H, Schurmann C, Kettunen J, et al. Systematic identification of trans eQTLs as putative drivers of known disease associations. Nat Genet. 2013;45(10):1238–43.
321. Mehta D, Heim K, Herder C, Carstensen M, Eckstein G, Schurmann C, et al. Impact of common regulatory single-nucleotide variants on gene expression profiles in whole blood. Eur J Hum Genet. 2013;21(June 2012):48–54.
322. Joehanes R, Zhang X, Huan T, Yao C, Ying SX, Nguyen QT, et al. Integrated genome-wide analysis of expression quantitative trait loci aids interpretation of genomic association studies. Genome Biol. 2017;18(1):16.
323. Raj T, Rothamel K, Mostafavi S, Ye C, Lee MN, Replogle JM, et al. Polarization of the effects of autoimmune and neurodegenerative risk alleles in leukocytes. 2014;344(May):519–24.
324. Zeller T, Wild P, Szymczak S, Rotival M, Schillert A, Castagne R, et al. Genetics and beyond--the transcriptome of human monocytes and disease susceptibility. PLoS One. 2010;5(5):e10693.
325. Huehn J, Polansky JK, Hamann A. Epigenetic control of FOXP3 expression: the key to a stable regulatory T-cell lineage? Nat Rev Immunol. 2009;9(2):83–9.
326. Huang J, Plass C, Gerhauser C. Cancer chemoprevention by targeting the

epigenome. Curr Drug Targets. 2011;12(13):1925–56.

327. Eccleston A, DeWitt N, Gunter C, Marte B, Nath D. Epigenetics. Nature. 2007;447(7143):395–395.

328. Conaway JW. Introduction to theme "Chromatin, epigenetics, and transcription". Annu Rev Biochem. 2012;81:61–4.

329. Bidwell J, Keen L, Gallagher G, Kimberly R, Huizinga T, McDermott, M F Oksenberg J, et al. Cytokine gene polymorphism in human disease: on-line databases. Genes Immun. 1999;1(1):3–19.

330. Hollegaard MV, Bidwell JL. Cytokine gene polymorphism in human disease: on-line databases, Supplement 3. Genes Immun. 2006;7(4):269–76.

331. Pravica V, Perrey C, Stevens A, Lee JH, Hutchinson I V. A single nucleotide polymorphism in the first intron of the human IFN-gamma gene: absolute correlation with a polymorphic CA microsatellite marker of high IFN-gamma production. Hum Immunol. 2000;61(9):863–6.

332. de Craen AJ, Posthuma D, Remarque EJ, van den Biggelaar AH, Westendorp RG, Boomsma DI. Heritability estimates of innate immunity: an extended twin study. Genes Immun. 2005;6(2):167–70.

333. Peresi E, Oliveira LRC, da Silva WL, da Costa EAPN, Araujo JP, Ayres JA, et al. Cytokine polymorphisms, their influence and levels in Brazilian patients with pulmonary tuberculosis during antituberculosis treatment. Tuberc Res Treat. 2013;2013:285094.

334. Banerjee M, Saxena M. Genetic polymorphisms of cytokine genes in type 2 diabetes mellitus. World J Diabetes. 2014;5(4):493–504.

335. Tseng FC, Brown EE, Maiese EM, Yeager M, Welch R, Gold BD, et al. Polymorphisms in cytokine genes and risk of Helicobacter pylori infection among Jamaican children. Helicobacter. 2006;11(5):425–30.

336. Garg PR, Saraswathy KN, Kalla AK, Sinha E, Ghosh PK. Pro-inflammatory cytokine gene polymorphisms and threat for coronary heart disease in a North Indian Agrawal population. Gene. 2013;514(1):69–74.

337. Yu Z, Liu Q, Huang C, Wu M, Li G. The interleukin 10 -819C/T polymorphism and cancer risk: a HuGE review and meta-analysis of 73 studies including 15,942 cases and 22,336 controls. OMICS. 2013;17(4):200–14.

338. Li F, Xu J, Zheng J, Sokolove J, Zhu K, Zhang Y, et al. Association between interleukin-6 gene polymorphisms and rheumatoid arthritis in Chinese Han

population: a case-control study and a meta-analysis. Sci Rep. 2014;4:5714.

339. Dubois PC, Trynka G, Franke L, Hunt KA, Romanos J, Curtotti A, et al. Multiple common variants for celiac disease influencing immune gene expression. Nat Genet. 2010;42(4):295–302.
340. Larsen MH, Albrechtsen A, Thørner LW, Werge T, Hansen T, Gether U, et al. Genome-wide association study of genetic variants in LPS-stimulated IL-6, IL-8, IL-10, IL-1ra and TNF-α cytokine response in a Danish cohort. PLoS One. 2013;8(6):e66262.
341. Matteini AM, Li J, Lange EM, Tanaka T, Lange L a, Tracy RP, et al. Novel gene variants predict serum levels of the cytokines IL-18 and IL-1ra in older adults. Cytokine. 2014;65(1):10–6.
342. Ayele FT, Doumatey A, Huang H, Zhou J, Charles B, Erdos M, et al. Genome-wide associated loci influencing interleukin (IL)-10, IL-1Ra, and IL-6 levels in African Americans. Immunogenetics. 2012;64(5):351–9.
343. Debette S, Visvikis-Siest S, Chen MH, Ndiaye NC, Song C, Destefano A, et al. Identification of cis-and trans-acting genetic variants explaining up to half the variation in circulating vascular endothelial growth factor levels. Circ Res. 2011;109(5):554–63.
344. Ahola-Olli A V, Würtz P, Havulinna AS, Aalto K, Pitkänen N, Lehtimäki T, et al. Genome-wide association study identifies 27 Loci influencing concentrations of circulating cytokines and growth factor. Am J Hum Genet. 2017;100(1):40–50.
345. Shah SH, Kraus WE, Newgard CB. Metabolomic profiling for the identification of novel biomarkers and mechanisms related to common cardiovascular diseases form and function. Circulation. 2012;126(9):1110–20.
346. Ellis DI, Dunn WB, Griffin JL, Allwood JW, Goodacre R. Metabolic fingerprinting as a diagnostic tool. Pharmacogenomics. 2007;8:1243–66.
347. Nicholson JK, Lindon JC. Systems biology: Metabonomics. Nature. 2008;455(7216):1054–6.
348. Nicholson JK, Holmes E, Kinross J, Burcelin R, Gibson G, Jia W, et al. Host-gut microbiota metabolic interactions. Science. 2012;336(6086)(6086):1262–7.
349. Soininen P, Kangas AJ, Würtz P, Tukiainen T, Tynkkynen T, Laatikainen R, et al. High-throughput serum NMR metabonomics for cost-effective holistic studies on systemic metabolism. Analyst. 2009;134(9):1781–5.

350. Nicholson JK, Lindon JC, Holmes E. Metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. Xenobiotica. 1999;29(11):1181–9.
351. Soininen P, Kangas a. J, Wurtz P, Suna T, Ala-Korpela M. Quantitative Serum Nuclear Magnetic Resonance Metabolomics in Cardiovascular Epidemiology and Genetics. Circ Cardiovasc Genet. 2019;8:192–206.
352. Dona AC, Kyriakides M, Scott F, Shephard EA, Varshavi D, Veselkov K, et al. A guide to the identification of metabolites in NMR-based metabonomics/metabolomics experiments. Comput Struct Biotechnol J. 2016;14:135–53.
353. Zheng C, Zhang S, Ragg S, Raftery D, Vitek O. Identification and quantification of metabolites in (1)H NMR spectra by Bayesian model selection. Bioinformatics. 2011;27(12):637–44.
354. Aebersold R, Mann M. Mass spectrometry-based proteomics. Nature. 2003;422(6928):198–207.
355. Bartel J, Krumsiek J, Schramm K, Adamski J, Gieger C, Herder C, et al. The Human blood metabolome-transcriptome interface. PLoS Genet. 2015;11(6):e1005274.
356. Gowda GA, Djukovic D. Overview of mass spectrometry-based metabolomics: opportunities and challenges. Methods Mol Biol. 2014;1198:3–12.
357. Suhre K, Meisinger C, Döring A, Altmaier E, Belcredi P, Gieger C, et al. Metabolic footprint of diabetes: a multiplatform metabolomics study in an epidemiological setting. PLoS One. 2010;5(11):e13953.
358. Peng S, Zhang J, Liu L, Zhang X, Huang Q, Alamdar A, et al. Newborn meconium and urinary metabolome response to maternal gestational diabetes mellitus: a preliminary case-control study. J Proteome Res. 2015;14(4):1799–809.
359. Lau SK, Lee KC, Curreem SO, Chow WN, To KK, Hung IF, et al. Metabolomic profiling of plasma from patients with tuberculosis by use of untargeted mass spectrometry reveals novel biomarkers for diagnosis. J Clin Microbiol. 2015;53(12):3750–9.
360. Wurtz P, Havulinna AS, Soininen P, Tynkkynen T, Prieto-Merino D, Tillin T, et al. Metabolite profiling and cardiovascular event risk: a prospective study of

3 population-based cohorts. Circulation. 2015;131(9):774–85.

