Biological and Aetiological Inference from the Statistical Genetic Analyses of Blood Cell Traits

Parsa Akbari

Downing College

University of Cambridge
November 2019

This dissertation is submitted for the degree of Doctor of Philosophy

Abstract

Blood cells are crucial to human physiology, with functions in oxygen transport, infection control, and wound healing. Molecular mechanisms endogenous to blood cells have been implicated in the aetiologies of cancer, infection and inflammatory and immune disorders. The genetic determinants of blood cell function have not been comprehensively characterised, because it is too difficult to perform direct assays of cell function in large population samples. High-throughput flow cytometry can be used to measure functionally relevant phenotypes such as cell granulation, nucleic acid content, and cell size. Many of these phenotypes are important for the diagnosis of diseases such as sepsis, Szary disease, toxic granulation, and myelodysplastic syndromes, or correlate with assessments of cell morphology from blood smear images. Here, I report the results of my genome-wide association study of 63 previously genetically unstudied blood cell flow cytometry phenotypes. I have identified associated variants in loci containing genes coding for established drug targets with known roles in white cell function and immunity. I have colocalised the association signals with blood cell transcriptomic, blood proteomic, and disease risk, identifying possible causal roles for molecular mechanisms endogenous to white cells in the aetiology of a range of immune disorders, including atopic dermatitis, multiple sclerosis and celiac disease. My results have utility in drug design and therapeutic target selection, demonstrated by examples including the replication of the mechanism of action of Daclizumab, a treatment for multiple sclerosis, and evidence for the role of *IL-18R1* in aetiology of celiac disease. Furthermore, mendelian randomisation analyses suggest a causal role for blood cell flow cytometry phenotypes in the aetiology of coronary artery disease, lung cancer, and asthma. In addition to my work on flow cytometry traits, I report a major contribution to the largest ever GWAS meta-analysis of routine clinical haematological phenotypes, including 563,085 individuals. I performed primary and conditional analyses, identifying parsimonious sets of independently associated variants. This is the largest genome-wide association study study of clinical haematological phenotypes to date and identifies 7,122 association signals.

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text. It is not substantially the same as any that I have submitted, or am concurrently submitting, for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or is being concurrently submitted, for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. This dissertation does not exceed the prescribed limit of 60,000 words.

Parsa Akbari
November, 2019

Acknowledgements

I would like to thank my supervisors Dr William Astle, Professor Nicole Soranzo, Dr Adam Butterworth, and Professor Willem Ouwehand for their support and guidance over the previous years. Their mentorship has been fundamental to the writing of this thesis and their guidance has enabled me to develop as an independent researcher. I would also like to thank Dr Alice Mann, Dr Bram Prins, and Dr Emily Kawabata for their advice regarding the organisation and writing of this thesis.

Chapter 2 I would like to acknowledge Dr William Astle for performing the technical and environmental correction of the full blood count parameters in the UK Biobank cohort and both Dr William Astle and Dr Tao Jiang for performing the GWAS study of these phenotypes and performing analysis to identify related individuals between the UKBIOBANK and INTERVAL studies. I would also like to acknowledge Dr Abdou Mousas, Dr Ming Heui, Dr Guillaume Lettre, and colleagues at the Blood cell genetics consortium for their meta-analysis of blood cell phenotypes and subsequent conditional analysis of these results.

Chapter 3 I acknowledge Dr Heather Elding for phasing and imputation of the genotype data from the INTERVAL utilised in my analysis. I would also like to acknowledge Dr William Astle for the development of the *UKBBFBC* R package which I adapted to enable technical and environmental adjustment of Sysmex parameters studied in my analysis. Furthermore, I also acknowledge Dr William Astle and Dr Tao Jiang for their insight regarding conditional analysis and LD clumping.

Chapter 4 I thank Ms Joanna Westmoreland for her illustration of the graphics presented in Figure 4.7.

Chapter 5 I thank Dr James Peters for his insight regarding ANCA vasculitis and the involvement of *PRTN3* and *MPO* in the aetiology of this disease. I would also like to thank Dr Kousik Kundu for collecting summary statistics for immune related disorders studied in my colocalisation.

I would like to acknowledge my brothers Amir and Masoud for their encouragement. Finally, I give special thanks to my parents for their support which began early in my life and continues to this day, without them this thesis would not exist.

Contents

Acronyms

AM additive model.

APC antigen presenting cell.

AS-LYMP antibody synthesising lymphocytes.

BCR B cell receptor.

BCX Blood Cell Genetics Consortium.

BPI bactericidal permeability increasing protein.

CAD coronary artery disease.

CBC complete blood count.

CD cluster of differentiation.

CHD coronary heart disease.

ECP eosinophil cationic protein.

EDTA ethylenediaminetetraacetic acid.

eQTL expression quantitative trait loci.

FBC full blood count.

FDC follicular dendritic cell.

fMLP formyl-methionyl-leucyl-phenylalanin.

FSC forward scatter.

GAM generalised additive model.

GCTA genome-wide complex trait analysis.

GRM genetic relationship matrix.

GWAS genome wide association study.

H-IPF highly fluorescent immature platelet fraction.

HFR high fluorescence reticulocytes.

HGMD human gene mutation database.

HLA human leukocyte antigen.

HPC high performance compute.

HRC haplotype reference consortium.

HSC haematopoietic stem cell.

HTML hypertext markup language.

HWE Hardy-Weinberg equilibrium.

IBD identity by descent.

IG immature granulocyte.

IG# immature granulocyte count.

IGE immunoglobulin E.

INFO information.

IPF immature platelet fraction.

IV instrumental variable.

IVW inverse variance weighted.

JSNP Japanese single-nucleotide polymorphisms.

LD linkage disequilibrium.

LDL low density lipoprotein.

LFR low fluorescence reticulocytes.

LGL large granular lymphocyte.

LMM linear mixed model.

LOO leave one out analysis.

LPS lipopolysaccharides.

MAF minor allele frequency.

MBP major basic protein.

MCV mean corpuscular volume.

MFR medium fluorescence reticulocytes.

MHC major histocompatibility complex.

MPO myeloperoxidase.

MR Mendelian randomisation.

MRI magnetic resonance imaging.

MS multiple sclerosis.

MSCV mean sphered cell volume.

NET neutrophil extracellular trap.

NK natural killer.

NRBC nucleated red blood cells.

PBWT positional Burrows-Wheeler transform.

PC principal components.

PCA principal component analysis.

PP posterior probability.

pQTL protein quantitative trait loci.

QC quality control.

QTL quantitative trait loci.

RBC red blood cells.

RDW red cell distribution width.

RE-LYMP reactive lymphocytes.

SD standard deviation.

SFL side fluorescence.

SNP single nucleotide polymorphism.

SNV single nucleotide variant.

SO sequence ontology.

SSC side scatter.

TCR T cell receptor.

TGN toxic granulation neutrophils.

UTR untranslated region.

VEP variant effect predictor.

VWF von Willebrand factor.

WDF white blood cell differential channel.

WES whole-exome sequencing.

WTCHG Wellcome Trust Centre for Human Genetics.

Chapter 1

Introduction

Cells are a primary unit of biology and cell behaviour such as cell count, protein production and exocytosis, mitosis, and signalling are important biological functions. Underpinning these cell phenotypes are proteins which are encoded by translation from the genome. Changes in cellular behaviour, protein structure, or expression of proteins can lead to downstream and knock-on effects throughout the organism. The motivation for my thesis is to understand the association between variation in the genome and variation in blood cell phenotypes. This analysis can implicate potential genes, transcripts, or proteins in blood cell behaviour and inform further biological experimentation and drug development.

In this chapter I will give an introduction to blood cell types, blood cell function, and the fundamentals of a genome wide association study (GWAS) which can identify genetic variations that influence phenotypes of interest. Following this, I will discuss automated haematology analysers and explain how flow cytometry and electrical impedance may be used to derive blood cell phenotypes from whole blood samples. Finally, I will review previous work in the study of haematological genetics and present my contribution to the field, the first ever GWAS of functionally relevant haematological phenotypes, and the largest ever GWAS of previously studied haematological measurements.

1.1 Haematology and the study of blood cells

Blood cells permeate most tissues and organs in the human body and are implicated in the aetiology of many rare and common diseases. All blood cells originate from haematopoietic stem cells (HSCs), differentiation of HSCs resulting in the generation of new blood cells is termed *haematopoiesis*. Haematopoiesis occurs in the medullary cavity of the bone which contains bone marrow. Bone marrow is semi-solid tissue composed of non-cellular connective tissue and cells such as adipose tissue and haematopoietic cells. HSCs are multipotent meaning they are able to differentiate into multiple specialised cell types and self-renew. Haematopoiesis begins with generation of myeloid or lymphoid progenitor

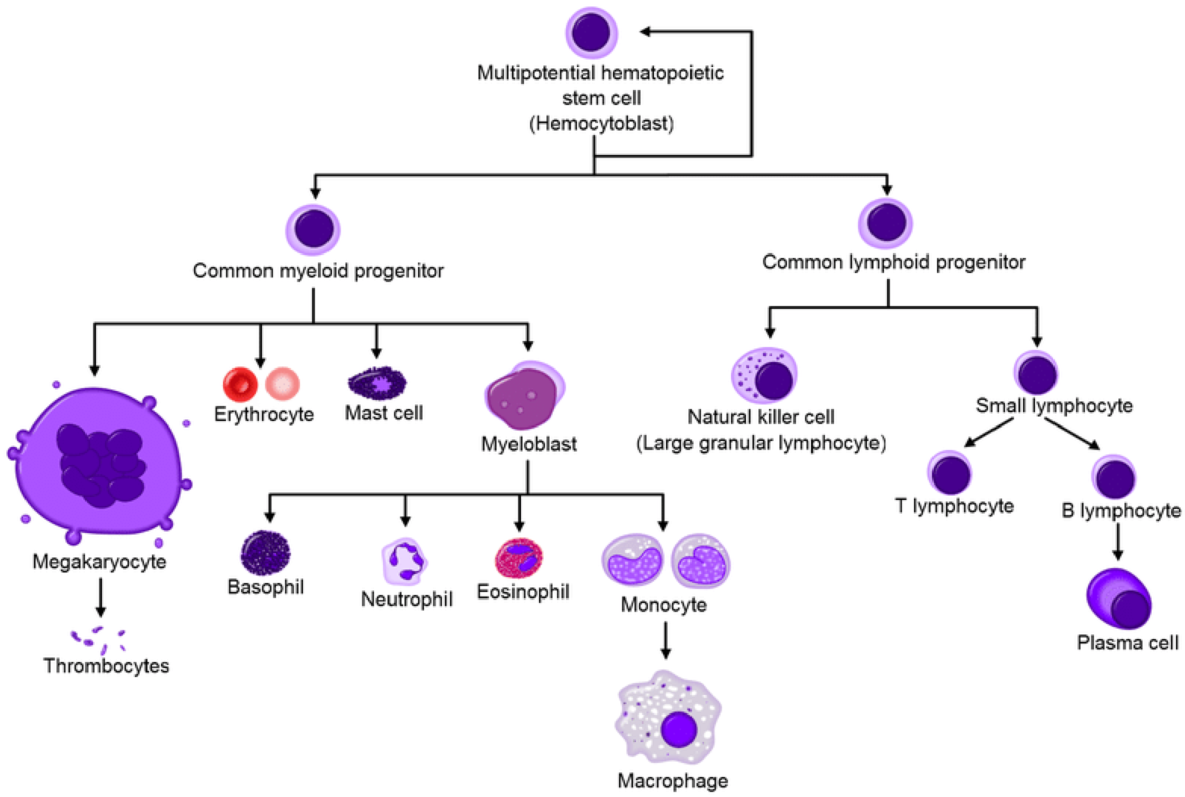
Multipotential hematopoietic stem cell (Hemocytoblast)
Common myeloid progenitor
Common lymphoid progenitor
Megakaryocyte
Thrombocytes
Erythrocyte
Mast cell
Myeloblast
Basophil
Neutrophil
Eosinophil
Monocyte
Macrophage
Natural killer cell (Large granular lymphocyte)
Small lymphocyte
T lymphocyte
B lymphocyte
Plasma cell

Figure 1.1: Haematological cell lineage and differentiation into cell types. Pluripotent stem cells in the bone marrow produce haematopoietic cells which differentiate into progenitors to create increasingly specialised cells. Immature erythrocytes develop into red cells, thrombocytes or platelets are generated from megakaryocytes. White cells include basophils, neutrophils, eosinophils, monocytes, and lymphocytes which includes natural killer cells and T or B lymphocytes (Figure source: [140]).

cells from the division of HSCs. Progenitor cells differentiate into specialised blood cells: platelets, reticulocytes, and white cells which are released into circulation (Fig. 1.1). Cells can develop further following release from the medullary cavity, for example, circulating reticulocytes mature in circulation to become red blood cells.

Blood cells are broadly categorised into platelets, red blood cells, and white blood cells. These categories are based on structural differences observable with microscopy. Unsurprisingly, categorisation based on structure also delineates functional differences between these cell categories:

- Platelets also known as thrombocytes are cytoplasmic fragments of megakaryocyte cells. Being up to 2 - 3 μm in diameter, platelets are smaller in comparison to other blood cells [135]. Platelets are responsible for coagulation preventing blood loss from damaged vessels.

- Red blood cells are the most common blood cell, responsible for oxygen transport through the circulatory system to tissues. Red blood cells do not contain a nucleus, and their cytoplasm is rich in haemoglobin, an oxygen binding molecule. Observed with microscopy these cells appear as discoid shaped distinguishing them from other cell types.

- White blood cells are also known as leukocytes and include neutrophils, monocytes, eosinophils, basophils, and lymphocyte cell types. These cells are distinguished from red blood cells and platelets by the presence of a nucleus. This category includes a broad range of immune cells responsible for clearing infectious agents and pathogens. Staining and microscopy of white blood cell nuclei and granules results in further sub-categorisation (Section 1.1.3).

The phenotypes studied in my analysis are often blood cell type specific, platelets, red blood cells, and each of the five white blood cell types discussed above. These blood cell types were historically delineated with staining and microscopy, in particular, differences in nuclear structure (polynuclear or mononucleuar cells) and presence of granules (granulocyte or agranulocyte cells) (Section 1.1.3). I utilise these categories because high throughput blood cell assay technology based on flow cytometry or electrical impedance can differentiate between the aforementioned categories. However, within this categorisation more specific cell types have been identified, such as T or B lymphocytes. T lymphocytes perform cell-mediated immunity activating phagocytes and releasing cytokines. B lymphocytes perform humoural immunity generating macromolecules such as antibodies. Such sub-categories are not currently easily identified by available high throughput blood cell assay technology. This is a major drawback of high-throughput assay technology for the study of blood cells.

1.1.1 Platelets

Platelets (also known as thrombocytes) are blood cells produced in the bone marrow from cytomplasmic fragments of megakaryocytes. Platelet formation begins with cytoplasmic extensions on megakaryocytes which fragment to form platelet cells, each megakaryocyte can generate 1000 - 5000 platelets [81, p. 316]. Platelets have no nucleus but contain subcellular components such as mitochondria and granules which contribute to the primary function of platelets: to generate a haemostatic plug or thrombus to prevent loss of blood through a perforation in the vessel wall (Fig. 1.2). Dormant platelets circulate in the blood and show a dramatic response when activated following vessel injury. Platelet response to vessel injury begins with adhesion to the site of perforation, followed by activation and aggregation of platelet cells. Finally aggregated platelet cells are bound by a fibrin mesh, this body of aggregated platelets is termed a thrombus (Fig. 1.2). The fibrin mesh is created by catalytic conversion of fibrinogen to fibrin by the enzyme thrombin. Thrombin is activated by the blood coagulation cascade, which is initiated by release of tissue factor or exposure of collagen (described below) and amplified by a positive feedback loop involving the sequential activation of a number of protease enzymes resulting in cleavage of prothombin to thrombin.

Internally, blood vessels are lined by endothelial cells attached to subendothelial collagen which under normal circumstances is not exposed to blood. Attachment of endothelial cells to collagen is maintained by von Willebrand factor (VWF). VWF is a glycoprotein which plays a number of roles in hemostasis, VWF is also found in the granules of platelets and endothelial cells being released during thrombus formation. Normally, endothelial cells provide a non-adhesive surface to platelets. However, if the endothelial layer is damaged or the blood vessel is perforated, collagen fibrils and VWF are exposed to platelets. Platelet cell membranes contain a number of receptors such as GPIb (Glycoprotein Ib) and GPIIb/IIIa which bind to VWF and GPIa which binds to collagen. Numerous platelets bind to long chains of VWF and collagen, localising platelets to the site of rupture, this initial adhesion is the first step of thrombus formation [81, p. 318].

Adhesion also initiates platelet activation, firstly, platelets will undergo a dramatic shape change from smooth discoid cells to spheres with extending filopodia [122, p. 446]. Adhered platelets will also release contents of their α and dense granules [122, p. 443]. Both granule types contain a number of important molecules which promotes thrombus formation. α-granules contain VWF and fibrinogen, VWF promotes platelet adhesion as previously described [122, p. 443]. Similarly, fibrinogen localises pairs of activated platelets by binding to the GPIIb/IIIa receptor on the platelet cell membrane. In addition fibrinogen can be converted to fibrin, a key component for thrombus formation which is described later [122, p. 448]. Dense granules contain molecules such as ADP and serotonin which promote further platelet activation [122, p. 443]. This creates a feedback loop recruiting further platelets which adhere to the growing thrombus and are activated. A crucial step in the positive feedback loop of platelet activation is the generation of thromboxane A2 by activated platelets [122, p. 448]. Thromboxane A2 binds to the thromboxane receptor on nearby platelets promoting activation and further thromboxane A2 production [122, p. 448].

Thrombin is a key component in thrombus formation, thrombin catalyses the conversion of fibrinogen to fibrin. Fibrin molecules crosslink creating a binding mesh which holds aggregated platelets together creating a thrombus. Thrombin is produced by cleavage of prothrombin, this is the last step in the coagulation cascade which is initiated by the intrinsic and extrinsic pathways [122, p. 448]. The intrinsic pathway begins with formation of an activating complex initiated by collagen, the extrinsic pathway is activated by tissue factor, a protein present on subendothelial tissue [130]. In addition to thrombus formation vasoconstriction reduces blood flow through the injured vessel [122, p. 443]. Vasoconstriction is promoted by thromboxane A2 produced catalytically by activated platelets and serotonin which is released from platelet dense granules.

1.1.2 Red Blood Cells

Red blood cells (also known as erythrocytes) form up to 45% of blood volume [81, p. 25] being the most frequently observed cells and appear red under microscopy following application of Wright's stain (Fig 1.3). Red blood cells are described as having a dougnut shape and have structural flexibility allowing them to pass through narrow capillaries which permeate tissues. Up to 10^{12} red blood cells are generated each day through a process called erythropoiesis [81, p. 16]. Differentiation begins with HSC, as with the generation of all blood cells (Fig. 1.1). Red blood cells differentiate in two stages, firstly in the bone marrow leading to generation of reticulocytes which are released and undergo final maturation in circulation. Reticulocytes originate from erythroid precursor cells, which differentiate into pronormoblasts which generate early normoblasts leading to late normoblasts which differentiate into reticulocytes that exit the bone marrow (Fig. 1.4). Given the large numbers of red blood cells which need to be produced, a substantial degree of amplification occurs from the differentiation of a HSC to the generation of a red blood cell. Mitosis of intermediate and late normoblasts increases the number of unipotent stem cells reducing the requirement for division of multipotent HSCs (Fig. 1.4). Unipotent stem cells are those which can differentiate into only one lineage, in contrast to multipotent stem cells which can differentiate into many.

Once reticulocytes are released from bone marrow they will gradually mature into red blood cells. The absence of a nucleus allows extra space in the cytoplasm for additional haemoglobin molecules which enables the primary function of red blood cells which is transport oxygen throughout the organism. Haemoglobin consists of four globular proteins each with a haem iron metalloprotein complex, which can bind oxygen molecules. Oxyhaemoglobin is formed by binding of oxygen to haemoglobin molecules which occurs in pulmonary capillaries of the lungs. When red blood cells flow to the periphery of the organism, a lower oxygen concentration encourages a dissociation of oxygen from haemoglobin and diffusion of oxygen into tissues where oxygen molecules contribute to metabolism. Thus, oxyhaemoglobin is converted to deoxyhaemoglobin. Red blood cells containing deoxyhaemoglobin then circulate back through to the pulmonary capillaries of the lungs where they are once again oxygenated.

1.1.3 White Blood Cells

The primary function of white blood cells (also known as leukocytes) is to clear infection by pathogens. White cells can be categorised based on the presence or absence of granules (granulocytes or agranulocytes), by lineage (myeloid or lymphoid) (Fig. 1.1), and by nuclear

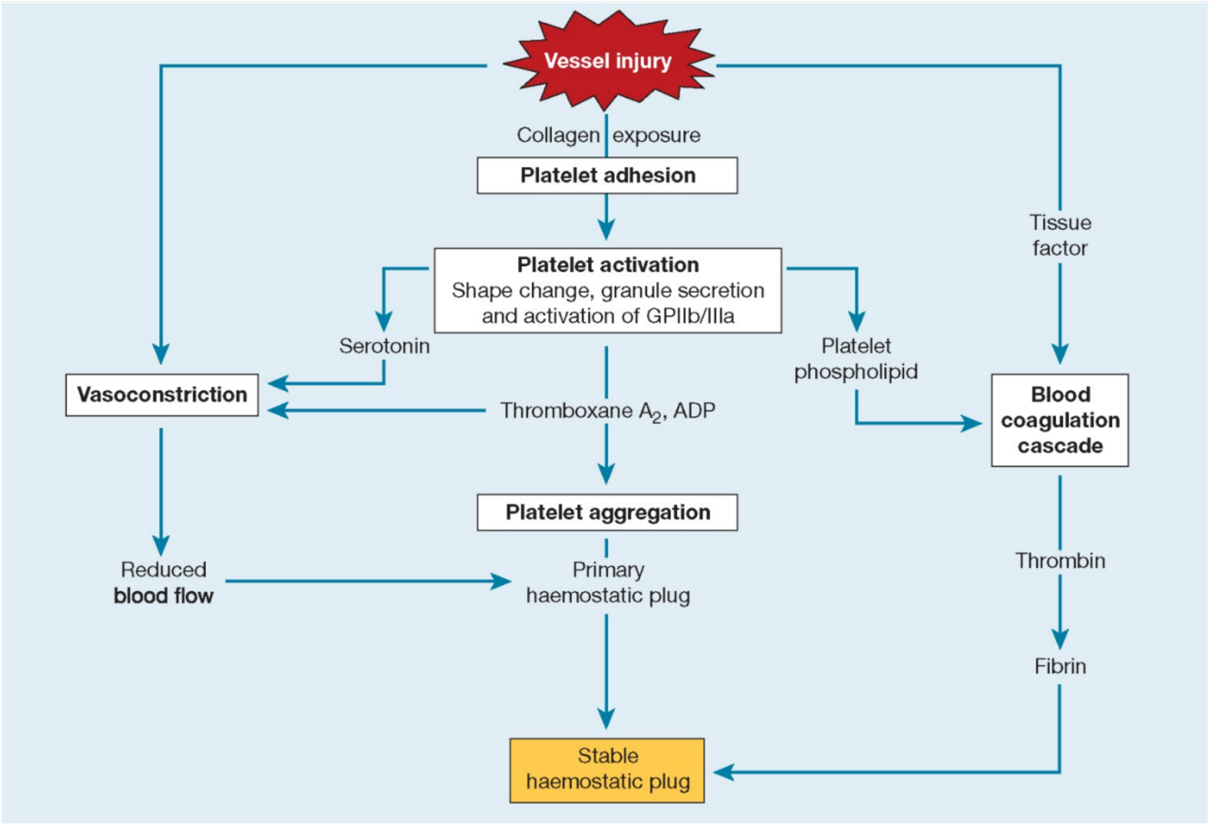

Figure 1.2: Platelet generation of haemostatic plug (thrombus) following vessel injury.

Multiple factors contribute to the generation of a stable haemostatic plug (thrombus). Firstly, a primary haemostatic plug is generated by aggregated platelets. Collagen and VWF exposure promotes platelet adhesion to the site of injury and subsequent activation. Activated platelets release thromboxane A2 recruiting further platelets to the site of injury forming a primary haemostatic plug. The primary haemostatic plug is bound by a fibrin mesh to create a stable haemostatic plug. Fibrin is generated by thrombin produced by the blood coagulation cascade. The blood coagulation cascade is initiated by tissue factor release following vessel injury and platelet phospholipid release. Finally, activated platelets release serotonin resulting in vasoconstriction reducing blood flow through the injured vessel (Figure source [81, p. 315]).

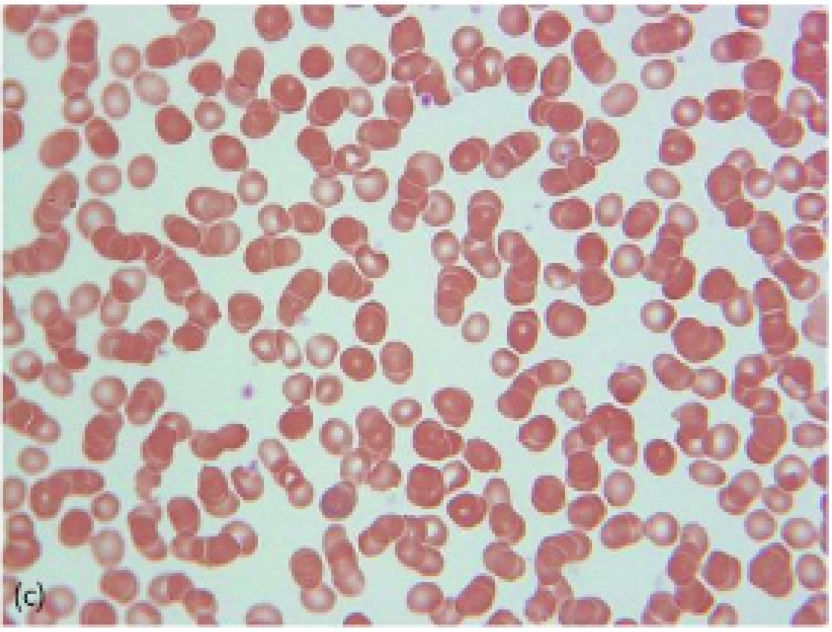

Figure 1.3: Stained red blood cell observed by microscopy. Red blood cells are visibly distinct from other blood cells by their smaller size and doughnut shape, they form the largest proportion of blood by volume and are the most numerous when observed with Wright-Gimesa stain and microscopy. (Figure source [122, p. 26]).

structure (polynuclear or mononuclear). Blood cells were initially studied by extraction of blood samples and microscopy following application of Wright-Giemsa stain, a mixture of red and methylene blue dyes [57]. This leads to the delineation of five white blood cell type categories, listed in order of abundance: neutrophils, lymphocytes, monocytes, eosinophils, and basophils (Table 1.1). These categories are based on differences easily observable with microscopy which also correspond to functional differences. However, this method of classification is limiting. Firstly, the staining of white cells to observe granules results in a agranulocyte classification for monocytes. This is incorrect, as monocytes do contain granules at lower quantities not easily observable under microscopy. Furthermore, as the study of haematology has progressed, it is now clear that functionally important subtypes exist within the previously defined categories. This is especially true for lymphocyte cells which contain a number of functionally heterogeneous subclasses which are discussed later.

Neutrophils

Neutrophils are short lived cells with a lifespan of only 6 - 10 hours in circulation and are part of the innate immune system which is the first to react in response to pathogenic assault [81, p. 110]. Neutrophils are highly abundant forming up to half the population of white blood cells (Table 1.1) being distinguished by their multilobed nucleus and granulated cytoplasm (Fig. 1.5). Neutrophils flowing through the circulatory system are recruited to a site of infection by endothelial cells. This allows neutrophils to leave the blood vessel and

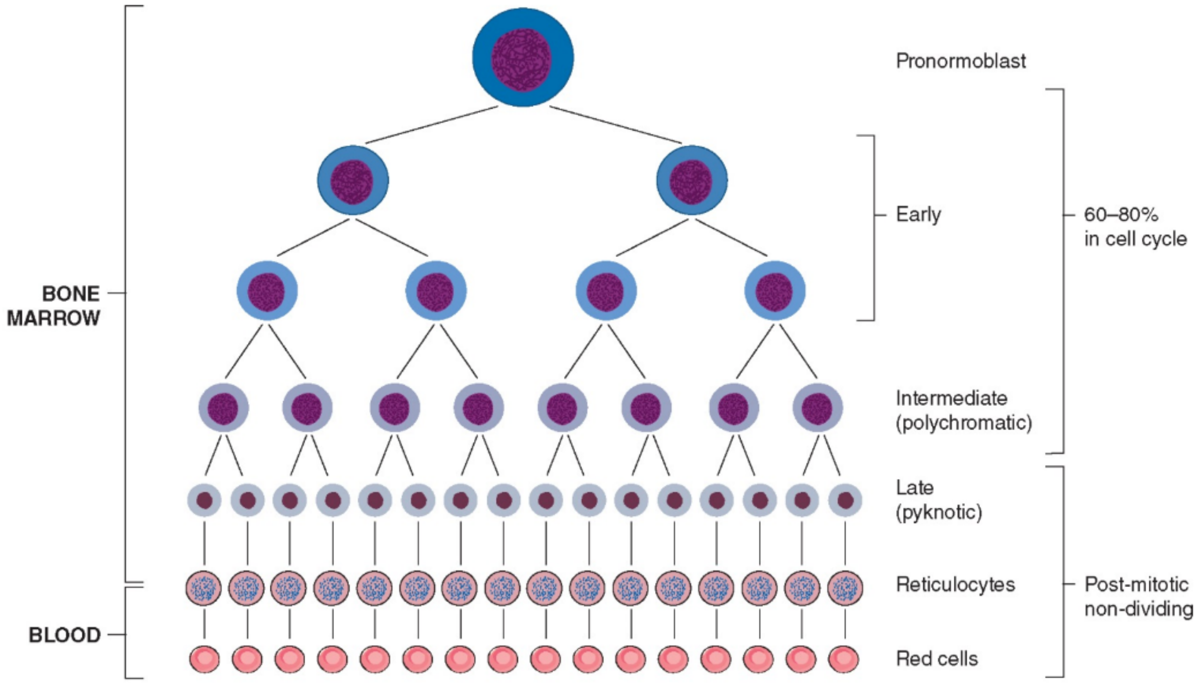

Figure 1.4: Erythropoiesis of red blood cells from pronormoblasts.
Red blood cells originate from the common myeloid progenitor which differentiates into pronormoblast cells, then early, intermediate, and late normoblasts (precursor cells of red blood cells) finally becoming reticulocytes which mature into red blood cells. Unipotent early and intermediate normoblasts can undergo mitosis resulting in a greater number of red cells produced from a single pronormoblast (Figure source [81, p. 17]).

Category	Abundance (Cells/Litre blood)	Progenitor	Nuclear Structure	Granulation
Neutrophils	4.00 - 11.00 x 10^9	Myeloid	Multilobed	Granulocyte
Lymphocytes	1.5 - 3.5 x 10^9	Lymphoid	Mononuclear (round)	Agranulocyte
Monocytes	0.2 - 0.8 x 10^9	Myeloid	Unilobed	Agranulocyte*
Eosinophils	0.04 - 0.4 x 10^9	Myeloid	Bilobed	Granulocyte
Basophils	0.01 - 0.1 x 10^9	Myeloid	Bilobed	Granulocyte

Table 1.1: White blood cell categories, their abundance in circulation and features of categorisation.
Neutrophils are by far the most abundant white cell type in circulation with basophils rarely found in circulation (see table). Cell types are categorised based on progenitor, or structural differences in nuclear structure, and granulation. Myeloid cells differentiate from myeloid progenitors, and lymphoid cells from lymphoid progenitor cells. * Monocyte cells contain granules, however these granules are fine and not easily observable by staining and microscopy - thus leading to the agranulocyte classification. Abundance statistics from [81, p. 109].

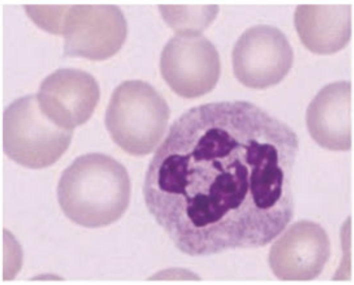

Figure 1.5: Neutrophil cell observed by microscopy following staining. Neutrophils are differentiated by their multilobed nucleus, in this case containing roughly 4 lobes and the presence of granules (pink stain) in the cytoplasm. The cell is surrounded by discoid shaped red blood cells which are highly abundant in blood plasma (Figure source [81, p. 109]).

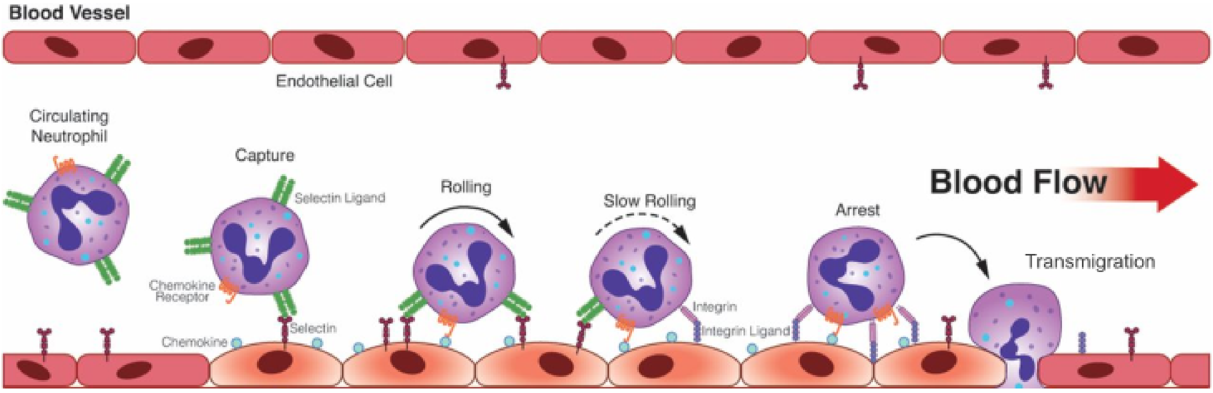

Figure 1.6: Neutrophil cell recruitment to the site of infection Steps in neutrophil recruitment begin with circulating neutrophils adhering to endothelial cells presenting the selectin receptor which binds to the selectin ligand on the neutrophil cell membrane. Following recruitment, neutrophils roll towards the source of chemoattractant molecules, rolling is arrested by binding of integrins leading to eventual extraversion through the endothelial wall (Figure source [40]).

begin migrating to the site of infection. Recruitment begins by interaction of P-selectin ligand on neutrophil cells with P or E-selectin receptors presented by endothelial cells (Fig. 1.6) [99]. Following recruitment, neutrophils will begin to migrate towards the source of chemokines produced at the site of infection. This migration along the concentration gradient of attractant molecules is termed chemotaxis and is performed by rolling along the endothelial wall. Eventually, rolling is arrested by binding of the integrin ligand on neutrophil cells to integrin receptor on endothelial cells. Leading to passing of neutrophils through the endothelial wall towards the site of infection in the surrounding tissue (Fig. 1.6) [185].

Once circulating neutrophils are recruited, adhere, and move to the site of infection they engage and destroy pathogens. This can occur by degranulation, phagocytosis, and generation of neutrophil extracellular traps (NETs). Key to all these responses are neutrophil granules which contain cytotoxic compounds such as defensin peptides and bactericidal permeability increasing protein (BPI). Straightforward degranulation results in the release of granule contents onto pathogenic cells in order to induce cell death. This may occur, for example, by release of defensin peptides which permeabalise the pathogen membrane. Phagocytosis engulfs and digests pathogen cells which have been bound by antibodies. During phagocytosis neutrophils generate cytotoxic reactive oxygen species by a process termed respiratory burst, and granules enable digestion by releasing their contents as the pathogenic cell is engulfed. Finally, generation of NETs entangles pathogens with a fibre of chromatin from neutrophil DNA and serine proteases such as neutrophil elastase and cathepsin G which are released from granules.

Neutrophil granules consist of primary granules, more common secondary granules, and a smaller number of gelatinase granules. The contents of these granules has been studied extensively by separation on a density gradient and proteome profiling with mass spectrometry [144]:

- Primary (azurophilic) granules contain antibacterial compounds such as defensin peptides and serine proteases. Primary granules are distinguished by the presence of myeloperoxidase (MPO) protein.

- Secondary (specific) granules are most numerous and contain further cytotoxic compounds such as lysozyme and lactoferrin.

- Gelatinase granules contain matrix metalloproteinase proteins allowing neutrophils to pass the endothelial wall [123].

In addition to cytotoxic action, neutrophil signalling following pathogen detection can further activate the immune system. Presentation of antigens by neutrophils to lymphocytes informs the adaptive immune system, and release of cytokines activates nearby macrophages which assist in phagocytosis [178].

Eosinophils

Similar to neutrophils, eosinophils are derived from myeloid progenitor cells and contain a granulated cytoplasm when observed with staining and microscopy. Unlike neutrophil cells, eosinophils are far less abundant in circulation (Table. 1.1), have a maximum of two nuclear lobes, and display orange-red granules following application of Wright-Giemsa stain [122, p. 239]. Wright-Giemsa stain consists of a mixture of red and methylene blue dyes. The red dyes, also known as eosin compounds preferentially bind to eosinophils due to the high amount of basic arginine rich proteins in their granules [122, p. 239]. Arginine is an amino acid which is negatively charged and described as 'basic' in contrast to positively charged 'acidic' amino acids. The purpose of the arginine rich proteins packaged in eosinophil granules is to destroy pathogenic cells, particularly parasite cells.