361. Goek ON, Döring A, Gieger C, Heier M, Koenig W, Prehn C, et al. Serum metabolite concentrations and decreased GFR in the general population. Am J Kidney Dis. 2012;60(2):197–206.

362. Wang TJ, Larson MG, Vasan RS, Cheng S, Rhee EP, McCabe E, et al. Metabolite profiles and the risk of developing diabetes. Nat Med. 2011;17(4):448–53.

363. Fischer K, Kettunen J, Würtz P, Haller T, Havulinna AS, Kangas AJ, et al. Biomarker profiling by nuclear magnetic resonance spectroscopy for the prediction of all-cause mortality: an observational study of 17,345 persons. PLoS Med. 2014;11(2):e100160.

364. Langley RJ, Tsalik EL, van Velkinburgh, J C Glickman SW, Rice BJ, Wang C, Chen B, et al. An integrated clinico-metabolomic model improves prediction of death in sepsis. Sci Transl Med. 2013;5(195):195ra95.

365. Karlíková R, Široká J, Friedecký D, Faber E, Hrdá M, Mičová K, et al. Metabolite Profiling of the Plasma and Leukocytes of Chronic Myeloid Leukemia Patients. J Proteome Res. 2016;15(9):3158–66.

366. Daemen A, Peterson D, Sahu N, McCord R, Du X, Liu B, et al. Metabolite profiling stratifies pancreatic ductal adenocarcinomas into subtypes with distinct sensitivities to metabolic inhibitors. Proc Natl Acad Sci U S A. 2015;112(32):E4410–E4417.

367. Roe B, Kensicki E, Mohney R, Hall WW. Metabolomic profile of hepatitis C virus-infected hepatocytes. PLoS One. 2011;6(8):e23641.

368. Wahl S, Yu Z, Kleber M, Singmann P, Holzapfel C, He Y, et al. Childhood obesity is associated with changes in the serum metabolite profile. Obes Facts. 2012;5(5):660–70.

369. Wahl S, Vogt S, Stückler F, Krumsiek J, Bartel J, Kacprowski T, et al. Multi-omic signature of body weight change: results from a population-based cohort study. BMC Med. 2015;13:48.

370. Auro K, Joensuu A, Fischer K, Kettunen J, Salo P, Mattsson H, et al. A metabolic view on menopause and ageing. Nat Commun. 2014;5:4708.

371. Wang Q, Würtz P, Auro K, Morin-Papunen L, Kangas AJ, Soininen P, et al. Effects of hormonal contraception on systemic metabolism: cross-sectional and longitudinal evidence. Int J Epidemiol. 2016;45(5):1445–57.

372. Würtz P, Mäkinen VP, Soininen P, Kangas AJ, Tukiainen T, Kettunen J, et al. Metabolic signatures of insulin resistance in 7,098 young adults. Diabetes. 2012;61(6):1372–80.
373. Pearce EL, Pearce EJ. Metabolic pathways in immune cell activation and quiescence. Immunity. 2013;38(4):633–43.
374. Loftus RM, Finlay D. Immunometabolism: Cellular metabolism turns immune regulator. J Biol Chem. 2016;291(1):1–10.
375. Inouye M, Kettunen J, Soininen P, Silander K, Ripatti S, Kumpula LS, et al. Metabonomic, transcriptomic, and genomic variation of a population cohort. Mol Syst Biol. 2010;6:441.
376. Ferrara CT, Wang P, Neto EC, Stevens RD, Bain JR, Wenner BR, et al. Genetic networks of liver metabolism revealed by integration of metabolic and transcriptional profiling. PLoS Genet. 2008;4(3):e1000034.
377. Connor SC, Hansen MK, Corner A, Smith RF, Ryan TE. Integration of metabolomics and transcriptomics data to aid biomarker discovery in type 2 diabetes. Mol Biosyst. 2010;6(5):909–21.
378. Nath AP, Ritchie SC, Byars SG, Fearnley LG, Havulinna AS, Joensuu A, et al. An interaction map of circulating metabolites, immune gene networks, and their genetic regulation. Genome Biol. 2017;18(1):146.
379. Hotamisligil GS. Inflammation, metaflammation and immunometabolic disorders. Nature. 2017;542(7640):177–85.
380. Hotamisligil GS, Shargill NS, Spiegelman BM. Adipose expression of tumor necrosis factor-alpha: direct role in obesity-linked insulin resistance. Science. 1993;259(5091):87–91.
381. Weisberg SP, Mccann D, Desai M, Rosenbaum M, Leibel RL, Ferrante AW. Obesity is associated with macrophage accumulation. J Clin Investig. 2003;112(12):1796–1808.
382. Senn JJ, Klover PJ, Nowak IA, Mooney RA. Interleukin-6 induces cellular insulin resistance in hepatocytes. Diabetes. 2002;51(12):3391–9.
383. Maedler K, Sergeev P, Ris F, Oberholzer J, Joller-jemelka HI, Spinas GA, et al. Glucose-induced beta cell production of IL-1 beta contributes to glucotoxicity in human pancreatic islets. J Clin Invest. 2002;110(6):851–60.
384. Böni-Schnetzler M, Boller S, Debray S, Bouzakri K, Meier DT, Prazak R, et al. Free fatty acids induce a proinflammatory response in islets via the abundantly

expressed interleukin-1 receptor I. Endocrinology. 2009;150(12):5218–29.

385. Böni-Schnetzler M, Thorne J, Parnaud G, Marselli L, Ehses JA, Kerr-Conte J, et al. Increased interleukin (IL)-1beta messenger ribonucleic acid expression in beta -cells of individuals with type 2 diabetes and regulation of IL-1beta in human islets by glucose and autostimulation. J Clin Endocrinol Metab. 2008;93(10):4065–74.
386. Hansson G. Inflammation, atherosclerosis, and coronary artery disease. N Engl J Med. 2005;21(352):1685–95.
387. Rajamäki K, Lappalainen J, Öörni K, Välimäki E, Matikainen S, Kovanen PT, et al. Cholesterol crystals activate the NLRP3 inflammasome in human macrophages: a novel link between cholesterol metabolism and inflammation. PLoS One. 2010;5(7):e11765.
388. Li Y, Schwabe RF, DeVries-Seimon T, Yao PM, Gerbod-Giannone MC, Tall AR, et al. Free cholesterol-loaded macrophages are an abundant source of tumor necrosis factor-alpha and interleukin-6: model of NF-kappaB- and map kinase-dependent inflammation in advanced atherosclerosis. J Biol Chem. 2005;280(23):21763–72.
389. Raitakari OT, Juonala M, Rönnemaa T, Keltikangas-Järvinen L, Räsänen L, Pietikäinen M, et al. Cohort profile: The cardiovascular risk in Young Finns Study. Int J Epidemiol. 2008;37(6):1220–6.
390. Nuotio J, Oikonen M, Magnussen CG, Jokinen E, Laitinen T, Hutri-Kähönen N, et al. Cardiovascular risk factors in 2011 and secular trends since 2007: The cardiovascular risk in Young Finns Study. Scand J Public Health. 2014;42(7):563–71.
391. Smith EN, Chen W, Kähönen M, Kettunen J, Lehtimäki T, Peltonen L, et al. Longitudinal genome-wide association of cardiovascular disease risk factors in the Bogalusa heart study. PLoS Genet. 2010;6(9):e1001094.
392. Teo YY, Inouye M, Small KS, Gwilliam R, Deloukas P, Kwiatkowski DP, et al. A genotype calling algorithm for the Illumina BeadArray platform. Bioinformatics. 2007;23(20):2741–6.
393. Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet. 2009;5(6):e1000529.
394. Raitoharju E, Seppälä I, Oksala N, Lyytikäinen LP, Raitakari O, Viikari J, et al.

Blood microRNA profile associates with the levels of serum lipids and metabolites associated with glucose metabolism and insulin resistance and pinpoints pathways underlying metabolic syndrome. The cardiovascular risk in Young Finns Study. Mol Cell Endocrinol. 2014;391(1–2):41–9.

395. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008;9:559.
396. Smyth GK, Phipson B. Permutation P-values should never be zero: calculating exact P-values when permutations are randomly drawn. Stat Appl Genet Mol Biol. 2010;9:e39.
397. Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z. GOrilla : a tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC Bioinformatics. 2009;10(1):48.
398. Supek F, Bošnjak M, Škunca N, Šmuc T. REVIGO summarizes and visualizes long lists of gene ontology terms. PLoS One. 2011;6(7):e21800.
399. Breuer K, Foroushani AK, Laird MR, Chen C, Sribnaia A, Lo R, et al. InnateDB: systems biology of innate immunity and beyond—recent updates and continuing curation. Nucleic Acids Res. 2013;41(1):D1228–D1233.
400. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience. 2015;4:7.
401. Willer CJ, Li Y, Abecasis GR, Overall P. METAL : fast and efficient meta-analysis of genomewide association scans. Bioinformatics. 2010;26(17):2190–1.
402. Freytag S, Gagnon-Bartsch J, Speed TP, Bahlo M. Systematic noise degrades gene co-expression signals but can be corrected. BMC Bioinformatics. 2015;24(16):309.
403. Gibson G. The environmental contribution to gene expression profiles. Nat Rev Genet. 2008;9(8):575–81.
404. Morrell CN, Aggrey AA, Chapman LM, Modjeski KL. Emerging roles for platelets as immune and inflammatory cells. Blood. 2016;123(18):2759–68.
405. Hidalgo LG, Einecke G, Allanach K, Halloran PF. The transcriptome of human cytotoxic T cells: Similarities and disparities among allostimulated CD4+ CTL, CD8+ CTL and NK cells. Am J Transplant. 2008;8(3):627–36.
406. Zhang X, Wang Q, Shen Y, Song H, Gong Z, Wang L. Compromised natural

killer cells in pulmonary embolism. Int J Clin Exp Pathol. 2015;8(7):8244–51.

407. Wu N, Zhong M, Roncagalli R, Guo H, Zhang Z, Lenoir C, et al. A hematopoietic cell – driven mechanism involving SLAMF6 receptor , SAP adaptors and SHP-1 phosphatase regulates NK cell education. Nat Immunol. 2016;17(4):387–96.
408. Serafini N, Vosshenrich CAJ, Di Santo JP. Transcriptional regulation of innate lymphoid cell fate. Nat Rev Immunol. 2015;15(7):415–28.
409. Chiang YJ, Kole HK, Brown K, Naramura M, Fukuhara S, Hu RJ, et al. Cbl-b regulates the CD28 dependence of T-cell activation. Nature. 2000;403(6766):216–20.
410. Au-Yeung BB, Deindl S, Hsu LY, Palacios EH, Levin SE, Kuriyan J, et al. The structure, regulation, and function of ZAP-70. Immunol Rev. 2009;228(1):41–57.
411. Maghazachi AA. Role of chemokines in the biology of natural killer cells. Curr Top Microbiol Immunol. 2010;341:37–58.
412. Schoggins JW, Rice CM. Interferon-stimulated genes and their antiviral effector functions. Curr Opin Virol. 2012;1(6):519–25.
413. Zhou X, Michal JJ, Zhang L, Ding B, Lunney JK, Liu B, et al. Interferon induced IFIT family genes in host antiviral defense. Int J Biol Sci. 2013;9(2):200–8.
414. Choi UY, Kang JS, Hwang YS, Kim YJ. Oligoadenylate synthase-like (OASL) proteins: dual functions and associations with diseases. Exp Mol Med. 2015;47:e144.
415. Borden EC, Sen GC, Uze G, Silverman RH, Ransohoff RM, Foster GR, et al. Interferons at age 50: past, current and future impact on biomedicine. Nat Rev Drug Discov. 2007;6(12):975–90.
416. Cheon H, Holvey-Bates EG, Schoggins JW, Forster S, Hertzog P, Imanaka N, et al. IFNβ-dependent increases in STAT1, STAT2, and IRF9 mediate resistance to viruses and DNA damage. EMBO J. 2013;32(20):2751–63.
417. Hashimoto S, Chiorazzi N, Gregersent PK, Human B. Alternative splicing of CD79a (Ig-alpha/mb-1) and CD79b (Ig-beta/B29) RNA transcripts in human B cells. Mol Immunol. 1995;32(9):651–9.
418. Tedder TF, Tuscano J, Sato S, Kehrl JH. CD22, a B lymphocyte-specific adhesion molecule that regulates antigen receptor signaling. Annu Rev

Immunol. 1997;15:481–504.

419. Ferrer G, Hodgson K, Montserrat E, Moreno C. B cell activator factor and a proliferation-inducing ligand at the cross-road of chronic lymphocytic leukemia and autoimmunity. Leuk Lymphoma. 2009;50(7):1075–82.
420. Klein U, Dalla-Favera R. Germinal centres: role in B-cell physiology and malignancy. Nat Rev Immunol. 2008;8(1):22–33.
421. Wiede F, Fromm PD, Comerford I, Kara E, Bannan J, Schuh W, et al. CCR6 is transiently upregulated on B cells after activation and modulates the germinal center reaction in the mouse. Immunol Cell Biol. 2013;91(5):335–9.
422. Breloer M, Kretschmer B, Lüthje K, Ehrlich S, Ritter U, Bickert T, et al. CD83 is a regulator of murine B cell function in vivo. Eur J Immunol. 2007;37(3):634–48.
423. Poluektov YO, Kim A, Sadegh-Nasseri S. HLA-DO and its role in MHC class II antigen presentation. Front Immunol. 2013;4:260.
424. Stegner D, Nieswandt B. Platelet receptor signaling in thrombus formation. J Mol Med. 2011;89(2):109–21.
425. Protty MB, Watkins NA, Colombo D, Thomas SG, Heath VL, Herbert JMJ, et al. Identification of Tspan9 as a novel platelet tetraspanin and the collagen receptor GPVI as a component of tetraspanin microdomains. Biochem J. 2009;417(1):391–400.
426. Kato K, Martinez C, Russell S, Nurden P, Nurden A, Fiering S, et al. Genetic deletion of mouse platelet glycoprotein Ibbeta produces a Bernard-Soulier phenotype with increased alpha-granule size. Blood. 2004;104(8):2339–44.
427. Ganz T. Angiogenin: an antimicrobial ribonuclease. Nat Immunol. 2003;4(3):213–4.
428. Ganz T. Defensins: antimicrobial peptides of innate immunity. Nat Rev Immunol. 2003;3(9):710–20.
429. Levy O. A neutrophil-derived anti-infective molecule: bactericidal/permeability-increasing protein. Antimicrob Agents Chemother. 2000;44(11):2925–31.
430. Xu X, Su S, Wang X, Barnes V, De Miguel C, Ownby D, et al. Obesity is associated with more activated neutrophils in African American male youth. Int J Obes (Lond). 2015;39(1):26–32.
431. Kuroki M, Abe H, Imakiirei T, Liao S, Uchida H, Yamauchi Y, et al.

Identification and comparison of residues critical for cell-adhesion activities of two neutrophil CD66 antigens, CEACAM6 and CEACAM8. J Leukoc Biol. 2001;70(4):543–50.

432. Wu Z, Sawamura T, Kurdowska AK, Ji HL, Idell S, Fu J. LOX-1 deletion improves neutrophil responses, enhances bacterial clearance, and reduces lung injury in a murine polymicrobial sepsis model. Infect Immun. 2011;79(7):2865–70.

433. Shashidharamurthy R, MacHiah D, Aitken JD, Putty K, Srinivasan G, Chassaing B, et al. Differential role of lipocalin 2 during immune complex-mediated acute and chronic inflammation in mice. Arthritis Rheum. 2013;65(4):1064–73.

434. Alalwani SM, Sierigk J, Herr C, Pinkenburg O, Gallo R, Vogelmeier C, et al. The antimicrobial peptide LL-37 modulates the inflammatory and host defense response of human neutrophils. Eur J Immunol. 2010;40(4):1118–26.

435. Cruse G, Kaur D, Leyland M, Bradding P. A novel FcepsilonRIbeta-chain truncation regulates human mast cell proliferation and survival. FASEB J. 2010;24(10):4047–57.

436. Kraft S, Kinet JP. New developments in FcepsilonRI regulation, function and inhibition. Nat Rev Immunol. 2007;7(5):365–78.

437. Melbye H, Hvidsten D, Holm A, Nordbø SA, Brox J. The course of C-reactive protein response in untreated upper respiratory tract infection. Br J Gen Pr. 2004;54(506):653–8.

438. Sasaki K, Fujita I, Hamasaki Y. Differentiating between bacterial and viral infection by measuring both C-reactive protein and 2 -5 -oligoadenylate synthetase as inflammatory markers. J Infect Chemother. 2002;8(1):76–80.

439. Nakayama T, Sonoda S, Urano T, Yamada T, Okada M. Monitoring both serum amyloid protein A and C-reactive protein as inflammatory markers in infectious diseases. Clin Chem. 1993;39(2):293–7.

440. Jenabian MA, El-Far M, Vyboh K, Kema I, Costiniuk CT, Thomas R, et al. Immunosuppressive tryptophan catabolism and gut mucosal dysfunction following early HIV infection. J Infect Dis. 2015;212:355–66.

441. Mehraj V, Routy J. Tryptophan Catabolism in Chronic Viral Infections : Handling Uninvited Guests. Int J Tryptophan Res. 2015;8:41–8.