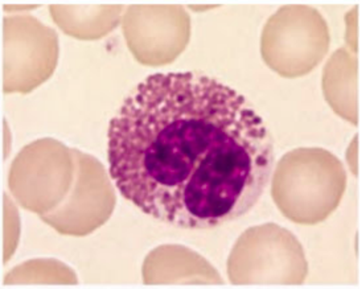

Figure 1.7: Eosinophil cell observed by microscopy following staining. Eosinophils are differentiated by their bilobed nucleus and the presence of granules (pink stain) in the cytoplasm. The cell is surrounded by discoid shaped red blood cells which are highly abundant in blood plasma (Figure source [81, p. 109]).

Examples include the negatively charged major basic protein (MBP) which permeabilises cell membranes of parasites and other targets, and eosinophil cationic protein (ECP) a cytotoxic protein with ribonuclease activity which also signals to nearby immune cells [122, p. 239]. Eosinophils are notable for their role in the destruction of parasites which are differentiated from other pathogens by being multicellular pathogenic organisms [93]. Furthermore, eosinophil function has been implicated in patients with allergic disease and related immune disorders such as asthma [15] [93]. Mild eosinophilia (an abundance of eosinophils) is observed in patients with allergic diseases such as asthma and allergic rhinitis [93].

Basophils

Basophils are the least common of all leukocyte cell categories, identified by their bilobed nucleus and a cytoplasm rich in granules which are stained dark purple by Wright-Gimesa dye and can often conceal the nucleus itself [122, p. 241]. Functionally, basophil cells are known for their high affinity for immunoglobulin E (IGE), a class of antibody primarily targeted to antigens present on parasite cells [161]. The binding of basophils to IGE is facilitated by the high affinity IGE receptor FcεRI, where expression of this receptor correlates with circulating IGE concentration [161]. IGE binding results in phosphorylation of tyrosine kinase Syk, leading to intracellular calcium release (Syk mediated signalling cascade), resulting in exocytosis of granules and their contents [108]. Basophils known for their highly granular cytoplasm will degranulate secreting large amounts of cytokines (IL-4, IL-13) and histamine [161]. Release of chemokines by basophils leads to characteristic allergic responses such as increased blood flow, itching of the skin (puritis), and sneezing (if in the respiratory tract) [161]. These are immune responses which aim to expel the parasite which cannot be easily destroyed by other immune functions such as phagocytosis or release cytotoxic proteins. Basophil activation contributes to allergic responses to allergens such as pollen when antibodies are binding these antigens [161].

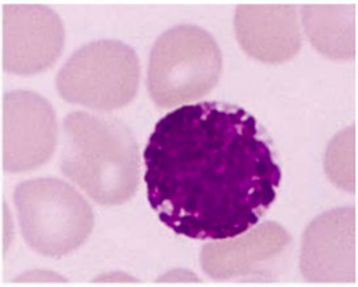

Figure 1.8: Basophil cell observed by microscopy following staining
Basophils are differentiated by their bilobed nucleus and the abundance of granules stained dark purple in the cytoplasm. The cell is surrounded by discoid shaped red blood cells which are highly abundant in blood plasma (Figure source [81, p. 109]).

Monocytes

Monocytes are short lived cells with a lifespan of 2 - 3 days in blood [122, p. 242] and are distinguished by their c-shaped nucleus and fine vacuoles in the cytoplasm which often stain blue [81, p. 111] (Fig. 1.9). Granules are present at a lower abundance than that of neutrophils, eosinophils, or basophils [122, p. 242]. In particular monocytes harbour primary (azurophil) granules [122, p. 242]. Most circulating monocytes are known to be classical monocytes which are differentiated by high expression of the CD14 cell surface receptor, a smaller number of monocytes are termed non-classical monocytes with higher CD16 and lower expression of CD14 receptor [187]. It is difficult to determine a specific boundary between the classical and non-classical population because non-classical monocytes develop from classical monocytes which change their expression of cell surface receptors [187]. Furthermore, an intermediate monocyte subpopulation has also been proposed with high CD14 and low CD16 expression [187]. Classical monocytes are the abundant subpopulation and modulate inflammatory responses at the site of infection. Non-classical monocytes are less characterised and thought to interact with endothelial cells in the vascular lumen [90]. The intermediate monocyte population is known to expand during infection, however their function is not well characterised [187].

Neutrophils are usually the first cells responding to infection followed by monocytes which support neutrophil cells in removal of pathogens [122, p. 242]. Furthermore, once monocytes enter the tissue they also convert into tissue resident macrophages which undergo a limited number of cell divisions resulting in mature macrophages. Macrophages can respond to pathogenic attack by phagocytosis of pathogens and release of signalling proteins. Macrophages also help return the tissue to homoeostasis following resolution of infection by clearance of cellular debris and contribution to wound closure [53]. Most macrophages in the human organism originate from a pool of self-renewing tissue resident macrophages. However, monocytes enable further production of macrophage cells at sites of infection [138].

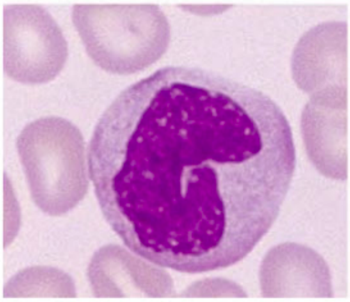

Figure 1.9: Monocyte cell observed by microscopy following staining. Monocytes appear larger than other blood leukocytes with a large oval or indented nucleus, the cytoplasm contains many small vacuoles and is stained blue by Wright-Gimesa stain with a 'ground-glass' appearance (Figure source [81, p. 109]).

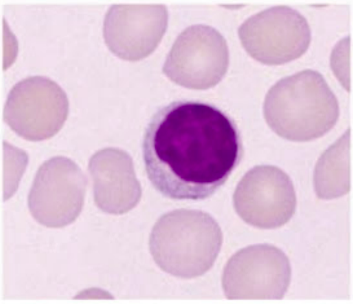

Figure 1.10: Lymphocyte cell observed by microscopy following staining. Lymphocytes are distinguished by their large rounded nucleus surrounded by a thin cytoplasm. Some lymphocyte subtypes (not pictured above) may also be granular (Figure source [81, p. 109]).

Lymphocytes

Lymphocyte cells are the second most abundant white cell after neutrophils, they are distinguished by a round nucleus and can be variable in size (Fig. 1.10) [81, p. 127] . Lymphocytes are the most heterogeneous of white cell categories and include, listed by order of abundance: T cells, B cells, and large granular lymphocytes (LGLs) (which includes natural killer (NK) cells). Roughly 75% of lymphocytes are T cells with the remaining 25% consisting of a roughly equal proportion of B cells and LGLs [122, p. 244]. Within the lymphocyte subtypes further heterogenity emerges during lymphocyte maturation which begins in the primary lymphoid organs: the medulla of the bone marrow and the thymus [81, p. 127].

Fundamental to lymphocyte maturation is the major histocompatibility complex (MHC) cell surface protein encoded in the human leukocyte antigen (HLA) locus. MHC presents antigens on the cell surface. These can be self antigens which are peptide fragments encoded in the genome, or foreign antigens, peptide fragments from pathogenic cells or viruses. There are two classes of MHC, class I presents antigens existing inside the cell, class II presents antigens collected from pathogens by antigen presenting cells (APCs) which include neutrophils and macrophages. Healthy cells expressing MHC class I will present self-antigens, however MHC class I on virus infected cells may present viral antigens - thus signalling that cell for destruction. During maturation lymphocyte cells are guided to ensure that they interact with the MHC molecule generally, but crucially that they do not react in situations where MHC is presenting self-antigens.

T cells mature in the thymus from precursor cells which originate from the bone marrow [81, p. 129]. A crucial part of maturation is development of the T cell receptor (TCR) which is expressed on the cell surface of T cells and should have a high affinity for MHC, but low affinity if MHC is presenting self antigens [81, p. 129]. The TCR gene contains a hypervariability region, in this region point mutations occur often leading to heterogeneous TCR proteins even between T cells in the same individual [122, p. 246]. Maturation occurs in the thymus by positive and negative selection. Positive selection promotes expansion of T cell populations with TCR which have affinity for MHC, this is important to enable T cells to bind to MHC in order to carry out their immune function. Subsequently negative selection removes T cells which bind to native antigens presented by MHC [163] (Fig. 1.11). Negative selection is achieved by presentation of self-antigens to maturing T cells, ensuring that T cells which react to these antigens are not permitted to survive [122, p. 249]. T cells which develop with affinity to MHC class I become cytotoxic T cells and those with affinity to MHC class II become T helper cells [163]. Cytotoxic T cells are responsible for lysis of virus infected cells which are presenting foreign antigens by MHC class I. T helper cells are responsible for detection of foreign antigens when presented by APCs by MHC class II [122, p. 249]. If a helper or cytotoxic T cell is activated by a foreign peptide it will proliferate and signal for further immune responses. Cytotoxic T cells contain granules which release perforin peptides and serine proteases upon contact with a virus infected cell [122, p. 249].

Unlike T cells expressing TCR with general specificity to recognise foreign antigens presented by MHC, B cells produce antibodies which recognise a specific antigen directly [79]. B cell maturation begins in the bone marrow where cells initiate expression of B cell receptor (BCR), a dimer of immunoglobulin molecules. Immunoglobulins are membrane bound surface molecules with constant Fc regions and variable antigen binding Fab regions [81, p. 131]. In the bone marrow, B cells undergo positive and negative selection for BCR molecules that bind to MHC, but not self-antigens [122, p. 246]. Following selection, naive B cells leave the bone marrow and move to the secondary lymphoid organs, the spleen or lymph nodes where pathogenic antigens are collected by immune cells such as T cells or by the passive flow of lymphatic fluid [79]. In the secondary lymphoid organs antigens are stored and concentrated by follicular dendritic cells (FDCs) located in germinal centres [79]. Naive B cells move into the germinal centre and proliferate undergoing somatic mutations of the BCR encoding genes (Fig. 1.12). B cells compete for binding with an

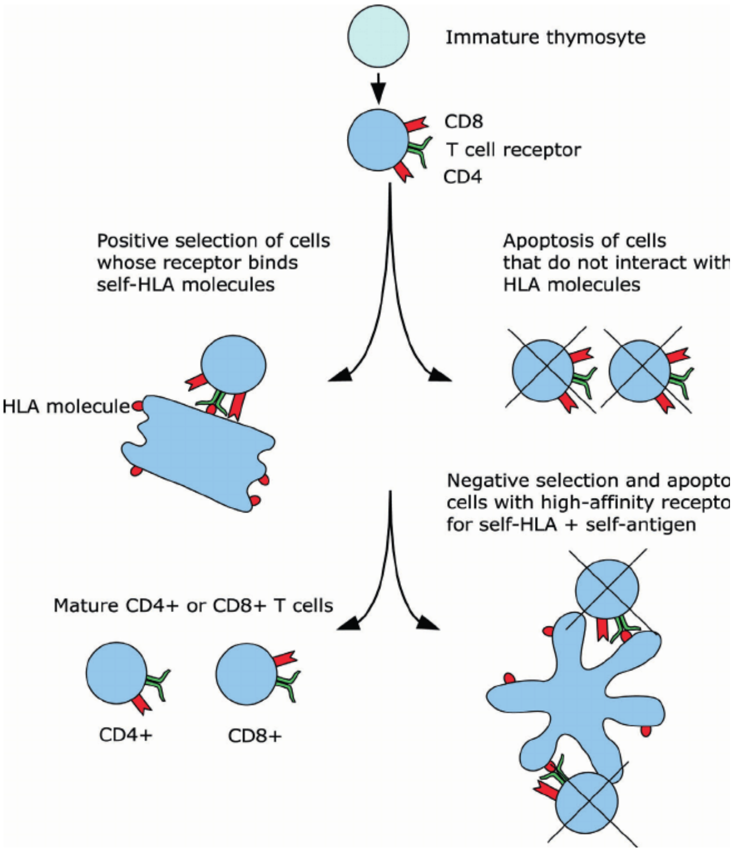
Immature thymosyte
CD8
T cell receptor
CD4
Positive selection of cells whose receptor binds self-HLA molecules
Apoptosis of cells that do not interact with HLA molecules
HLA molecule
Negative selection and apopto cells with high-affinity recepto for self-HLA + self-antigen
Mature CD4+ or CD8+ T cells
CD4+
CD8+

Figure 1.11: Positive and negative selection during T lymphocyte maturation. This figure shows maturation of T lymphocytes which are selected such that their TCR receptor binds HLA and secondly such that they do not bind to HLA presenting self-antigens, this is termed positive and negative selection respectively (Figure source [74]).

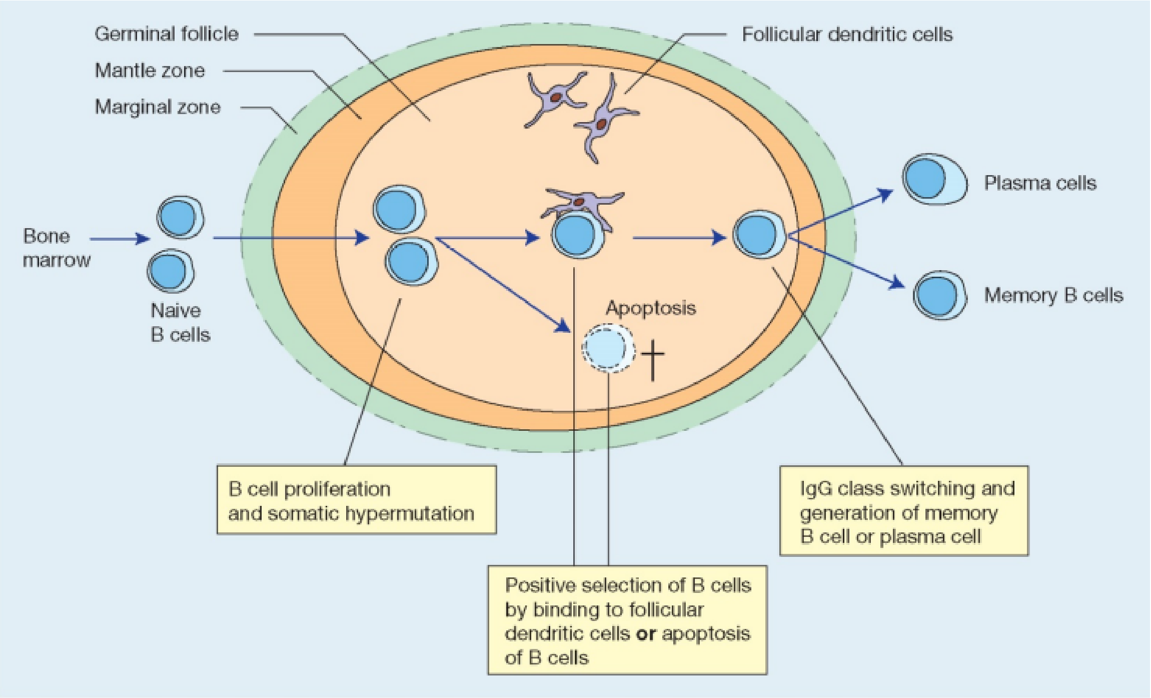
Germinal follicle
Mantle zone
Marginal zone
Follicular dendritic cells
Bone marrow
Naive B cells
Plasma cells
Memory B cells
Apoptosis
B cell proliferation and somatic hypermutation
Positive selection of B cells by binding to follicular dendritic cells **or** apoptosis of B cells
IgG class switching and generation of memory B cell or plasma cell

Figure 1.12: B cell maturation in the germinal centre
B cell maturation occurs in the germinal centres within secondary lymphoid organs. B cells move into the germinal centre and proliferate with somatic hypermutation which results in heterogeneous BCR. B cells which bind to dendritic cells presenting antigens are promoted to proliferate, this results in generation of BCR which binds to antigens. (Figure source [81, p. 137]).

antigen presented by the FDC, those that outcompete are encouraged to proliferate [81, p. 137]. This leads to production of B cells which can generate antibodies with specificity for an antigen. Surviving B cells differentiate into plasma cells responsible for high volume antibody secretion, or memory B cells which have a long lifespan and when activated by presentation of their antigen will rapidly proliferate and instantiate an immune response [113]. Plasma cells produce and secrete soluble antibodies based on their BCR generated by the previously described positive and negative selection.

LGL cells consist of two major classes, the previously described cytotoxic T cells and NK cells which are known for their response to virus infected and tumour cells. NK cells recognise their targets both by binding to the constant Fc region of antibodies or by detecting a lack of MHC class I expression, common in virus infected or malignant cells [81, p. 130]. Similar to cytotoxic T cells NK cells initiate destruction of their target by exocytosis of granule contents which includes proteases and other cytotoxic compounds [122, p. 249].

1.2 Haematological analysers

GWAS studies require collection of phenotype and genotype data in a cohort of individuals large enough to power identification of associations with statistical confidence (Section 1.5). This requires high-throughput measurement of phenotypes with limited measurement error. Automated haematological analysers are designed to rapidly count cell types and measure other cell properties from a blood sample. The measurements routinely obtained by haematological analysers can be condensed as follows: cell counts for the seven primary blood cell types (platelets, red cells, neutrophils, eosinophils, basophils, monocytes, and lymphocytes), red cell and platelet volumes, and red cell haemoglobin measurements. I will introduce the two major techniques for automated haematological analysis: impedance electrometry and flow cytometry and show that additional blood cell properties can be derived by flow cytometry.

1.2.1 Impedance flow cytometry (Coulter principle)

The origin of automated haematology analysers began in October 1953 with a patent for automated cell counting by flow cytometry and impedance electrometry also known as the Coulter principle, filed by Wallace H. Coulter and his brother Joseph R. Coulter, Jr [72] [50]. Historically, erythrocyte counting was routinely performed with manual microscopy, a process taking a haematologist up to 30 minutes per sample [72]. There was a pressing need for the development of an automated solution. A publication in 1934 by Canadian Andrew Moldavan proposed automated counting of 'microscopical cells' in solution using a capillary tube [120]. Here, cells in solution are forced to process through a capillary tube observed by a microscope and photoelectric apparatus which registers the passing of cells. Intriguingly this concept seems to have originally been proposed in a French publication by Marcandier, Bideau, and Dubreuil in 1928 [111]. Regardless, the publication by Andrew Moldavan is noted as being the inspiration for the development of the first flow cytometry and electrical impedance based automated cell counter by Wallace H. Coulter and his brother Joseph R. Coulter, Jr [72], their work being published in the aforementioned patent and referred to as the 'Coulter principle'. The Coulter based counter obtains an electrical contrast between cells and surrounding solution which is ten times higher than those identified by the photoelectric method proposed by Andrew Moldavan [72].

The Coulter principle relies on the changes in conductivity when cells pass through an aperture and displace surrounding solution. A container filled with fluid holds an electrode, inside this container a smaller tube is placed inside which an electrode of opposite charge is placed (Fig. 1.13). A microscopic aperture is made in the smaller tube allowing electrical current to flow from one electrode to the next. When fluid is drawn through the aperture, the observed current will not change as the conductance through the aperture is

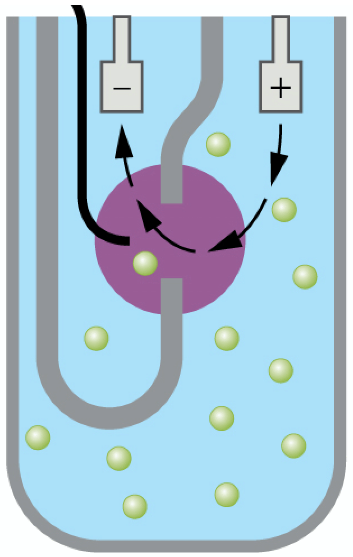

Figure 1.13: Schematic of the Coulter principle of electrical impedance. Schematic showing the Coulter principle for counting cells in solution. A conductive solution flows through an aperture with an electrical current also passing through the aperture. As a cell passes through the aperture the electrical current is disrupted because the conductive solution is displaced by a cell. The magnitude of the disruption of electrical current will depend on the size of the cell. Analysis of the disruptions in electrical current allow blood cells to be identified (Figure source [125]).

not changing. However, if a cell is drawn through the aperture the current is disrupted, because the conductance of a cell is different to that of fluid [72]. By studying the waveform of current flowing through the aperture the concentration and size distribution of cells can be measured. The Coulter principle is utilised in many modern automated haematological analysers, such analysers have also been utilised to phenotype large population cohorts empowering a number of GWAS studies of blood cell count [15] [162] [66].

1.2.2 Fluorescence flow cytometry

Fluorescence flow cytometry began with a patent filed by Göhde and Dittrich in 1968 [177]. The use of light for cell counting is based on work by George Oliver in 1896 who proposed counting blood cells by measuring the loss of light passing through a test tube caused by scattering and absorption by suspended cells [128]. This loss of light will be correlated not just to cell count, but also with cell size and haemoglobin content. To make this method of fluorescence flow cytometry a reality, it was required to separate individual cells. In 1965 Mark Fulwyler published an article in Science utilising the Coulter principle to separate cells by volume which proved to be the enabling factor which paved the way to fluorescence flow cytometry [63] [73]. Göhde and Dittrich utilised this technique with a laser beam to create the first fluorescence cytometer [177].

Flow cytometry relies on the scatter of laser light incident upon cells which flow

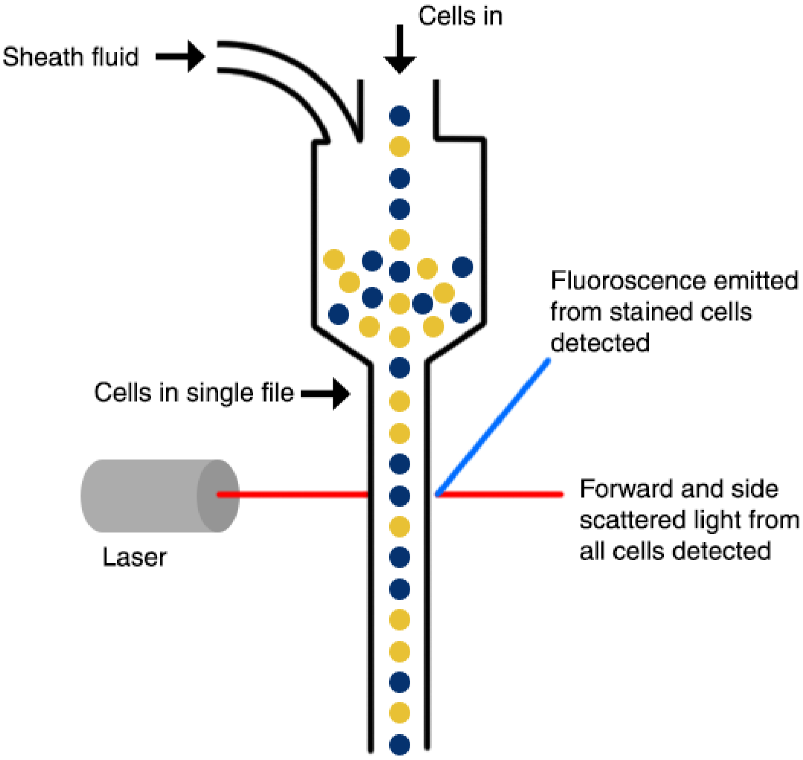
Cells in
Sheath fluid
Fluoroscence emitted from stained cells detected
Cells in single file
Laser
Forward and side scattered light from all cells detected

Figure 1.14: Schematic of fluorescence flow cytometry. Cells flow single file through the Sysmex flow cytometry channel and are hit by a laser beam, light is scattered or fluoresced by dyes in the cell and this is recorded resulting in three readings (SSC, SFL, and FSC) per cell. * SFL is an index of nucleic acid content also influenced by membrane composition of cells which affects the rate of absorption of nucleic acid staining dye into the cell (Figure source [4]).

through the cytometer in single file and the fluorescence of dyes which stain the cell (Fig. 1.14). In most modern flow cytometers fluorescent nucleic acid stains are applied to cells based on nucleic acid binding dyes such as acridine orange and membrane perforating agents, although the exact formulation of stains usually remains proprietary [73]. Three parameters are derived from cells flowing through a standard flow cytometer: side scatter (SSC), forward scatter (FSC), side fluorescence (SFL) (Table 1.2). These parameters represent cellular properties, cell structure or granularity, cell size, and cell nucleic acid content respectively. Plotting these parameters results in a scattergram with clusters of cells corresponding to different cell types (Fig. 1.15). Not only can individual cell types be identified from the scattergram, position of cells within a certain cell type in the three axes (SSC, FSC, SFL) represents important properties about the state of the haematological system which varies between individuals. Further discussion of flow cytometry parameters, their derivation from scattergram information, and clinical relevance is made in Section 3.2.2.

Property	Abbreviation	Description
Side Scatter	SSC	Cell internal structure and granularity
Forward Scatter	FSC	Cell size
Side Fluorescence	SFL	Fluorescence of stain, usually nucleic acid stain. Measurement also influenced by membrane composition which influences progression of stain into cell.

Table 1.2: Properties measured by standard cell flow cytometery.
A standard flow cytometer measures: SSC, FSC, and SFL. These measurements are representative of important cell properties such as cell size, cell structure or granularity, and cell nucleic acid content.

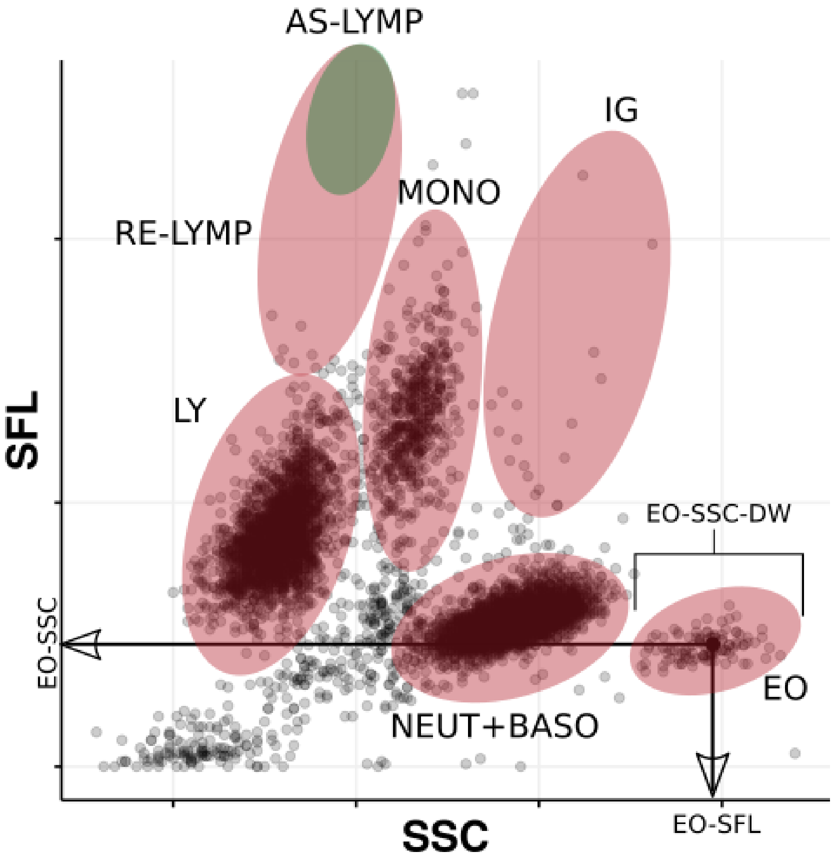

Figure 1.15: White blood cell differential channel scattergram.
This plot represents results obtained from the white blood cell differential channel (WDF) from analysis of a blood sample. Each data point represents a blood cell for which SSC, SFL, and FSC values are derived - only SSC and SFL axes are drawn. Cell types are identified using bounds set across the axes, for each cluster of cells representing a cell type cells are counted. Furthermore, phenotypes are derived from the median position of each cell type in the axes. The example drawn represents the SSC and SFL values for eosinophil cells and distribution width values are also calculated from the width of the cluster. Cell types observed from the WDF scattergram are highlighted. LY: lymphocytes, RE-LYMP: reactive lymphocytes, AS-LYMP: antibody synthesising lymphocytes, MONO: monocytes, IG: immature granulocytes, NEUT: neutrophils, BASO: basophils, EO: eosinophils.

1.3 Variation in the human genome

The human genome consists of 23 pairs of chromosomes located in the nucleus where each chromosome is a DNA molecule containing genetic information of the organism. DNA is formed of base pairs of 'nucleotides' of which there are four types: adenine (A), cytosine (C), guanine (G), and thymine (T). Across all 23 chromosomes, DNA consists of 6 billion base pairs of nucleotides. Genetic variations are changes in DNA sequence which exist between individuals. Fundamentally, it is genetic variation which underpins almost all the heritable differences in phenotypes between individuals. The total number of observable variations in the human genome has not been determined and depends greatly on the reference sequence used and population being studied. Genetic variants include single nucleotide polymorphisms (SNPs), the substitution of a single nucleotide in the genome, or inversions, copy number variations, insertions, and deletions. Copy number variations are regions larger than 1000 base-pairs which appear a variable number of times within the genome [141]. Copy number variations are a type of structural variation which includes other alterations such as inversion or translocation of segments of DNA larger than 1000 base-pairs [62]. Variants can be genotyped directly using microarray or sequencing technology, or imputed using a scaffold of nearby genotyped variations. Imputation relies on linkage disequilibrium (LD) calculations from a reference population to infer genetic variations which have not been genotyped and exist near to genotyped variations in the genome (Section 3.1.3). Genotyping arrays are designed to detect genetic variations target smaller variations such as SNPs, or small insertions and deletions and cannot identify larger structural variants such as copy number variations.

The inheritance of genetic variation was first determined by Mendel who described the 'law of segregation' and 'independent assortment' in inheritance of a phenotype from parent to offspring [176]. Mendel observed segregation of alleles on different chromosomes from parent to offspring, and independent assortment of those chromosomes. The laws of inheritance as described by Mendel cannot easily describe the inheritance of polygenic (or complex) traits which depend on multiple genetic loci across the genome [143]. Therefore, to characterise the genetics of a complex trait we must systematically test genetic variants across the genome for association with the phenotype. This type of analysis, a systematic search for genetic associations across the genome is termed a genome wide association study (GWAS), and requires a large population sample on which genetic and phenotypic data has been recorded.

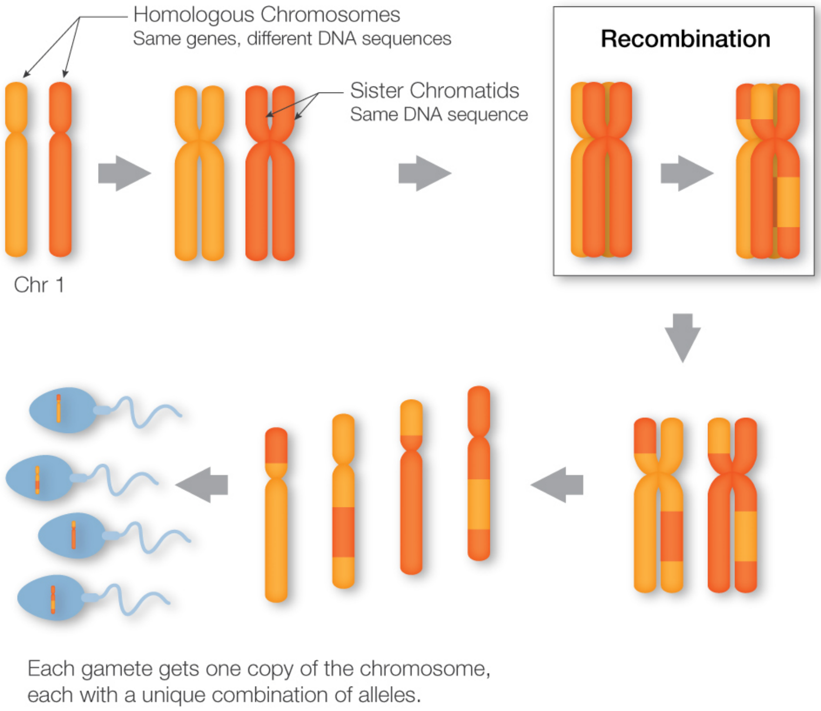
Homologous Chromosomes
Same genes, different DNA sequences
Sister Chromatids
Same DNA sequence
Recombination
Chr 1
Each gamete gets one copy of the chromosome,
each with a unique combination of alleles.

Figure 1.16: Recombination exchanges of information between chromosomes. Exchange of information between chromosomes during homologous recombination. This effect results in chromosomes inherited by offspring being different from those parental chromosomes. Recombination means that variants located physically close on chromosomes are more likely to be inherited together (Figure source: [127]).

1.4 Genetic recombination and linkage disequilibrium

Genetic inheritance relies on chromosomes, which are continuous coiled segments of DNA that are passed from parent to offspring. Every human individual normally inherits 23 chromosomes from each parent, resulting in a genome of 46 chromosomes. However, chromosomes are not an indivisible unit of genetic information. Recombination can occur during formation of gametes which pass genetic information to offspring, thus allowing exchange of DNA segments between each chromosome in the pair, this is termed homologous recombination (Fig. 1.16).

The result of genetic recombination is that alleles which are physically located close together on a chromosome are more likely to be inherited together. Therefore, in the population sample, the correlation between two alleles on the same chromosome reduces as the physical distance between their variants in the genome increases (Fig. 1.17). This also creates challenges in the context of GWAS as it is difficult to distinguish whether a variant is associated with a phenotype due to a true mechanistic effect, or because the variant is located close to, and is therefore often inherited with a variant which is the true mechanistic variant. In Figure 1.18 a single association signal with the blood cell

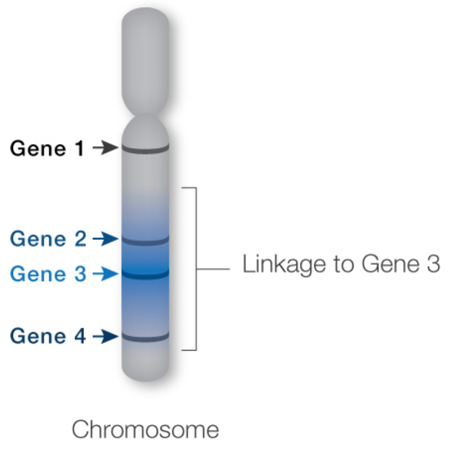

Figure 1.17: Linkage to a point in the genome across a chromosome. Alleles of genes which are physically closer to gene three are more likely to be inherited with gene three, this is defined as a greater degree of linkage. Similarly, proximal alleles are also more likely to be inherited together (Figure source: [127]).

phenotype H-IPF a measure of immature platelets (a platelet parameter described further Section 3.2.2) is identified on chromosome 8, and there are multiple highly correlated variants with low P-values for association with the phenotype. It is not immediately clear which (if any) of these highly correlated and associated variants is the causal variant for the observed signal. Therefore, GWAS analyses are often followed by methods to identify the number of independent association signals identified accounting for correlation or LD between alleles. Methods to address this are discussed including multiple stepwise conditional analysis (Section 2.2.8.1) and LD based clumping methods (Section 2.2.9).

1.5 Genome wide association study

A GWAS analysis can identify genetic variants which are associated with changes in a phenotype, for example: transcript level, protein concentration, or some other biological property. The phenotype is measured in a sample of individuals which are also genotyped to determine genetic variations. A GWAS analysis estimates the magnitude of the effects variants have on the phenotype and the standard error of this estimation. This analysis highlights regions of the genome which influence the phenotype being studied. GWAS analyses are subject to confounding factors such as population stratification which are discussed in more detail in Section 2.2.6.

GWAS analyses are categorised into those studying case-control outcomes such as diagnosis of disease, or quantitative outcome such as height, weight, or the value of a haematological measurement. If the outcome in question is binary (case-control) a logistic model will most often be used to generate test statistics, alternatively a linear regression is used to model quantitative outcomes. Each SNP is tested individually with adjustment for covariates which can include participant factors such as age, weight or height which could also effect the phenotype. A quantitative phenotype is studied as follows, where y is a vector of phenotype values, x_c represents a matrix of covariate values across individuals, x represents the genotype of the individual coded as 0, 1, or 2 for homozygous reference allele, heterozygous, or homozygous alternate allele respectively, and α, β, β_c represent the intercept and effect sizes of the genotype and covariates to be estimated:

$$\mathbb{E}[y] = \alpha + x_c\beta_c + x\beta \tag{1.1}$$

$$y \sim \mathcal{N}(\mu, \sigma^2) \tag{1.2}$$

Alternatively, working with a case-control GWAS where phenotype values are binary outcomes $y \in \{0, 1\}$, Equation 2.1 is modified as follows using a sigmoid function to fit a logisitc regression:

$$\mathbb{E}[y] = \frac{1}{1 + e^{-}(\alpha + x_c\beta_c + x\beta + \epsilon)} \tag{1.3}$$

$$y \ \sim B(p) \tag{1.4}$$

Population stratification and relatedness are effects which may lead to false positive results and are discussed in more detail in Section 2.2.6. Briefly, covariates are included in the model such as population principal componentss (PCs) or an additional genetic relationship matrix (GRM) term which accounts for relatedness amongst the sample population to help account for population stratification. A GRM requires modelling of random effects, in which case the linear model above is modified to a linear mixed model (LMM) (Section 2.2.7). At its basis, a GWAS study utilises a linear model or a LMM. Therefore, the primary assumptions of linear regression still apply which are discussed with more detail in Section 2.2.4.

1.5.1 The history of GWAS analysis

The history of GWAS necessarily begins with that of the human genome project which formally began in October 1990. The goal of the human genome project was to determine the sequence of nucleotide base pairs in DNA and map all the genes within the human genome. The first genome to be sequenced was that of bacterial virus $\phi X174$ with 5400 nucleotides by Fred Sanger in 1977 [147]. For many years the prospect of sequencing the human genome was met with incredulity and disbelief [158]. In May 1985 Robert Sinsheimer organised a workshop to propose sequencing the human genome and wrote the following about the response of his audience: “The sources of hesitation ranged from concerns over the introduction of Big Science into biology to arguments that most of the human DNA is junk" [158]. In response to arguments about "junk DNA" Sinsheimer would retort that "one man's garbage is another's treasure" and used the example of large projects in physics and astronomy such as the Hubble space telescope to argue for the benefits of "Big science" [158]. Unfortunately, Sinsheimers proposal to sequence the human genome was not pursued at that time. However, subsequent workshops proposing sequencing of the human genome by Charles DeLisi [54], James Watson [75], and a publication in Science by Renato Dulbecco [56] began to turn the tide. Of particular importance was New Mexico US Senator Peter Domenici, a friend of DeLisi who offered political support to proposals to dedicate money from various government organisations to fund the prospective genome project. Efforts culminated with approval for funding from the Senate Committee on Energy and Natural Resources chaired by Senator Domenici and an act of Congress in January 1987, both of which committed money for the purpose of sequencing the human genome. The act was a budget submission signed by US President Ronald Reagan himself, and thus began the era of 'Big Science' proposed by Sinsheimer which continues to this day. The remaining history of the human genome project is well documented in a number of sources [1] [54]. Laboratories in the United States, United Kingdom, Germany, Japan, and China would eventually contribute to the project in addition to (and often competing with) private organisations in particular Celera Genomics finally leading to publication of the draft human genome sequence in 2001.

Publication of the human genome sequence enabled creation of the Japanese single-nucleotide polymorphisms (JSNP) database of 190,652 variants in the Japanese population. It is from this database that the first GWAS for myocardial infarction in Japanese participants was performed [86]. From JSNP, 92,788 variants were selected which the authors assumed would account for each of the 100,000 genes expected in the genome. Of course, we now know that the estimation of 100,000 genes in the genome is incorrect and the number is closer to 30,000. The chosen variants were genotyped in 1,133 individuals affected by myocardial infarction and 1,006 controls [129]. This analysis, which is now known to be the first ever GWAS identified only a single association locus composed of 5 SNPs on chromosome six [129] [86]. Although, it could be argued that 92,788 variants studied in this analysis hardly cover the entire genome. In the same regard even modern GWAS studies do not cover the entire genome often missing regions which are difficult to genotype or impute such as the MHC locus. Regardless, since the first study in 2002 [129], GWAS Catalog, a repository for results of GWAS analyses has collected over 3,000 GWAS results [34]. Almost all major common diseases have been subject to GWAS, often in sample sizes of hundreds of thousands of individuals.

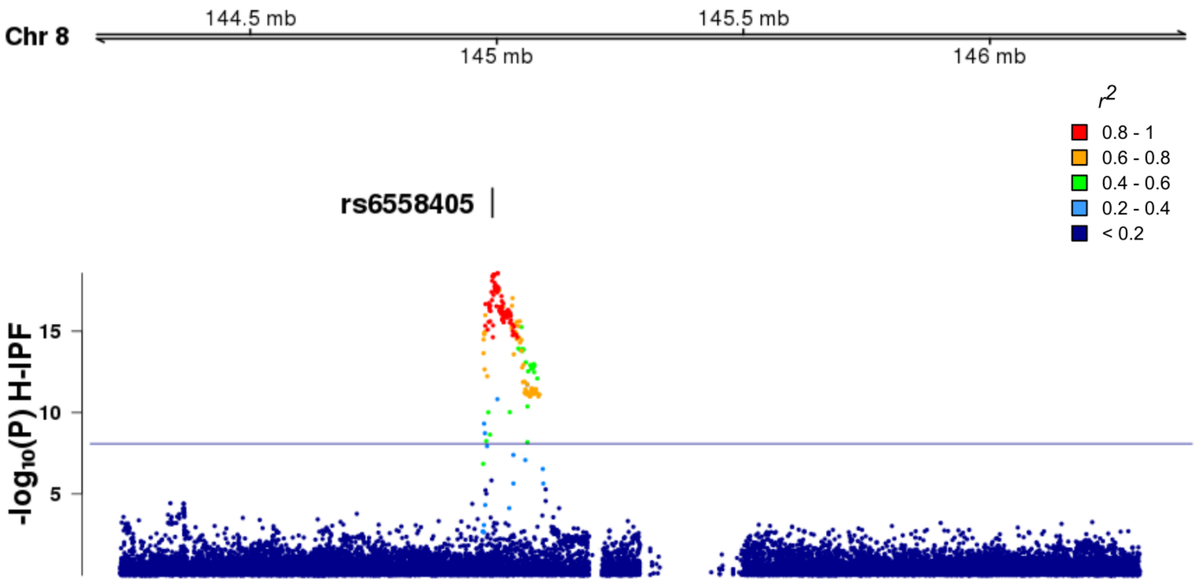

Figure 1.18: GWAS identifies a large number of significant variants but only one signal at this locus.

This plot represents an association signal on chromosome eight with haematological trait H-IPF, a measure of immature platelets (Section 2.1.3). Each data point represents a genetic variant with the position of the data point on the x axis corresponding to the physical location of that variant on chromosome eight. The position of each variant in the y axis is the -$\log_{10}(P)$ for association with H-IPF. Variants are coloured based on their LD to the conditionally significant variant (rs6558405) which is identified with multiple stepwise conditional analysis to be the best statistical candidate for this association signal, although this does not imply that rs6558405 is the causal variant for this association signal. Variants in high LD to the conditionally significant variant show significant association with the phenotype H-IPF, however conditional analysis shows only one independent signal at this locus, suggesting only one variant is mechanistically associated with changes in the phenotype.

1.5.2 Genotyping of genetic variants

Genotyping is a biological assay which allows the determination of alleles in DNA. The two primary classes of genotyping technology utilised in GWAS analysis are DNA microarrays and whole genome sequencing. In my analysis I utilise DNA microarray based technology which at the time of study recruitment and genotyping (2012 - 2015) was affordable enough to enable genetic analysis of a large cohort of individuals as required for GWAS.

The most common DNA microarray technique for GWAS genotyping is one-channel detection utilised by the Affymetrix Gene Chip and Illumina Bead Chip systems. This system allows determination of hundreds of thousands of genetic variants in a single sample. Here, a DNA microarray contains a number of probes, each of which can assay for the presence of an allele in a sample. A probe is a fragment of single stranded DNA encoded to hybridise to a piece of DNA which contains the variant in question. The probe is fixated to the silicon wafer by a covalent bond which forms the body of the microarray. Probes can be designed as required, selection of probes is discussed in more detail in Section 2.2.1. A sample of DNA is fragmented and washed over the silicon wafer containing the probes. If fragments contain the genetic variant of interest it will form a strong hybridisation with the corresponding probe. The silicon wafer is washed to remove DNA fragments which are not hybridised with an appropriate probe. The DNA microarray is designed such that specific coordinates on the array are dedicated to detecting a particular variant. A fluorescent dye is added which binds to double stranded DNA and a laser is used to query coordinates along the DNA microarray to identify probes which have hybridised. Thus it is possible to deduce which alleles exist in the DNA sample by identifying which probes are fluorescent. However, only a limited number of variants can be genotyped. In the case of my analysis a Affymetrix Axiom array including 820,967 probes was utilised. It is known that variants across the genome are correlated depending on physical distance between them (Section 1.4). Utilising this linkage (correlation) structure which can be determined by whole genome sequencing of a reference population, it is possible to further impute the existence of millions more genetic variants.

Imputation of genetic variants

Imputation allows prediction of genotypes which were not directly measured. Imputation is performed using the correlation structure (LD) between variants in the genome estimated from a reference population. To ensure similar LD structure, the reference population must have similar ancestry as the genotyped individuals. All imputation methods begin by phasing genotyped variants to estimate ‘haplotype blocks’ along the genome (Fig. 1.19). Haplotype blocks are contiguous regions of DNA which show little evidence of recombination in the reference population thus the genetic variants in a haplotype block are very likely to be inherited together. Once haplotypes have been estimated, haplotypes from the reference population are used to impute missing variants in the sample population (Fig. 1.19). Imputed alleles are estimated with a degree of uncertainty, this is often represented with the information (INFO) score metric which is between 0 and 1. The INFO score for an imputed variant multiplied by the total sample size represents the equivalent effective sample size for the power of an association test, a perfectly imputed variant will have an INFO score of 1 [25]. Given this uncertainty, if genotyped variants are encoded as follows: 0 for homozygote reference, 1 for heterozygote, and 2 homozygote alternate alleles, imputed variants will have values closer to 0 if the likelihood of homozygote reference alleles is high and closer to 2 in the reverse scenario. Examples of reference populations include the European ancestry populations from the UK10K [46] and haplotype reference consortium (HRC) which combines data from multiple cohorts including UK10K and the 1000 Genomes project [167]. These data were utilised by the UK Biobank consortium to impute nearly 96 million variants in their cohort of 500,000 British volunteers [38].

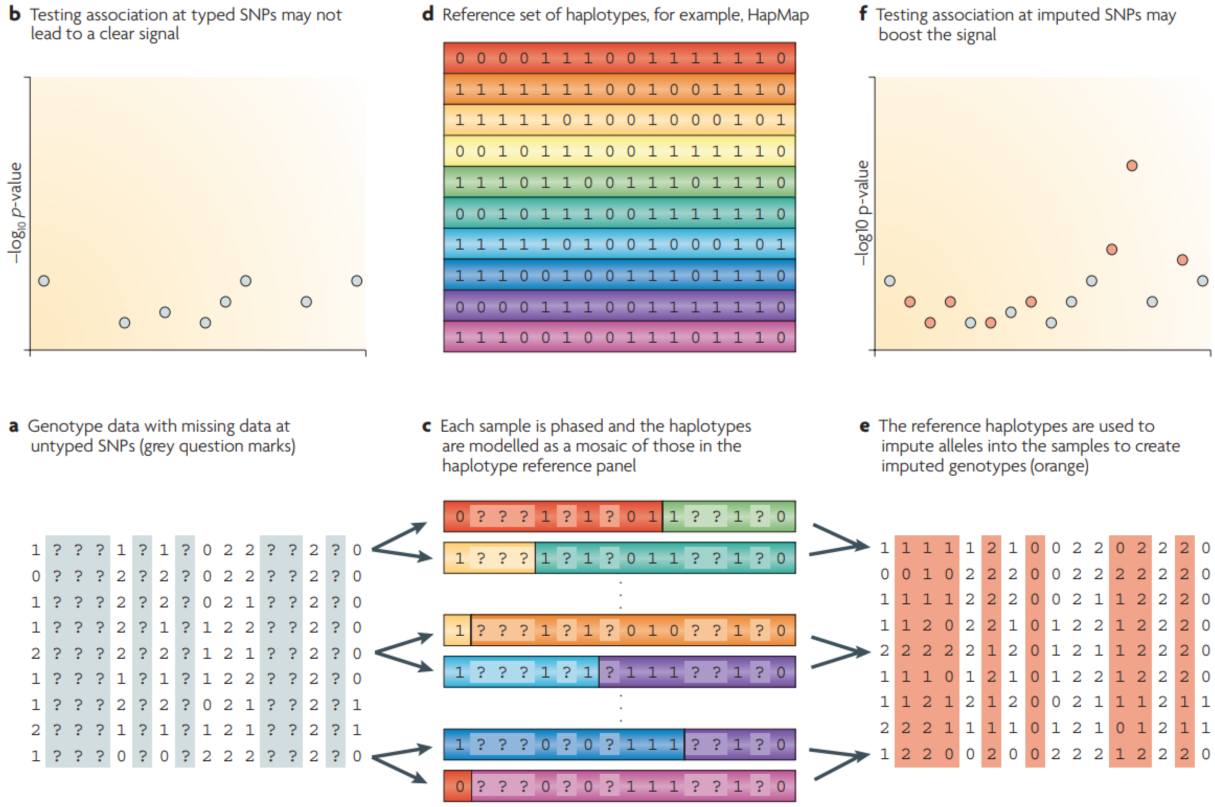
b Testing association at typed SNPs may not lead to a clear signal
−log10 p-value
d Reference set of haplotypes, for example, HapMap
f Testing association at imputed SNPs may boost the signal
−log10 p-value
a Genotype data with missing data at untyped SNPs (grey question marks)
c Each sample is phased and the haplotypes are modelled as a mosaic of those in the haplotype reference panel
e The reference haplotypes are used to impute alleles into the samples to create imputed genotypes (orange)

Figure 1.19: Schematic of steps for imputation of genotype data to estimate missing variants

a) Genotype data from the sample population with missing genotypes represented by question marks. **b)** Testing for an association signal using genotype data alone results in no association peak. **c)** Using a set of reference haplotypes from d) genotype data is phased to determine haplotypes present at each position along the genome. Three phased individuals are represented in the figure, each genome is a mosaic of haplotypes from the reference population. **d)** Reference haplotypes are defined from whole genome sequencing of a population with similar ancestry to the sample population. **e)** Missing variants in the genotyped sample population are estimated using the imputation procedure, with imputed variants highlighted in orange. **f)** In this example, testing for association of genotyped and imputed SNPs results in an association signal which was not identified before (Figure source [112]).

1.6 Interpretation of GWAS results

Since the first GWAS of myocardial infarction in 2002 which found an association locus on chromosome six [129], the number of identified associations has been increasing rapidly [173]. As more associations are identified, there is a greater need for interpretation of these results to generate relevant scientific insights. Large amounts of public money from government agencies such as the National Institute of Health and charities such as the Wellcome Trust have been and continue to be committed to GWAS studies on the promise of great advances in our understanding of biology and disease. As we approach two decades since the first GWAS study, questions still remain about how to use the results of GWAS analysis to inform biological experimentation and clinical development. I will attempt to address these questions by discussing the tools and techniques which have been developed to query GWAS summary statistics and how these techniques aim to answer fundamental biological questions. Some of the primary challenges regarding inference from GWAS results can be categorised as follows:

- Confident identification of the genes mediating each genetic association, a starting point for further inference.
- Understanding the mechanisms of biology which lead to emergence of a genetic association and the tissue specificity of those mechanisms.
- Inferring a causal relationship between two measurements, for example a risk factor and disease risk, and the implications of this for the consideration of the risk factor as a target for therapeutic modulation.

Fundamentally, all these questions are proposed in context of the same overarching goal: "how can we interpret the results of GWAS analysis to infer causal underlying biological mechanisms". However, inferring causal mechanisms is difficult, analysis such as annotation of associations to genes can get closer to this goal. In an attempt to make biological and aetiological inferences from GWAS analysis I utilise many tools and techniques. Here I will provide an overview of these techniques within the context of broader biological questions, with more detailed methodological reviews in the relevant chapters.

1.6.1 From genetic associations to genes

GWAS can provide insight into the genes which contribute to the studied phenotype. However, inferring a mediating gene from an associated variant is not trivial. In some cases, the associated variant may be near or overlapping several genes, in other cases the associated variant is located megabases away from the nearest gene. This problem is encapsulated in the wider challenge of identifying a causal biological mechanism which is leading to associations identified by a GWAS study. For many associated variants it is not even clear which tissues the mechanism of action occurs in and how this leads to changes in the phenotype which is detected by GWAS. Understanding the biological mechanism causal for the genetic associations must necessarily include the gene which is being modulated. Broadly, two solutions exist to annotate genetic variants to probable mediating genes:

- Physical overlap or physical distance of the variant with genes, this is implemented by software packages such as variant effect predictor (VEP).
- By integrating (colocalising) loci of associations from GWAS with gene or protein expression data.

Studying which genes and genetic elements the variant in question is overlapping can be misleading, a variant may overlap with multiple genes some of which may not even be expressed in the relevant tissue. Colocalisation of genetic associations with gene or protein expression GWAS results is a more reliable method for annotation, discussed in detail below (Section 1.6.2).

1.6.2 Understanding genetic associations with colocalisation analysis

Genetic colocalisation analysis can determine if different phenotypes with a genetic association in the same locus are being mediated by the same underlying causal variant. The context of my work is to better understand genetic associations with haematological phenotypes derived from a haematological analyser. I utilise colocalisation to explore the following questions about genetic associations with haematological measurements:

- Which blood cell transcripts are modulated by the genetic association? This will suggest genes which may be participating in the biological mechanism leading to changes in the haematological measurement.
- The concentration of which blood plasma proteins are modulated by the genetic association? This analysis could identify which haematological cells are producing particular blood plasma proteins, or which haematological cells are modulated by particular blood plasma proteins.
- Does this genetic association, which is modulating a haematological phenotype, also influence disease risk? This is a starting point for more detailed analysis to study the effect of haematological cells on disease aetiology.

It is crucial to emphasise, colocalisation analysis cannot determine causal relationships between the phenotypes being studied, and neither can it determine the direction of causality between the two phenotypes. In the case were two phenotypes both have an association signal in a locus, colocalisation analysis determines if those signals are caused by a common variant. The interpretation of colocalisation analysis depends on the specific nature of the phenotypes being colocalised. Haematological phenotypes studied in my analysis are discussed in Section 2.1.3 and 3.2.2, colocalising expression quantitative trait loci (eQTL), protein quantitative trait loci (pQTL), and disease risk phenotypes are described further in Chapter 5.

1.6.3 Causal inference with mendelian randomisation

Mendelian randomisation (MR) allows determination of a causal relationship between an exposure and outcome using genetic variants across the population sample. A population level causal association is a relationship where if across the entire population the exposure is modulated, this will cause a concomitant change in the outcome. An example of this is low density lipoprotein and heart disease, low density lipoprotein is known to cause heart disease and is modulated in the population by statin medication to reduce the risk of heart disease. It is known that there is a causal association between this risk factor and outcome. In this context, genetic associations are instrumental variables (IVs) used to assess the influence of the exposure on the outcome. This helps avoid the pervasive problem of confounding in epidemiological studies. Confounding factors influence both the exposure and outcome and can therefore induce a correlation between them. Often, confounding factors are unlikely to be measured or even known, therefore epidemiologists may never know if identified correlations are driven by confounding factors or result from a true causative mechanism. In MR studies the confounding effect is greatly ameliorated as it is not likely for confounding factors to affect genetic variants which are randomised at birth. Therefore, we are able to assess causality without the risk of our inferences being unduly influenced by confounding factors.

MR studies are often termed ‘naturally randomised trials’, as properly chosen exposures can be used as proxies for clinically relevant biomarkers or modifiable exposures. Genetic evidence for efficacy of drug targets can be tested using MR (Fig. 1.20). For example, instrumental variables in LDLR have demonstrated LDL-c is a risk factor for coronary heart disease (CHD) (Fig. 1.21). However, MR relies on a large number of assumptions which must be met for the analysis to be reliable. These assumptions and suggestions for sensitivity analysis are described in Section 5.2.2.

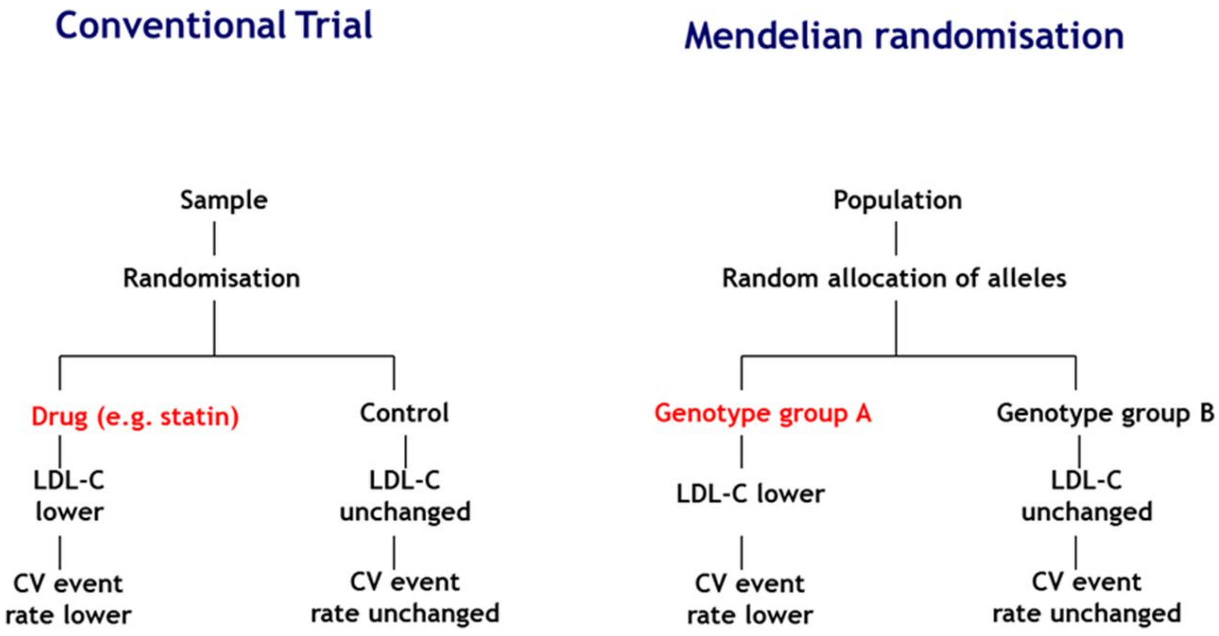

Figure 1.20: Mendelian Randomisation and a ‘naturally randomised control trial’.

MR analysis as an analogy of a conventional randomised control trial, in this case the relationship between LDL-C and cardiovascular (CV) events is studied by genetic instrumental variables associated with a reduction in LDL-C. Assuming the genetic instruments meet assumptions of validity (Section 5.2.2), MR can be used to assess the causal relationship between LDL-C and cardiovascular events and thus predict whether LDL-C is a causal factor in modulation of cardiovascular risk. This is relevant to clinical trials of LDL-C lowering therapies such as statins (Figure source: [22]).

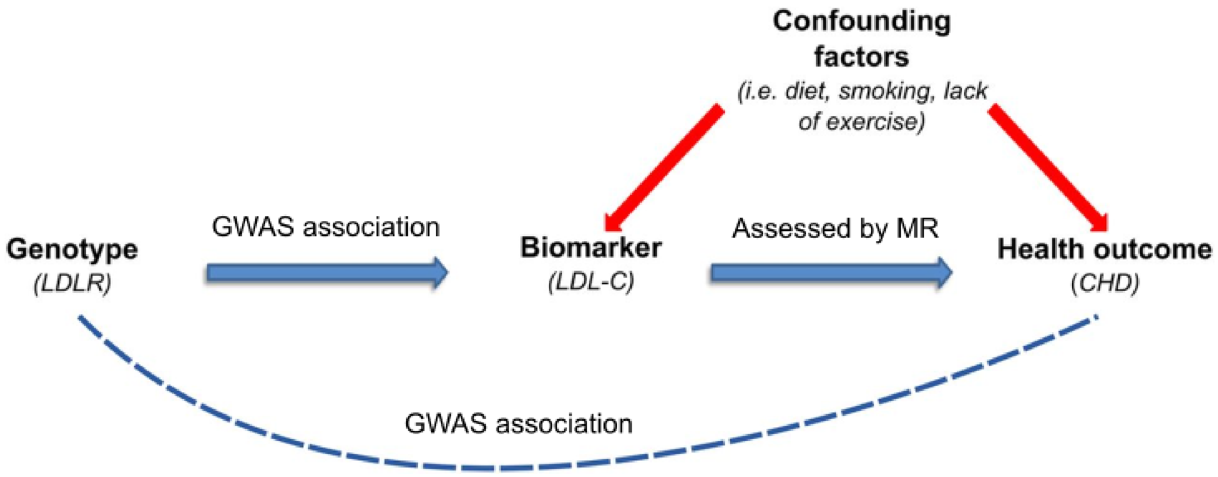

Figure 1.21: Mendelian Randomisation to test causal association between LDL-c and chronic heart disease.
A diagram representing the framework of a MR analysis, instrumental variables are selected from genetic variants located in the LDL Receptor (LDLR) gene which raises LDL-C levels. Instrumental variables are used to assess for causal association between LDLR (exposure) and CHD (outcome), allowing estimation of a causal effect between LDL-C and CHD without influence of confounding factors (Figure adapted from: [119]).

1.6.4 Intermediate traits for the study of disease aetiology

GWAS of disease outcomes identifies many associated variants, however a primary challenge is understanding which genes, proteins, and cellular behaviours are influenced by disease associations. An example of this is identifying which genes are modulated by an association, a problem which is described in detail above (Section 1.6.1). Measurements of gene expression, protein concentration, and cellular properties are often termed intermediate traits. GWAS of intermediate traits identifies associations which influence those phenotypes. Colocalisation analysis can identify cases where an intermediate trait association and a disease association is driven by the same underlying causal variant. Furthermore, a causal relationship between the intermediate trait and disease risk can be identified with MR. Individuals with asthma often have increased eosinophil count, does asthma cause increases in eosinophil count or are individuals with higher eosinophil count at greater risk of asthma? A MR study using GWAS of eosinophil count and GWAS of asthma suggests the causal relationship is driven by higher eosinophil count which in turn increases the risk of asthma [15]. In this way my GWAS of haematological phenotypes including functionally relevant phenotypes can help increase understanding of disease aetiology.

1.6.5 Genetic analysis to inform drug development

In recent decades there has been rapid progress in scientific advances including sequencing of the human genome, emergence of new therapeutic modalities such as antibodies and

RNAi, and advances in combinatorial chemistry allowing synthesis of thousands of small molecule compounds. Given these developments it could be expected that the rate of approval of new drugs to be higher relative to historical averages. However, approval of new drugs is not only taking longer and getting more expensive, candidates are also now more likely to fail at late stage clinical trials than in the past [159] [91]. In response to this unfavourable outlook AstraZeneca undertook a "major revision of its R&D strategy" in 2011. Part of this revision was the publication of a systematic longitudinal analysis of its drug development portfolio for all projects between 2005 - 2010 [47]. Notably, their analysis identified that that failure at late stages of clinical trails (Phase II and III) were more likely to be due to drug efficacy than any other factor, reaching up to 88% in Phase IIb trials (Fig. 1.22), where smaller Phase IIa trials (<200 patients) are distinguished from larger Phase IIb trials (<400 patients). In 40% of cases, the reason for failure due to efficacy was cited to be "target linkage to disease not established or no validated models available", this is in contrast to other reasons such as the dose of drug being limited by compound characteristics (Fig. 1.23).

These results show that many drug compounds pass all the early milestones of drug discovery only to fail at the latest stages of clinical development due to no efficacious effect in man (Fig. 1.22). A study of the costs of drug development shows that from a cost per new launch of $1.78 billion, roughly $1 billion of costs are incurred prior to Phase II clinical trials, the first real opportunity to assess clinical efficacy in man (Fig. 1.25) [134]. These analyses demonstrate a need to assess drug efficacy as early as possible, and that experiments with tissue and mouse models are not providing accurate enough insights into human biology. However, testing compounds in a clinical trial prior to passing all the early development milestones which contribute to the aforementioned $1 billion of cost would be deemed highly unethical.

Role of genetics in drug development

It has been shown that genes which are drug targets in a database of drug approvals in the United States and European Union are significantly enriched with genetic variants associated with human traits compared to other genes [124], and candidates with genetic linkage evidence are 30% more likely to succeed (Fig. 1.24). Here, associated genetic variants were assigned to genes using physical proximity and evidence that the genetic association influences gene expression, for more details see Nelson *et al* [124]. A promise of modern genetics is to use observational data collected from cohort level analysis to make inferences about the aetiology of disease. In this context, the benefits of genetic analysis are clear. GWAS of disease outcomes can identify associated variants which may implicate a gene in disease aetiology and therefore indicate a potential drug target (Section 1.5). Furthermore, using intermediate phenotypes (Section 1.6.4) one can generate a multiomic

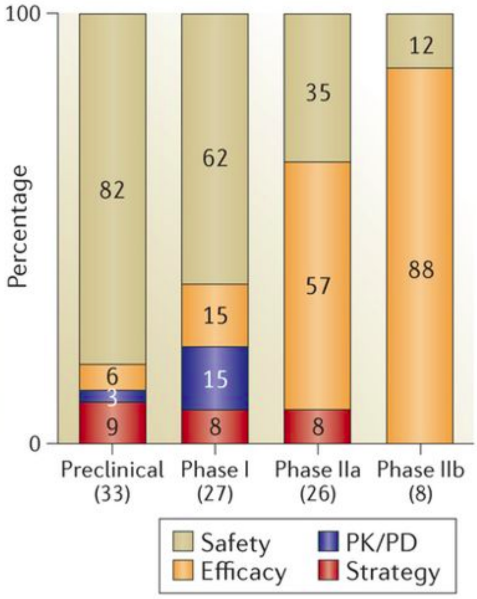

Figure 1.22: Primary reasons for project closure in AstraZeneca pipeline 2005 - 2010.

Project closures were classified into the following categories: safety (toxicology or clinical safety), efficacy (failure to achieve sufficient efficacy), pharmacokinetics/pharmacodynamics (PK/PD) or closure due to oranisation strategic reasons. The percentage of projects failing due to each category for each phase of development is indicated in the plot with the total number of projects shown in brackets below the bars (Figure source: [47]).

picture of the consequences of variations using colocalisation analysis between multiple phenotypes such as transcript levels, biomarker levels, and multiple disease outcomes (Section 1.6.2). Finally, MR analysis can test purported causal relationships between risk factors and disease outcomes (Section 1.6.3). However, such analyses must be performed with awareness about the limitations of the ability of genetics and the challenges of drug discovery as a whole. As previously described, the first GWAS study of myocardial infarction was performed in 2002, since that time many complex common diseases have been subject to GWAS. This has not lead to a substantial increase in the number of therapeutic compounds being developed or approved. Indeed, very few genes implicated by GWAS analysis as being associated with disease risk are also subject to therapeutic intervention.

Limitations of genetics in drug development

In addition to examining reasons why drugs fail in clinical trials (see above), it may also be helpful to ask how the pool of candidate drug targets is initially selected. Based on current technology it is estimated that roughly 10% of proteins in the human organism can be targeted by small molecule drugs, an additional 10% can be targeted with biologics such as antibodies [171]. This estimation of the total pool of ‘druggable’ proteins shows that the

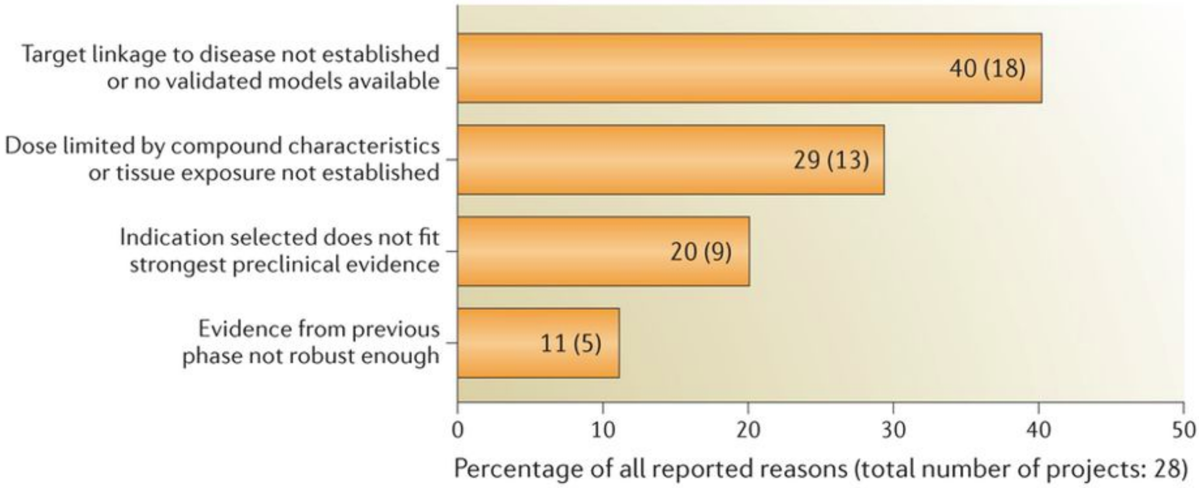

Figure 1.23: Reasons for lack of efficacy in clinical trials in AstraZeneca pipeline 2005 - 2011.
Project teams were surveyed to identify reasons for project failure due to lack of clinical efficacy and answers were classified into one of four categories shown. Teams could report more than one reason for lack of clinical efficacy. Percentages are shown in the bars with the total number of projects failing due to the listed reason in brackets (Figure source: [47]).

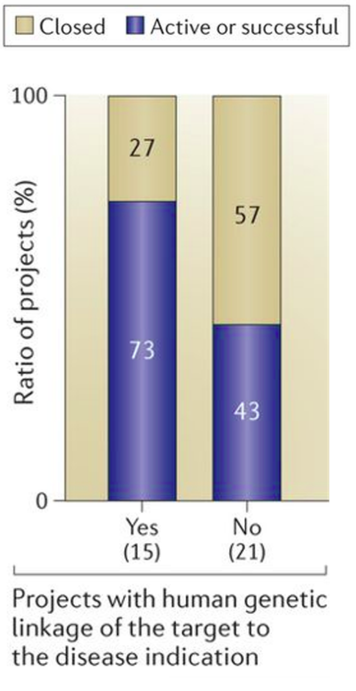

Figure 1.24: Success rate of projects in Phase IIa stratified as those with or without human genetic linkage evidence linking the target to disease.
Phase IIa is classified as Phase II projects with fewer than 200 patients. Projects in Phase IIa were classified as those with or without human genetic linkage of the target to disease. Projects were also classified as those still active or successful or closed. A higher closure rate is observed for projects without human genetic evidence for linkage to disease. Percentages are shown in the bars and total number of projects in the brackets below (Figure source: [47]).

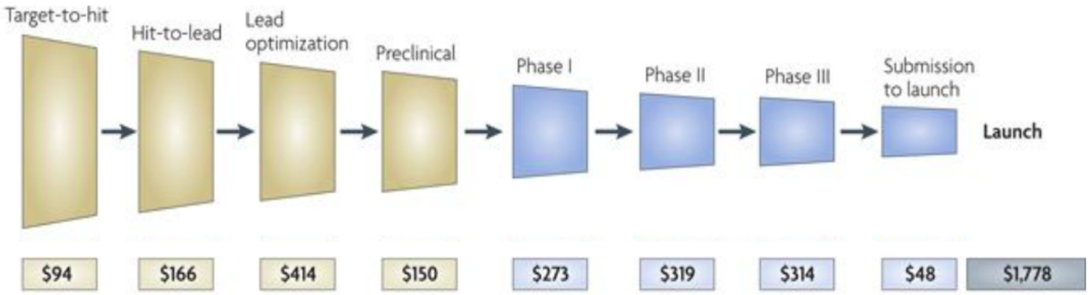

Figure 1.25: Flow chart of drug development showing costs at each stage of development.

The size of the trapezium at each stage indicates the higher number of projects required to be initiated for one successful project to reach launch with cost listed in US dollars. The phases of drug discovery are defined as follows: 'target-to-hit': an initial screen to identify compounds which perturb the target, 'hit-to-lead': high throughput process by which lead compounds which bind the target are generated, 'lead optimisation': leads compounds are optimised for favourable pharmacokinetic properties, 'preclinical': study to further understand pharmacokinetics of leads, potential side effects of the lead, and determination of dose in man. Phases I - III: standard phases for drugs in clinical trial (Figure source: [134]).

vast majority of proteins are not druggable with current technology. Many proteins which are known to play critical roles in disease aetiology and are expected to be efficacious drug targets are not pursued due to the technical difficulty perturbing those targets. Examples include c-Myc, K-Ras and BCL-2 [171], where perturbation is limited by druggability or fear of side effects, although efforts in this respect are ongoing [17] [104] [68].

Criticism of clinical pipelines that include drug targets for which genetic evidence does not exist is possibly beyond the point in many cases. Standout targets which are known to be involved in disease aetiology and for which genetic evidence exists are likely either already subject to therapeutic intervention, or undruggable with current technology. It could be argued that searching for genetic evidence linking genes to disease aetiology will chiefly lead to targets for which therapeutic agents have already been developed or targets for which therapeutic agents have not been developed due to the difficulty in perturbing them in a safe way.

Therefore, perhaps simply increasing the sample size of GWAS studies and performing more MR to find ever more 'causal' associations between risk factors and outcomes, or utilising whole genome sequencing techniques to replace microarray based technology will not yield to downstream advances in drug discovery and patient outcomes. As explained by Cook *et al* who performed the aforementioned review of the drug pipeline at AstraZeneca "industrialization of R&D" has lead to poor outcomes by encouraging "quantity-based metrics to drive productivity" [47]. Simply increasing the number of drug candidates in clinical development did not increase the number of successful outcomes. Optimising on quantity based metrics suppresses the ability of scientific investigators to perform research which answers real questions about biology and disease aetiology.

1.7 Previous GWAS of haematological phenotypes

The first GWAS of a haematological phenotype published in 2007 and 2008 explained nearly half the variation in fetal haemoglobin, identifying three major associated loci [116] [169] [149]. This phenotype was considered important due to the ability of increased fetal haemoglobin to ameliorate symptoms of sickle cell disease and β-thalassemia [148]. Since this time, GWAS of a number of other haematological phenotypes have been performed, including the count of the major types of blood cells (platelets, red blood cells, neutrophils, eosinophils, basophils, monocytes, and lymphocytes), and other measurements such as mean corpuscular volume (MCV), hematocrit (percentage of blood volume made of erythrocytes), red cell distribution width (RDW), and mean sphered cell volume (MSCV) [15, 66, 162, 42, 23, 61, 71]. These studies with progressively increasing sample sizes have identified more rare variants with higher effect sizes, and an increasing number of common variants. Furthermore, there has been a growing complement of research into the genetics of haematology in non-European ancestry individuals, including African American, Hispanic, and east Asian individuals [170]. All such analyses have identified yet greater numbers of genetic associations with haematological phenotypes. In tandem, there have been genetic studies of other phenotypes such as disease risk [34], blood plasma proteins [164], and blood or blood cell transcript levels [183]. Further analysis should not only search additional genetic space: rare variants, or common variants with weaker effect sizes to find new associations with haematological phenotypes, but also integrate information from GWAS of other phenotypes to offer a more complete picture of the influence of genetic variants. I aimed to address these challenges with my work, performing GWAS of previously unstudied blood cell phenotypes, and contributing to a large meta-analysis of full blood count (FBC) phenotypes in a meta-analysis of 563,085 individuals. Furthermore, I performed broad integration of my association results by colocalisation with disease risk, blood plasma proteome, and blood cell transcript GWAS results.