442. Ishida, H. Kato, T. Takehana K. Valine, the branched-chain amino acid,

suppresses hepatitis C virus RNA replication but promotes infectious particle formation. Biochem Biophys Res Commun. 2013;437(1):127–33.

443. Klassen P, Fürst P, Schulz C, Mazariegos M, Solomons NW. Plasma free amino acid concentrations in healthy Guatemalan adults and in patients with classic dengue 1 , 2. Am J Clin Nutr. 2001;73(3):647–52.
444. Zangerle R, Kurz K, Neurauter G, Kitchen M, Sarcletti M, Fuchs D. Increased blood phenylalanine to tyrosine ratio in HIV-1 infection and correction following effective antiretroviral therapy. Brain Behav Immun. Elsevier Inc.; 2010;24(3):403–8.
445. Yang B, Wang X, Ren X. Amino acid metabolism related to immune tolerance by MDSCs. Int Rev Immunol. 2012;31(November):177–83.
446. Li P, Yin Y-L, Li D, Kim SW, Wu G. Amino acids and immune function. Br J Nutr. 2007;98:237–52.
447. Cragg MS, Claude Chan HT, Fox MD, Tutt A, Smith A, Oscier DG, et al. The alternative transcript of CD79b is overexpressed in B-CLL and inhibits signaling for apoptosis. Blood. 2002;100(9):3068–76.
448. Clark MR, Campbell KS, Kazlauska A, Johnson SA, Hertz M, Potter TA, et al. The B cell antigen receptor complex: association of Ig-a and Ig-b with distinct cytoplasmic effectors. Sci. 1992;258(October):123–6.
449. Königsberger S, Prodöhl J, Stegner D, Weis V, Andreas M, Stehling M, et al. Altered BCR signalling quality predisposes to autoimmune disease and a pre-diabetic state. EMBO J. 2012;31(15):3363–74.
450. Hardy IR, Anceriz N, Rousseau F, Seefeldt MB, Hatterer E, Irla M, et al. Anti-CD79 antibody induces B cell anergy that protects against autoimmunity. J Immunol. 2014;192(4):1641–50.
451. Li Y, Chen F, Putt M, Koo YK, Madaio M, Cambier JC, et al. B cell depletion with anti-CD79 mAbs ameliorates autoimmune disease in MRL/lpr mice. J Immunol. 2008;181(5):2961–72.
452. Brühl H, Cihak J, Talke Y, Gomez MR, Hermann F, Goebel N, et al. B-cell inhibition by cross-linking CD79b is superior to B-cell depletion with anti-CD20 antibodies in treating murine collagen-induced arthritis. Eur J Immunol. 2015;45(3):705–15.
453. Shattil SJ, Newman PJ. Integrins: dynamic scaffolds for adhesion and signaling in platelets. Blood. 2004;104(6):1606–15.

454. Shattil SJ, Kashiwagi H, Pampori N. Integrin signaling: the platelet paradigm. Blood. 1998;91(8):2645–57.
455. van der Stoep M, Korporaal SJA, Van Eck M. High-density lipoprotein as a modulator of platelet and coagulation responses. Cardiovasc Res. 2014;103(3):362–71.
456. van Willigen G, Goiter G, Akkerman JN. LDLs increase the exposure of fibrinogen binding sites on platelets and secretion of dense granules. Arter Thromb. 1993;14(1):41–6.
457. Surya II, Gorter G, Mommersteeg M, Akkerman JW. Enhancement of platelet functions by low density lipoproteins. Biochim Biophys Acta. 1992;1165(1):19–26.
458. Pedreño J, de Castellarnau C, Cullaré C, Sánchez J, Gómez-Gerique J, Ordóñez-Llanos J, et al. LDL binding sites on platelets differ from the "classical" receptor of nucleated cells. Arterioscler Thromb Vasc Biol. 1992;12(11):1353–62.
459. Koller E, Koller F, Binder BR. Purification and identification of the lipoprotein-binding proteins from human blood platelet membrane. J Biol Chem. 1989;264(21):12412–8.
460. Bisoendial RJ, Kastelein JJ, Levels JH, Zwaginga JJ, van den Bogaard, B Reitsma PH, Meijers JC, et al. Activation of inflammation and coagulation after infusion of C-reactive protein in humans. Circ Res. 2005;96(7):714–6.
461. Danenberg HD, Kantak N, Grad E, Swaminathan R V, Lotan C, Edelman ER. C-reactive protein promotes monocyte-platelet aggregation: an additional link to the inflammatory-thrombotic intricacy. Eur J Haematol. 2007;78(16):246–52.
462. Yaron G, Brill A, Dashevsky O, Yosef-Levi IM, Grad E, Danenberg HD, et al. C-reactive protein promotes platelet adhesion to endothelial cells: a potential pathway in atherothrombosis. Br J Haematol. 2006;134(4):426–31.
463. Sawada A, Takihara Y, Kim JY, Matsuda-Hashii Y, Tokimasa S, Fujisaki H, et al. A congenital mutation of the novel gene LRRC8 causes agammaglobulinemia in humans. J Clin Invest. 2003;112(11):1707–13.
464. Kumar L, Chou J, Yee CSK, Borzutzky A, Vollmann EH, von Andrian UH, et al. Leucine-rich repeat containing 8A (LRRC8A) is essential for T lymphocyte development and function. J Exp Med. 2014;211(5):929–42.

465. Higazi AA, Nassar T, Ganz T, Rader DJ, Udassin R, Bdeir K, et al. The alpha-defensins stimulate proteoglycan-dependent catabolism of low-density lipoprotein by vascular cells: a new class of inflammatory apolipoprotein and a possible contributor to atherogenesis. Blood. 2000;96(4):1393–8.
466. Abu-Fanne R, Maraga E, Abd-Elrahman I, Hankin A, Blum G, Abdeen S, et al. α-Defensins induce a post-translational modification of LDL that promotes atherosclerosis at normal levels of plasma cholesterol. J Biol Chem. 2015;3966.
467. Hotamisligil GS, Arner P, Caro JF, Atkinson RL, Spiegelman BM. Increased adipose tissue expression of tumor necrosis factor-alpha in human obesity and insulin resistance. J Clin Invest. 1995;95(5):2409–2415.
468. Kern PA, Saghizadeh M, Ong JM, Bosch RJ, Deem R, Simsolo RB. The expression of tumor necrosis factor in human adipose tissue. Regulation by obesity, weight loss, and relationship to lipoprotein lipase. J Clin Invest. 1995;95(5):2111–9.
469. Solomon DH, Massarotti E, Garg R, Liu J, Canning C, Schneeweiss S. Association between disease-modifying antirheumatic drugs and diabetes risk in patients with rheumatoid arthritis and psoriasis. JAMA. 2011;305(24):2525–31.
470. Burska AN, Sakthiswary R, Sattar N. Effects of tumour necrosis factor antagonists on insulin sensitivity/resistance in rheumatoid arthritis: A systematic review and meta-analysis. PLoS One. 2015;10(6):e0128889.
471. Kothari V, Galdo JA, Mathews ST. Hypoglycemic agents and potential anti-inflammatory activity. J Inflamm Res. 2016;9:27–38.
472. Lancaster GI, Febbraio MA. The immunomodulating role of exercise in metabolic disease. Trends Immunol. 2014;35(6):262–9.
473. Scheen AJ, Esser N, Paquot N. Antidiabetic agents: Potential anti-inflammatory activity beyond glucose control. Diabetes Metab. 2015;41(3):183–94.
474. Liston A, Carr EJ, Linterman MA. Shaping Variation in the Human Immune System. Trends Immunol. 2016;37(10):637–46.
475. Oral EA, Reilly SM, Gomez A V, Meral R, Butz L, Ajluni N, et al. Inhibition of IKKε and TBK1 improves glucose control in a subset of patients with Type 2 diabetes. Cell Metab. 2017;26(1):157–170.e7.
476. Rafiq S, Stevens K, Hurst AJ, Murray A, Henley W, Weedon MN, et al.

Common genetic variation in the gene encoding interleukin-1-receptor antagonist (IL-1RA) is associated with altered circulating IL-1RA levels. Genes Immun. 2007;8(4):344–51.

477. He M, Cornelis MC, Kraft P, Van Dam RM, Sun Q, Laurie CC, et al. Genome-wide association study identifies variants at the IL18-BCO2 locus associated with interleukin-18 levels. Arterioscler Thromb Vasc Biol. 2010;30(4):885–90.
478. Ferreira MA, Purcell SM. A multivariate test of association. Bioinformatics. 2009;25(1):132–3.
479. Yang Q, Wang Y. Methods for analyzing multivariate phenotypes in genetic association studies. J Probab Stat. 2012;2012:652569.
480. Kim S, Xing EP. Statistical estimation of correlated genome associations to a quantitative trait network. PLoS Genet. 2009;5(8):e1000587.
481. van der Sluis S, Posthuma D, Dolan C V. TATES: efficient multivariate genotype-phenotype analysis for genome-wide association studies. PLoS Genet. 2013;9(1):e1003235.
482. O'Reilly PF, Hoggart CJ, Pomyen Y, Calboli FCF, Elliott P, Jarvelin MR, et al. MultiPhen: joint model of multiple phenotypes can increase discovery in GWAS. PLoS One. 2012;7(5):e34861.
483. Zhou X, Stephens M. Efficient multivariate linear mixed model algorithms for genome-wide association studies. Nat Methods. 2014;11(4):407–9.
484. Marttinen P, Pirinen M, Sarin AP, Gillberg J, Kettunen J, Surakka I, et al. Assessing multivariate gene-metabolome associations with rare variants using Bayesian reduced rank regression. Bioinformatics. 2014;30(14):2026–34.
485. Inouye M, Ripatti S, Kettunen J, Lyytikäinen LP, Oksala N, Laurila PP, et al. Novel loci for metabolic networks and multi-tissue expression studies reveal genes for atherosclerosis. PLoS Genet. 2012;8(8):e1002907.
486. Borodulin K, Vartiainen E, Peltonen M, Jousilahti P, Juolevi A, Laatikainen T, et al. Forty-year trends in cardiovascular risk factors in Finland Katja. Int J Epidemiol. 2015;39:1–8.
487. Josse J, Husson F. missMDA: A package for handling missing values in multivariate data analysis. J Stat Softw. 2016;70(1):1–31.
488. Whitcomb BW, Schisterman EF. Assays with lower detection limits: Implications for epidemiological investigations. Paediatr Perinat Epidemiol. 2008;22(6):597–602.

489. Aulchenko YS, Ripke S, Isaacs A, van Duijn CM. GenABEL: An R library for genome-wide association analysis. Bioinformatics. 2007;23(10):1294–6.
490. Whitlock MC. Combining probability from independent tests: the weighted Z-method is superior to Fisher's approach. J Evol Biol. 2005;18(5):1368–73.
491. Zaykin D V. Optimally weighted Z-test is a powerful method for combining probabilities in meta-analysis. J Evol Biol. 2011;24(8):1836–41.
492. Zhu J, Paul W. Heterogeneity and plasticity of T helper cells. Cell Res. 2010;20(1):4–12.
493. Dong C. TH17 cells in development: an updated view of their molecular identity and genetic programming. Nat Rev Immunol. 2008;8(April):337–48.
494. Ahola-Olli A, Würtz P, Havulinna AS, Aalto K, Pitkänen N, Lehtimäki T, et al. Genome-wide association study identifies 17 new loci influencing concentrations of circulating cytokines and growth factors. Am J Hum Genet. 2017;100(1):40–50.
495. Choi SH, Ruggiero D, Sorice R, Song C, Nutile T, Vernon S, et al. Six novel loci associated with circulating VEGF levels identified by a meta-analysis of genome-wide association studies. PLoS Genet. 2016;12(2):e1005874.
496. Bouton M, Boulaftali Y, Richard B, Michel J. Emerging role of serpinE2 / protease nexin-1 in hemostasis and vascular biology. Blood. 2012;119(11):2452–7.
497. Tufarelli C, Frischauf AM, Hardison R, Flint J, Higgs DR. Characterization of a widely expressed gene (LUC7-LIKE; LUC7L) defining the centromeric boundary of the human alpha-globin domain. Genomics. 2001;71(3):307–14.
498. Li Y, Oosting M, Deelen P, Ricaño-Ponce I, Smeekens S, Jaeger M, et al. Inter-individual variability and genetic influences on cytokine responses to bacteria and fungi. Nat Med. 2016;22(8):952–60.
499. Couper KN, Blount DG, Riley EM. IL-10: the master regulator of immunity to infection. J Immunol. 2008;180(9):5771–7.
500. Oriss TB, McCarthy SA, Morel BF, Campana MA, Morel PA. Crossregulation between T helper cell (Th)1 and Th2: inhibition of Th2 proliferation by IFN-gamma involves interference with IL-1. J Immunol. 1997;158(8):3666–72.
501. Baranovski BM, Freixo-Lima GS, Lewis EC, Rider P. T Helper Subsets, Peripheral Plasticity, and the Acute Phase Protein, α1-Antitrypsin. Biomed Res Int. 2015;2015:184574.

502. Muranski P, Restifo NP. Essentials of Th17 cell commitment and plasticity. Blood. 2013;121(13):402–14.
503. Li Y, Oosting M, Smeekens SP, Jaeger M, Aguirre-Gamboa R, Le KT, et al. A functional genomics approach to understand variation in cytokine production in humans. Cell. 2016;167(4):1099–110.
504. Solovieff N, Cotsapas C, Lee PH, Purcell SM, Smoller JW. Pleiotropy in complex traits: challenges and strategies. Nat Rev Genet. 2013;14(7):483–95.
505. Jiang C, Zeng ZB. Multiple trait analysis of genetic mapping for quantitative trait loci. Genetics. 1995;140(3):1111–27.
506. Galesloot TE, van Steen K, Kiemeney LALM, Janss LL, Vermeulen SH. A comparison of multivariate genome-wide association methods. PLoS One. 2014;9(4):e95923.
507. Santoro A, Conde J, Scotece M, Abella V, Lois A, Lopez V, et al. SERPINE2 inhibits IL-1α-induced MMP-13 expression in human chondrocytes: Involvement of ERK/NF-κB/AP-1 pathways. PLoS One. 2015;10(8):e0135979.
508. Vaughan PJ, Cunningham DD. Regulation of protease nexin-1 synthesis and secretion in cultured brain cells by injury-related factors. J Biol Chem. 1993;268(5):3720–7.
509. Guttridge DC, Lau AL, Cunningham DD. Protease nexin-1, a thrombin inhibitor, is regulated by interleukin-1 and dexamethasone in normal human fibroblasts. J Biol Chem. 1993;268(25):18966–74.
510. Li X, Zhao D, Guo Z, Li T, Qili M, Xu B, et al. Overexpression of SerpinE2/protease nexin-1 Contribute to Pathological Cardiac Fibrosis via increasing Collagen Deposition. Sci Rep. 2016;6:37635.
511. Wu X, Liu W, Duan Z, Gao Y, Li S, Wang K, et al. The involvement of protease Nnexin-1 (PN1) in the pathogenesis of intervertebral disc (IVD) degeneration. Sci Rep. 2016;6:30563.
512. Demeo DL, Mariani TJ, Lange C, Srisuma S, Litonjua AA, Celedon JC, et al. The SERPINE2 gene is associated with chronic obstructive pulmonary disease. Am J Hum Genet. 2006;78(2):253–64.
513. Stanley ER, Violeta C. CSF-1 Receptor Signaling in Myeloid Cells. Cold Spring Harb Perspect Biol. 2014;6(6):a021857.
514. Kamdar SJ, Fuller JA, Nishikawa S, Evans R. Priming of mouse macrophages

with the macrophage colony-stimulating factor (CSF-1) induces a variety of pathways that regulate expression of the interleukin 6 (Il6) and granulocyte-macrophage colony-stimulating factor (Csfgm) genes. Exp Cell Res. 1997;235(1):108–16.

515. Roberts, W M, Look AT, Roussel MF, Sherr CJ. Tandem linkage of human CSF-1 receptor (c-fms) and PDGF receptor genes. Cell. 1988;55(4):655–61.
516. Magoulas C, Fried M. The Surf-6 gene of the mouse surfeit locus encodes a novel nucleolar protein. DNA Cell Biol. 1996;15(4):305–16.
517. McLachlan S, Giambartolomei C, White J, Charoen P, Wong A, Finan C, et al. Replication and characterization of association between ABO SNPs and red blood cell traits by meta-analysis in Europeans. PLoS One. 2016;11(6):e0156914.
518. Nikpay M, Goel A, Won HH, Hall LM, Willenborg C, Kanoni S, et al. A comprehensive 1,000 Genomes-based genome-wide association meta-analysis of coronary artery disease. Nat Genet. 2015;47(10):1121–30.
519. Moyerbrailean GA, Kalita CA, Harvey CT, Wen X, Luca F, Pique-Regi R. Which genetics variants in DNase-Seq footprints are more likely to alter binding? PLoS Genet. 2016;12(2):e1005875.
520. Grundberg E, Small KS, Hedman ÅK, Nica AC, Buil A, Keildson S, et al. Mapping cis- and trans-regulatory effects across multiple tissues in twins. Nat Genet. 2012;44(10):1084–9.
521. Breen EC, Reynolds SM, Cox C, Jacobson LP, Magpantay L, Mulder CB, et al. Multisite comparison of high-sensitivity multiplex cytokine assays. Clin Vaccine Immunol. 2011;18(8):1229–42.
522. Browne RW, Kantarci A, LaMonte MJ, Andrews CA, Hovey KM, Falkner KL, et al. Performance of multiplex cytokine assays in serum and saliva among community-dwelling postmenopausal women. PLoS One. 2013;8(4):e59498.
523. de Jager W, Bourcier K, Rijkers GT, Prakken BJ, Seyfert-Margolis V. Prerequisites for cytokine measurements in clinical trials with multiplex immunoassays. BMC Immunol. 2009;10:52.
524. Barnes PJ. The cytokine network in asthma and chronic obstructive pulmonary disease. J Clin Invest. 2008;118(11):3546–3556.
525. Germain M, Chasman D, de Haan H, Tang W, Lindström S, Weng LC, et al. Meta-analysis of 65,734 individuals identifies TSPAN15 and SLC44A2 as two

susceptibility loci for venous thromboembolism. Am J Hum Genet. 2015;96(4):532–42.