1.8 Aims and structure of thesis

Haematological cells are known to be important in the aetiology of disease including cardiovascular and immune disorders. The aim of this thesis is use statistical genetic analysis to derive biological insight into haematology and the role of blood cells in aetiology of disease. Using a GWAS analysis I identify genetic determinants which influence blood cell properties including cell count, volume, size, and other flow cytometric properties. I use these results in a hypothesis generating approach to identify genes and proteins which contribute to haematological cell function and evidence for linkage with disease risk. Finally, I perform MR analyses to identify causal relationships between haematological cell properties and disease. The analysis outlined in this thesis contributes to the following scientific outcomes:

1. Chapter 2: contribute to the largest ever GWAS of FBC haematological properties. The summary statistics generated by this analysis will be shared with the scientific community to enable further work.

2. Chapter 3: obtain and extract functionally relevant blood phenotypes (termed Sysmex parameters) from Sysmex XN-1000 analysers used to study participants in the INTERVAL cohort.

3. Chapter 3: adjust Sysmex parameters for environmental and technical variation thus increasing power to detect association signals.

4. Chapter 4: Identify novel genetic determinants of haematological function using Sysmex parameters which index functionally relevant haematological properties.

5. Chapter 5: better understand the biological implications of associations with haematological properties on gene expression and blood plasma protein concentration.

6. Chapter 5: identify which associations with haematological properties also influence disease risk. From these results generate biological hypotheses about the role of haematological cells in aetiology of disease.

7. Chapter 5: study the direction of causality between haematological properties and disease aetiology.

Chapter 2

Discovery of genetic associations with FBC haematological phenotypes

2.1 Introduction

In this chapter I discuss my analysis of haematological phenotypes to identify new associations with FBC haematological parameters. I perform conditional analysis on results of a GWAS of 28 FBC haematological parameters from the UK Biobank cohort and a meta-analysis of FBC phenotypes including 563,085 individuals. This is the largest GWAS of haematological phenotypes compared to the previous largest including 173,480 individuals by Astle *et al.*, [15]. The increased sample size enables identification of variants with lower effect sizes than previous possible and rare variants with lower minor allele frequency (MAF), the minimum MAF studied by Astle *et al.*, was 0.01% compared to 0.005% in this analysis.

In this chapter I will firstly expand on my introduction to GWAS (Section 1.5) and discuss challenges regarding multiple testing, population stratification and relatedness, genotype and phenotype quality control (QC). Secondly, I discuss a protocol for conditional analysis in meta-analysis and single cohort frameworks, in particular, my software enables the largest ever exact conditional analysis of GWAS results with 500,000 individuals which is a significant computational challenge. Finally, I present my results from the combined meta-analysis of the UK Biobank cohort and 26 haematological GWAS studies collected by Blood Cell Genetics Consortium (BCX). This meta-analysis allows a further increase in sample size to 563,085 individuals.

2.1.1 UK Biobank cohort

UK Biobank is a cohort of 500,000 individuals living in the United Kingdom recruited aged between 40 and 69 at the time of recruitment. A large number of phenotypic outcomes have been recorded including haematological measurements, lifestyle factors, biomarkers in urine, and magnetic resonance imaging (MRI) imaging in a subset of individuals. Participants have also been genotyped followed by phasing and imputation resulting in a total of 96 million variants [39]. The first release of genotype data from UK Biobank occurred in May 2015 including 150,000 individuals [15], this was followed in 2018 by a full release of 500,000 individuals of which 403,112 were utilised in my genetic analysis. Haematological phenotyping was performed with a Coulter full blood count analyser (Chapter 1.2) [39] generating 28 blood cell phenotypes derived from red blood cells (RBC), platelets, and white cells (Table 2.1.3). One primary advantage of the large UK Biobank cohort is the ability to model associations of rare genetic variants reliably. Following the precedent set by Astle et al., 2016, I excluded variants which did not have atleast 40 minor heterozygote alleles in the dataset [15], this results in a MAF threshold of 0.005% in comparison to 0.04% in the smaller INTERVAL study.

2.1.2 Blood Cell Genetics consortium

The BCX is an international collaboration of geneticists, haematologists, and statisticians with the goal of utilising genetic analyses to study and blood cell genetics. This large scale collaborative effort has allowed sharing of data from 26 blood cell GWAS studies allowing a meta-analysis of 14 blood cell phenotypes in 563,085 European ancestry individuals, work which I contributed to and results which I present in this chapter.

2.1.3 Full blood count haematological phenotypes

FBC reports are used routinely in a clinical setting being one of the most common laboratory tests [37]. The derivation of FBC parameters are described in detail in Section 1.2.1 and 1.2.2 from Coulter based impedance or fluorescence based measurements respectively. In Table 2.1 I provide a description of the haematological FBC parameters which were studied in analysis of phenotypes from the UK Biobank cohort or associated meta-analysis. Phenotypes are determined in four ways: measurement from gating and counting of cells following flow cytometry, impedance, light absorbance, or calculation from a combination of the aforementioned directly measured phenotypes.

Abbreviation	Unit	Description	Determination
PLT#	per nL	Count of platelets per unit volume of blood.	Impedance
MPV	fL	Mean volume of platelets.	(PCT/PLT#)10000
PDW	fL	The spread of the platelet volume distribution.	Impedance
PCT	%	Volume fraction of blood occupied by platelets.	Impedance
RBC#	per pL	Count of red blood cells per unit volume of blood.	Impedance
MCV	fL	Mean volume of red blood cells.	(HCT/RBC#)10
HCT	%	Volume fraction of blood occupied by red cells.	Impedance
MCH	pg	Average mass of hemoglobin per red cell.	(HGB/RBC#)10
MCHC	g/dL	Concentration of hemoglobin with respec to unit of volume occupied by red cells.	(HGB/HCT)100
HGB	g/dL	Concentration of hemoglobin with respect to unit of volume of blood.	Light absorbance
RDW	fL	Coefficient of variation of red cell volume distribution.	Impedance
MRV	fL	Mean volume of reticulocyte cells.	Impedance
RET#	pL	Count of reticulocytes per unit volume of blood.	(RET%RBC#)/100
RET%	%	Percentage of red blood cells that are reticulocytes.	Flow cytometry/impedance gates
IRF	-	Fraction of reticulocytes with high RNA content, as measured by light scatter.	HLSR#/RET#
HLSR#	per pL	Count of high RNA content (immature) reticulocytes per unit volume of blood.	(HLSR%RBC#)/100%
HLSR%	%	Immature reticulocyte count as a percentage of red blood cell count.	Flow cytometry/impedance gates
MONO#	per nL	Count of monocytes per unit volume of blood.	(MONO%WBC#)/100%
NEUT#	per nL	Count of neutrophils per unit volume of blood.	(NEUT%WBC#)/100%
EO#	per nL	Count of eosinophils per unit volume of blood.	(EO%WBC#)/100%
BASO#	per nL	Count of basophils per unit volume of blood.	(BASO%WBC#)/100%
LYMPH#	per nL	Aggregate count of lymphoid cells per unit volume of blood.	(LYMPH%WBC#)/100%
WBC#	per nL	Aggregate count of white cells per unit volume of blood.	Impedance
MONO%	%	Percentage of white cells that are monocytes.	Flow cytometry gates
NEUT%	%	Percentage of white cells that are neutrophils.	Flow cytometry gates
EO%	%	Percentage of white cells that are eosinophils.	Flow cytometry gates
BASO%	%	Percentage of white cells that are basophils.	Flow cytometry gates
LYMPH%	%	Percentage of white cells that are lymphocytes.	Flow cytometry gates

Table 2.1: Table of phenotypes studied from the UK Biobank cohort.
In total 28 haematological phenotypes derived from red blood cells, platelets, or white blood cells were studied by GWAS, conditional analysis, and further downstream analysis.

2.2 Methods

2.2.1 Genotyping and quality control

Collection and QC of genotype data for the UK Biobank cohort which is utilised in my analysis was collected by Bycroft *et al.,* [38], this includes genotyping, imputation, and QC on the UK Biobank cohort. Genotyping was performed on a total of 488,377 participants, a subset of 49,950 individuals were genotyped with a UK BiLEVE Axiom array containing 807,411 probes, and the remaining 438,427 participants were genotyped using a custom UK Biobank Axiom array with 825,927 probes [38]. The UK Biobank Axiom array was designed with additional markers to assay more variants, in particular insertion deletion variations, the two arrays share 95% of their probes. Probes were specifically selected to assay for both common and low frequency variants and also variants previously suggested to be important in other phenotypes such as autoimmune disease, cancer or blood phenotypes [38]. Blood samples were collected from participants visiting a UK Biobank assessment centre and samples shipped to Affymetrix for genotyping, sample retrieval and DNA extraction which are described in Welsh *et a.,* 2017 [175]. Of genotyped individuals 94% reported their ancestry as 'White' with the remaining 6% as Asian, Black, Chinese, mixed or unknown ancestry. Given the heterogeneous ancestry of the cohort many standard QC tools will not be effective for this dataset. For example, deviations from Hardy-Weinberg equilibrium (HWE), which in a cohort of homogeneous ancestry normally occurs due to poor genotyping will be expected in a cohort of mixed ancestry [38]. QC was divided into marker or probe based and sample based, I will discuss these separately.

Probe based quality control

In order to avoid the complications of heterogeneous ancestry, Bycroft *et al* performed probe based QC only on participants with European ancestry [38]. European ancestry individuals were identified by projecting samples on the two major PCs from the 1000 Genomes cohort [3] and selecting samples which fall in the European cluster (CEU) identified by sequencing of European ancestry individuals by the 1000 Genomes project. This analysis resulted in identification of 463,844 European ancestry individuals. The following tests were performed to identify and exclude poorly genotyped variants:

- Test for batch and plate effects to check if the allele frequency of genotyped variants significantly differs between genotype batch or sample plate.
- Test for departure from HWE on somatic chromosomes only to identify markers which have been genotyped poorly.
- Test to see if variants on chromosome X have a consistently different allele frequency

between males and females. This shows technical bias for that variant caused by the use of different calling algorithms on autosomal chromosomes between males and females.

- Test for array effects to exclude variants which show systematic differences depending on if they were genotyped by the UK Biobank or UK BiLEVE Axiom array.
- Two wells on each plate were dedicated to two control samples which were loaded on all plates. A discordance metric was designed [38] to exclude markers which are significantly discordant between controls across different wells.

Sample based quality control

Poor quality samples were excluded using a set of 605,876 high quality autosomal markers which were genotyped on both UK Biobank and UKBiLEVE arrays. High quality markers were defined such that they meet the following criteria:

- Marker is a SNP and not an insertion / deletion variant.
- Marker passed QC in all genotyped batches.
- Marker has a MAF across all samples higher than 0.01%.
- Not in the list of SNPs listed by Affymetrix to be affected by an artefact in a small subset of 300 individuals.

Tests to identify poor quality samples were performed using only the aforementioned high quality markers. Heterozygosity was calculated as the ratio of heterozygous genotypes divided by the total number of non-missing genotypes. Heterozygosity was then adjusted for population ancestry effects by regressing out the first six PCs calculated from the genotype data. Following this, individuals with with outlying heterozygosity or higher than 5% missing genotype data were excluded [38]. Heterozygosity outliers were identified using the R package *abberant* and a lambda value of 120, where lambda represents ratio of the standard deviations of outlying and normal individuals [20]. Furthermore, samples with mismatch between self reported and genotypic sex or with potential aneuploidy in sex chromosomes were flagged, these individuals are excluded from my analysis.

2.2.2 Phasing and imputation of variants

As previously described, imputation utilises LD structure from a reference population to enable identification of variants which have not been genotyped (Section 1.5.2). To ensure reliable imputation across all samples Bycroft *et al.,* phased a subset of genotyped variants, chosen to ensure a high proportion are good quality in all individuals (for a description of phasing and imputation see Section 1.5.2). Variants were excluded if they were not genotyped in both arrays, failed QC in more than one batch, had a MAF of smaller than 0.01% across samples, or had a missingness of greater than 5%. Phasing was performed using the SHAPEIT3 software [126].

The accuracy of imputation is also influenced by the reference panel, accuracy is higher if the reference panel contains a higher number of haplotypes and is a close match to the ancestry of the sample population [38]. Bycroft *et al.,* used a combination of HRC [167] and UK10K [46] reference panels, selected as they contain a high percentage of European ancestry individuals and a small subset of individuals with diverse ancestry, thus having a similar ancestry distribution to the UK Biobank cohort [38]. To perform imputation Bycroft *et al.,* modified the IMPUTE2 package [84] to perform only haploid imputation on the pre-phased samples, their new software was termed IMPUTE4 and executes the same hidden markov model (HMM) as IMPUTE2 and obtains identical results to IMPUTE2 [38]. Imputation estimated 92,693,895 variants in 487,442 individuals [38], it is this genotype dataset which forms the basis of further analysis in this chapter.

2.2.3 Adjustment of phenotype values for influencing covariates

The outcome of a GWAS study is highly dependent on the quality of both phenotype and genotype data used in the analysis. Technical and environmental factors which influence phenotype values increase variation in phenotype values and decrease power to detect associations. Therefore, phenotypes values are adjusted to account for the influence of environmental and technical factors. Technical variables include: seasonal effects, time dependent drift of equipment, sample decay, centre of sample collection, systematic differences in equipment, and systematic changes resulting from calibration of equipment. Adjustment is also made for participant environmental variables such as participant sex, menopause status, age, height, weight, and lifestyle factors including smoking, alcohol consumption, and diet. For more details regarding the adjustment of haematological phenotypes for technical and environmental covariates refer to Section 3.2.7 and 3.2.8.

Adjustment for participant phenotypes such as weight and menopause status will prevent detection genetic determinants which influence haematological traits mechanistically through these participant phenotypes. On the other hand, adjusting for participant phenotypes such as weight and menopause status which cause a large degree in variation of haematological trait values allows greater power to detect other genetic association signals. In the case of participant phenotypes such as sex or menopause status, adjustment is further required to ensure a similar distribution of phenotype values across all participants, an assumption required for the linear modelling of GWAS (Section 2.2.4).

2.2.4 Genome wide association study

As introduced in Section 1.5, a GWAS identifies genetic variants associated with changes in a phenotype. In the context of this study GWAS analysis was utilised to test for association of genetic variants in the UK Biobank cohort with recorded haematological measurements. The analysis for N individuals was modelled as follows:

$$\mathbb{E}[y] = \alpha + x_{pc}\beta_{pc} + x\beta + g \tag{2.1}$$

Where x_{pc} is a $(N \times 10)$ matrix including the top ten PCs, x is a $(N \times 1)$ genotype matrix coded as described in Section 1.5, and g models genetic effects which contribute to population stratification (Section 2.2.6) with a GRM matrix which is described in more detail in Section 2.2.7. At its basis, a GWAS study utilises a linear model or a LMM. Therefore, the primary assumptions of linear regression still apply. I will discuss these individually and explain how these assumptions could be broken in the context of a GWAS study.

Limited multicolinearity

I assume there is not a high correlation between the independent variables in the model, if this is the case it will lead to poor estimates for the effect size of the correlated independent variables. Therefore covariates should be selected to ensure they are not strongly correlated with genetic variants which are being tested. Furthermore, multicolinearity can also occur when identifying independent variants by multivariable analysis where multiple variants are put in the same linear model, this is discussed in more detail in Chapter 2.2.8.

Samples drawn from an independent distribution

A core assumption of a linear regression is that under the null hypothesis samples are independent and identically distributed given the covariates and the model. This assumption is broken if there are related individuals within the sample population (relatedness). Relatedness between can be estimated with the identity by descent (IBD) parameter and related samples filtered out. In addition, a GRM can be used as a random effect covariate to help account for relatedness and reduce the influence of this effect on the estimated variant effect sizes. This is discussed in more detail in Chapter 2.2.6.

Homoscedasticity

The homoscedasticity assumption states that the variance of the outcome variable is constant across the range of values for the independent variables (genotype and covariates). Given the relatively low proportion of variance in the outcome explained by any single variant or covariate being modelled, GWAS is not likely to break the homoscedasticity assumption.

Normal distribution of residuals

The residuals of the model are assumed to be normally distributed, in the context of my analysis the dependent variable (phenotype to be tested) is transformed to be normally distributed by an inverse-normal quantile transformation. This helps ensure normal distribution of residuals assuming there is not a serious deviation from the homoscedasticity assumption, which as previously explained is unlikely to be the case due to the low variance explained by any one variant being tested.

2.2.5 Multiple testing

In a frequentist paradigm, statistical tests of a null hypothesis are considered to be significant if the P-value of association falls below a threshold usually set to 5%. If the assumptions of the statistical model are correct this procedure will incorrectly reject the null hypothesis (false positive result) in 5% of cases. In a GWAS analysis each SNP is tested separately for association. Therefore I perform a very large number of parallel tests thus increasing the total number of false positives if I maintain the 5% P-value threshold. To reduce the number of false positive results, I adjusted the 5% P-value threshold by dividing the threshold by the number of effective independent tests, this is also known as a bonferroni correction. A GWAS analysis testing a large number of imputed variants will contain many variants which are highly correlated (Section 1.4). Therefore, the number of effective independent tests is less than the total number of variants tested for association. The number of effective independent tests has been found to vary greatly depending on the MAF threshold. Studies which include many rare variants will perform more independent tests as rare variants are less likely to be in LD with nearby variants (Table 2.2) [180]. My analysis utilised a MAF threshold of 0.005%, thus according to the simulations performed by Xu *et al.,* it is appropriate to use the same P-value threshold of 8.31×10^{-9} as that utilised by Astle *et al.,* 2016. This MAF threshold was set to ensure at least 40 minor alleles per variant in the sample population, this is inline the threshold set by other studies [15].

Alternatively it is possible to limit false positive findings using permutation to obtain an empirical null distribution. With this approach the phenotype is permuted to ensure that there is no true association between genotype and phenotype. All variants are tested with the permuted data and the smallest P-value is recorded. This shuffling and testing procedure is repeated to obtain an empirical null distribution of the smallest P-values calculated by chance [155]. P-values calculated from analysis of the un-shuffled dataset are

MAF Threshold	Range of Predicted Independent Tests	Range of Appropriate GWAS thresholds
0.05%	$2,746,888 - 4,306,272$	$1.16 \times 10^{-8} - 1.82 \times 10^{-8}$
0.01%	$4,412,096 - 6,019,458$	$1.13 \times 10^{-8} - 8.31 \times 10^{-9}$
0.005%	$5,933,687 - 8,547,380$	$5.85 \times 10^{-9} - 8.43 \times 10^{-9}$

Table 2.2: Range of appropriate GWAS thresholds.
Range of predicted GWAS thresholds depend on the MAF filter applied to the variants being studied, calculated by Xu et al., 2014 who assume participants of European ancestry [180].

then compared to this empirical null distribution. This method relies on the assumption that samples within the population are independent, this assumption can be broken by relatedness in the population [105]. The influence of relatedness can be reduced by filtering related individuals (Section 2.2.4), or using a permutation procedure which accounts for relatedness in the sample population [105]. I did not employ the permutation procedure due to the computational burden of calculating the null distribution.

2.2.6 Population stratification and relatedness

As described above, application of an appropriate P-value threshold will help limit false positive results from GWAS studies. Confounding factors may also lead to inflated false positive or false negative results. An example is population stratification, the presence of correlated ancestry within a stratum of the population. This is problematic as a stratum of the population may also have common environmental or genetic exposures. Population stratification makes it difficult to distinguish between a variant is associated with an outcome because of a genetically mediated mechanism, or because an allele of that variant happens to be common in a stratum of the population where the presence of the measured phenotypic outcome is common by chance. In addition to population stratification, closely related individuals in the sample population (cryptic relatedness) can lead to similar inflation in false positive or negative results. Linear models assume phenotype values are independent given association with the test variant, relatedness in the population cohort breaks this assumption (Section 1.5). There are multiple ways to account for population stratification and relatedness which are used in combination as no single method can sufficiently account for these confounding factors:

- Bycroft *et al,* filtered all samples to ensure to only include those of British ancestry [38]. This was performed using self reported ancestry of individuals and PCs generated from genotype data.

- Bycroft *et al,* removed all samples with a high degree of relatedness, as determined by IBD analysis [38].

- Regress the effect of clinic, a variable recording the location of blood donation against the phenotype values and use the residuals for GWAS analysis.

- Include PCs as covariates in the GWAS model.

- Use a LMM model which allows including a GRM random-effect covariate in the GWAS model (Section 2.2.7).

- Filter tested variants by MAF>0.005%.

Generation of PCs from genotype data provides information about participant ancestry and is used in calculation of relatedness, heterozygosity and other sample quality metrics. However, PCs should ideally be calculated from high-quality unrelated samples. To address this problem Bycroft *et al* performed two rounds of principal component analysis (PCA), firstly to identify unrelated high-quality samples, and secondly to compute PC adjusted heterozygosity and measures of relatedness [38].

First calculation of principal components

Firstly, Bycroft *et al* estimated kinship coefficients up to third degree of relation between all samples using the software package *KING* [109]. The kinship coefficient is calculated pairwise between samples and is the probability that two randomly sampled alleles are identical due to shared descent between the samples. A parent-offspring pair is expected to have a kinship coefficient of 0.5, decreasing by a multiple of 0.5 for every additional degree of relatedness, grandparent-grandchild pairs have a kinship coefficient of 1/8 [38]. From the set of kinship coefficients calculated across the samples a maximal set of unrelated individuals is calculated by pruning the relatedness graph of individuals using the *i-graph* (v1.0.1) package [10] in R. Samples were then excluded based on the following properties:

- Missing rate of autosomes > 0.02.

- Mismatch between inferred and self-reported sex.

- Not in the set of unrelated individuals.

SNPs were also excluded based on the following properties before being pruned into a set of independent markers by pairwise $r^2 > 0.1$:

- Missing rate > 0.015.

- MAF $> 0.01\%$.

- In regions of long-range LD such as regions of inversion, these are defined in Bycroft *et al* [38].

Application of filters to the genotype data resulted in a set of 147,551 SNPs and 406,257 samples which were used to compute PCs which are then utilised to generated adjusted QC metrics as described below. Generation of PCs was performed with *fastPCA* [65] and the top 8 components were extracted.

PC adjusted QC metrics

Heterozygosity and kinship metrics are sensitive to participant ancestry effects. For example recent admixture can lead to inflation of kinship and heterozygosity estimates. Therefore, kinship metrics are recalculated and heterozygosity using the first set of PCs. Exclusion of samples was recalculated from the adjusted kinship and heterozygosity metrics and a second round of PCs was calculated resulting in 40 components.

Kinship was recalculated as described but with a subset of SNPs which contribute loads less than 0.0003 in the first three PCs. Bycroft *et al* chose this threshold to meet the trade-off where inclusion of SNPs with high loads of contribution to the PCs will inflate the kinship matrix due to recent admixture, but a stringent threshold would lead to exclusion of too few SNPs would result in kinship estimates with high variance. In total 93,511 SNPs were used for final kinship inference and from the kinship estimates unrelated individuals identified as described above.

Heterozygosity is a ratio of genotypes which are not homogeneous in the population:

$$h = \frac{N_{nm} - N_{hom}}{N_{nm}} \tag{2.2}$$

Where N_{nm} is the number of non-missing genotypes and N_{hom} is the number of homozygous genotypes. However, heterozygosity is influenced by ancestry effects, therefore heterozygosity was adjusted for the top 6 PCs $x = (x_1, x_2, x_3, x_4, x_5, x_6)$, features which correlate with ancestry:

$$h(x) = h_0 + \beta(x) \tag{2.3}$$

Where h(x) is the raw heterozygosity, $\beta(x)$ is a function of the bias due to population structure, and h_0 which is the ancestry adjusted heterozygosity. $\beta(x)$ has a quadratic form and includes all linear and quadratic terms x_i and x_i^2 and cross terms $x_i x_j$ where i and j index over the six features in x [38]. The fitted value for ancestry adjusted heterozygosity $\hat{h_0}$ identified with ordinary least squares is utilised in sample QC to exclude samples with outlying heterozygosity. Samples with outlying heterozygosity are identified and excluded with the *aberrant* R package [21] from the logit transformed missing rate and ancestry adjusted heterozygosity $\hat{h_0}$ and a λ value of 120. *Aberrant* is a clustering algorithm which uses a mixture model to identify outlying data point where outliers are defined as those which have standard deviation (SD) λ times higher than that of the sample distribution which is inferred from the data by the *aberrant* package.

With the QC metrics adjusted for ancestry effects with PCs as described above, exclusion of samples were performed with the following criteria:

- Missing rate on autosomes > 0.02.
- Not in a set of unrelated individuals as identified by kinship estimates.
- In the list of outliers based on heterozygosity and missing rates.
- Mismatch between inferred and self-reported sex.

This exclusion resulted in a set of 147,606 SNPs and 407,599 samples which were used to compute a second round of PCs. The top 40 PCs were computed with *fastPCA* [65], the top 10 PCs were included in the LMM model to account for population ancestry effects (Section 2.2.7).

2.2.7 Linear mixed model GWAS

GWAS analyses are traditionally modelled using linear regression with a set of fixed effect independent variables such as PCs and the genotype of the variant (Section 1.5). Relatedness in the sample population can be represented with a GRM and included as a covariate (Section 2.2.6). The GRM covariate is necessarily a random effect as it will not have a consistent effect size across all individuals in the sample population, some individuals may have more relatedness to model than others [106]. Therefore, a fixed effects linear model will not suffice, inclusion of a GRM requires construction of a LMM. Using the BOLT-LMM application by Loh *et al.,* I constructed a LMM with GRM and the top ten PCs as covariates [106]:

$$y = x_{PCs}\beta_{PCs} + x\beta + u + e \tag{2.4}$$

Where y is a vector of phenotype values across all individuals N, x_{PCs} is a matrix where columns are one of ten PCs calculated from the genotype data ($N \times 10$), and x is a genotype vector for the variant being modelled as a fixed effect with coefficient β_{test}. The effect of relatedness is included in the model with the u term and environmental effects with $e \sim \mathcal{N}(0, \sigma_e^2 I)$. The effect of relatedness is modelled with a GRM matrix containing a subset of M_{GRM} SNPs across all N individuals: X_{GRM} ($N \times M_{GRM}$):

$$u = X_{GRM}\beta_{GRM} \tag{2.5}$$

Here β_{GRM} is a vector (length M_{GRM}) of random effect sizes drawn a normal distribution thus resulting in the requirement for a LMM to estimate these effects.

$$u \sim \mathcal{N}(0,\, \sigma_g^2 X_{GRM} X'_{GRM}/M_{GRM}) \tag{2.6}$$

It is evident that the SNP being tested and its proxies in the genotype vector x defined above should not be included in u, as this would lead to deflation of the test statistic for that SNP. This is effect is also known as 'proximal contamination' and is avoided in the BOLT-LMM implementation by removing all SNPs on the same chromosome as that being tested, this defines the 'M_{GRM} subset of SNPs' mentioned above.

A LMM analysis is computationally expensive to fit with time complexity $O(MN^2)$ or $O(M^2N)$, where N is the number of samples and M the number of tests [182]. The BOLT-LMM algorithm uses a series of approximations to achieve a $O(MN^{1.5})$ time complexity [182]. BOLT-LMM achieves this by estimating variance parameters using a stochastic approximation algorithm which avoids 'spectral decomposition', a time expensive operation where the matrix of genotype values is represented in terms of its eigenvalues and eigenvectors.

BOLT-LMM also allows modelling of associations with a non-infinitesimal prior on the SNP effect size coefficient β, which is in contrast to the infinitesimal prior used in standard LMM. The infinitesimal model assumes that all variants have effect sizes drawn from a Gaussian (or normal) distribution. In reality traits usually have a few associated variants with large effect sizes compared to many associations with smaller effect sizes. Therefore empirically, effect sizes are not Gaussian distributed. To enable reductions in computational time complexity BOLT-LMM uses a 'spike-and-slab' mixture of two Gaussian distributions. One to model the few causally associated variants with large effect sizes, and a second Gaussian distribution to model the higher number of more weakly associated variant [106].

2.2.8 Conditional analysis to identify independent associations

Due to LD between variants it is not clear from a GWAS analysis alone how many independent association signals are present in a locus. Conditional analysis determines how many association signals are present in a locus and which variants are the good statistical representatives for those signals. However, conditional analysis cannot determine which variants are casual for the association signal. Causality only be truly determined with a downstream follow up experiment, although statistical methods such as FINEMAP can build a credible set of variants which are likely to contain the causal variant.

Methods for conditional analysis are split between those which utilise summary statistics to perform analysis such as genome-wide complex trait analysis (GCTA) [181], and those which rely on availability of participant level genotype and phenotype data. Summary statistics based methods rely on LD calculated from a reference population resulting in less accurate results. Therefore, I utilise the multiple step-wise conditional analysis using phenotype and genotype data from the study population (Section 2.2.8.1).

2.2.8.1 Multiple stepwise conditional analysis algorithm

Stepwise multiple regression is an algorithm which identifies a parsimonious set of conditionally independent genetic variants, which represent the underlying association signals for each phenotype. Principally, a series of joint models are utilised to test for independence of genetic variants in association to the phenotype, constructed as follows:

$$y = x_{PCs}\beta + x_{\mathrm{VAR1}}\beta_{\mathrm{VAR1}} + x_{\mathrm{VAR2}}\beta_{\mathrm{VAR2}} + ...x_{\mathrm{VAR}n}\beta_{\mathrm{VAR}n} \tag{2.7}$$

Where y is the phenotype value across individuals, $x_{\mathrm{VAR}i}$ is a vector of genotypes for the ith variant being tested across individuals, and $\beta_{\mathrm{VAR}i}$ represents the effect size for that variant. If the estimated effect size for a particular variant has a P-value below that of genome-wide significance, that variant is considered to be independently significant from the other variants in the model. To avoid collinearity between predictors a variant is never put in the model if it has a squared correlation higher than 0.9 with any other variant in the model. In these cases it is assumed that the entering variant would not be independent from its correlated variant.

In order to find a parsimonious set of independent variants from all the genome-wide significant variants, a multiple-stepwise regression algorithm is run which tests many combinations of variants in a joint model. This is Efroymsons stepwise regression algorithm which has been shown to be convergent in most cases to a global minimum across the search space of variant combinations which would best explain the association signal [117]. To limit the search space and make the algorithm computationally tractable the genome is initially subsetted into blocks. Variants within those blocks are tested separately using the multiple-stepwise regression algorithm and independently associated variants are put forward into a larger chromosome wide pool on which a second multiple-stepwise regression algorithm is executed.

For each phenotype, I split the genome into blocks of variants associated at genome-wide significance threshold with the phenotype so that each block is not larger than 2,500 variants, and there are no genome-wide significant variants 5 Mb on either side of each block. I then performed the multiple stepwise conditional analysis procedure on genome-wide significant variants within each block (Fig. 2.1):

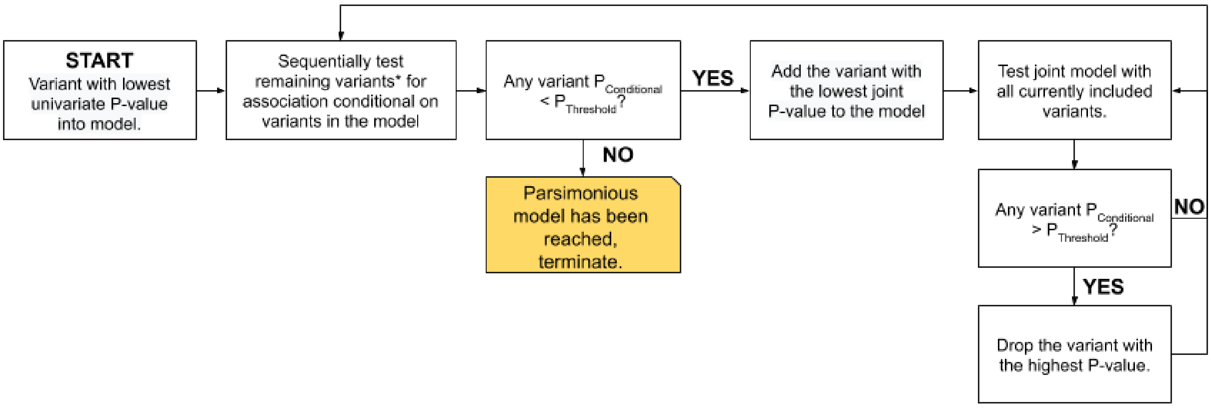
START
Variant with lowest univariate P-value into model.
Sequentially test remaining variants* for association conditional on variants in the model
Any variant $P_{Conditional}$ < $P_{Threshold}$?
YES
NO
Parsimonious model has been reached, terminate.
Add the variant with the lowest joint P-value to the model
Test joint model with all currently included variants.
Any variant $P_{Conditional}$ > $P_{Threshold}$?
NO
YES
Drop the variant with the highest P-value.

Figure 2.1: Flowchart of multiple-stepwise conditional analysis algorithm. Protocol for the conditional analysis algorithm where P is the joint P-value of the variant and $P_{Threshold}$ is the genome-wide significance threshold. The multiple-stepwise conditional analysis algorithm creates multiple joint models including different subsets of variants to identify a parsimonious set of independently associated variants. The algorithm begins by addition of variant with the smallest P-value and proceeds to test all other variants sequentially in a joint model. Here the algorithm alternates between 'adding' and 'dropping'. Adding: sequentially test all remaining variants in the joint model, then add the variant with lowest joint P-value. Dropping: run the joint model with all currently included variants and drop the variant with the lowest joint P-value if it is below the threshold. The algorithm will terminate if it cannot add or drop any variants to the model. * Uncorrelated variants are defined as those not in LD $r^2 > 0.9$ with any variants already in the joint model.

1. START: The variant with the lowest univariate P-value in the block is put in the linear model.

2. All other variants in the block are sequentially tested in the model if they are not $r^2 > 0.9$ with any other variant in the model.

3. Of the variants tested in Step 2, the variant with the lowest conditional P-value is put in the linear model.

4. The joint model is fitted again, and variants in the model with conditional P-value above the genome-wide association threshold are dropped. Dropped variants continue to be tested in Step 2 and could re-enter the model.

5. Repeat Step 4 until no more variants can be dropped from the model.

6. Iterate through Steps 2 - 4 until there are no other variants which can be added or dropped from the model.

Given the computational challenges of executing a multiple-stepwise regression algorithm on a dataset of up to 403,112 individuals an additional filter was utilised to reduce the number of iterations. If a variant reaches -$\log_{10}(P) < 2$ for association in the addition phase, this variant is excluded from all further analysis. Once a parsimonious set of conditionally significant variants is identified for each block, those variants are brought forward into a chromosome-wide multiple-stepwise conditional analysis procedure including all blocks. The resultant set of variants are labelled as 'conditionally significant' and are identified independently for each phenotype. Practical implementation of the block level and chromosome-wide conditional analysis procedure is explained further in Section 2.2.8.2.

2.2.8.2 Software pipeline for large-scale individual level conditional analysis

The storage and manipulation of genetic data, including up to 90 million variants in 403,112 individuals is a significant computational challenge. Therefore, I developed a software pipeline which subsets BGEN files and converts them into smaller and more easily readable HD5 format (Fig. 2.2) [8]. Genetic data for the UK Biobank cohort is stored in compressed BGEN format, reading this dataset into memory in order to perform computation is not feasible due to their large size. Therefore, I subsetted BGEN files into less compressed and more accessible file formats. Firstly, a GEN file for each block containing the dosage genotype data for all variants within that block, this was generated by subsetting the BGEN files using QCTOOL [16], following this the GEN files are converted to HD5 format [14]. GEN format is a less compressed alternative to BGEN which can be more easily read by software. HD5 is a common file format for which packages and libraries exist in *R* for the manipulation of this data. HD5 files were subsequently read by an R script using the *rhdf5* package. The R script fit linear regression models using the *fastLm* package iterating through the steps previously described (Section 2.2.8.1) and identified a parsimonious set of conditionally significant variants for each block. Once all blocks were analysed using the conditional analysis procedure, conditionally significant variants identified from blocks were combined to perform a genome-wide level conditional analysis. These variants were again extracted from the BGEN files into a separate GEN file (one for each chromosome) which was then converted to HD5 file and analysed by an R script which performed a final round of conditional analysis.