526. Du T, Tan Z. Relationship between deep venous thrombosis and inflammatory cytokines in postoperative patients with malignant abdominal tumors. Braz J Med Biol Res. 2014;47(11):1003–1007.
527. Clarke JM, Hurwitz HI. Understanding and targeting resistance to anti-angiogenic therapies. J Gastrointest Oncol. 2013;4(3):253–63.
528. Hagstrom SA, Ying GS, Maguire MG, Martin DF, CATT Research Group, Gibson J, Lotery A, et al. VEGFR2 gene polymorphisms and response to anti-vascular endothelial growth Factor therapy in age-related macular degeneration. Ophthalmology. 2015;122(8):1563–8.
529. Anderson CA, Boucher G, Lees CW, Franke A, D'Amato M, Taylor KD, et al. Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47. Nat Genet. 2011;43(3):246–52.
530. Hu G, Chen J. A genome-wide regulatory network identifies key transcription factors for memory $CD8^+$ T-cell development. Nat Commun. 2013;4:2830.
531. Sallusto F, Geginat J, Lanzavecchia A. Central memory and effector memory T cell subsets: function, generation, and maintenance. Annu Rev Immunol. 2004;22(1):745–63.
532. Masopust D, Vezys V, Wherry EJ, Barber DL, Ahmed R. Cutting edge: gut microenvironment promotes differentiation of a unique memory CD8 T cell population. J Immunol. 2006;176(4):2079–83.
533. Wakim LM, Woodward-Davis A, Liu R, Hu Y, Villadangos J, Smyth G, et al. The molecular signature of tissue resident memory CD8 T cells isolated from the brain. J Immunol. 2019;189(7):3462–71.
534. Heath WR, Carbone FR. The skin-resident and migratory immune system in steady state and memory: innate lymphocytes, dendritic cells and T cells. Nat Immunol. 2013;14(10):978–85.
535. Gebhardt T, Whitney PG, Zaid A, Mackay LK, Brooks AG, Heath WR, et al. Different patterns of peripheral migration by memory CD4+ and CD8+ T cells. Nature. 2011;477(7363):216–9.
536. Masopust D, Choo D, Vezys V, Wherry EJ, Duraiswamy J, Akondy R, et al. Dynamic T cell migration program provides resident memory within intestinal epithelium. J Exp Med. 2010;207(3):553–64.

537. Boyman O, Hefti HP, Conrad C, Nickoloff BJ, Suter M, Nestle FO. Spontaneous development of psoriasis in a new animal model shows an essential role for resident T cells and tumor necrosis factor-alpha. J Exp Med. 2004;199(5):731–6.

538. Conrad C, Boyman O, Tonel G, Tun-Kyi A, Laggner U, de Fougerolles, A Kotelianski V, et al. Alpha1beta1 integrin is crucial for accumulation of epidermal T cells and the development of psoriasis. Nat Med. 2007;13(7):836–42.

539. Boyman O, Conrad C, Tonel G, Gilliet M, Nestle FO. The pathogenic role of tissue-resident immune cells in psoriasis. Trends Immunol. 2007;28(2):51–7.

540. Zhu J, Koelle DM, Cao J, Vazquez J, Huang ML, Hladik F, et al. Virus-specific CD8+ T cells accumulate near sensory nerve endings in genital skin during subclinical HSV-2 reactivation. J Exp Med. 2007;204(3):595–603.

541. Zhu J, Peng T, Johnston C, Phasouk K, Kask AS, Klock A, et al. Immune surveillance by CD8αα+ skin-resident T cells in human herpes virus infection. Nature. 2013;497:494–7.

542. Schiffer JT, Abu-Raddad L, Mark KE, Zhu J, Selke S, Koelle DM, et al. Mucosal host immune response predicts the severity and duration of herpes simplex virus-2 genital tract shedding episodes. Proc Natl Acad Sci U S A. 2010;107(44):18973–8.

543. Turner DL, Bickham KL, Thome JJ, Kim CY, D'Ovidio F, Wherry EJ, et al. Lung Niches for the Generation and Maintenance of Tissue-resident Memory T cells. Mucosal Immunol. 2014;7(3):501–10.

544. Piet B, de Bree GJ, Smids-Dierdorp BS, van der Loos, C M Remmerswaal EB, von der Thüsen JH, van Haarst JM, et al. $CD8^+$ T cells with an intraepithelial phenotype upregulate cytotoxic function upon influenza infection in human lung. J Clin Invest. 2011;121(6):2254–63.

545. Li J, Olshansky M, Carbone FR, Ma JZ. Transcriptional analysis of T cells resident in human skin. PLoS One. 2016;11(1):e0148351.

546. Shiohara T, Mizukawa Y, Teraki Y. Pathophysiology of fixed drug eruption: the role of skin-resident T cells. Curr Opin Allergy Clin Immunol. 2002;2(4):317–23.

547. Teraki Y, Shiohara T. IFN-gamma-producing effector CD8+ T cells and IL-10-producing regulatory CD4+ T cells in fixed drug eruption. J Allergy Clin

Immunol. 2003;112(3):609–15.

548. Watanabe R, Gehad A, Yang C, Scott LL, Teague JE, Schlapbach C, et al. Human skin is protected by four functionally and phenotypically discrete populations of resident and recirculating memory T cells. Sci Transl Med. 2015;7(279):279ra39.

549. Clark RA, Watanabe R, Teague JE, Schlapbach C, Tawa, M C, Adams N, et al. Skin effector memory T cells do not recirculate and provide immune protection in alemtuzumab-treated CTCL patients. Sci Transl Med. 2012;4(117):117ra7.

550. Ariotti S, Beltman JB, Chodaczek G, Hoekstra ME, van Beek AE, Gomez-Eerland R, et al. Tissue-resident memory CD8+ T cells continuously patrol skin epithelia to quickly recognize local antigen. Proc Natl Acad Sci U S A. 2012;109(48):19739–44.

551. Mackay LK, Stock AT, Ma JZ, Jones CM, Kent SJ, Mueller SN, et al. Long-lived epithelial immunity by tissue-resident memory T (TRM) cells in the absence of persisting local antigen presentation. Proc Natl Acad Sci U S A. 2012;109(18):7037–42.

552. Zens KD, Chen JK, Farber DL. Vaccine-generated lung tissue-resident memory T cells provide heterosubtypic protection to influenza infection. JCI Insight. 2016;1(10):e85832.

553. Gilchuk P, Hill TM, Guy C, McMaster SR, Boyd KL, Rabacal WA, et al. A distinct lung-interstitium-resident memory CD8(+) T cell subset confers enhanced protection to lower respiratory tract infection. Cell Rep. 2106;16(7):1800–9.

554. Gebhardt T, Mueller SN, Heath WR, Carbone FR. Peripheral tissue surveillance and residency by memory T cells. Trends Immunol. 2013;34(1):27–32.

555. Gaide O, Emerson RO, Jiang X, Gulati N, Nizza S, Desmarais C, et al. Common clonal origin of central and resident memory T cells following skin immunization. Nat Med. 2015;21(6):647–53.

556. Casey KA, Fraser KA, Schenkel JM, Moran A, Abt MC, Beura LK, et al. Antigen-independent differentiation and maintenance of effector-like resident memory T cells in tissues. Immunology. 2012;188(10):4866–75.

557. Hofmann M, Oschowitzer A, Kurzhals SR, Krüger CC, Pircher H. Thymus-resident memory CD8+ T cells mediate local immunity. Eur J Immunol.

2013;43(9):2295–304.

558. Wakim LM, Gupta N, Mintern JD, Villadangos JA. Enhanced survival of lung tissue-resident memory $CD8^+$ T cells during infection with influenza virus due to selective expression of IFITM3. Nat Immunol. 2013;14(3):238–45.

559. Bergsbaken T, Bevan MJ. Proinflammatory microenvironments within the intestine regulate the differentiation of tissue-resident $CD8^+$ T cells responding to infection. Nat Immunol. 2015;16(4):406–14.

560. Schenkel JM, Fraser KA, Masopust D. Cutting edge: resident memory CD8 T cells occupy frontline niches in secondary lymphoid organs. J Immunol. 2014;192(7):2961–4.

561. Woon HG, Braun A 2, Li J, Smith C, Edwards J, Sierro F, et al. Compartmentalization of total and virus-specific tissue-resident memory CD8+ T cells in human lymphoid organs. PLoS Pathog. 2016;12(8):e1005799.

562. Steinert EM, Schenkel JM, Fraser KA, Beura LK, Manlove LS, Igyártó BZ, et al. Quantifying memory CD8 T cells reveals regionalization of Iimmunosurveillance. Cell. 2015;161(4):737–49.

563. Cepek KL, Shaw SK, Parker CM, Russell GJ, Morrow JS, Rimm DL, et al. Adhesion between epithelial cells and T lymphocytes mediated by E-cadherin and the alpha E beta 7 integrin. Nature. 1994;372(6502):190–3.

564. Higgins JM, Mandlebrot DA, Shaw SK, Russell GJ, Murphy EA, Chen YT, et al. Direct and regulated interaction of integrin alphaEbeta7 with E-cadherin. J Cell Biol. 1998;140(1):197–210.

565. Pauls K, Schön M, Kubitza RC, Homey B, Wiesenborn A, Lehmann P, et al. Role of integrin alphaE(CD103)beta7 for tissue-specific epidermal localization of CD8+ T lymphocytes. J Invest Dermatol. 2001;117(3):569–75.

566. Schön MP, Schön M, Parker CM, Williams IR. Dendritic epidermal T cells (DETC) are diminished in integrin alphaE(CD103)-deficient mice. J Invest Dermatol. 2002;119(1):190–3.

567. Schön MP, Arya A, Murphy EA, Adams CM, Strauch UG, Agace WW, et al. Mucosal T lymphocyte numbers are selectively reduced in integrin alpha E (CD103)-deficient mice. JImmunol. 1999;162(11):6641–9.

568. Feng Y, Wang D, Yuan R, Parker CM, Farber DL, Hadley GA. CD103 expression is required for destruction of pancreatic islet allografts by CD8(+) T cells. J Exp Med. 2002;196(7):877–86.

569. Lee YT, Suarez-Ramirez JE, Wu T, Redman JM, Bouchard K, Hadley GA, et al. Environmental and antigen receptor-derived signals support sustained surveillance of the lungs by pathogen-specific cytotoxic T lymphocytes. J Virol. 2011;85(9):4085–94.
570. Zhang N, Bevan M. Transforming growth factor-b signaling controls the formation and maintenance of gut-resident memory T cells by regulating migration and retention. Immunity. 2013;39(4):687–96.
571. Ziegler SF, Ramsdell F, Hjerrild KA, Armitage RJ, Grabstein KH, Hennen KB, et al. Molecular characterization of the early activation antigen CD69: A type II membrane glycoprotein related to a family of natural killer cell activation antigens. Eur J Immunol. 1993;23(7):1643–8.
572. Kelley J, Walter L, Trowsdale J. Comparative genomics of natural killer cell receptor gene clusters. PLoS Genet. 2005;1(2):129–39.
573. Shiow LR, Rosen DB, Brdicková N, Xu Y, An J, Lanier LL, et al. CD69 acts downstream of interferon-alpha/beta to inhibit S1P1 and lymphocyte egress from lymphoid organs. Nature. 2006;440(7083):540–4.
574. Mackay LK, Braun A, Macleod BL, Collins N, Tebartz C, Bedoui S, et al. Cutting edge: CD69 interference with sphingosine-1-phosphate receptor function regulates peripheral T cell retention. J Immunol. 2015;194(5):2059–63.
575. Bankovich AJ, Shiow LR, Cyster JG. CD69 suppresses sphingosine 1-phosophate receptor-1 (S1P1) function through interaction with membrane helix 4. J Biol Chem. 2010;285(29):22328–37.
576. Takamura S, Yagi H, Hakata Y, Motozono C, McMaster SR, Masumoto T, et al. Specific niches for lung-resident memory CD8+ T cells at the site of tissue regeneration enable CD69-independent. J Exp Med. 2016;213(13):3057–73.
577. Laidlaw BJ, Zhang N, Marshall HD, Staron MM, Guan T, Hu Y, et al. CD4+ T cell help guides formation of CD103+ lung-resident memory CD8+ T cells during influenza viral infection. Immunity. 2014;41(4):633–45.
578. Sowell RT, Rogozinska M, Nelson CE, Vezys V, Marzo AL. Cutting edge: generation of effector cells that localize to mucosal tissues and form resident memory CD8 T cells is controlled by mTOR. J Immunol. 2014;193(5):2067–71.
579. Mackay LK, Wynne-Jones E, Freestone D, Pellicci DG, Mielke LA, Newman

DM, et al. T-box transcription factors combine with the cytokines TGF-β and IL-15 to control tissue-resident memory T cell fate. Immunity. 2015;43(6):1101–11.

580. Hu Y, Lee YT, Kaech SM, Garvy B, Cauley LS. Smad4 promotes differentiation of effector and circulating memory CD8 T cells but is dispensable for tissue-resident memory CD8 T cells. J Immunol. 2015;194(5):2407–14.
581. Schenkel JM, Fraser KA, Casey KA, Beura LK, Pauken KE, Vezys V, et al. IL-15–independent maintenance of tissue-resident and boosted effector memory CD8 T cells. J Immunol. 2016;196(9):3920–6.
582. Mackay LK, Minnich M, Kragten NA, Liao Y, Nota B, Seillet C, et al. Hobit and Blimp1 instruct a universal transcriptional program of tissue residency in lymphocytes. Science (80-). 2016;352(6284):459–63.
583. Hombrink P, Helbig C, Backer RA, Piet B, Oja AE, Stark R, et al. Programs for the persistence, vigilance and control of human CD8+ lung-resident memory T cells. Nat Immunol. 2016;17(12):1467–78.
584. Intlekofer AM, Takemoto N, Wherry EJ, Longworth SA, Northrup JT, Palanivel VR, et al. Effector and memory CD8+ T cell fate coupled by T-bet and eomesodermin. Nat Immunol. 2005;6(12):1236–44.
585. Boddupalli CS, Nair S, Gray SM, Nowyhed HN, Verma R, Gibson JA, et al. ABC transporters and NR4A1 identify a quiescent subset of tissue-resident memory T cells. J Clin Invest. 2016;126(10):3905–16.
586. Kadow S, Jux B, Zahner SP, Wingerath B, Chmill S, Clausen BE, et al. Aryl hydrocarbon receptor is critical for homeostasis of invariant gammadelta T cells in the murine epidermis. J Immunol. 2011;187(6):104–10.
587. Zaid A, Mackay LK, Rahimpour A, Braun A, Veldhoen M, Carbone FR, et al. Persistence of skin-resident memory T cells within an epidermal niche. Proc Natl Acad Sci U S A. 2014;111(14):5307–12.
588. Teijaro JR, Turner D, Pham Q, Wherry EJ, Lefrançois L, Farber DL. Cutting edge: tissue-retentive lung memory CD4 T cells mediate optimal protection to respiratory virus infection. J Immunol. 2011;187(11):5510–4.
589. Collins N, Jiang X, Zaid A, Macleod BL, Li J, Park CO, et al. Skin CD4(+) memory T cells exhibit combined cluster-mediated retention and equilibration with the circulation. Nat Commun. 2016;7(11514).

590. Li J, Olshansky M, Carbone FR, Ma JZ. Transcriptional analysis of T cells resident in human Skin. PLoS One. 2016;11(1):e0148351.
591. Cipolletta D, Feuerer M, Li A, Kamei N, Lee J, Shoelson SE, et al. PPAR-γ is a major driver of the accumulation and phenotype of adipose tissue Treg cells. Nature. 2012;486(7404):549–53.
592. Panduro M, Benoist C, Mathis D. Tissue Tregs. Annu Rev Immunol. 2016;34:609–33.
593. Rossjohn J, Pellicci DG, Patel O, Gapin L, Godfrey DI. Recognition of CD1d-restricted antigens by natural killer T cells. Nat Rev Immunol. 2012;12(12):845–857.
594. Ng SY, Yoshida T, Zhang J, Georgopoulos K. Genome-wide lineage-specific transcriptional networks underscore Ikaros-dependent lymphoid priming in hematopoietic stem cells. Immunity. 2009;30(4):493–507.
595. Elo LL, JJärvenpää H, Oresic M, Lahesmaa R, Aittokallio T. Systematic construction of gene coexpression networks with applications to human T helper cell differentiation process. Bioinformatics. 2007;23(16):2096–2103.
596. Nitsch D, Tranchevent LC, Thienpont B, Thorrez L, Van Esch H, Devriendt K, et al. Network analysis of differential expression for the identification of disease-causing genes. PLoS One. 2009;4(5):e5526.
597. Gill R, Datta S, Datta S. A statistical framework for differential network analysis from microarray data. BMC Bioinformatics. 2010;11:95.
598. Heng TSP, Painter MW. The Immunological Genome Project: networks of gene expression in immune cells. Nat Immunol. 2008;9(10):1091–4.
599. Shay T, Kang J. Immunological Genome Project and systems immunology. Trends Immunol. 2013;34(12):602–9.
600. Gautier L, Cope L, Bolstad BM, Irizarry RA. affy—analysis of Affymetrix GeneChip data at the probe level. Bioinformatics. 2004;20(3):307–15.
601. Galili T. dendextend: An R package for visualizing, adjusting and comparing trees of hierarchical clustering. Bioinformatics. 2015;31(22):3718–20.
602. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. Limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43(7):e47.
603. Yuan Y, Li CT, Windram O. Directed partial correlation: inferring large-scale gene regulatory network through induced topology disruptions. PLoS One.