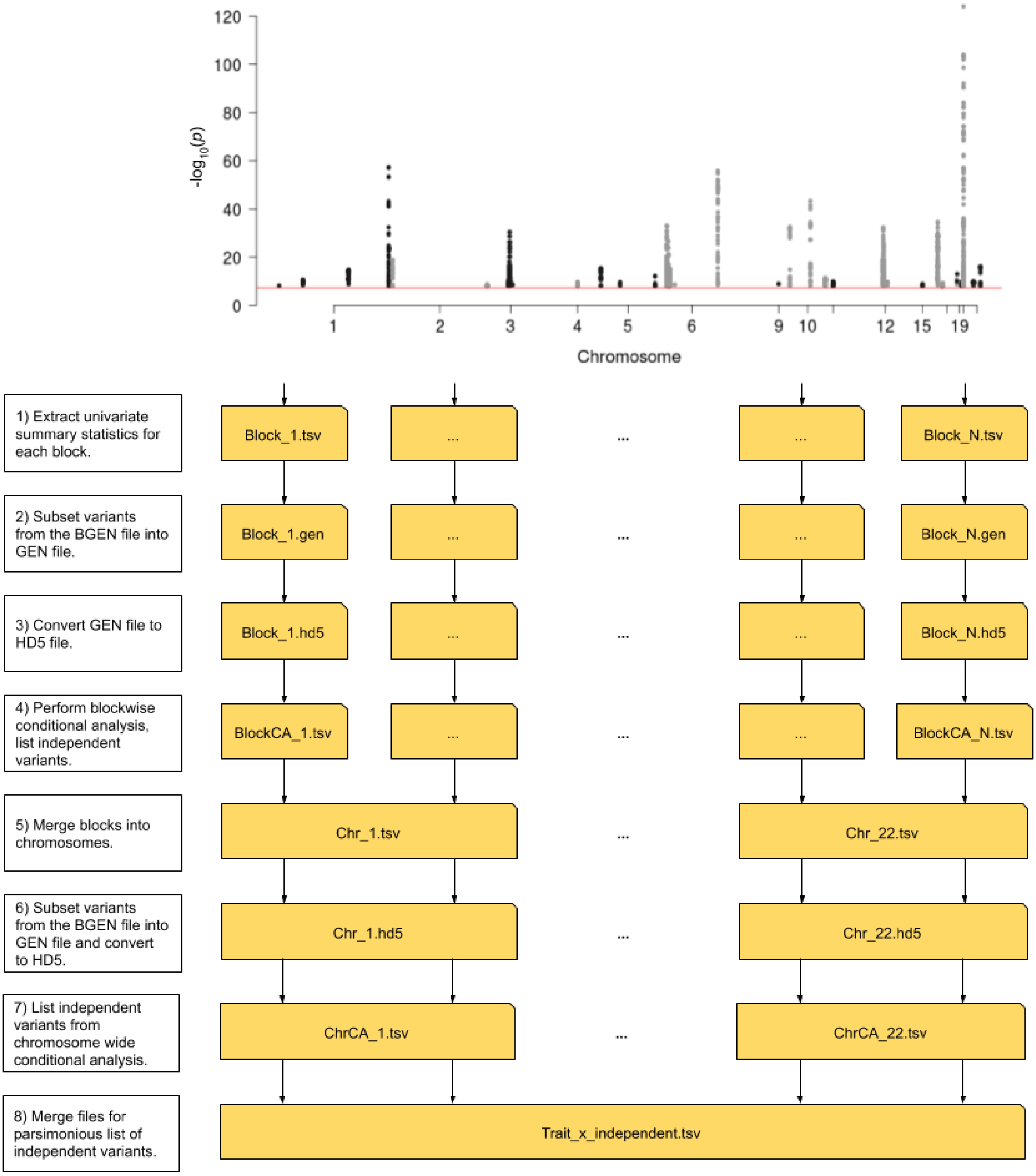
120
100
80
60
40
20
0
$-\log_{10}(p)$
1
2
3
4
5
6
9
10
12
15
19
Chromosome
1) Extract univariate summary statistics for each block.
2) Subset variants from the BGEN file into GEN file.
3) Convert GEN file to HD5 file.
4) Perform blockwise conditional analysis, list independent variants.
5) Merge blocks into chromosomes.
6) Subset variants from the BGEN file into GEN file and convert to HD5.
7) List independent variants from chromosome wide conditional analysis.
8) Merge files for parsimonious list of independent variants.
Block_1.tsv
...
Block_N.tsv
Block_1.gen
Block_N.gen
Block_1.hd5
Block_N.hd5
BlockCA_1.tsv
BlockCA_N.tsv
Chr_1.tsv
Chr_22.tsv
Chr_1.hd5
Chr_22.hd5
ChrCA_1.tsv
ChrCA_22.tsv
Trait_x_independent.tsv

Figure 2.2: Flowchart of software pipeline for conditional analysis algorithm. Initially univariate summary statistics for variants in each block were extracted, and the genotype data for these variants were extracted from the BGEN file into a GEN file and converted to HD5 format. The HD5 files were read by the multiple-stepwise conditional analysis script which generated a list of independent variants for each block. The independent variants for each block were collated chromosome wide and these variants were extracted again from the BGEN file into a GEN file and converted to HD5 format. A chromosome wide conditional analysis was executed generating a final list of conditionally significant variants.

2.2.9 Linkage disequilibrium grouping to identify total number of signals

Conditional analysis identifies the number of independently associated variants for each phenotype. LD grouping can identify the number of independent signals identified across multiple phenotypes. In order to assess the number of independent association signals across the 28 phenotypes studied in the analysis, I used a LD clumping procedure with a threshold of $r^2 > 0.8$ to assign conditionally significant variants to independent sets. Where the correlation between variants in the same set is high, but variants between sets have a correlation or LD lower than the $r^2 > 0.8$ threshold. The LD clumping protocol begins with generation of a correlation matrix between all conditionally significant variants using PLINK and execution of the following steps:

1. Iterate over conditionally significant variants in order of chromosome and position, terminate once iterated over all conditionally significant variants.

2. Populate set F with the conditionally significant variant chosen in Step 1) and all conditionally significant variants that are in LD r^2 0.8 or higher with this variant.

3. If none of the variants in set F are in a pre-existing LD set then create a new LD set with these variants, return to step 1).

4. If variants in set F exist in a single LD set which already exists, then assign all variants in set F to that LD set, return to step 1).

5. If variants in set F exist in more than one pre-existing LD sets, then merge all variants in those LD sets, and variants in set F, into a new larger LD set, return to step 1).

6. Terminate once the algorithm has sequentially iterated over all conditionally significant variants.

The result of this algorithm is assignment of all conditionally significant variants to LD sets based on a threshold of $r^2 > 0.8$ pairwise LD. It is important to distinguish between LD sets as a measure of distinct genetic association signals and determination of genetic signals based on locus or physical distance. Genetic loci are defined based on genomic location, however LD sets are defined based on variant LD. This means that two LD sets could be physically overlapping but distinct signals if the conditionally significant variants which constitute the LD sets are in low LD. This is often the case with rare variants which may be allocated to distinct LD set amongst a preexisting association signal of common variants in a locus.

Meta-analysis	UK Biobank Only
BASO#, EO#, HCT, HGB, LYMPH#, MCHC, MCH, MCV, MONO#, MPV, NEUT#, PLT#, RBC#, RDW	MRV, PDW, LYMPH%, RET# EO%, PCT, HLSR#, HLSR%, IRF, MSCV, NEUT%, RET% MONO%, BASO%

Table 2.3: Traits studied in the meta-analysis (including UK Biobank) and traits studied in UK Biobank only.
The meta-analysis (including the UK Biobank study) was restricted to fewer blood cell traits compared to the UK Biobank study. Of 28 traits studied in UK Biobank, 14 were analysed in the meta-analysis, this is due to the absence of traits within many studies consisting of the meta-analysis. A description of traits is presented in Table 2.1.3.

2.2.10 Meta-analysis conditional analysis

Initial analysis was performed on 28 blood cell phenotypes from the UK Biobank cohort, this was followed by a meta-analysis of 14 blood cell traits (Table 2.3) adding an additional 159,973 participants of European ancestry from 25 studies (Table 2.4). The purpose of the meta-analysis was to achieve a large sample size to discover additional association signals. Conditional analysis of the meta-analysis summary statistics data was performed using the GCTA-COJO algorithm which approximates a variance-covariance matrix for the genotype values due to the lack of individual level participant data in a meta-analysis setting (Section 2.2.11.2). Furthermore, GCTA-COJO uses a protocol alternative to the conditional analysis defined above (Section 2.3.1) which calculates joint models with subsets of all variants in the dataset not just the genome-wide significant subset as utilised in the multiple stepwise conditional analysis protocol. I hypothesised that many sub-threshold variants which are not genome-wide significant may be conditionally significant when placed in a joint model. To study this I performed analysis to test if variants identified by GCTA-COJO were jointly significant when accounting for conditionally significant variants already identified by conditional analysis of UK Biobank (Section 2.3.2). However, the GCTA-COJO algorithm relies on a reference sample approximation in the absence of exact genotype and phenotype data. If the approximation for the variance-covariance matrix of the genotype values made by GCTA is not accurate, this analysis may lead to false positive results (Section 2.2.11.1). Therefore, I tested the conditionally significant variants from the meta-analysis identified by application of GCTA-COJO, in the UK Biobank population using a joint model utilising individual level genotype and phenotype data. Firstly, I calculated the LD between conditionally independent associations as identified by GCTA-COJO to identify highly correlated variants which are proposed as independent by GCTA-COJO. Secondly, I performed statistical tests to identify which variants obtained from GCTA-COJO conditional analysis of the meta-analysis summary statistics are independent from conditionally significant variants identified from the UK Biobank cohort alone.

Study Name	Samples in analysis	References (PMID)
Airwave	13,113	25194498
BioME	802	21573225
CaPS	1181	1999035 21043637 11395343
CHD	3249	1669507
Estonia (Chip)	22417	28031487
Estonia (WGS)	2242	28031487
Framingham Heart Study	6451	14025561
FINCAVAS	924	16515696
GERA EA Chip	53822	26092716
GERA AFR Chip	1363	26092716
GERA LAT Chip	1504	26092716
Health2006	3177	23615486
Health2008	752	22587629
Health2010	1474	25113139
INTERVAL	39260	28941948
MESA (EA)	1172	12397006
MHIphase1	1991	24777453
MHIphase2	3436	24777453
RS-I	1455	29064009
RS-II	1269	29064009
RS-III	2378	29064009
SHIP	3159	20167617
SHIP-TREND	940	20167617
UKBB_EA	456785	30305743
WHI (EA)	17682	24777453
YFS	1889	-

Table 2.4: Studies which contribute their summary statistics to the meta-analysis of FBC haematological phenotypes

In total 26 studies contributed to the meta-analysis of haematological traits, the largest study was UK Biobank which contributed 456,786 individuals with the smallest being BioME with 802 individuals. The primary prublication for the 'YFS' has not yet been published and a reference is not present in the table.

2.2.11 GCTA conditional and joint association

The GCTA-COJO package allows the construction of useful joint models using GWAS summary data. This method is effective in meta-analyses where individual level genotype and phenotype data cannot collated due to data sharing restrictions. Conditional and joint analysis using summary-level statistics are estimated using LD from a reference sample which is similar to the population cohort. The method for estimation of joint effects of multiple SNPs using GWAS summary statistics and a reference population is described by Yang et al [181], who then extend their derivation to perform step-wise conditional analysis. Their methodology is implemented in the GCTA-COJO software package. There are limitations associated with the GCTA-COJO approach due to the assumption that allele frequencies and LD between variants in the reference sample population are the same as the study sample population. Due to the much larger population size of our meta-analysis compared to any available reference population we are able to model the association of very rare variants which are not likely to be well represented in the reference population. Therefore, a reference LD set using 100,000 individuals from the UK Biobank dataset was generated.

In the next section, I explain the implications of meta-analysis in providing additional insight beyond the GWAS of the UK Biobank cohort alone (Section 2.2.11.1). Following this, I work through the approximation underlying the GCTA-COJO package (Section 2.2.11.2) and finally, describe the GCTA-COJO conditional analysis protocol (Section 2.2.11.3).

2.2.11.1 Implications for the meta-analysis of blood cell traits

The GCTA-COJO algorithm is very similar to the forward and backward stepwise regression as defined previously in my analysis of Sysmex parameters (Section 2.2.8.1). The primary difference being the ability to test all SNPs across the genome due to an approximation which avoids direct use of the genotype and phenotype data (Eqn. 2.18). This is compared to the previously described multiple-stepwise regression conditional analysis algorithm which only tests genome-wide significant variants for conditional significance (Section 2.2.8.1). It can be hypothesised that the benefit of testing all variants, rather than just the genome-wide significant subset of variants is that many sub-threshold signals could reach conditional significance when analysed in a joint model. This is because, as independent variants are added to the model, a greater proportion of variance in the phenotype is explained, thus increasing the calculated statistical significance of variants in the model. I attempt to identify such variants and present the results below in Section 2.3.2. Furthermore, GCTA-COJO allows conditional analysis on the meta-analysis results providing an extra 159,973 samples to the conditional analysis which were not accessible in the UK Biobank conditional analysis alone. However, the GCTA-COJO approximation

(Eqn. 2.18) relies on the reference population providing a good estimate of LD between variants in the sample population. The reference population must be large enough to reliably calculate LD of tested rare variants and must also have close ancestry to the sample population.

2.2.11.2 Estimation of joint effects by Yang *et al.*, 2012

Yang *et al,* begin by considering a multi-SNP model as follows:

$$y = Xb + e \tag{2.8}$$

Where $y = \{y_i\}$, a $n \times 1$ vector of phenotype values. $X = \{x_{ij}\}$, a $n \times N$ genotype matrix where each element of X is a function of the allele count of SNP or variant j in individual i. The number of individuals and number of SNPs is n and N respectively, $e = \{e_j\}$, a $N \times 1$ vector of residuals, and finally $b = \{b_j\}$, a $N \times 1$ vector of SNP effects. Yang *et al* also centre phenotype values (y) to removing the requirement for an intercept term. Given (Eqn. 2.8) joint effects can be estimated using the least-squared approach as follows:

$$\hat{b} = (X^TX)^{-1}X^Ty, var(\hat{b}) = \sigma_J^2(X^TX)^{-1} \tag{2.9}$$

Where σ_J^2 represents the residual variance of the joint model (the capital J subscript represents the joint model). However, in many cases individual level phenotype or genotype data is not accessible, therefore the data y and X are unavailable. GCTA approximates $\hat{b}$ and $var(\hat{b})$ with a reference population and a set of univariate summary statistics defined as follows where each variant (indexed by j) is tested for association with the phenotype:

$$y = x_j\beta_j + e_j \tag{2.10}$$

Where x_j is the column j in X and β_j is the effect of variant j on the phenotype, this is the marginal effect of SNP or variant j on the phenotype, as before e_j is the remaining residual. As we are not taking into account covariances between the variants (because Equation 2.10 is not a joint model), the diagonal of $X'X$ is represented by diagonal matrix D, where $D_j = \sum_i^n x_{ij}^2$, such that the marginal effects for multiple variants is represented as follows:

$$\hat{\beta} = D^{-1}X^Ty, var(\hat{\beta}) = \sigma_M^2D^{-1} \tag{2.11}$$

Here σ_M^2 is the residual variance in the univariate model (Eqn. 2.10). Of course, obtaining D_j requires individual level genotype data, which is unavailable in this context, an approximation to obtain D_j is described later (Eqn 2.17). I have previously discussed the drawbacks of single-SNP or univariate analysis at length, namely LD between variants resulting in proxies of a variant causally associated with a change in phenotype also appearing significantly associated (Section 1.4), thus making it difficult to determine the true number of associated signals in a locus. Joint models of SNPs can identify the total number of independent signals in a locus. It is such joint models which are approximated by GCTA-COJO.

It was previously shown by Yang et al., that $X^T y = D\hat{\beta}$ (Eqn. 2.11). Therefore *Yang et al* re-wrote the joint model (Eqn. 2.9) in terms of $D\hat{\beta}$ which includes $\hat{\beta}$ obtainable in the summary data shared from univariate GWAS analysis (Eqn. 2.11). An approximation for D, the diagonal matrix of $X^T X$ using a reference population is described later (Eqn. 2.17).

$$\hat{b} = (X^T X)^{-1} D\hat{\beta}, var(\hat{b}) = \sigma_J^2 (X^T X)^{-1} \tag{2.12}$$

The coefficient of determination of a multiple or joint regression model (represented by subscript J) is a calculation of the total phenotypic variance explained by the covariates (in this case the SNPs) is as follows:

$$R_J^2 = \frac{\hat{b}^T X^T y}{y^T y} = \frac{\hat{b}^T D\hat{\beta}}{y^T y} \tag{2.13}$$

Which can be used to calculate residual variance of the joint model $\hat{\sigma}_J^2$ and residual variance of the single-SNP analysis $\hat{\sigma}_{M(j)}^2$ as follows:

$$\hat{\sigma}_J^2 = \frac{(1 - R_J^2) y^T y}{n - N} = \frac{y^T y - \hat{b}^T D\hat{\beta}}{n - N} \tag{2.14}$$

$$\hat{\sigma}_{M(j)}^2 = \frac{y^T y - D_j \hat{\beta}_j^2}{n - 1} \tag{2.15}$$

Where N is the number of variants and n the number of samples. Given the squared standard error of the estimate of the effect size for each variant (indexed by j) is $S_j^2 = \hat{\sigma}_{M(j)}^2 / D_j$ we deduce that: $y'y = D_j S_j^2 (n-1) + D_j \hat{\beta}_j^2$, and this can be calculated for each SNP from data readily available in GWAS summary data.

In order to perform the calculations listed above in the absence of individual level data, the matrix D must be approximated. D is a diagonal matrix of the variance-covariance matrix of variant genotypes $X'X$. As shown by Yang et al., 2012 variances can be calculated from allele frequencies and covariances from LD in a suitable reference population. The genotype matrix of the reference sample is defined as $W = \{w_{ij}\}$ where j is an index over the SNPs in the reference sample of size m. Furthermore, D_w is the diagonal matrix of $W'W$ with $D_{W(j)} = \sum_i^m w_{ij}^2$ defined using the allele frequencies available from GWAS summary data. Assuming the reference sample is drawn from the same population as the meta-analysis cohort, LD correlations between variants will be approximately similar. This approximation is defined as follows, with the LD between two variants i and j in the genotype matrix X defined on the left hand side:

$$\frac{\sum_j^n x_{ij}x_{ik}}{\sqrt{\sum_i^n x_{ij}^2 \sum_i^n x_{ik}^2}} \approx \frac{\sum_j^n w_{ij}w_{ik}}{\sqrt{\sum_i^n w_{ij}^2 \sum_i^n w_{ik}^2}} \tag{2.16}$$

Following, we can denote the variance-covariance matrix $X'X$ to be approximately equal to B defined as such:

$$B_{jk} \approx \sqrt{\frac{D_j D_k}{D_{W(j)} D_{W(k)}}} \sum_i^m w_{ij}w_{ik} \tag{2.17}$$

It was defined above that $D_j = \sum_i^n x_{ij}^2$, as x_ij is not available allele frequencies from the summary statistics are used: $D_j \approx 2p_j(1 - p_j)n$. Therefore, Yang et al., 2012 can approximate the joint analysis of multiple SNPs as follows, with $\tilde{b}$ represents the approximated joint effect sizes of SNPs in the model:

$$\tilde{b} = B^{-1}D\hat{\beta}, var(\tilde{b}) = \sigma_J^2 B^{-1} \tag{2.18}$$

Yang et al., make further adjustments to their estimates of the variance-covariance matrix to account for changes in sample size between different SNPs within the meta-analysis. Conditional analysis is performed as an extension of the joint model defined above, using the same approximation to allow estimation in the absence of individual level genotype and phenotype data, for more details see [181].

2.2.11.3 GCTA-COJO conditional analysis

Yang et al., apply the step-wise conditional analysis algorithm across the entire genome using their approximate joint regression models as follows:

1. Begin the model including the most significant SNP across the entire genome.
2. Calculate the P-values of all remaining SNPs conditional on SNPs already in the model. Do not test SNPs which are highly correlated with SNPs already in the model. Highly correlated SNPs are defined as those with an r^2 threshold usually set to 0.9.
3. Select the SNP from Step 2 with the lowest conditional P-value, assuming this is below the set significance threshold.
4. Fit all SNPs in the model in a single joint model to test for significance dropping variants which are below the significance threshold.

5. Repeat Steps 2, 3, and 4 until no SNPs are added or removed from the model.
6. Perform a final joint model test to ensure all are conditionally significantly associated with the phenotype.

This algorithm results in sequential testing of all variants in the joint model. Similar to the multiple stepwise regression procedure, this type of conditional analysis can only identify sentinel or representative variants and cannot necessarily identify the variant mechanistically causal for the signals of interest. However, determination of the number of independent association signals in a locus could inform setting of priors in a fine-mapping procedure which can determine credible sets of variants likely to be causal for the association signal.

2.3 Results

2.3.1 Conditional analysis of UK Biobank identifies novel signals

A primary motivation for my work was to identify new genetic associations with full blood count haematological measurements. Multiple-stepwise conditional analysis identified 16,900 associations across 7,122 LD sets representing independent association signals in 23 chromosomes (Section 2.2.9). This is almost three-fold greater than the 6,736 associations across 2,706 LD sets identified in the previous largest GWAS of the same phenotypes in 173,480 individuals performed by Astle *et al.*, 2016 [15]. I identified 7,122 novel sets, defined as those which do not contain any variants which are in LD $r^2 > 0.8$ with any variants identified by Astle *et al* (Fig. 2.3). A full list of conditionally significant associations and their comparison with Astle *et al.,* can be found in Section A.2.

2.3.2 Distinct associations identified by GCTA-COJO

The meta-analysis includes 23 studies (including UK Biobank) with a total sample size of 563,085 individuals, versus 403,112 individuals in UK Biobank alone. The higher power afforded by the meta-analysis allows discovery of additional association signals. Because a large proportion of samples in the meta-analysis are from the UK Biobank cohort, I sought to determine which of the association signals identified by the meta-analysis are distinct to those already identified by analysis of the UK Biobank cohort alone (Section 2.3.1). To test this I used genotype and phenotype data from the UK Biobank cohort to create exact joint models instead of relying on conditional analysis that was performed on the meta-analysis summary statistics using the GCTA-COJO module. From the phenotypes, I regressed out the effect of all conditionally independent variants identified from analysis of the UK

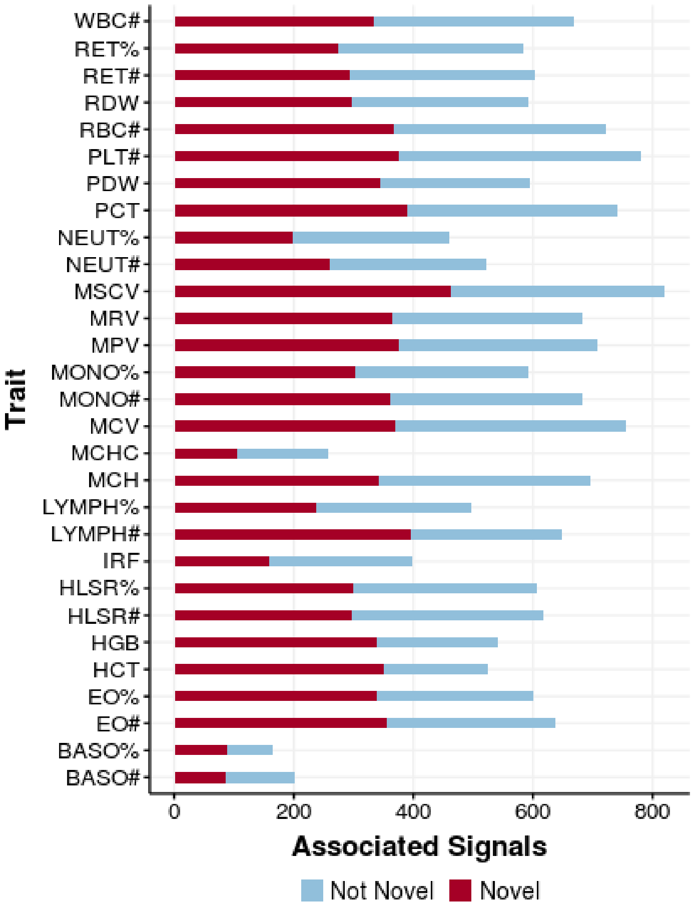

Figure 2.3: Bar plot showing number of novel signals identified per trait. A plot of conditionally independent associations across the 28 studied haematological traits assigned as 'not-novel' if they exist in an LD clump which contains conditionally significant variants with higher than $r^2 > 0.8$ pairwise LD with any trait identified by Astle *et al.*, 2016 [15] or assigned as 'novel' otherwise. Of all associations 52.0% as designated as novel compared to Astle *et al.*, 2016 [15]. This result shows that my conditional analysis of up to 403,112 individuals in UK Biobank makes new findings compared to the previous largest study of the same haematological phenotypes [15].

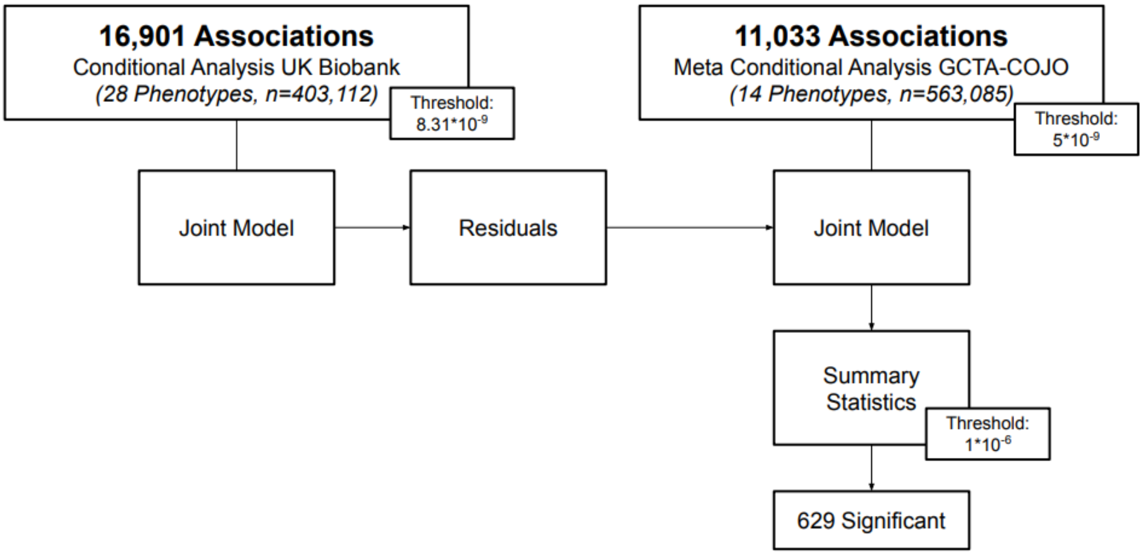

Figure 2.4: Flowchart showing meta-analysis conditional analysis pipeline. Variants identified by GCTA-COJO following meta-analysis are put in a joint model to test if they are significantly associated given residuals calculated by regressing out conditionally significant variants identified by conditional analysis of the UK Biobank dataset. The threshold of significance was made less stringent account for the smaller sample size in the UK Biobank dataset compared to the meta-analysis.

Biobank cohort alone. Then, I created a joint model of conditionally independent variants identified from the GCTA-COJO analysis to test for their significance of association beyond what was already discovered from the UK Biobank cohort alone (Fig. 2.4).

This analysis identified 626 associations which are significantly associated with their respective phenotypes distinct to the conditionally significant variants identified from analysis of UK Biobank. Notably, most of these associations are near to previously defined UK Biobank conditionally significant variants. Only 193 variants exist more than 1MB from a UK Biobank conditionally significant variant (Fig. 2.5). Labelling each of the 626 distinct meta-analysis associations with its best LD proxy from the UK Biobank variants shows that most are in very low LD with previously discovered UK Biobank variants. From the total of 626 distinct meta-analysis associations, all but one association is in LD $r^2 > 0.8$ and 454 associations in LD $r^2 < 0.02$ with the set of UK Biobank conditionally independent associations for their respective traits (Fig. 2.6).

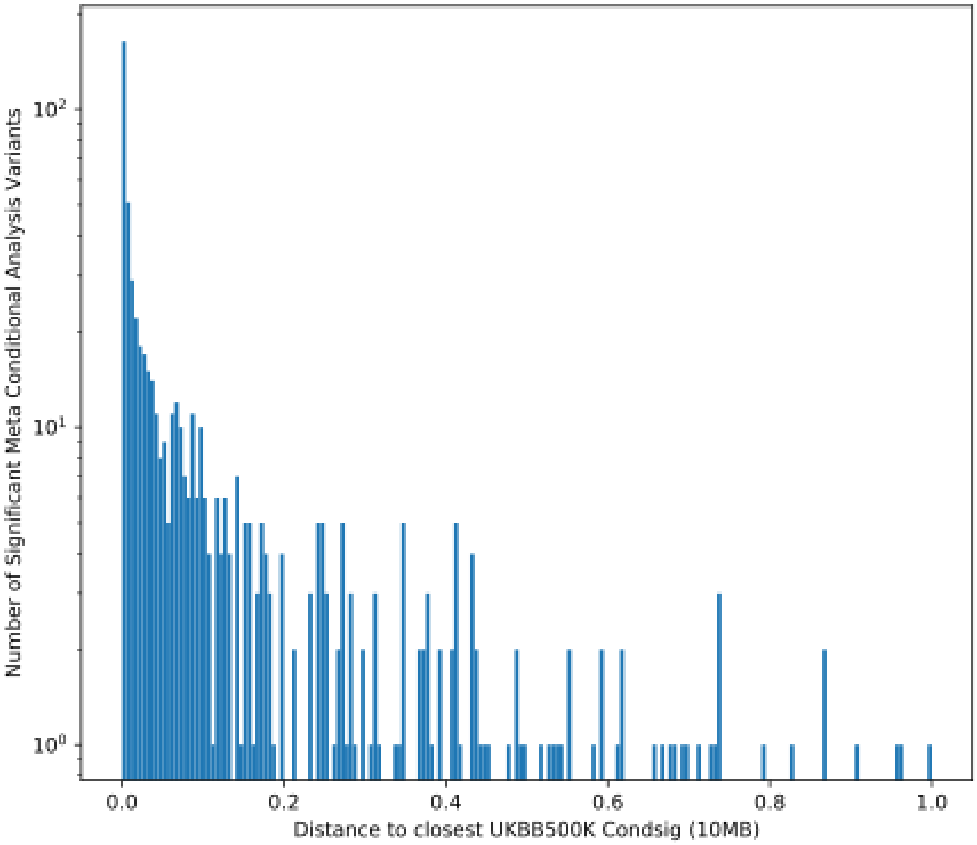

Figure 2.5: Histogram showing the absolute distance of distinct meta-analysis variants to the nearest UK Biobank association.
The 629 new associations identified by meta-analysis of 23 studies by GCTA-COJO are plotted in a histogram (y axis log scaled) depending on their distance to the nearest variant associated with the same trait identified by conditional analysis of up to 403,112 individuals in UK Biobank. Only 13 identified significant associations exist more than 10MB (range x axis) and only 193 less than 1 MB from the nearest UK Biobank conditionally significant variant. This plot shows that discovery of new association signals is more likely to be near to already discovered signals.

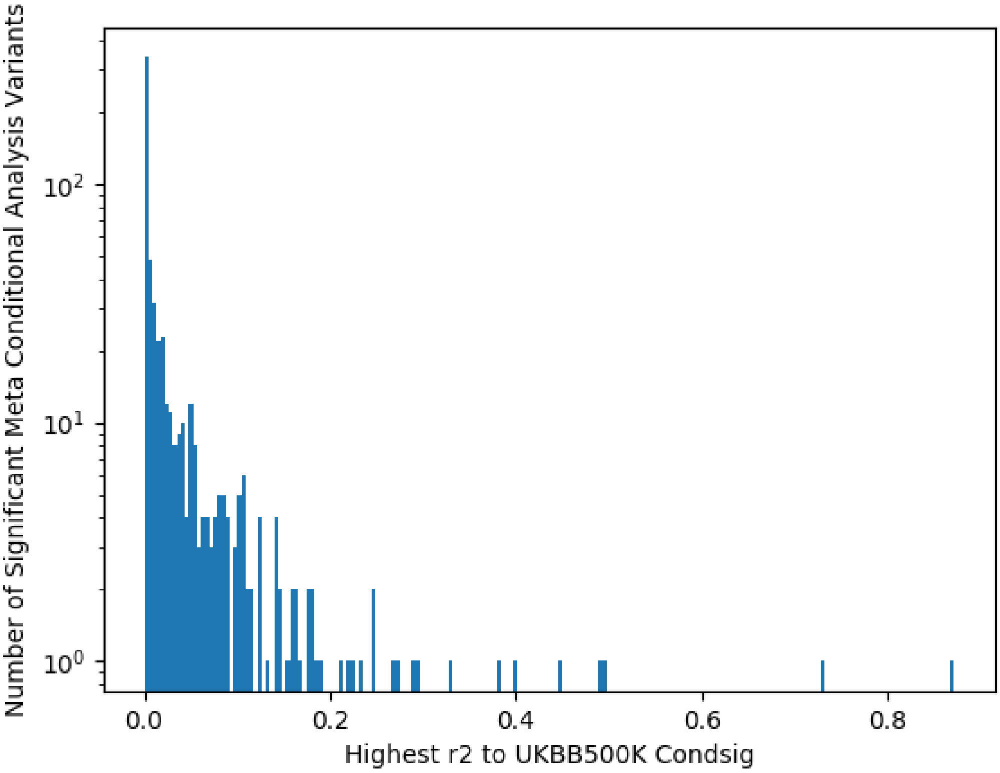

Figure 2.6: Histogram showing the highest pairwise LD meta-analysis variants to UK Biobank associations.

The 629 new associations identified by meta-analysis of 23 studies by GCTA-COJO are plotted in a histogram (y axis log scaled) depending on their highest LD to any variant associated with the same trait identified by conditional analysis of up to 403,112 individuals in UK Biobank. Only 1 identified significant association has an LD $r^2 > 0.8$ (range X axis) and 454 have $r^2 < 0.02$. This result shows that almost all of the 629 new associations are in very low LD with previously discovered signals.

As previously explained the phenotype is adjusted by regressing out the effect of conditionally significant variants identified by UK Biobank. Following this I classified associations identified by GCTA-COJO of meta-analysis results into four categories:

- Not significant (9,834): The P-value for the test of genetic association with the adjusted phenotype is not significant given a P-value threshold of -$\log_{10}(P) > 6$.
- Threshold (507): This variant is below the UK Biobank significance threshold -$\log_{10}(P) > 8.08$, but above the meta-analysis threshold -$\log_{10}(P) > 8.30$, and loosened conditional meta-analysis significance threshold -$\log_{10}(P) > 6$.
- Jumper (68): The tested variant is not associated at genome-wide significant threshold in the meta-analysis summary statistics, but becomes significant in the joint model.
- Faller (47): This variant is associated with a lower effect size in the joint model than it is in the univariate meta-analysis summary statistics.
- Other (7): Does not fit into any of the previous categories, these variants exist within the loosened conditional meta-analysis significance threshold.

These results show that most distinct associations identified by the meta-analysis are identified due to the larger sample size afforded in the meta-analysis GWAS. Intriguingly, I identified a set of 1,227 associations identified by GCTA-COJO which are in LD $r^2 > 0.9$ with associations with the same trait. Given the high LD between these associations it is unlikely that they are truly distinct signals and could represent false positive signals caused by the reference sample approximation used by GCTA-COJO.

2.4 Summary

In this chapter I discuss my contribution to the largest GWAS of complete blood count (CBC) haematological phenotypes including 14 phenotypes in a meta-analysis of 563,085 individuals across 23 studies, and a single cohort analysis of 28 phenotypes in 403,112 individuals. My subsequent exact conditional analysis using a stepwise regression protocol of the UK Biobank cohort identified 16,900 associations across 7,112 LD sets of which 5,106 are novel signals compared to the previously largest GWAS of haematological phenotypes performed by Astle *et al.*, [15]. I also present an additional set of 629 associations identified from the meta-analysis study determined to be independent from signals in the aforementioned 7,112 LD sets by exact meta-analysis conditional analysis.

Chapter 3

Data collection and quality control of cytometry parameters

3.1 Introduction

I previously discussed the ability of flow cytometry based methods to make measurements from blood cells including SSC, SFL, and FSC (Section 1.2.2). In this chapter I show that these additional parameters of blood cells can be clinically and functionally informative and identify association signals distinct to the results of FBC GWAS which largely measure blood count and volume phenotypes (Section 2.1.3). I show that GWAS of these novel and functionally relevant phenotypes is able to identify distinct associations compared to the largest previous GWAS of FBC phenotypes by Astle *et al.*, with a roughly 3.8 times larger sample size [15]. The results generated in this chapter will inform further work in Chapter 4, to better understand the genetic architecture of the functional properties of blood cells by performing the first ever GWAS and downstream analysis of SSC, SFL, and FSC blood cell parameters.

I begin by discussing the INTERVAL study and the extraction of blood phenotypes from the Sysmex XN-1000 analyser, following this I discuss genotyping of participants, including quality control and variant imputation. Then I provide a review of the literature regarding the clinical and functional relevance of Sysmex parameters. Finally, I describe phenotype and genotype QC of data which is prepared for the GWAS and downstream analyses in described Chapter 4.

3.1.1 INTERVAL study

INTERVAL is a randomised clinical trial of 45,263 healthy blood donors who have been assigned to blood donation schedules of 8, 10, or 12-weeks for male participants and 12, 14, or 16-weeks for female participants [11]. Donors were recruited from 25 National Health

Service Blood and Transplant (NHSBT) static donor centres across England [121].

The purpose of the trial was to identify factors which influence the safe interval for blood donation. In order to assess donor health, and identify factors which may predict the safe optimum interval for blood donation a range of haematological and genetic measurements have been performed. Participants also answer an extensive questionnaire to assess their mental and physical health, lifestyle, and diet, prior to beginning the trial and finally upon completion after two years of participation. Informed consent was obtained from all participants, and the INTERVAL study was approved by Cambridge East Research Ethics Committee. Participants who have subsequently withdrawn from the study were removed at the time of analysis. I studied phenotype data collected at baseline of the trial from participants prior to being randomised to a blood donation schedule. However in the case of missing measurement data from the second time point measured upon completion of the trial was used (Section 3.1.2).

3.1.2 Extraction of extended Sysmex cytometry traits

The INTERVAL study used two Sysmex XN-1000 haematological analysers, the hardware and software for the analysers were provided by the company Sysmex. Following flow cytometry of a sample, the Sysmex analyser internally computes thousands of variables which are used to produce a haematological report. In my study, I was not only interested in studying parameters which are output in the standard Sysmex haematological report, but also accessing variables which the software calculates that are not directly accessible to the user. In many cases these hidden variables have become accessible in later versions of the Sysmex software, examples include RE-LYMP and AS-LYMP measures of the reactive and antibody synthesising sub-population of lymphocytes respectively. For each analysis performed, the Sysmex analyser also saves an encrypted binary file containing data calculated internally by the software which contains the aforementioned hidden variables and is used to produce the haematological report accessible to the user. Following negotiation with Sysmex we were given a decryption key to access this encrypted binary file, and I searched through the thousands of variables to identify relevant markers of haematological function. I communicated directly with representatives of Sysmex (J. Saker) to confirm the variables I had identified represented the parameters which I was intending to analyse. Parameters which were extracted in this way include reactive lymphocytes (RE-LYMP), and all SSC, SFL, FSC, and distribution width parameters associated with eosinophil, basophil, red blood, or platelet cells. Unfortunately, for these extracted parameters, the first 20% of binary files were overwritten during the course of the original study, this meant that baseline time point measurements for some Sysmex parameters were missing. In order to address this I replaced this measurements with final time point measurement at two years whenever possible. This is reflected by the $I(i)$

variable when adjusting for environmental factors (Section 3.2.8). A full description of all the Sysmex parameters extracted and analysed in my study is available at Chapter 3.2.2.

3.1.3 Genotyping and quality control

Genotyping of participants and subsequent imputation and QC was performed in a previous study [15]. I will give an overview of sample collection, genotyping, imputation, and QC steps, for more detail please refer to Astle *et al.,* 2016 [15].

3.1.3.1 Sample collection and genotyping

Blood samples were shipped in buffy coat aliquots to LGC Genomics (UK) where DNA was extracted using a Kleargene method. Subsequently samples were shipped to Affymetrix (Santa Clara, California, USA) in 96-well barcoded wells including two empty wells for Affymetrix control samples. Genotyping was performed with an Affymetrix GeneTitan Multi-Channel Instrument implementing the Affymetrix Axiom 2.0 Assay Automated Workflow. A customised UK BIOBANK Affymetrix Axiom array with 820,967 probes was employed to assay SNPs and short insertion deletion variations (Section 3.1.3.2). Genotypes were called using Affymetrix Power Tools software which implements the Axiom GT1 algorithm [15]. For more details please refer to the paper by Astle *et al.,* 2016 who performed and described this work in their study of the genetics of standard haematological measurements.

3.1.3.2 Genotyping array

Genotyping for this study utilised a customised UK Biobank Affymetrix Axiom array with 845,485 probesets assaying 820,967 single nucleotide variants (SNVs) and short insertion/deletions. More probes exist than number of genotyped SNVs, this is because some SNV exist in regions with high sequence homology making these variants difficult to genotype. In such cases multiple probes are sometimes designed to target these variations. Probesets were selected to target variants which meet the following criteria:

- A genome wide scaffold which provides good coverage of common (MAF<5%) or low frequency variation (1%<MAF<5%) in the European population. This is the basis of later imputation (Section 3.1.3.3).

- Rare variants which exist in exomic regions and are likely to have transcriptional consequences (non-synonymous, splice altering, truncating).

- Rare variations known to increase the risk of cardiac disease, cancer, or listed on the human gene mutation database (HGMD) database.

The genome wide scaffold was designed using a custom algorithm based on the 1000 Genomes CEU population, European 1000 Genomes population, and a further tranche of variants selected to boost imputation of low frequency variants [15].

3.1.3.3 Quality control and variant imputation

Imputation allows variants which have not been directly genotyped to be inferred from genotyped variants. Imputation is fundamentally based on the principle of LD between regions of the genotype (Section 1.4). To enable reliable imputation, it is important to establish an initial set of high quality genotyped variants which serve as a scaffold. To ensure this is the case genotyped variants were filtered based on the following criteria:

- HWE filter of P-value $< 5 \times 10^{-6}$.
- Call rate filter of 99% in batches where the variant did not fail.
- Variant must have passed in atleast eight of the ten batches where it was genotyped.
- All monomorphic, non-autosomal, and multi-allelic variants were removed.
- Variants must have a MAF>0.04%.

Based on the following criteria QC was also performed on samples, excluding those which met any one of the following criteria:

- Samples with more than 10% sample contamination [89].
- Samples with 3 - 10% sample contamination and ten or more first or second-degree relatives in the study.
- Duplicate samples.
- Samples with heterozygosity three standard deviations from the mean.
- Samples who were missing or had mismatch sex information.
- Samples who are not of European ancestry.
- Samples with poor genotype signal intensity (<82%) and low call rate (<97%) based on roughly 20,000 probes to be known of high quality.

Following genotype and sample based QC, genotyped data was phased using SHAPEIT3 with chunks of 5,000 variants and an overlap of 250 variants per chunk [126]. The genotype data was used for imputation using IMPUTE3 [85] in chunks of 2mb with a 250kb buffer region. A combined 1000 Genomes Phase 3-UK10K panel was used for both phasing and imputation. Imputation was implemented using the positional Burrows-Wheeler transform (PBWT) imputation algorithm [58] on the Sanger imputation server. No imputation quality or variant frequency filters were applied at this stage, in total 87,696,910 variants were imputed or genotyped [15]. Following imputation of the INTERVAL genotype data, Astle *et al.,* 2016 used whole-exome sequencing (WES) data for 3,976 INTERVAL study participants who were also included in the imputation dataset to confirm a very high concordance between sequencing and genome imputation. Concordance ranged with a median precision from 99.5% for common variants (MAF>5%) to 98.5% for rare variants (MAF<1%). In addition IMPUTE3 will also calculate an info score representing the certainty in the value of imputed variant. INFO score ranges from 0 to 1 representing poorly to well imputed variants respectively. The INFO score metric represents a ratio for calculation of effective sample size, at total sample size N, an imputed SNP will have effective sample size dependent on it's INFO score: $INFO * N$. GWAS studies tend to use an INFO score filter of around 0.3 or 0.4 [38] [186] [15], with an INFO score of greater than 0.4 being defined as well-imputed [118]. I utilised an INFO score filter of 0.4 leaving 26.8 million variants to be tested in my GWAS.

3.2 Methods

3.2.1 Blood sample collection

Participants in the INTERVAL study were assessed for a range of health and lifestyle factors which included a questionnaire and a full blood haematological analysis at recruitment and upon completion of the trial, this data is the subject of my analysis. Samples for haematological analysis were collected from a pouch attached to the standard blood collection unit during blood donation. Blood samples were collected in, 3 ml or 6 ml ethylenediaminetetraacetic acid (EDTA), and 6 ml serum tubes. After collection the tubes were inverted three times and transported at ambient temperature to NHSBT sample holding sites at Manchester, Colindale (London), and Bristol. EDTA is not noted to cause a difference in the mean of activation effects of white cells, although some differences in cytokine production have been observed [103]. Following collection, samples were transported to the UK Biocentre facility in Stockport, UK for analysis. Almost all samples (98%) were processed within 48 hours of venipuncture, and 72% within 24 hours [15]. Analysis of samples was performed with a Sysmex XN-1000 analyser from which haematological indices utilised in my analysis were derived.

3.2.2 Sysmex flow cytometry channels

The Sysmex XN blood cell analyser is a modular system for analysing blood samples containing 4 flow cytometry, one electrical impedance channel, and a photometric channel for haemoglobin measurements. Each sample is aliquoted into six channels responsible for measuring different haematological cell types and properties. Reactants particular to each channel lyse and stain aliquots to target cell types and to allow polymethine and oxazine dyes to bind to nucleic acids in organelles and the nucleus. The composition of the reagents is kept commercially confidential by Sysmex. The Sysmex haematological analyser contains the following channels of measurement:

- The WNR (white count and nucleated red blood cells) channel detects nucleated red blood cells (NRBC) and provide an accurate count of basophil cells.

- The WDF (white cell differential channel by fluorescence) channel is responsible for analysis of white blood cells including counting lymphocytes, neutrophils, and monocyte cells.

- The PLT-F channel performs platelet measurements, including counting mature and immature platelets.

- The RET (reticulocyte) channel provides measures of erythropoiesis including reticulocyte count and reticulocyte maturity.

- The Photometric channel analyses haemoglobin content in red blood cells using sodium lauryl sulphate staining agent.

- The Impedance channel measures passing of blood cells through an aperture between electrodes allows measurement of red cell phenotypes such as cell volume, cell count, haematocrit, and platelet count.

Within each channel light from a stable red diode laser is incident on cells passing in single file through the flow cytometer. This results in three bands of light being recorded from each cell. Two light sources of different wavelength are separated by a dichroic mirror to obtain SSC and SFL light intensity measurements. Light passing through cells is recorded in a separate third direction FSC (Fig. 3.1). Sysmex parameters are recorded from these three sources of light. I also assess distribution width of each of the measurements (SSC, SFL, FSC) for each cell type in the sample. Distribution width is determined as the width of each peak of light at 20% of the peak height. Each cell is plotted on a three dimensional 'scattergram' based on its SSC, SFL, FSC values (Fig. 3.4). Cells are classified into cell types using thresholds based on the position of cells on the three dimensional scattergram. The Sysmex parameters studied in my analysis

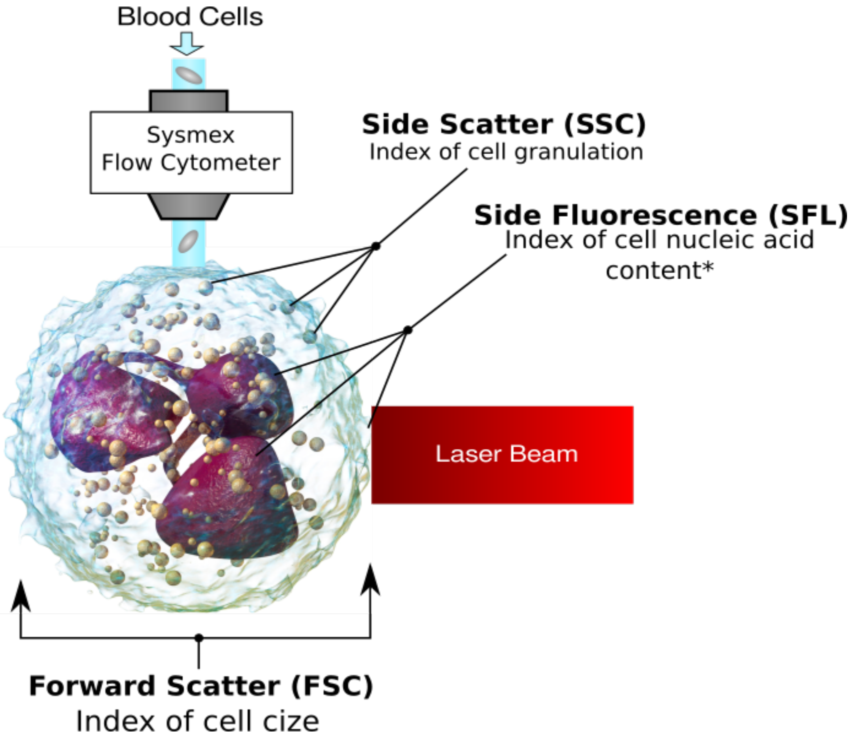

Figure 3.1: Flow cytometry of haematological cells.
Cells flow single file through the Sysmex flow cytometry channel and are hit by a laser beam, light is scattered or fluoresced by dyes in the cell and this is recorded resulting in three readings (SSC, SFL, and FSC) per cell. * SFL is an index of nucleic acid content also influenced by membrane composition of cells which affects the rate of absorption of nucleic acid staining dye into the cell (Figure adapted from [45]).

involve the median position and distribution width of each cell type in the SSC, SFL, FSC axis, and also counts of cells which are outliers from the primary clusters of cells in their respective scattergram, examples of which are given below. These cellular properties have been shown to be relevant for diagnosis of disease and to measure important physiological properties (Section 3.2.3).

3.2.2.1 PLT-F Channel Parameters

The PLT-F channel provides a count of mature and immature platelets (Table 3.1). Cells are lysed and stained using a fluorescent nucleic acid marker which also helps remove interfering particles such as RBC fragments [165] (Fig. 3.2).

3.2.2.2 RET Channel Parameters

The RET channel is responsible for measuring circulating red blood cell maturity and measures of reticulocyte and red blood cell haemoglobin content (Table 3.2). As reticulocytes mature they lose their nucleus and their cellular nucleic acid content drops. The SFL measurement is used to classify reticulocytes as high fluorescence reticulocytes

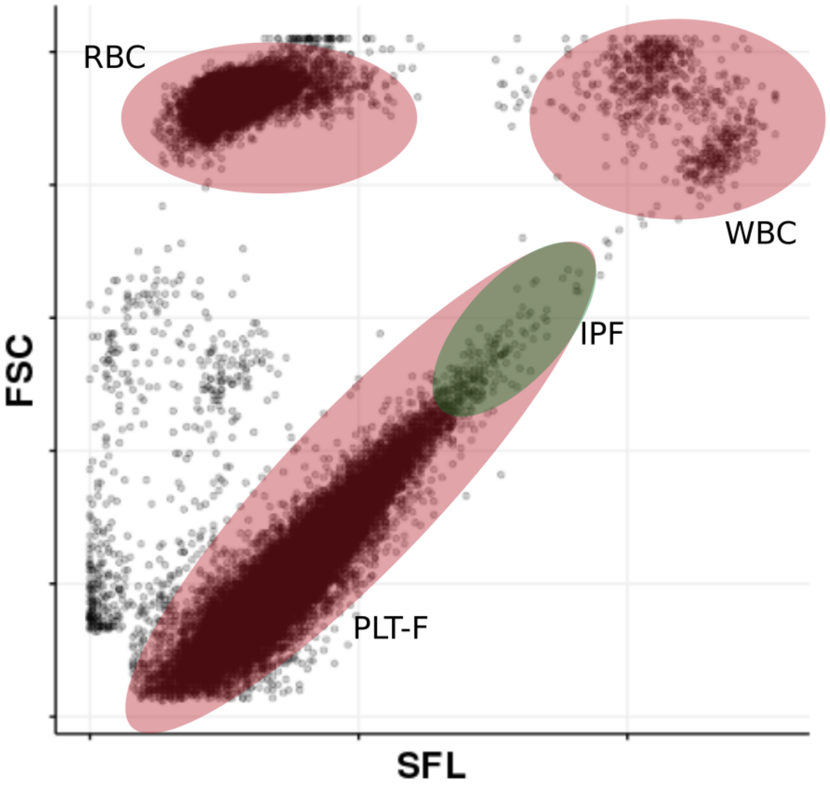

Figure 3.2: PLT-F channel scattergram.
The PLT-F channel with SFL and FSC measurements plotted, the channel allows good separation within the platelet population allowing identification of the immature platelet fraction (IPF) and highly fluorescent immature platelet fraction (H-IPF) fraction. Furthermore, the PLT-F channel provides good separation between platelet, white blood cell, and red blood cell types.

(HFR), medium fluorescence reticulocytes (MFR) to low fluorescence reticulocytes (LFR), representing increasing stages of maturity, eventually cells join the red blood cell (RBC) population (Fig 3.3). Cells are perforated by a lysis reagent which allows a fluorescent marker to pass into the cell staining nucleic acids [165].

3.2.2.3 WDF Channel Parameters

The WDF channel is responsible for counting of lymphocyte, neutrophil, and monocyte cells. This channel also measures counts of cells which are outlying from their primary cluster of cells, such as immature granulocytes (IG), RE-LYMP, and antibody synthesising lymphocytes (AS-LYMP) (Table 3.3). In total 29 of the 63 Sysmex parameters studied in my analysis originate from the WDF channel. Lysis reagents perforate the cell membrane allowing a fluorescent dye to stain nucleic acids. Reagents are designed to keep the cells largely intact and to ensure that the rate of fluorescent dye uptake is proportional to nucleic acid content which is recorded by the SFL measurement [165]. Immature cells such

Sysmex Parameter	Description
H-IPF	Highly Fluorescent Immature Platelet Fraction: A measure of highly immature platelets
IPF & IPF#	Immature Platelet Fraction Immature Platelet count
P-LCR	Platelet Large Cell Ratio
PLT-F-SSC, SFL, FSC, -DW	Platelet Scatter and distribution width

Table 3.1: PLT-F channel parameters.
Table of the 10 parameters studied from the PLT-F channel, DW represents the intra-individual distribution width for each of the SSC, SFL, and FSC measurements.

Sysmex Parameter	Description
IRF	Immature reticulocyte fraction
Hyper-He	Hyper haemoglobinised red cells
RBC-He	Red blood cell haemoglobin
RET-SFL, FSC	Reticulocyte scatter parameters
RET-He	Reticulocyte haemoglobin
LFR, MFR, HFR	Low, Medium, High fluorescent reticulocytes
IRF-FSC	Immature reticulocyte fraction forward scatter
RET-RBC SSC, SFL and -DW	Red blood cell RET scatter
Delta-He	Difference between RBC and Reticulocyte haemoglobin

Table 3.2: RET channel parameters.
I analysed 14 parameters from the RET channel, where '-DW' represents distribution width for each of the SSC and SFL measurements. The measurements are derived from the red blood cell and reticulocyte cell types.

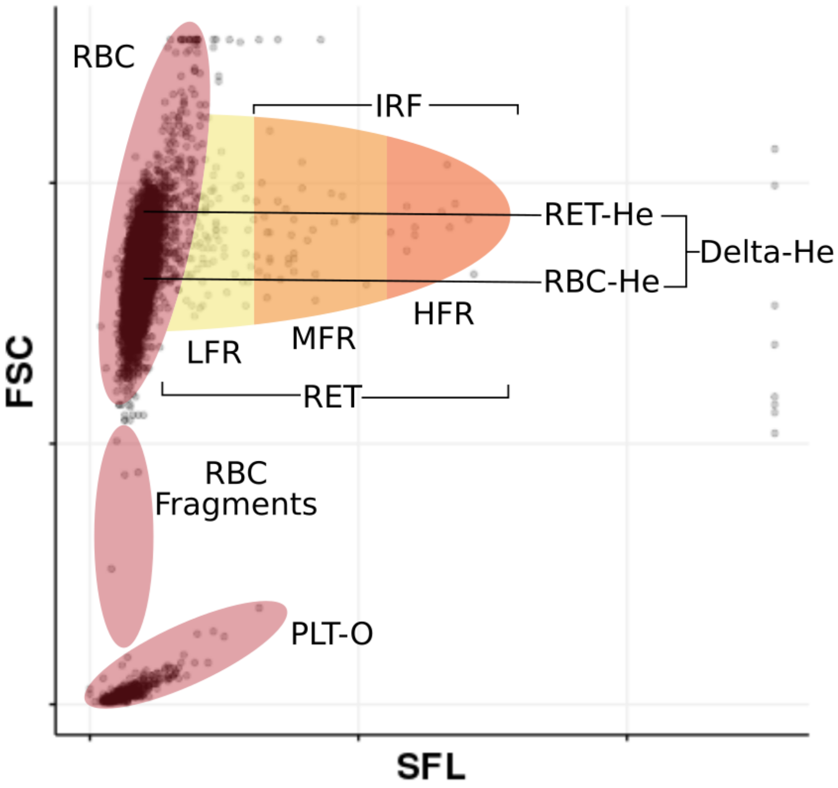

Figure 3.3: RET channel scattergram.
The RET scattergram, each cell is plotted with SFL and FSC on the x and y axes respectively, the z axis (SSC) is hidden. Measurements are assigned per cell-type based on the median position of the cell cloud in each axis of the scattergram. Cell types are highlighted in the three axes, RBC: red blood cells, PLT-O: optical observation of platelet count (compared to impedance), LFR: low fluorescence reticulocytes, MFR: medium fluorescence reticulocytes, HFR: high fluorescence reticulocytes, RBC Fragments: fragments of red blood cells this cluster is not analysed.

Sysmex Parameter	Description
NE-SSC, SFL, FSC, and -DW	Neutrophil scatter parameters
EO-SSC, SFL, FSC, and -DW	Eosinophil scatter parameters
MO-SSC, SFL, FSC, and -DW	Monocyte scatter parameters
LY-SSC, SFL, FSC, and -DW	Lymphocyte scatter parameters
RE-LYMP# (count) RE-LYMP% (of white blood cells) RE-LYMP(L)% (of lymphocytes)	Reactive lymphocytes
IG# (count) IG% (of granulocytes)	Immature Granulocytes

Table 3.3: WDF channel parameters

I analysed 29 parameters from the WDF scattergram, '-DW' represent the intra-individual distribution width for each of the SSC, SFL, and FSC measurements. Parameters are stratified across the five primary white blood cell types and also immature granulocytes which largely consist of immature neutrophils. Parameters are split across neutrophil, eosinophil, monocyte, lymphocyte, and immature granulocyte (mostly consisting of immature neutrophils) cell types.

Sysmex Parameter	Description
BASO-SFL, FSC, and -DW	Basophil Scatter

Table 3.4: WNR channel parameters.

Four Basophil phenotypes were studied from the WNR channel, DW represents distribution width parameters for SFL and FSC.

as immature granulocyte count (IG#) or highly activated cells such as (RE-LYMP or AS-LYMP) tend to contain higher levels of nucleic acids and have higher SFL measurements (Fig. 3.4). Cells are also separated according to their SSC and FSC measurements, which are measures of cell structure which are indicative of cell granularity and cell size respectively (Fig. 3.4).

3.2.2.4 WNR Channel Parameters

The WNR channel only measures SFL and FSC, and is responsible for counting nucleated red blood cells, total white blood cell count, and basophil count (Table 3.4). The 'ghost' proportion of the scattergram is occupied by contaminants such as air bubbles and lipids. Similar to the WDF channel, cells are processed in a two stage reaction, starting with perforation of the white cell membranes keeping the cells largely intact, following this, nucleic acids in the cells are stained with a fluorescent dye to allow detection by the flow cytometer (Fig. 3.5) [165]. The cell membrane of NRBC is lysed and the nuclei are stained [165]. NRBCs exist in circulation of newborn infants and can be diagnostic of myelodysplastic syndromes when observed in adults [153], given that participants in the INTERVAL study are healthy adults (18 years of age or older) these cells are not observed in the scattergram (Fig. 3.5).

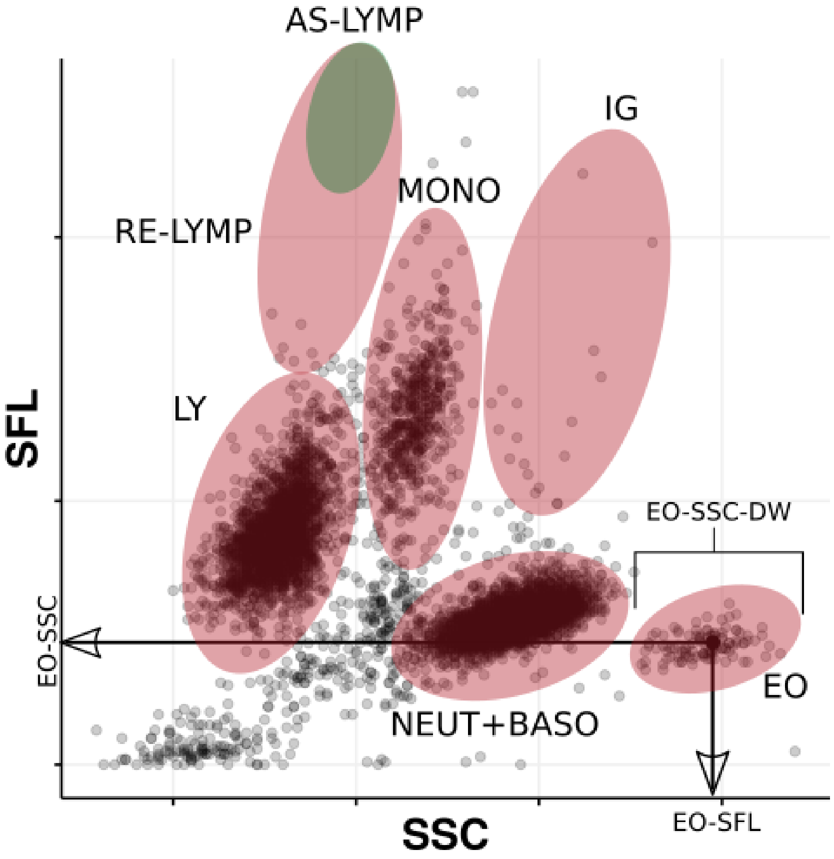

Figure 3.4: WDF channel scattergram
A plot of the WDF scattergram from an individual in the INTERVAL study, each cell is plotted with SSC and SFL on the x and y axis respectively, and the z axis (FSC) hidden. Measurements are assigned per cell-type based on the median position of the cell cloud in each axis of the scattergram. Cell types are highlighted in the three axes, LY: lymphocytes, RE-LYMP: reactive lymphocytes, AS-LYMP: antibody synthesising lymphocytes, MONO: monocytes, IG: immature granulocytes, NEUT: neutrophils, BASO: basophils, EO: eosinophils.

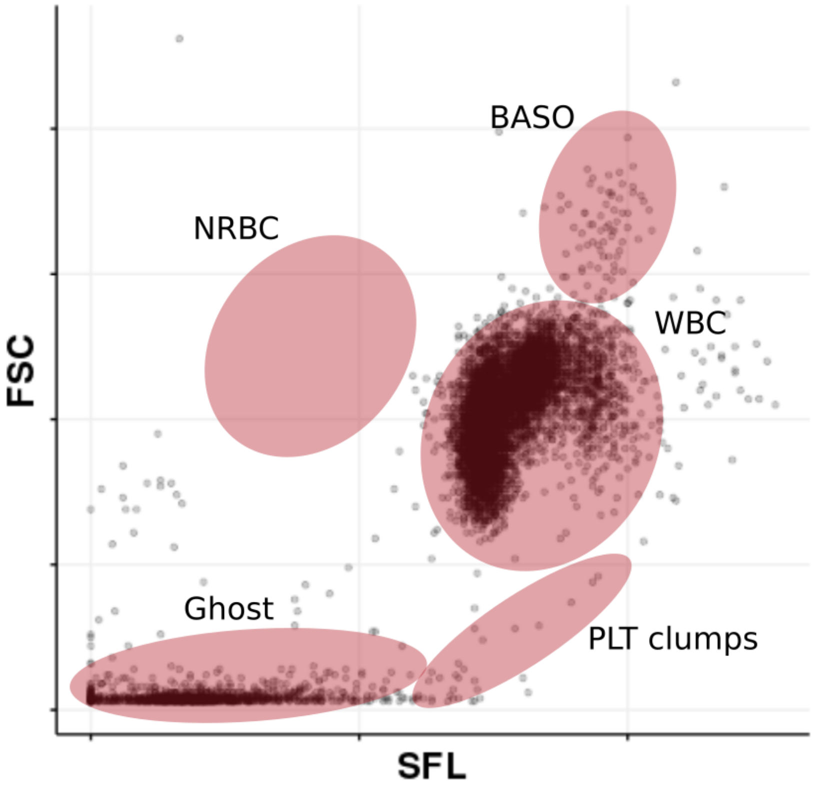

Figure 3.5: WNR channel scattergram.
A plot of the WNR scattergram from an individual in the INTERVAL study, each cell is plotted with SSC and SFL on the x and y axis respectively. Measurements are assigned per cell-type based on the median position of the cell cloud in each axis of the scattergram. WNR separates basophil cells from the white blood cell population. NRBC are not observed as this cell population does not occur in healthy adults. Approximate cell types are highlighted based on thresholds set in the two axes.

Sysmex Parameter	Description
MicroR / MacroR	Microcytic and Macrocytic RBCs (as percentage of all RBCs)
RDW-SD	Red cell size distribution width
Delta-HGB	Cell free haemoglobin
RPI	Reticulocyte Production Index Calculated: $RET\% * HCT/(2 * 0.45)$

Table 3.5: Other parameters.
I studied six parameters which were derived from electrical impedance, photometric analysis, or calculated from flow cytometry measurements, or calculated from a combination of flow cytometry and impedance measurements.

3.2.2.5 Other Parameters

In addition to flow cytometry the Sysmex analyser employs an electrical impedance channel (Section 1.2) to measure RBC and platelet parameters, such as mean cell volume in the blood or median size of individual blood cells, and a photometric channel which measures red blood cell haemoglobin content. I utilised six of these parameters in my analysis, including parameters which are derived from a combination of impedance and flow cytometry (RPI) and parameters which are calculated from a combination of flow cytometry measurements (Table 3.5).

3.2.3 Flow cytometry, immune cell function, and disease

I sought to understand the clinical and functional relevance of Sysmex measurements of blood cells, SSC, SFL, and FSC (Section 3.2.2). Previous studies show that Sysmex parameters of white cells correlate with changes in cell function, activation, morphology, and disease status including: myelodysplastic syndromes, toxic granulation, sepsis, septic shock, and Szary disease [13, 64, 188, 103, 131, 145]. The functional and clinical relevance of Sysmex parameters makes these phenotypes important intermediate traits, the significance of intermediate traits is discussed in Section 1.6.4. The ability to make automated high-throughput measurements using Sysmex enables GWAS study of such phenotypes, where the interpretation of GWAS results could provide important insights into human biology a detailed in Section 1.6.

Comparison to manual assessment of blood smear images shows that Sysmex parameters capture clinically important changes in white cell morphology [188]. Blood smear samples from 158 patients were scored by neutrophil granularity on a scale of 1 to 4 by trained haematological medical technicians [188]. The Sysmex parameter NE-SSC correlated with manual measurements of granularity performed by smear test (Spearman's rank correlation coefficient: r_s=0.839, P-value<1×10^{-4}) [188] (Fig. 3.6). Automated Sysmex measurements provide an advantage as manual measurements of granularity are slower to

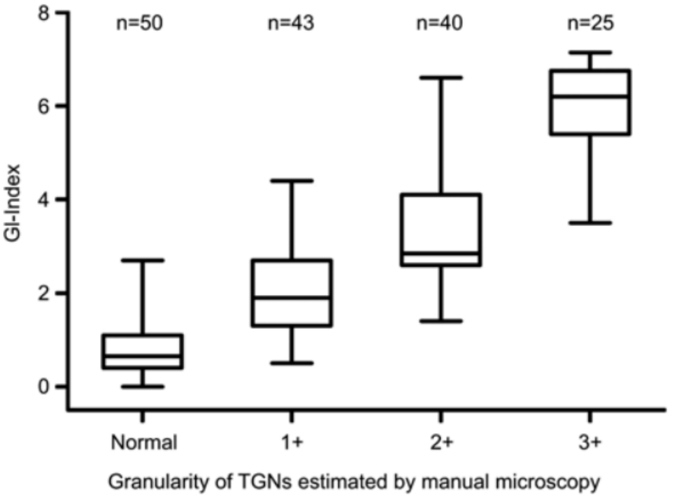

Figure 3.6: Comparison of granularity assessed by manual microscopy and Sysmex flow cytometer.
Assessments of toxic granulation neutrophils (TGN) granularity by GI-Index, a measure of SSC performed by Sysmex and by manual microscopy. The granularity index assessed by flow cytometry correlates with manual assessments of granularity ($r_s = 0.839$, p $<1 \times 10^{-4}$) (Figure reproduced from [188]).

perform and depend on interpretation of the individual haematologist.

Automated Sysmex measurements can also assess monocyte, neutrophil, and leukocyte activation following in vitro stimulation by activating compounds formyl-methionyl-leucyl-phenylalanin (fMLP) and lipopolysaccharides (LPS) [103]. Sysmex parameters have been compared to an automated image analysis pipeline of blood smear images to classify neutrophils, monocytes, and the combined leukocyte population into hypo or hyper-granulated categories [103]. In response to activation, neutrophils showed an initial hypo-granularity reaction and long term hyper-granularity which persists until approximately three hours after incubation with activating compounds. A statistically significant (P$<1 \times 10^{-4}$) correlation was seen between all neutrophil Sysmex measurements (NE-SSC, NE-SFL, NE-FSC) and automated measurements of neutrophil activation derived from microscope image analysis, including when cells were activated with LPS (P-value$<1 \times 10^{-4}$, $r_s = 0.693$) and fMLP (P-value$<1 \times 10^{-4}$, r_s: 0.641) [103].

An additional study to determine the utility of lymphocyte, monocyte, and neutrophil Sysmex parameters to diagnose sepsis found clinically significant correlation between these indices (except for NE-FSC, and LY-FSC) and occurrence of sepsis. MO-SSC and NE-SFL had the best diagnostic performance (AUC 0.75, and 0.72 respectively) [35]. Statistically significant relationships between Sysmex measurements (in particular neutrophil indices) and sepsis have been reported by a number of other authors [131] [13], and also other diseases including toxic granulation [188], and myelodysplastic syndromes [145] [64].

Sézary disease is a form of cutaneous T-cell lymphoma. Malignant lymphocytes cause inflamed, itchy, lesions on the skin of patients which develop into tumours. Sézary cells are characterised with irregular nuclei and condensed chromatin versus smaller cells with regular nuclei and clumped chromatin in chronic lymphocytic leukemia patients [31]. Automated analysis by Sysmex has been used to identify abnormal T cells typical of Sézary disease. LY-SSC was associated with the count of Sézary cells (classification threshold LY-SSC above 85, sensitivity 100%, specificity 94%) and LY-FSC (classification threshold LY-FSC above 67, sensitivity 89%, specificity 94%) to the presence of larger cells, both of which are diagnostic factors of Sézary disease [31]. This is consistent with the definition of these parameters and the known difference in morphology of neoplastic cells. Thus these Sysmex haematological measurements are consistent with expected differences in cell size, morphology, and nuclear structure. Brisou *et al*, suggested that these differences as identified by Sysmex parameters could be used to diagnose Sézary disease in a clinical setting [31].

However, there are limitations to the interpretability of these parameters. For example side-fluorescent light (SFL), a measure of DNA/RNA content cannot be used purely as a surrogate for the quantity of nucleic acids in the cell, because the dye does not saturate the cell due to short reaction time. SFL intensity depends on cell membrane composition and also nucleic acid content. Sysmex reports monocyte measurements as having higher average SFL values than lymphocytes, but a resting monocyte does not have a higher nucleic acid content than a resting lymphocyte (personal communication J. Saker, Sysmex) [151]. However, within a single cell type, Sysmex parameters are consistent with physiological changes within that cell type. It is this variation between individuals in the INTERVAL study which I utilise to perform a GWAS analysis.

3.2.4 Adjusting variables for scale

Sysmex parameters occur in varied scales of measurement, including percentages, ratios, and positively supported data for example cell counts which never hold negative values. In order to adjust for technical (Section 3.2.7) and environmental (Section 3.2.8) covariates, the parameters were transformed depending on their scale of measurement prior to adjustment. The transformations were performed as follows, where x is a vector of phenotype values.

- Percentages $0 < x < 100$, division by 100 and logit transformation: $a(x) = \text{logit}(\frac{x}{100})$.
- Ratios $0 < x < 1$, logit transformation: $a(x) = \text{logit}(x)$.
- Positively supported $x > 0$, log transformation: $a(x) = \log(x)$.

The logit transformation $\text{logit}(x) = \log \frac{x}{1-x}$ maps probability values from $[0, 1]$ to $(-, +)$. Logit transformed data with infinite bounds makes the distribution of the dependent variable more like a normal distribution thus enabling better specified linear regression (Section 1.5) [92]. As described previously in Section 2.2.4 linear regression is the statistical test utilised by GWAS to identify genetic associations.

3.2.5 Additive models and splines

Additive models (AMs) include both parametric and non-parametric predictors [76]. Non-parametric predictors allow additional flexibility as they do not require a specific function to be predefined. AMs can be used to model predictor variables which are cyclical (such as seasonal effects) or have otherwise non-linear effects on the dependent variable. AMs extend linear models where the linear elements $\sum \beta_j X_j$ are replaced by a sum of smoothing functions $\sum \beta_j m_j(X_j)$. Where $m_j()$ is a unspecified function not required to be defined by the user, thus the non-parametric form of AMs [76]. This is an extension of linear regression, instead of optimising the fit of a linear line to a set of data points we optimise the fit of an arbitrary function $m_j(X_j)$. However, in most cases it is prudent not just to optimise the fit of a curve to data points, but also ensure ‘smoothness’ or simplicity of that curve. This is called ‘smoothing’ and is discussed in more detail below. I utilise smoothing splines in my analysis. I will begin with a brief description of smoothing followed by splines including B-splines, P-splines, cyclic splines, and thin-plate splines.

Smoothing

As previously described AMs models optimise a curve to fit a set of data points (Section 3.2.5), however in a non-linear setting curve complexity can increase such that the resultant model is not representative of the true relationship between dependent and independent variables. This is called over-fitting and has been described by a number of authors [78] [77, p. 398]. Therefore, smoothing is utilised to reduce model complexity avoiding highly curved functions which contort to fit to every data point. Mathematically, curvature or smoothness at any point can be defined as the second derivative of the function $m(x)$ represented by $m''(x)$. Smoothness of $m(x)$ across its entire domain is defined by the following integral $\int (m''(x))^2 dx$, the squared operation is applied to avoid distinction between negative or positive curvature. In order to generate a regression, an objective function is created which states that a) we want a function $m(x)$ which fits as closely as possible to the data points and b) we want this function to be smooth [154, p. 177]:

$$L(m, \lambda) \equiv \frac{1}{n} \sum_{i=1}^{n} (y_i - m(x_i))^2 + \lambda \int (m''(x))^2 dx \tag{3.1}$$

Where x and y are the independent and dependent variables respectively, and λ is a hyper parameter which modulates the smoothness, a higher λ will prioritise smoothness over the fit of curve. The first term simply models the fit of the curve to the real values and the second term quantifies smoothness of the curve. A solution to this objective function will generate a function with a trade-off between maximising fit to the data while also maximising smoothness of the curve. The precise value of λ is set by cross validation over the dataset, for more details see [154, p. 179].