2011;6(4):e16835.

604. Johansson A, Løset M, Mundal SB, Johnson MP, Freed KA, Fenstad MH, et al. Partial correlation network analyses to detect altered gene interactions in human disease: using preeclampsia as a model. Hum Genet. 2011;129(1):25–34.

605. Zhang HM, Liu T, Liu CJ, Song S, Zhang X, Liu W, et al. AnimalTFDB 2.0: a resource for expression, prediction and functional study of animal transcription factors. Nucleic Acids Res. 2015;43:43:D76–D81.

606. Wang IM, Zhang B, Yang X, Zhu J, Stepaniants S, Zhang C, et al. Systems analysis of eleven rodent disease models reveals an inflammatome signature and key drivers. Mol Syst Biol. 2012;8:594.

607. Wherry EJ, Ahmed R. Memory CD8 T-cell differentiation during viral infection. J Virol. 2004;78(11)(11):5535–45.

608. Schenkel JM, Fraser K a, Vezys V, Masopust D. Sensing and alarm function of resident memory $CD8^+$ T cells. Nat Immunol. 2013;14(5):509–13.

609. Cheuk S, Schlums H, Gallais Sérézal I, Martini E, Chiang SC, Marquardt N, et al. CD49a expression defines tissue-resident CD8+ T cells poised for cytotoxic function in human skin. Immunity. 2017;46(2):287–300.

610. Björkström NK, Ljunggren HG, Michaëlsson J. Emerging insights into natural killer cells in human peripheral tissues. Nat Rev Immunol. 2016;16(5):310–20.

611. Mackay LK, Kallies A. Transcriptional regulation of tissue-resident lymphocytes. Trends Immunol. 2017;38(2):94–103.

612. Feuerer M, Herrero L, Cipolletta D, Naaz A, Wong J, Nayer A, et al. Fat Treg cells: a liaison between the immune and metabolic systems. Nat Med. 2009;15(8):930–9.

613. Vasanthakumar A, Moro K, Xin A, Liao Y, Gloury R, Kawamoto S, et al. The transcriptional regulators IRF4, BATF and IL-33 orchestrate development and maintenance of adipose tissue-resident regulatory T cells. Nat Immunol. 2015;16(3):276–85.

614. Gobbi A, Jurman G. A Null Model for Pearson Coexpression Networks. PLoS One. 2015;10(6):e0128115.

615. Iancu OD, Kawane S, Bottomly D, Searles R, Hitzemann R, S M. Utilizing RNA-Seq data for de novo coexpression network inference. Bioinformatics. 2012;28(12):1592–7.

616. Ballouz S, Verleyen W, Gillis J. Guidance for RNA-seq co-expression network construction and analysis: safety in numbers. Bioinformatics. 2015;31(13):2123–30.
617. Ma C, Zhang N. Transforming growth factor- β signaling is constantly shaping memory T-cell population. Pnas. 2015;2015(35):11013–7.
618. Huang F, Chen YG. Regulation of TGF-β receptor activity. Cell Biosci. 2012;2(1):9.
619. Kilshaw PJ, Murant SJ. Expression and regulation of beta 7(beta p) integrins on mouse lymphocytes: relevance to the mucosal immune system. Eur J Immunol. 1991;21(10):2591–7.
620. Cepek KL, Parker CM, Madara JL, Brenner MB. Integrin alpha E beta 7 mediates adhesion of T lymphocytes to epithelial cells. J Immunol. 1993;150(8):3459–70.
621. Pauls K, Schön M, Kubitza RC, Homey B, Wiesenborn A, Lehmann P, et al. Role of integrin alphaE(CD103)beta7 for tissue-specific epidermal localization of CD8+ T lymphocytes. J Invest Dermatol. 2001;117(3):569–75.
622. Wang D, Yuan R, Feng Y, El-Asady R, Farber DL, Gress RE, et al. Regulation of CD103 expression by CD8+ T cells responding to renal allografts. J Immunol. 2004;172(1):214–21.
623. Kane CJ, Knapp AM, Mansbridge JN, Hanawalt PC. Transforming growth factor-beta 1 localization in normal and psoriatic epidermal keratinocytes in situ. J Cell Physiol. 1990;144(1):144–50.
624. Koyama SY, Podolsky DK. Differential expression of transforming growth factors alpha and beta in rat intestinal epithelial cells. J Clin Investig. 1989;83(5):1768–73.
625. Yang L, Qiu CX, Ludlow A, Ferguson MWJ, Brunner G. Active Transforming Growth Factor-β in Wound Repair. Am J Pathol. 1999;154(1):105–11.
626. El-Asady R, Yuan R, Liu K, Wang D, Gress RE, Lucas PJ, et al. TGF-beta-dependent CD103 expression by CD8(+) T cells promotes selective destruction of the host intestinal epithelium during graft-versus-host disease. J Exp Med. 2005;201(10):1647–57.
627. Ishigame H, Mosaheb MM, Sanjabi S, Flavell RA. Truncated form of TGF β RII, but not its absence, induces memory CD8+ T cell expansion and lymphoproliferative disorder in mice. 2013;190(12):6340–50.

628. Mueller SN, Heath WR, Mclain JD, Carbone FR, Jones CM. Characterization of two TCR transgenic mouse lines specific for herpes simplex virus. Immunol Cell Biol. 2002;80(2):156–63.
629. Andrews S. FastQC: a quality control tool for high throughput sequence data.
630. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 2013;14(4):R36.
631. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9(4):357–9.
632. Schlickum S, Sennefelder H, Friedrich M, Harms G, Lohse MJ, Kilshaw P, et al. Integrin alpha E(CD103)beta 7 influences cellular shape and motility in a ligand-dependent fashion. Blood. 2008;112(3):619–25.
633. Kel JM, Girard-Madoux MJ, Reizis B, Clausen BE. TGF-beta is required to maintain the pool of immature Langerhans cells in the epidermis. J Immunol. 2010;185(6):3248–55.
634. Ohta T, Sugiyama M, Hemmi H, Yamazaki C, Okura S, Sasaki I, et al. Crucial roles of XCR1-expressing dendritic cells and the XCR1-XCL1 chemokine axis in intestinal immune homeostasis. Sci Rep. 2016;6:23505.
635. Yamazaki C, Sugiyama M, Ohta T, Hemmi H, Hamada E, Sasaki I, et al. Critical roles of a dendritic cell subset expressing a chemokine receptor, XCR1. J Immunol. 2103;190(12):6071–82.
636. Nakken B, Alex P, Munthe L, Szekanecz Z, Szodoray P. Immune-regulatory mechanisms in systemic autoimmune and rheumatic diseases. Clin Dev Immunol. 2012;2012:957151.
637. Pan Y, Tian T, Park CO, Lofftus SY, Mei S, Liu X, et al. Survival of tissue-resident memory T cells requires exogenous lipid uptake and metabolism. Nature. 2017;543(7644):252–6.
638. Yosef N, Regev A. Writ large: Genomic dissection of the effect of cellular environment on immune response. Science. 2016;354(6308):64–8.
639. Ravasi T, Suzuki H, Cannistraci C V, Katayama S, Bajic VB, Tan K, et al. An atlas of combinatorial transcriptional regulation in mouse and man. Cell. 2010;140(5):744–52.
640. Odom DT, Dowell RD, Jacobsen ES, Gordon W, Danford TW, MacIsaac KD, et al. Tissue-specific transcriptional regulation has diverged significantly

between human and mouse. Nat Genet. 2007;39(6):730–2.

641. Seok J, Warren HS, Cuenca AG, Mindrinos MN, Baker H V, Xu W, et al. Genomic responses in mouse models poorly mimic human inflammatory diseases. Proc Natl Acad Sci U S A. 2013;110(9):3507–12.

642. Monaco G, van Dam S, Casal Novo Ribeiro JL, Larbi A, de Magalhães JP. A comparison of human and mouse gene co-expression networks reveals conservation and divergence at the tissue, pathway and disease levels. BMC Evol Biol. 2015;15:259.

Appendices

The excel tables mentioned in text can be accessed through figshare using the link below:
https://figshare.com/s/8554ee02c49e6cb54622