Piece-wise polynomials and splines

In the previous section I left ambiguity regarding the definition of $m(x)$ (Eqn. 3.1) which depends on the context of implementation. One such implementation is that of smoothing splines, this is where spline functions are used to define $m(x)$. To begin I will describe piecewise polynomial functions and extend that definition to that of splines. Piecewise polynomial functions are best described visually (Fig. 3.7), in essence they split the domain of the input variable at into pieces, in each piece a function with differing constants is defined (Fig. 3.7) [77]. The points at which distinct pieces are bounded is termed 'knot points'. Please note the distinction between smoothness of each function within each piece and between pieces. Each piece in the displayed figure is maximally smooth (as they are linear or straight lines), however there is discontinuity between pieces. Discontinuity and smoothness at the knots points cannot be addressed by smoothing which is only applied to functions between knot points (Eqn. 3.1) and is instead addressed by enforcing that functions which meet at a knot point have equivalent values and also equivalent first and second order (or more) derivatives depending on the order of the piecewise functions (Fig. 3.8). Application of piecewise polynomials is limited due to the mathematical properties of the piecewise curve, an adaptation of this method is B-splines (basis splines), here a series of polynomials spanning the feature space is fit to the data (Fig. 3.9). This is explained further in the Appendix of Hastie and Tibshirani [77, p. 186]. P-splines are the application of smoothing to the fitting of B-splines to ensure that the spline is smooth as well as continuous (Section 3.2.5). An additional constraint would be to ensure that the spline takes the same value at the lower and upper limit of the domain, this is termed a cyclic spline and is effective in modelling cyclic data predictor variables such as day of year or day of the week. In contrast to B-splines which are generated on a single variable, thin plate splines allow fitting of splines to multiple dimensions of data. This involves multidimensional generalisation of the B-splines described above, for more detail see [77, p. 162].

3.2.6 Exclusion of outliers and erroneous measurements

Each Sysmex measurement is associated with an interpretive program (IP) message labelling the measurement as potentially 'abnormal', this could occur in participants with

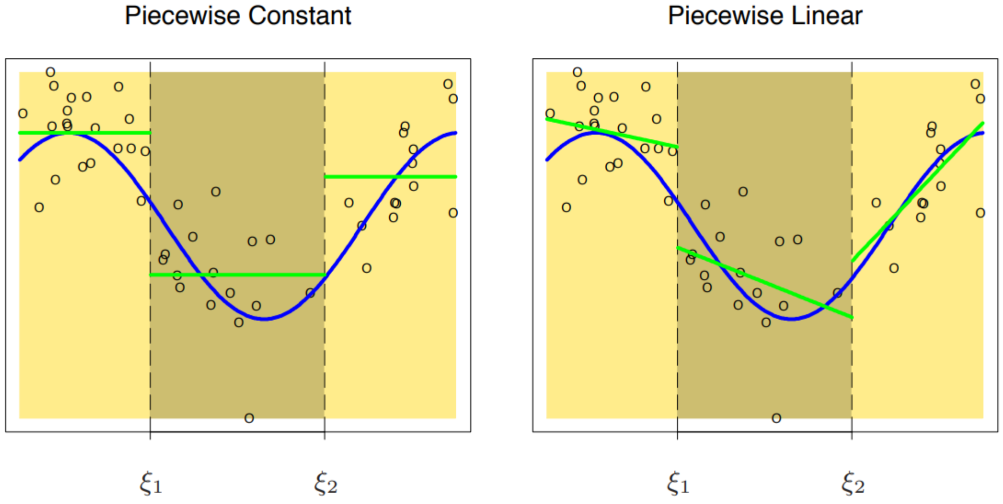

Figure 3.7: Examples of piecewise constant linear functions.
A plot of data points for which the piecewise functions have been fitted with independent and dependent variables on the x and y axes respectively. Three pieces are defined based on two knots indicated by ξ_1 and ξ_2, within each piece a different function is defined to best fit the data in that piece. Piece-wise functions are not enforced to meet at the knot points hence the discontinuity at knot points (Figure source [77]).

unusually low or high counts of cells, such as lymphopenia or lymphocytosis or containing unusual blood cell morphology, for example the presence of nucleated red blood cells or immature granulocytes. IP messages are generated by the haematological analyser and are classed into three categories, abnormal, suspect, or negative. I excluded all measurements associated with an abnormal or suspect IP flag. Following exclusion of measurements based on IP flags, I performed technical (Section 3.2.7) and environmental (Section 3.2.8) adjustment and utilised a PCA to identify outlying measurements which I removed from further analysis. Phenotypes were categorised into seven non-mutually exclusive classes, those related to platelets, red cells, reticulocytes, white cells, granulocytes, myeloid cells, and all phenotypes. For each of these categories the PCs were calculated and scaled by variance explained, squared, and summed to calculate the total deviation of each measurement from the population centre. An equal number of PCs were selected to the number of directly measured phenotypes in each class. Measurements which sufficiently deviated from the population centre were excluded, this was assessed with a χ^2 distribution (P-value$< 1 \times 10^{-7}$) (Fig. 3.10). If a measurement was an outlier within one category, the data for that sample was excluded across all phenotypes in further analyses.

3.2.7 Technical variation of Sysmex parameters

Technical factors which influence Sysmex parameters add noise to measurements which increases false negative findings in GWAS analyses and also reduces power to identify

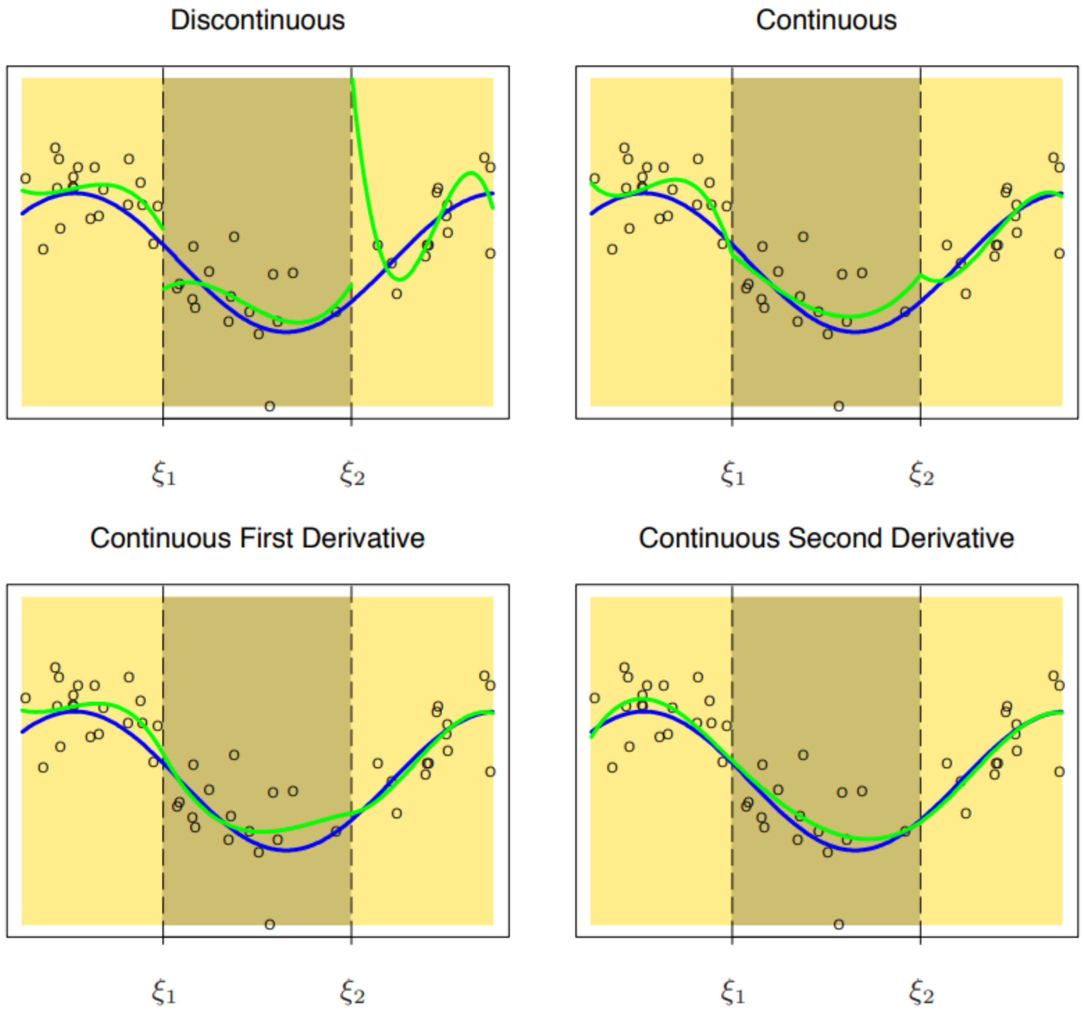

Figure 3.8: Examples of cubic polynomial piecewise functions.
A plot of data points for which cubic piecewise functions have been fitted with independent and dependent variables on the x and y axes respectively. Three pieces are defined based on two knots indicated by ξ_1 and ξ_2, A discontinuous function can be made continuous and smooth at the knot points by enforcing equal values for the function and first or second derivatives of the function at the knots (Figure source [77]).

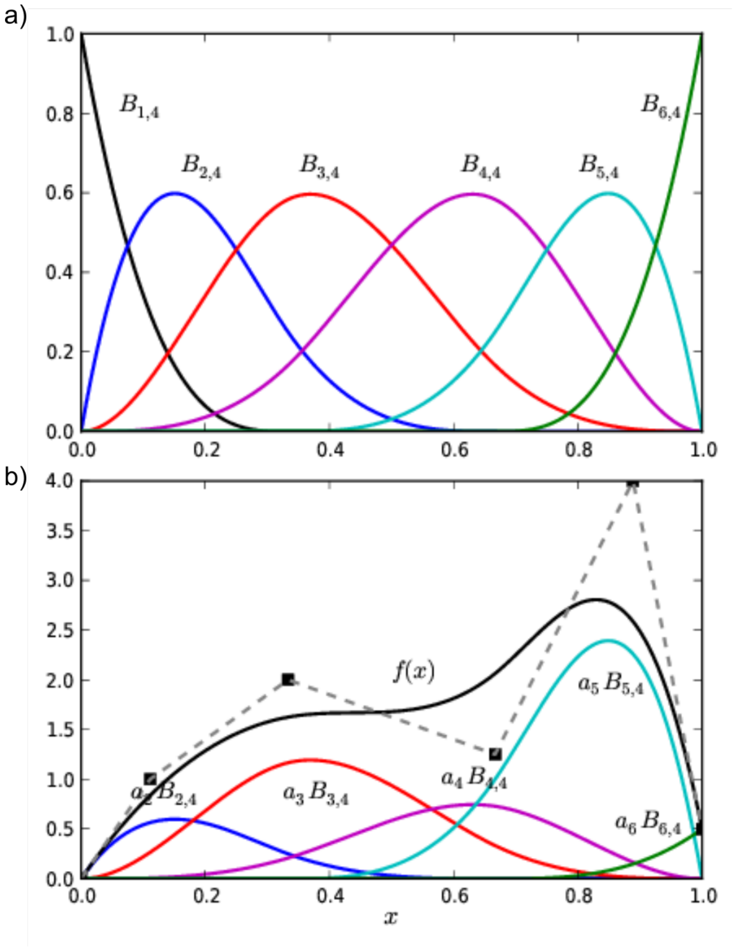

Figure 3.9: Fitting of cubic B-splines.

Data points are plotted with independent and dependent variables on the x and y axes respectively. **a)** A series of B-spline basis functions of third polynomial degree. **b)** Weighting of B-splines to enable the summation of B-splines to generate a function $f(x)$ to fit a dataset (Figure source [55]).

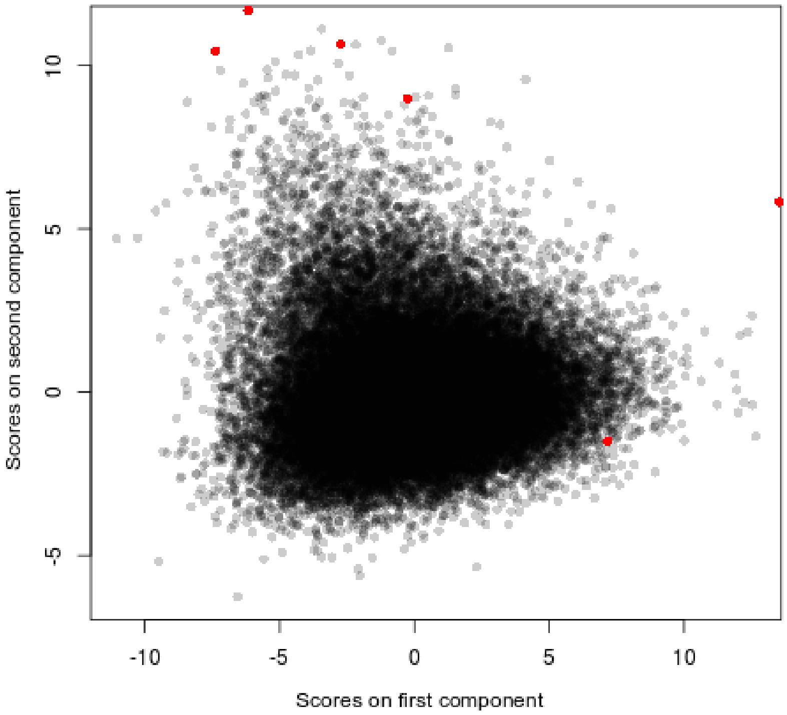

Figure 3.10: First two principal components of measurements in the platelet category.
Each data point represents an individual for which platelet measurements were recorded and the first two PCs for measurements in the platelet category are plotted, subsequent principal components are not visualised although they do contribute to the detection of outliers. Outlying data points are highlighted in red and defined by those in the upper tail of a null χ^2 distribution with P-value$< 1 \times 10^{-7}$. Platelet measurements for outlying individuals are excluded from further analysis.

influential genetic associations. Technical variation is not likely to increase false positive rates because there is no correlation between the technical variables and the genotype of the study participants. I modelled the effect of technical covariates on each blood trait using a AM and adjusted haematological indices to remove the effect of technical factors.

Measured traits were technically adjusted independently, the technical correction procedure began by excluding measurements from outlying days, defined as those with a Z score more than 8 from the daily mean. Following this, a generalised additive model (GAM) model was used to adjust measured indices for technical and seasonal variation, finally derived parameters were re-calculated accordingly from the measured parameters (Supp Table A.7). Defined parameters are blood traits which are calculated from ratios, percentages or other combinations of directly measured traits were re-calculated following technical adjustment of the measured traits. I corrected for a range of factors including time passed from start of the study, day of the year (ordinal from 1 to 365), time between venipuncture and sample analysis, day of the week, and instrument id. The adjustment was made across all measurements indexed by i using a GAM (R package *mgcv*) and smoothing terms with P-spline, cycling smoothing or thin plate splines:

$$\mathbf{E}(a(y_i)) = s[t(i) \otimes m(i)] + c[t_{\text{year}}(i)] + tp[(t_{\text{day}}(i), t_{\text{ven}}(i) \otimes (m(i), I(i))] + \sum_{D \in \{\text{mon .. sun}\}} 1_{D(i)=D} + \sum_{m} 1_{m(i)=m} \tag{3.2}$$

In equation 3.2:

- $a(x_i)$ represents the trait values x_i transformed as described in Section 3.2.4.
- $t(i)$ denotes the number of seconds between the first day of the study and measurement i.
- $m(i)$ is a categorical variable with two levels for each of the two machines used to record measurements.
- $D(i)$ is a categorical variable with 7 levels representing the day of the week on which the measurement was made.
- $t_{year}(i)$ is the number of seconds between January 1st and the time at which the observation was made.
- $t_{day}(i)$ is the number of seconds between measurement of observation i and midnight (am) on the day of observation.
- $t_{ven}(i)$ represents the number of seconds between midnight (am) on the day of observation and venipuncture.
- $I(i)$ is a binary variable which indicates whether the delay between measurement and observation was imputed $t_{ven}(i)$. Imputation was performed by a median calculation of real values for the $t_{venn}(i)$ variable.
- $s[]$ a P-spline smoothing term for univariate terms.
- $c[]$ cyclic smoothing term for seasonal data such as time of year, $t_{year}(i)$.
- $tp[]$ a thin plate spline smoothing term for bivariate data.
- $\otimes$ represents an interaction between variables.

The first term $s[t(i) \otimes m(i)]$ in the equation (Eqn. 3.2) models long term drift and calibration of the Sysmex analysers using a smooth P-spline with 50 knots. Drift is defined as systematic changes in recordings of the Sysmex analyser over the three year time period of analysis (Fig. 3.11). In contrast, calibration effects result in immediate and large change in the mean recordings of the Sysmex analyser as a result of periodic calibration of the analyser (Fig. 3.11). These effects are modelled independently by haematological analyser represented with categorical variable $m(i)$. The second term $c[]$ models seasonal effects using a cyclic smoothing term with 30 knots. To avoid an increasingly complex model, the second term assumes that the Sysmex instruments are influenced in the same way by seasonality effects. Thus seasonality is not modelled independently between the analysers. The third term jointly effectively models the effect of time delay between venepuncture and analysis, this effect is allowed to vary depending on the analyser and whether the $t_{venn}(i)$ variable was recorded or impute from the median of measured values (Section 3.1.2). The third term is modelled with a thin plate spline with 30 knots. Finally, I include dummy variables to model day of the week and the effect of instrument on the parameter measurements. Knots were chosen to be consistent with the similar data correction procedure performed by Astle *et al* [15]. The adjustment described was implemented in R code and is available on github [7].

Following application of the adjustment procedure on all 63 parameters I performed a manual inspection to assess the performance of the adjustment of all covariates. Each parameter was plotted against each covariate separately before and after correction (an example is Figure 3.11). It will be expected that these results will show less variation in the parameter values along the covariate x axis following correction compared to prior to correction. Inspecting all aforementioned plots indeed showed that this is the case and adds confidence to correction procedure which relies on the generation of splines.

3.2.8 Environmental variation of Sysmex Parameters

Following technical adjustment, Sysmex parameters were also adjusted for environmental covariates which are known to influence the values of blood measurements:

$$\begin{gathered}\mathbf{E}(env(y_i)) = s[\text{age}(i) \otimes \text{meno}(i)] + tp[(\log(\text{weight}(i)), \log(\text{height}(i)) \otimes \text{meno}(i)] + \\ \sum_{\text{drink}(i)} 1_{\text{drink}(i)=\text{drink}} + \sum_{\text{alc}(i)} 1_{\text{alc}(i)=\text{alc}} + \\ s[\text{pack\_yrs}(i)] + \sum_{\text{smoke}_s} 1_{\text{smoke}_s(i)=\text{smoke}_s} + \sum_{\text{smoke}_a} 1_{\text{smoke}_a(i)=\text{smoke}_a} + \\ \sum_{\text{int}} 1_{\text{int}(i)=\text{int}} + \sum_{\text{weight\_na}} 1_{\text{weight\_na}(i)=\text{weight\_na}} + \sum_{\text{height\_na}} 1_{\text{height\_na}(i)=\text{height\_na}} + \\ \sum_{\text{pack\_yrs\_na}} 1_{\text{pack\_yrs\_na}(i)=\text{pack\_yrs\_na}}\end{gathered} \tag{3.3}$$

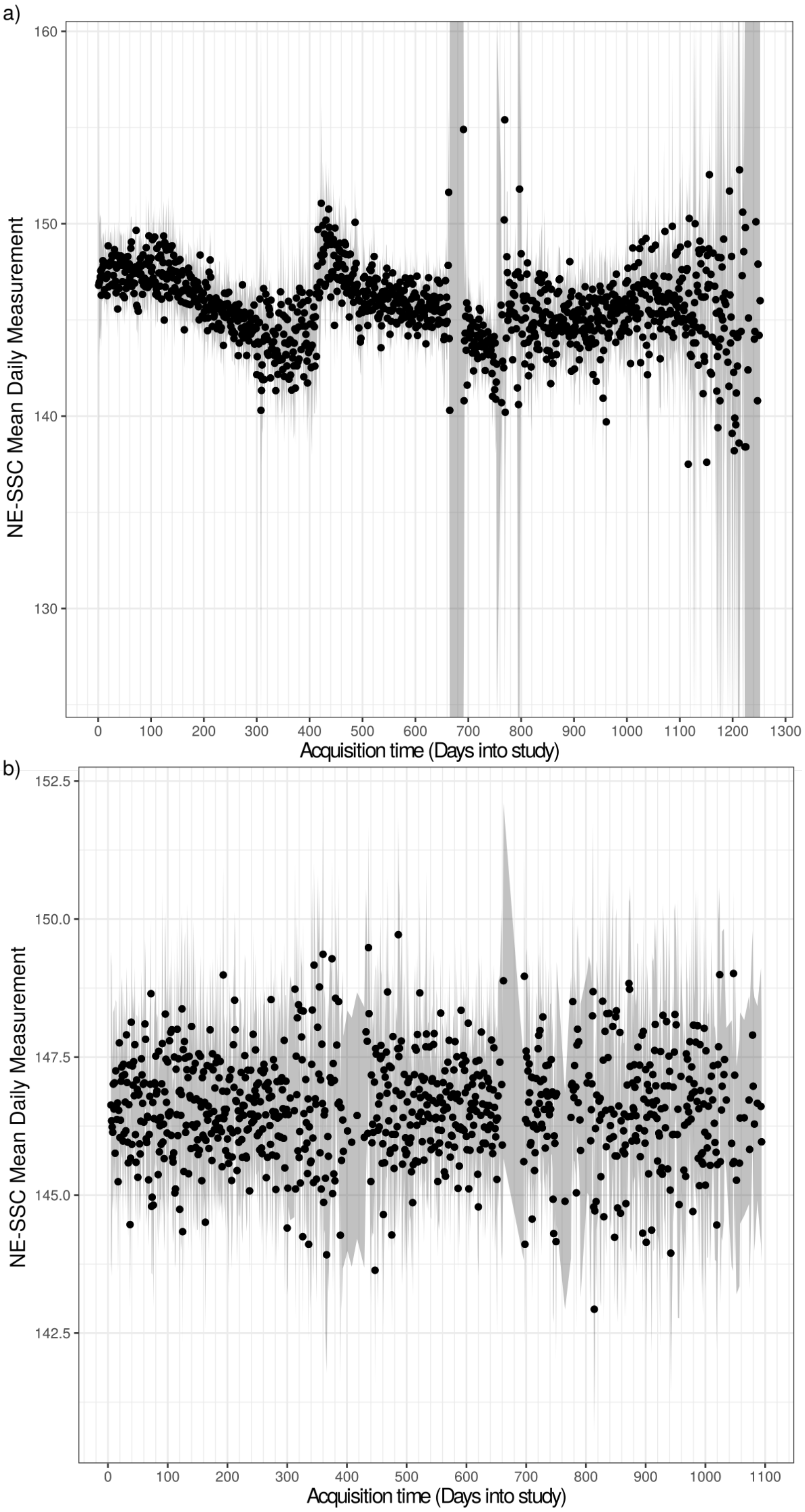

Figure 3.11: Variation in recorded mean daily value of the NE-SSC parameter over time before and after adjustment.

Each data point is the mean daily recorded value for the NE-SSC parameter over time course of the study, adjustment is performed by fitting the model described in Equation 3.2 and plotting the residuals. **a)** Raw mean daily recorded values from analyser for parameter NE-SSC shows considerable systematic drift and changes due to calibration effects. **b)** Mean daily recorded values post adjustment of data for technical covariates. This plot shows significant systematic drift of NE-SSC values and correction of a large proportion of this drift following adjustment.

In equation 3.3:

- e_i are the residuals for measurement i obtained from the technical adjustment model described in Equation 3.2.
- $s[]$ a P-spline smoothing term for univariate terms.
- $c[]$ cyclic smoothing term for seasonal data used here, time of year $(t_{year}(i))$.
- $tp[]$ a thin plate spline smoothing term for bivariate data.
- age(i) is the age of the participant.
- meno(i) is the menopausal status of the participant:
  meno$(i) \in \{$post, pre, hyst, male, NA$\}$.
- weight(i) is the weight of the participant.
- height(i) is the height of the participant.
- drink(i) is the drinking status of the participant:
  drink$(i) \in \{$never, previous, current, NA$\}$.
- alc(i) is the alcohol consumption of the participant:
  alc$(i) \in \{$rarely, 1 to 3 month, 1 to 2 weeks, 3 to 5 weeks, most days, never$\}$
- pack_yrs(i) a calculation of the pack years a participant has smoked.
- smoke$_s(i)$ is the smoking status of the participant: smoke$_s(i) \in \{$never, previous, current, NA$\}$
- smoke$_a(i)$ is the participants frequency of smoking:
  smoke$_a(i) \in \{$special occasions, rarely, occasional, most days, every day, never$\}$
- int(i) sex and assigned donation arm of the participant where the measurement was taken upon completion of the study:
  int$(i) \in \{$baseline, M8, M10, M12, F12, F14, F16$\}$
- $\text{weight\_na}(i), \text{height\_na}(i), \text{pack\_yrs\_na}(i)$ binary variables set to true if the participant has missing values for weight, height, or pack years smoked variables.

The first term in Equation 3.3 models the effect of age and menopause status on Sysmex parameters using a P-spline with 30 knots. The effect of log transformed height and weight and interaction with menopause status is modelled with a thin plate spline with 30 knots. Menopause status has been shown to influence blood cell measurements including red blood cell and platelet count [115, 41]. Dummy variables are used to model, drinking status, drinking frequency, smoking status, and smoking frequency and a P-spline

(19 knots) is used to model the effect of pack years smoked per participant on the Sysmex parameter. Pack years smoked is calculated based on participant responses to the health questionnaire at baseline of the study. Both smoking and drinking have been shown to influence blood cell measurements [160, 132]. Dummy variables are also used to model missing values for height, weight, and pack years smoked. The described adjustment was implemented in R code and is available on github [7]. Following adjustment for technical and environmental factors, PCs were constructed to identify and exclude outlying Sysmex parameter measurements as described in Section 3.2.6.

3.3 Summary

In this chapter I describe blood sample collection and genotyping in the INTERVAL study and provide a detailed description of steps for collection and QC of the Sysmex parameters and genotype data. I adjusted Sysmex parameters to remove technical and environmental factors which add variation to the phenotypes studied in my analysis. Removing such influencing factors reduces variability in the phenotypes and increases power to detect association signals. Following extraction and QC of SSC, SFL, FSC from Sysmex analysers in the INTERVAL study I performed the first ever GWAS analysis of these parameters which is discussed in Chapter 4.

Chapter 4

The genetic architecture of cytometry parameters

4.1 Introduction

GWAS of blood cell traits measured by automated FBC such as cell counts and cell volumes are a powerful approach to link disease-associated risk variants to the distinct types of blood cells and their molecular pathways [15]. However, these traits mostly provide information about genes and pathways regulating processes such as stem cell lineage-fating choices and blood cell survival. In particular for white blood cells, analytical methods to infer which variants identified by GWAS of CBC parameters influence cell function do not exist. To uncover this class of variants requires measurement of white cell function in thousands of individuals. This is not feasible because functional assays are laborious, often have poor reproducibility, and cannot be parallelised [110].

I report an alternative approach to obtain, in a large number of individuals, parameters which are proxies for immune cell function, particularly of granulocytes (neutrophils, eosinophils and basophils) and monocytes. As previously explained in Chapter 3, these parameters are obtained by exploiting flow cytometry measurements underlying the routine FBC by the Sysmex instrument. In this chapter, I present the first ever GWAS of SSC, SFL, FSC blood cell phenotypes where I identified novel association signals compared to the previous largest GWAS of haematological phenotypes performed by Astle *et al.* [15]. The comparison was performed with Astle *et al.*, 2016 as this was the largest published GWAS of blood cells available at the time of this analysis. My analysis identified 2,172 genetic associations annotated to genes by VEP known to be relevant in chemotaxis, adhesion, activation, degranulation and many types of immune responses. The results of this analysis was the basis of downstream analysis to further understanding of haematology and disease biology presented in Chapter 5.

4.2 Methods

4.2.1 Population stratification and relatedness

As previously described (Section 2.2.6), population stratification and relatedness within the population of a GWAS study can lead to inflation in false positive or negative associations. Multiple methods were utilised to account for broad population stratification within the population cohort and relatedness between individuals in the cohort:

- Filter all samples to only include those of European ancestry (Section 3.1.3.3).
- Remove samples with a high degree of relatedness, as determined by IBD analysis.
- Include PCs and *clinic*, a variable recording the location of blood donation as covariates in the GWAS model.
- Use a LMM model which allows potentially confounding polygenic effects to be modelled by a random effect in the GWAS regression (Section 2.2.7).
- Filter tested variants by MAF>0.04%.

PCs were generated from approximately 100,000 high quality variants which were selected by a number of factors including: the variant must be genotyped, MAF $\geq$ 2.5%, missingness $\leq$ 1.5%, variants which aren't insertions or deletions, and pruned to ensure low LD [15]. PCs were used to remove individuals who are outlying from the general population and therefore not likely to be of European ancestry. The purpose of excluding individuals with non-European ancestry is to create a more homogeneous sampling population and reduce the risk of false positive or false negative associations due to confounding population structure. Following this, PCs were calculated from the European ancestry genotype data and included as a covariate in the LMM enabling the model to account for population stratification effects. In this model the top 10 PCs were used consistent with other published GWAS studies of the INTERVAL dataset [15]. IBD analysis determines common stretches of nucleotide sequence between two individuals which indicate they share a common ancestor. If a large proportion of alleles between two individuals are identical by descent, this suggests those individuals are closely related. Of each pair of individuals who has higher than 98% of alleles IBD, one individual was removed from the study. Groups of individuals who were related (IBD>20%) were iteratively trimmed by removing the individual with the highest number of pairwise relationships and then the lowest call rate. This analysis was performed on the INTERVAL dataset by Astle *et al.,* [15].

4.2.2 GWAS of haematological phenotypes

GWAS was performed on phenotype values which were adjusted for technical and environmental factors, outlying measurements excluded, and residuals inverse quantile normalised (Section 3.2). Following QC, genotype data was imputed from a genome-wide scaffold of variants assayed by direct genotyping. Firstly the dataset was phased using SHAPEIT3 and then imputed using a combined 1000 Genomes and UK10K panel [15]. I filtered imputed variants by INFO score 0.4 and MAF>0.04% (Section 2.2.5). A LMM was used to test each imputed variant independently for marginal association with the phenotype. The following covariates were included in the LMM: the first ten principal components representing population structure, and the categorical 'clinic' variable which represents the centre at which the participant donated their blood sample (Section 4.2.1). The LMM was implemented using BOLT-LMM [106] which allows for computationally tractable LMM GWAS analyses of large datasets (Section 2.2.7). The analysis was performed on the Sanger high performance compute (HPC) cluster.

4.2.3 Comparison with GWAS of FBC phenotypes by Astle 2016

I compared the results of my GWAS and conditional analysis with the results obtained Astle *et al.,* in their GWAS of standard haematological traits [15]. The traits studied by Astle *et al.,* are measured in the same cell types from which the Sysmex parameters in my study are derived. In order to assess whether findings made by my analysis identifies signals unreported by Astle *et al.*, I extended the previously described LD clumping procedure (Section 2.2.9). Firstly, I performed LD clumping as previously described in Section 2.2.9. The LD clumping procedure assigns conditionally significant variants to sets of variants which exist in LD $r^2 > 0.8$ with each other. Following this, I generated a LD matrix between all the conditionally significant variants identified by my study or by Astle *et al.* I then used this matrix to sequentially label LD sets as already 'not novel' if that LD set contains any variants which are in LD $r^2 > 0.8$ with any conditionally significant variants identified by Astle *et al.*. Thus I annotated conditionally significant variants identified by my analysis as novel or not-novel in comparison to association signals identified by GWAS of FBC haematological phenotypes by Astle *et al.* [15]. This annotation was implemented in R code and is available in a github repository [6].

4.2.4 Genetic correlation between inherited components of variance by Bulik Sullivan *et al.*, 2015

The genetic correlation is the correlation in inherited components of variance between two phenotypes. I estimated genetic correlation using the summary statistics generated from a GWAS study [33]. LD score regression uses a pre-defined subset of common SNPs and assumes a polygenic model of association. The polygenic model assumes the genetic component of variation of a quantitative trait is determined by a set of genetic variants that have independent additive effects. For SNP j the expected value of the product z score of association with both traits is determined as follows [33]:

$$E[z_{1j}z_{2j}] = \frac{\sqrt{N_1 N_2}\rho_g}{M}\ell_j + \frac{\rho N_s}{\sqrt{N_1 N_2}} \tag{4.1}$$

Where ρ is the phenotypic correlation amongst the N_s overlapping samples, the purpose of this term is to adjust for correlation induced by a common sample population and avoid biasing the calculated genetic correlation. ρ_g is the genetic correlation, $N_1 N_2$ is the product of the number of samples in each study, M is the number of variants used in the estimation, and ℓ is the 'LD Score', a measure of the amount of genetic variation linked to SNP j, calculated where k is an index over all other variants [33]:

$$\ell_j = \sum_k r_{jk}^2 \tag{4.2}$$

To calculate the genetic correlation defined by ϱ_g, we perform a regression of $E[z_{1j}z_{2j}]$ against ℓ_j for each SNP. The slope of this regression line will be $\frac{\sqrt{N_1 N_2}\varrho_g}{M}$ from which the genetic correlation ϱ_g is calculated. My implementation of LD score regression is available in the following github repository [6]

Variant effect predictor

VEP is a software package which annotates genomic variants in coding and non-coding regions to genes by searching for variants which overlap or are close to (threshold of 5000 base pairs) known transcripts and regulatory regions [114]. The impact of a variant is classified into one of 48 sequence ontology (SO) terms which are then assigned into 'HIGH', 'MODERATE', 'LOW', or 'MODIFIER' terms in order of decreasing severity on the gene. For example, 'stop lost' variants are classified as a HIGH impact modification due to the disruption of a stop site, but 'synonymous' variants are classified as LOW impact as they do not lead to change in amino-acid structure of the protein [114]. MODIFIER terms are those where variants effect non-coding regions such as intergenic or intronic regions [114].

Display term	SO description	IMPACT
Transcript ablation	A feature ablation whereby the deleted region includes a transcript feature	HIGH
Splice acceptor variant	A splice variant that changes the 2 base region at the 3' end of an intron	HIGH
Splice donor variant	A splice variant that changes the 2 base region at the 5' end of an intron	HIGH
Stop gained	A sequence variant whereby at least one base of a codon is changed, resulting in a premature stop codon, leading to a shortened transcript	HIGH
Frameshift variant	A sequence variant which causes a disruption of the translational reading frame, because the number of nucleotides inserted or deleted is not a multiple of three	HIGH
Stop lost	A sequence variant where at least one base of the terminator codon (stop) is changed, resulting in an elongated transcript	HIGH
Start lost	A codon variant that changes at least one base of the canonical start codon	HIGH
Transcript amplification	A feature amplification of a region containing a transcript	HIGH
Inframe insertion	An inframe non synonymous variant that inserts bases into in the coding sequence	MODERATE
Inframe deletion	An inframe non synonymous variant that deletes bases from the coding sequence	MODERATE
Missense variant	A sequence variant, that changes one or more bases, resulting in a different amino acid sequence but where the length is preserved	MODERATE
Protein altering variant	A sequence_variant which is predicted to change the protein encoded in the coding sequence	MODERATE
Regulatory region ablation	A feature ablation whereby the deleted region includes a regulatory region	MODERATE
Splice region variant	A sequence variant in which a change has occurred within the region of the splice site, either within 1-3 bases of the exon or 3-8 bases of the intron	LOW
Incomplete terminal codon variant	A sequence variant where at least one base of the final codon of an incompletely annotated transcript is changed	LOW
Start retained variant	A sequence variant where at least one base in the start codon is changed, but the start remains	LOW
Stop retained variant	A sequence variant where at least one base in the terminator codon is changed, but the terminator remains	LOW
Synonymous variant	A sequence variant where there is no resulting change to the encoded amino acid	LOW

Table 4.1: Calculated consequence of sequence ontology terms annotated as HIGH, MODERATE, or LOW impact. Each variant is mapped to a SO term which is then categorised depending on impact on the transcript. This table shows SO terms with HIGH, MODERATE, or LOW impact.

Display term	SO description	IMPACT
Coding sequence variant	A sequence variant that changes the coding sequence	MODIFIER
Mature miRNA variant	A transcript variant located with the sequence of the mature miRNA	MODIFIER
5 prime UTR variant	A UTR variant of the 5' UTR	MODIFIER
3 prime UTR variant	A UTR variant of the 3' UTR	MODIFIER
Non coding transcript exon variant	A sequence variant that changes non-coding exon sequence in a non-coding transcript	MODIFIER
Intron variant	A transcript variant occurring within an intron	MODIFIER
NMD transcript variant	A variant in a transcript that is the target of NMD	MODIFIER
Non coding transcript variant	A transcript variant of a non coding RNA gene	MODIFIER
Upstream gene variant	A sequence variant located 5' of a gene	MODIFIER
Downstream gene variant	A sequence variant located 3' of a gene	MODIFIER
TFBS ablation	A feature ablation whereby the deleted region includes a transcription factor binding site	MODIFIER
TFBS amplification	A feature amplification of a region containing a transcription factor binding site	MODIFIER
TF binding site variant	A sequence variant located within a transcription factor binding site	MODIFIER
Regulatory region amplification	A feature amplification of a region containing a regulatory region	MODIFIER
Feature elongation	A sequence variant that causes the extension of a genomic feature, with regard to the reference sequence	MODIFIER
Regulatory region variant	A sequence variant located within a regulatory region	MODIFIER
Feature truncation	A sequence variant that causes the reduction of a genomic feature, with regard to the reference sequence	MODIFIER
Intergenic variant	A sequence variant located in the intergenic region, between genes	MODIFIER

Table 4.2: Calculated consequence of sequence ontology terms annotated as MODIFIER impact. Each variant is mapped to a SO term which is then categorised depending on impact on the transcript. This table shows SO terms with MODIFIER impact, representing variants which effect regulatory regions for the transcript, but not the coding sequences of the transcript itself.

4.3 Results

4.3.1 Genetic architecture of Sysmex parameters

GWAS and conditional analyses of 63 Sysmex parameters identified 2,172 conditionally independent associations which clustered into 849 LD sets (Section 2.2.9). LD sets are assigned to cell types depending on which phenotypes the conditionally significant variants constituting the LD set are associated with.

Although associations with red blood cell or platelet phenotypes form a larger proportion of the total findings, associations with these cell types are less likely to be distinct in comparison to previous haematological GWAS results. I performed a comparison between my study of Sysmex parameters and the study of FBC parameters by Astle *et al.* (Section 4.2.3) [15]. The study by Astle *et al.* is the largest GWAS of FBC traits and benefits from a roughly four fold larger sample size (173,480 versus 39,656 individuals). Despite this, of the 849 high LD ($r^2 > 0.8$) sets identified by my study of Sysmex parameters, 423 are novel in comparison to Astle *et al.* [15]. I find a far greater degree of overlap with LD sets containing variants associated with platelet and red cell traits than LD sets assigned to white cells (Fig. 4.1, 4.2). A total of 375 LD sets were associated solely with white cell Sysmex parameters of which 73.3% (275 sets) are not reported by Astle *et al.*. This is compared to a total of 410 LD sets assigned solely to platelet or red cell Sysmex parameters of which 30.5% are not reported by Astle *et al.* (Fig. 4.1, 4.2). This finding is consistent with low genetic correlation observed between white cell Sysmex parameters and traits studied by Astle *et al.* (Fig. 4.5) [15]. The striking difference in novel associations comparing white cell and red cell or platelet phenotypes can be explained by the intra-cellular complexity and heterogenity of white cells. Neutrophils, eosinophils, basophils can be highly granulated cells and associations with side scatter of these cells are often located in known granule genes. A full list of conditionally significant associations, their comparison with Astle *et al.*,, colocalisation with eQTL, pQTL, and disease GWAS can be found in Table A.1. Furthermore, white cells exhibit greater cellular heterogenity, especially the lymphocyte cell population which is an amalgamation of many lymphocyte sub-types, in particular reactive and antibody synthesising lymphocytes are known to vary in SFL, and FSC measurements (Section 3.2.2.3).

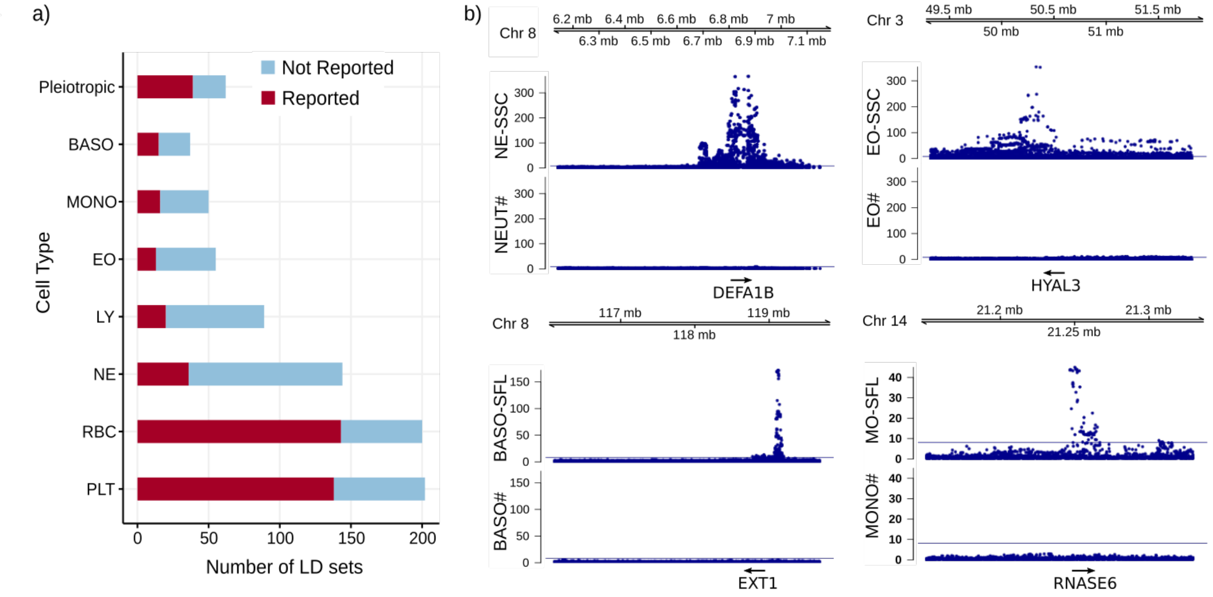

Figure 4.1: Assignment of LD sets to cell types and examples of sets labelled as 'not reported' by Astle *et al.*, 2016.

a) Conditionally significant variants were clustered into 849 high linkage disequilibrium sets which are in linkage disequilibrium $r^2 > 0.8$ representing distinct association signals. Sets were assigned to cell types depending on their association with Sysmex parameters. Sets were labelled as not reported if none of their constituent conditionally significant variants are in LD $r^2 > 0.8$ with any variants identified by Astle *et al.*, 2016 [15]. Most sets are specific to individual haematological cell types (not-pleiotropic), this demonstrates the specificity of Sysmex parameters. LD sets of white cell associations are more likely to be distinct in comparison to Astle et al.. **b)** Association plots showing the P-value (-log$_1$ 0(P)) for association of each genetic variant along the genome (x axis) with the phenotype (y axis). This plot shows statistically significant associations at genes with known roles contributing to white cell function: *DEFA1B*, *HYAL3*, *RNASE6* and *EXT1* with neutrophil, eosinophil, monocyte, and basophil cell type Sysmex parameters, and lack of genome wide significant association with respective count parameters as studied by Astle *et al.*., 2016.

An informal literature review finds association signals identified by my analysis of Sysmex parameters, but not from GWAS of FBC by Astle *et al*, are often annotated to genes by VEP which play fundamental roles in immune cell function. GWAS of white blood cell Sysmex parameters identified 767 associations mapped to 270 genes by VEP, of these genes, 185 were not identified by GWAS of FBC blood phenotypes by Astle *et al* [15]. Of these 185 newly identified genes 72 seem to have a plausible role in the immune system, as shown in Figure 4.7. An example includes the *DEFA* locus encoding α-defensin genes associated with NE-SSC. α-defensin genes at this locus have an established role in neutrophil immune function accumulating in the granules of neutrophil cells [60]. In response to pathogenic cells, the cysteine-rich cationic α-defensin peptides are released from granules of neutrophil cells and create perforations in the pathogen cell membrane [60]. Despite the established role of *DEFA* proteins in neutrophil function and physiology this locus has not been identified by GWAS of traditional neutrophil parameters. Other examples of new association signals compared to the findings of Astle *et al.*, identified by my analysis include:

- The *PRG2* gene encoding the *MBP* protein is a major component of eosinophil granules [136]. Similar to the α-defensin peptides discussed above, *MBP* is also a cationic protein which once released from eosinophil granules carries out anti-pathogenic function by perforating the cell membranes of pathogen cells [98].

- *RNASE6*, a known component of monocyte granules and a cationic ribonuclease antimicrobial protein contributing to urinary tract sterility [18].

- *EXT1*, encodes a glycosyltransferase protein contributing in heparin biosynthesis [102], heparin is known to be packaged in basophil granules [166] and has been proposed as an anticoagulant increasing blood flow to infected tissues [27].

- *HYAL3* encoding Hyaluronidase 3, a protease which degrades hyaluronan a major component of the extracellular matrix. The role of *HYAL3* in eosinophil function has not been fully elucidated. However hyaluronidases have been implicated in remodelling of the extracellular matrix and hyaluronan deposition has been shown to correlate with eosinophil infiltration of tissues [44].

As previously mentioned, assignment of LD sets to cell types shows there is little overlap between association signals across cell types (Fig. 4.1). This result indicates that the genetic determinants of Sysmex parameters are largely cell type specific. Furthermore, limited overlap is observed between Sysmex parameters of the same cell type - thus suggesting that Sysmex parameters assay genetically distinct phenomena within a given cell type (Fig. 4.2). This finding is supported by low phenotype and genetic correlation

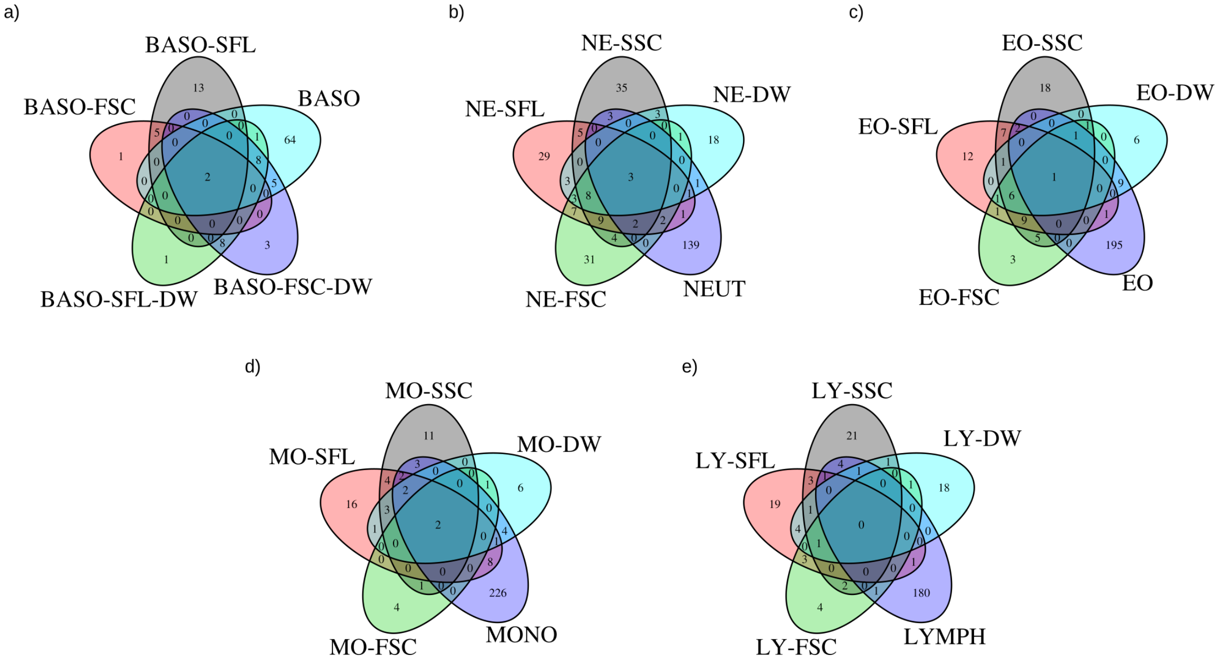

Figure 4.2: Overlap of LD sets of Sysmex parameters by cell type. **a-e)** LD sets associated with Sysmex parameters and those assigned to count traits studied by Astle *et al.*, 'DW' represents the three distribution width parameters, in the case of NE-DW: NE-SSC-DW, NE-SFL-DW, and NE-FSC-DW. Results show limited overlap between Sysmex parameters, in particular for neutrophil and lymphocyte cell types.

between Sysmex parameters across cell types and to a lesser extent within a cell type (Fig. 4.3 and 4.4).

4.3.1.1 Allelic spectra

Plotting conditionally significant variants in an allelic spectra with effect size on the y axis and MAF on the x axis shows general concordance with the expected trend: variants with high effect size having lower MAF, and variants with low effect size having higher MAF (Fig. 4.6). This trend is driven by natural selection, which eliminates variants with large effect sizes, as such the effect alleles of genetic variants with large effect sizes are generally deleterious to fitness [157]. In rare cases where arising genetic variation improves organism fitness, the mutant allele is driven to become more common in the population and is thus no longer the 'minor' allele the frequency of which is plotted in an allelic spectrum. However, there are notable exceptions to this trend in the data - these are variants which have a higher than expected effect size compared to their MAF. This phenomenon could arise due to balancing selection [157], for example, this could occur of a genetic variant is deleterious to fitness in some circumstances, but enhances the fitness of the organism in other circumstances. As an example, conditionally significant variants located in the α-defensin locus (chromosome 8, 6.78MB - 6.95MB) appear shifted from the expected

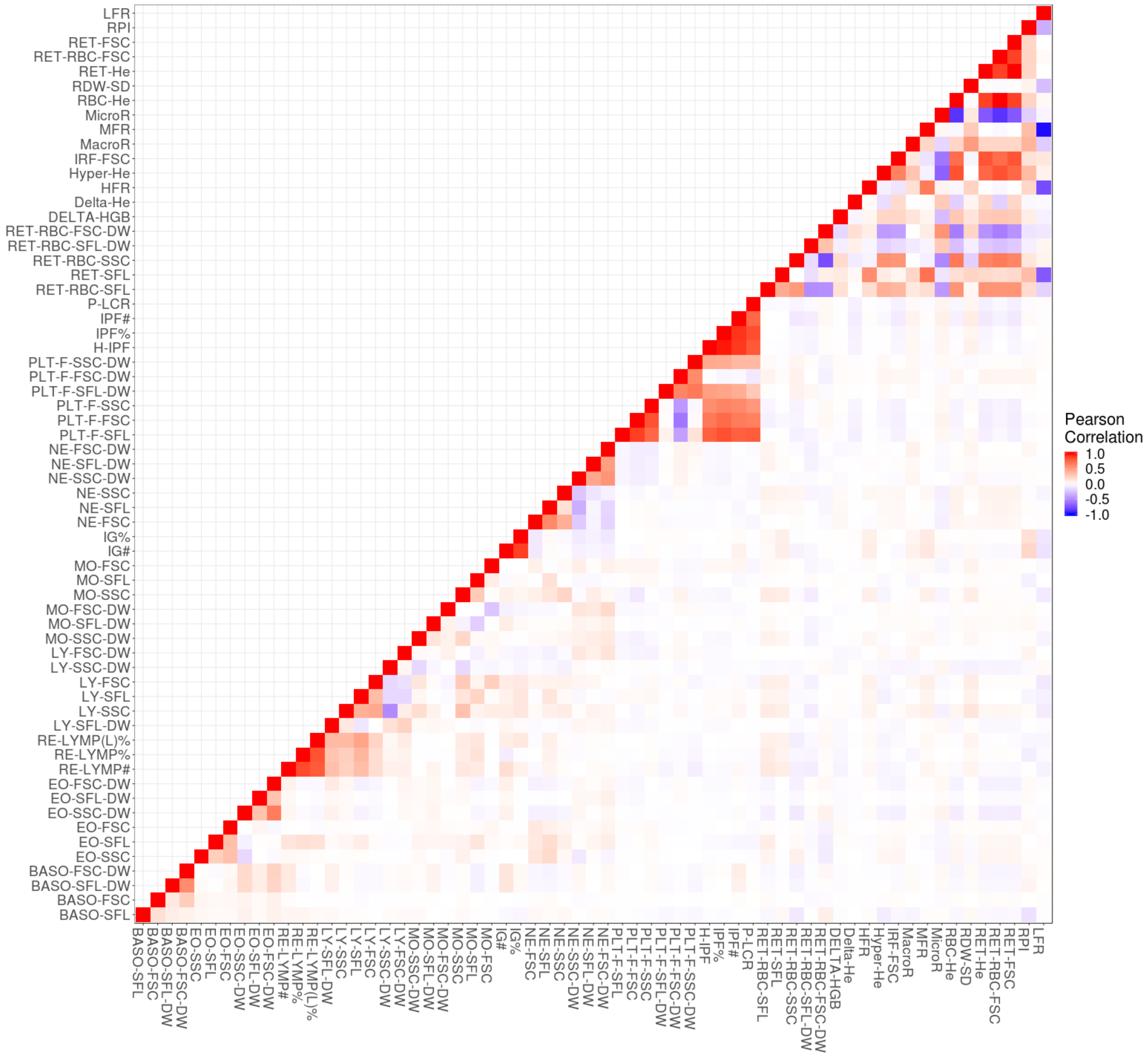

Figure 4.3: Pearson Correlation r^2 between Sysmex Parameter Phenotype Values

Pearson correlation r^2 between Sysmex parameters where the colour of each box indicates the magnitude of the correlation with red indicating positive, and blue negative correlation. Phenotypes are grouped by cell type, basophils, eosinophils, lymphocytes, monocytes, neutrophils, platelets, and red blood cells. High correlation is observed between red cell and platelet Sysmex parameters, limited correlation observed between Sysmex parameters of other cell types and very little correlation across Sysmex parameters of different cell types. Phenotype abbreviations are discussed in more detail in Chapter 3.2.2.

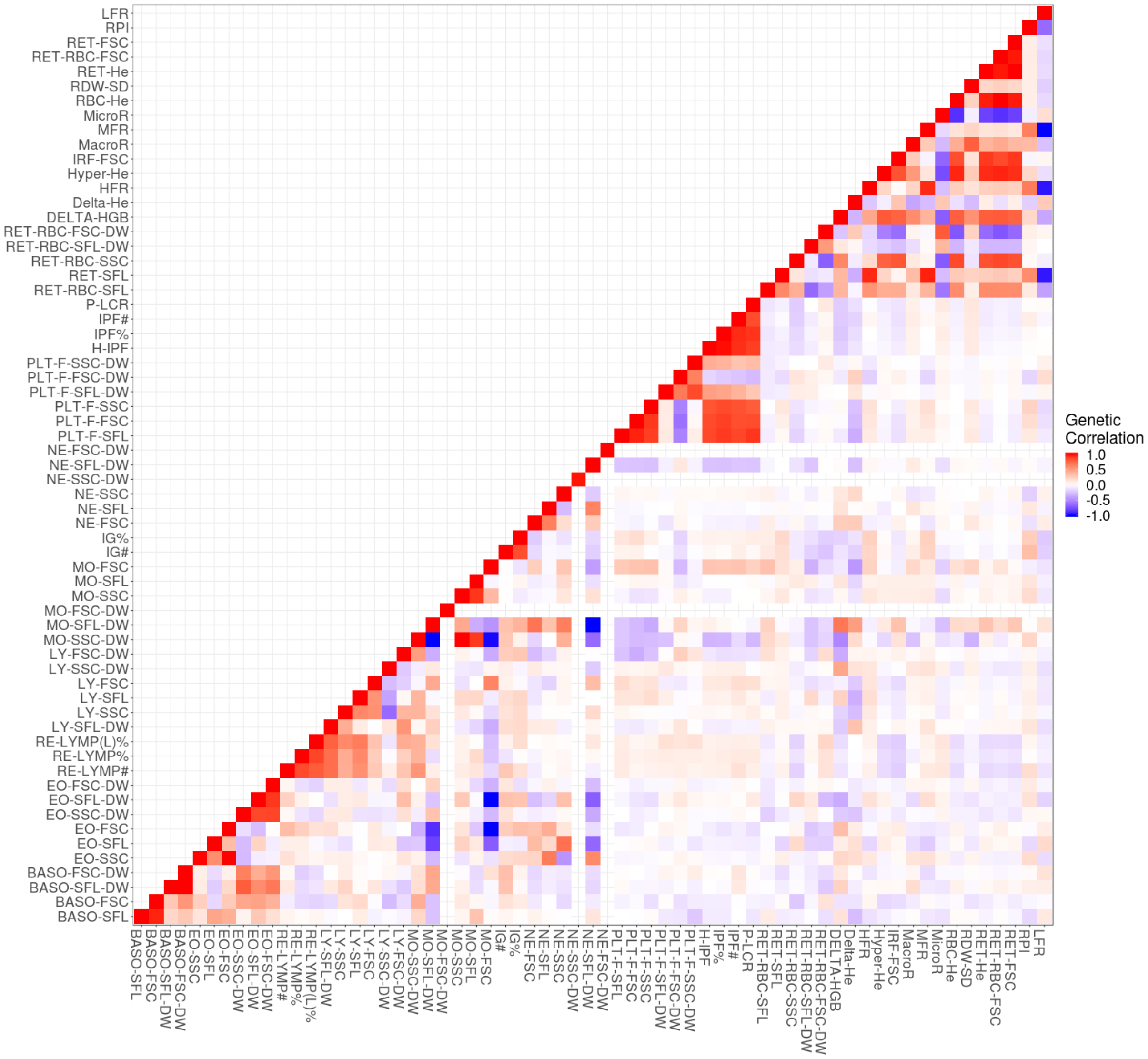

Figure 4.4: Pearson Correlation r^2 between Sysmex Parameter Phenotype Values

Genetic correlation between Sysmex parameters calculated by LD score regression where the colour of each box indicates the magnitude of the correlation with red indicating positive, and blue negative correlation. Phenotypes are grouped by cell type, basophils, eosinophils, lymphocytes, monocytes, neutrophils, platelets, and red blood cells. High correlation is observed between red cell and platelet Sysmex parameters, limited correlation observed between Sysmex parameters of other cell types and very little correlation across Sysmex parameters of different cell types. Phenotype abbreviations are discussed in more detail in Chapter 3.2.2.

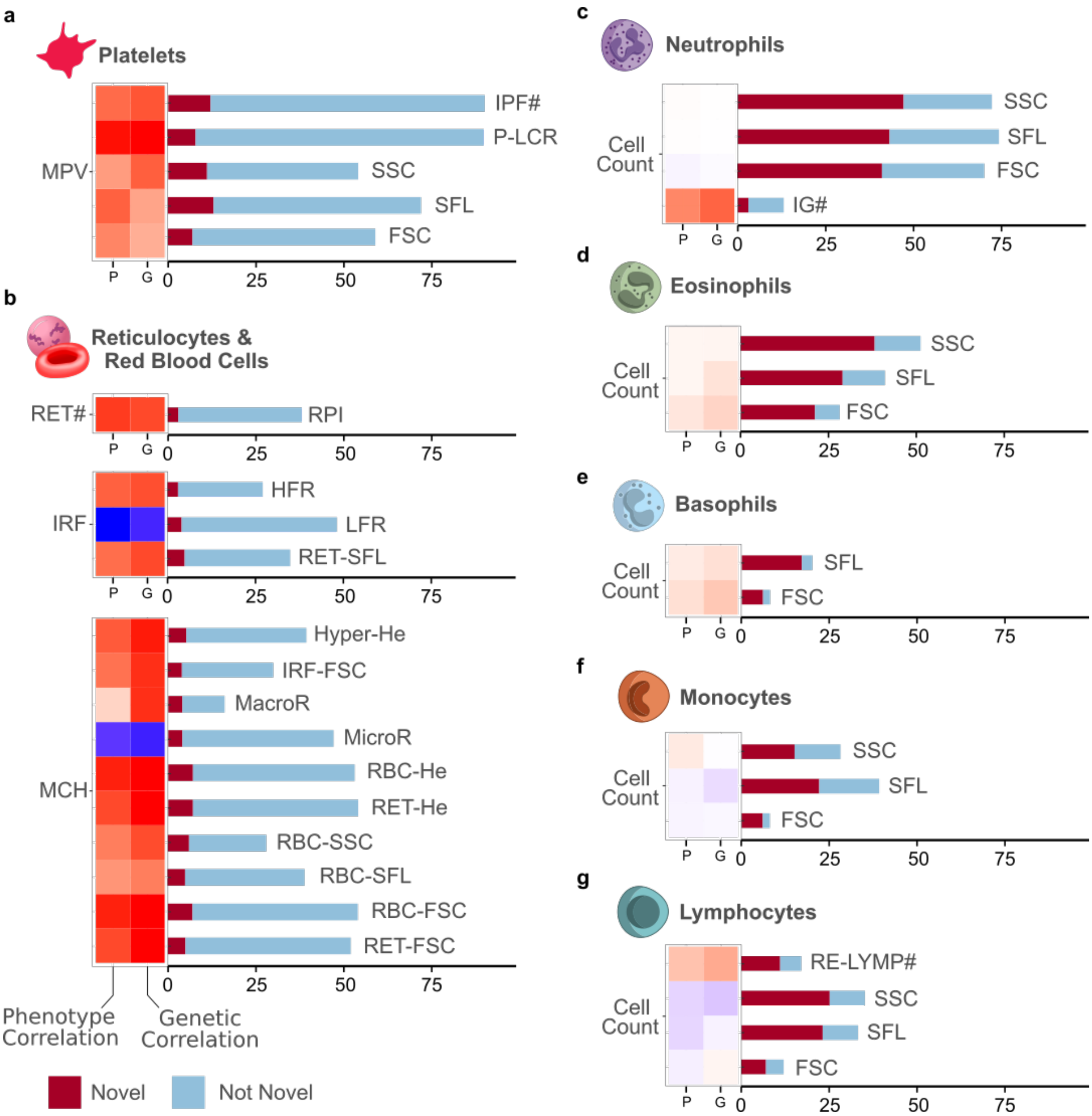

Figure 4.5: Overview of novel signals identified across Sysmex parameters and their correlation with traditional phenotypes.
The phenotypic and genetic correlation between each Sysmex parameter and a corresponding FBC measurement is represented in a heatmap and labelled by P and G respectively. For each phenotype a corresponding FBC measurement was selected for which there is the highest median correlation across all Sysmex parameters for that cell type. In addition, the number of clumps identified per parameter stratified by those novel, or not novel in comparison to Astle *et al.* are displayed. Sysmex parameters of white cells have lower correlation with related FBC blood cell measurements, red cell and platelet parameters have a higher correlation. This difference is reflected in the number of novel independent signals identified for each trait. Reticulocyte count (RET#), Mean cell haemoglobin (MCH), and Mean platelet volume (MPV). 'Cell count' represents the corresponding FBC cell count for that cell type. Phenotype abbreviations are discussed in more detail in Section 3.2.2.

relationship between effect size and MAF in an allelic spectrum (Fig. 4.6). α-defensin proteins are a component of the innate immune system encoding antimicrobial peptides released from granules to destroy pathogenic cells. Balancing selection to maintain a heterogeneous population of proteins has been long hypothesised [82, 52]. Phylogenic analysis of the α-defensin locus by comparison between primates (human, chimpanzee, orangutan, macaque, marmoset) shows divergence in the encoding of α-defensin genes, but also conservation of certain amino acid residues [52]. Evolution of the α-defensin locus seems to be influenced firstly by a need to conserve functionally important properties, but also the evolutionary advantage associated with encoding a diverse functionally diverse range of antimicrobial peptides [52]. The functional diversity of antimicrobial peptides is important as this allows activity against a range of pathogens and reduces the ability of pathogenic strains to overcome immune action [52]

My work adds further evidence to the hypothesis of balancing selection of α-defensin proteins. However, it must be noted that the α-defensin locus contains a high number of repeated genetic elements, repeated genetic elements can make the assignment of a genotyped variant to a location in the genome unreliable. This may be better addressed by genome sequencing of this locus, which may help to identify repeated elements by genome sequencing could also be an important genetic factor which modulates changes in neutrophil granularity and the NE-SSC parameter. However, in many scenarios classical genome sequencing technology can struggle to resolve regions where repeated elements constitute longer stretches of DNA [139]. In the future, next generation sequencing technology such as that provided by Oxford Nanopore may be provide a solution to better resolve repeat regions [87].

4.3.1.2 Identification of functionally relevant genes

GWAS analysis of Sysmex parameters identifies genetic signals which are annotated by VEP to be located in genes relevant to white cell function. I performed a literature review of genes identified by VEP annotation of conditionally significant variants associated with Sysmex parameters. My literature review found functional relevance of these genes in a number of blood cell functions such as haematopoiesis, cell adhesion and chemotaxis, cell activation, and others. The results of my literature review across the seven primary blood cell types (platelets, red blood cells, neutrophils, eosinophils, basophils, monocytes, and lymphocytes) are summarised in Figure 4.7 and corresponding references can be seen in Table A.1. These results show that conditionally significant variants often appear annotated to genes which perform known and functionally important roles in blood cells 4.7. Genes are further annotated by their colocalisation with eQTL, pQTL, or disease GWAS association signals which is labelled in Figure 4.7 and discussed further in Chapter 5.

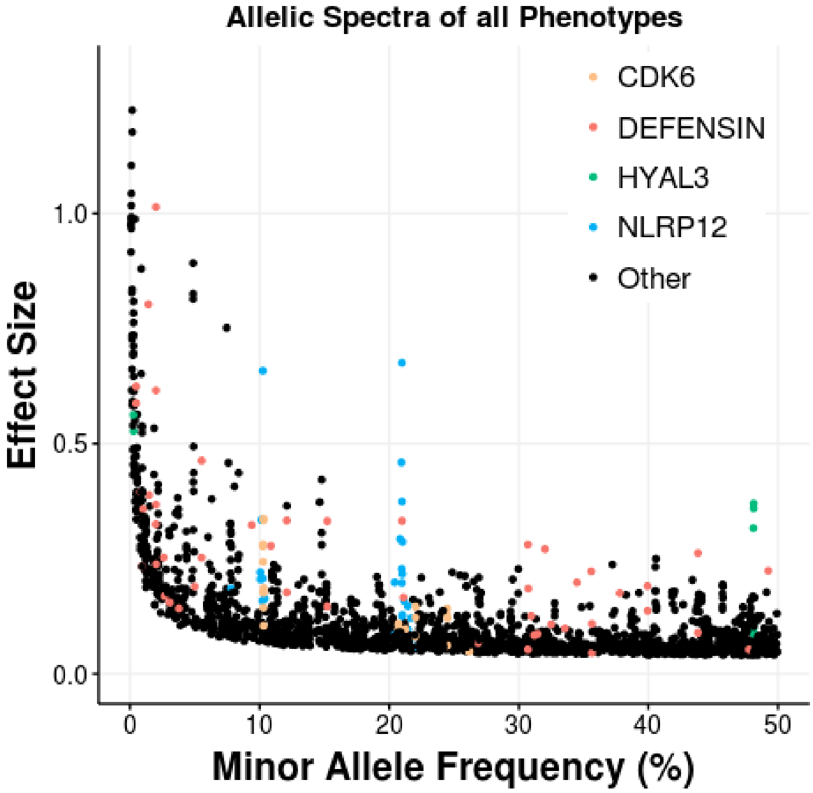

Figure 4.6: Allelic spectra of conditionally significant variants. Each conditionally significant association plotted with MAF on the x axis and effect size on the y axis. Since a conditionally significant variant may appear associated with multiple Sysmex parameters the same variant may appear more than once on the plot. Associations are coloured by their VEP annotation to *CDK6*, α-defensin, *HYAL3*, or *NLRP12* genes.

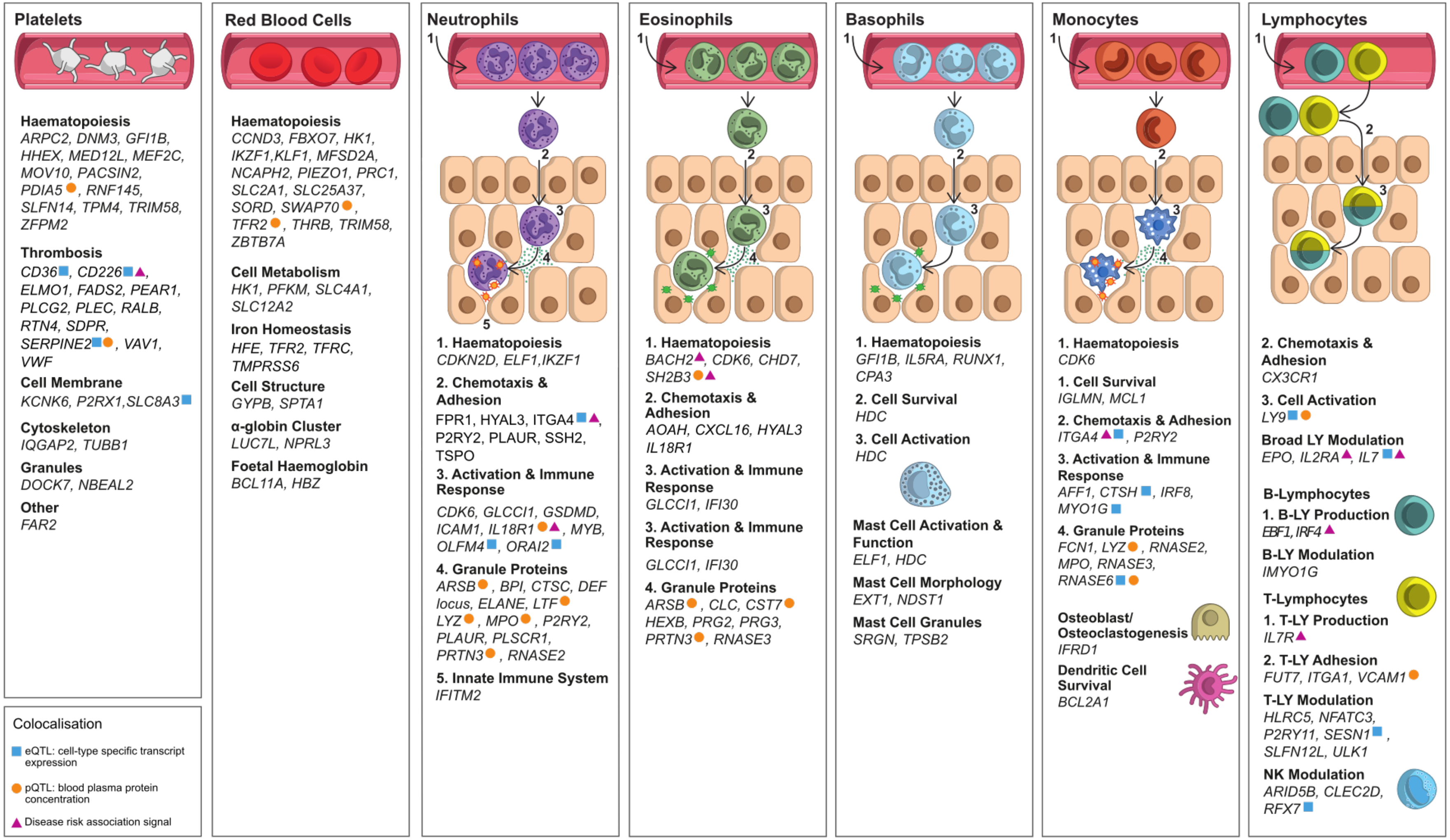

Figure 4.7: GWAS of Sysmex parameters identifies functionally important genes.
Genes identified by annotation of conditionally significant variants with VEP are assigned to functional categories with a review of the literature. A table of conditionally significant variants, VEP annotations, and relevant references to literature (Appendix A.1). eQTL (blue square), pQTL (orange circle), or disease (purple triangle) colocalisation is indicated by the relevant symbol in the figure. An eQTL or pQTL colocalisation may be with a cis signal for an distal gene not the gene name listed in the figure.

4.4 Summary

GWAS analyses can identify variants which are significantly associated with a phenotype of interest, however due to LD between variants in the genome the number of true genetic signals is not apparent from this analysis alone. I performed a conditional analysis to identify a parsimonious set of variants which represent the underlying genetic association signals. I show that the genetic determinants of Sysmex parameters are largely cell type specific, furthermore there is limited overlap between parameters for the same cell type. Finally using an LD clumping approach, I compared the total number of signals identified across all 63 Sysmex parameters in my study, with the genetic signals identified by Astle *et al.* [15]. VEP annotation of my conditionally significant variants to nearby genes and subsequent literature review shows identification of genes contributing to blood cell function including cell chemotaxis and adhesion, cell activation and immune response, and cell survival. The results of this work inform further downstream analysis to annotate genetic signals using corollary datasets from eQTL, pQTL, and disease risk GWAS studies (Section 5).

Chapter 5

Downstream analysis and biological inference

5.1 Introduction

In previous chapters I presented my GWAS analysis which has identified genetic variants associated with changes in blood cell phenotypes. However, I began this thesis by outlining my aims not only to identify new genetic associations with haematological phenotypes, but also the interpretation of GWAS results and the potential for this interpretation to inform biological and clinical experimentation. As previously explained in Section 1.6, the primary challenges in interpretation of GWAS are as follows:

- Confident identification of the genes mediating each genetic association, a starting point for further inference.
- Understanding the mechanisms of biology which lead to the emergence of a genetic association and understanding the tissue specificity of those mechanisms.
- Inferring a causal relationship between two measurements, for example a risk factor and disease risk, and the implications of this for the consideration of the risk factor as a target for therapeutic modulation.

I explored these questions by performing a number of analyses detailed in this chapter, including colocalisation and MR. In Section 1.6.2 I introduced genetic colocalisation analysis, a tool for interpretation of GWAS results which can determine the same variant is the common cause of associations with multiple phenotypes. With colocalisation I have identified genetic determinants of blood cell phenotypes which have concomitant effects on blood cell transcripts (Section 5.1.2), blood plasma proteins (Section 5.1.3), and disease risk (Section 5.1.4). I focused my analysis on cardiovascular and immune related disease outcomes due to the known role of blood cells in mediating these disease types

[83, 15, 97, 49, 150]. A broad discussion of the implications of colocalisation analysis in interpretation of GWAS results is given in Section 1.6.2.

However, colocalisation analysis is limited because it can only show that a genetic association is simultaneously influencing a set of phenotypes. Colocalisation cannot prove causal relationships between the associated phenotypes. MR can determine causal relationships between two phenotypes, often termed 'exposure' and 'outcome' (Section 1.6.3). I utilise MR to explore potential causal relationships between blood cell phenotypes and cardiovascular and immune disorders (Section 5.3.4).

I performed colocalisation between genetic associations from GWAS of 63 Sysmex parameters and GWAS of 5,995 blood cell type specific transcripts, 1,478 blood plasma proteins, and risk of 22 cardiovascular and autoimmune disorders. Blood cell type specific transcripts where chosen in order to assess the influence of associations not only on haematological parameters but expression of genes in the relevant blood cell-types. The dataset of 1,478 blood plasma proteins was chosen to study the influence of Sysmex parameters on the composition of the plasma proteome, because many clinically therapeutic drugs or candidate drugs target proteins in the plasma. We hypothesised that associations with blood cell properties (particularly granulation) as measured by Sysmex parameters would also influence composition of the blood plasma proteome. Furthermore, selection of the blood plasma proteome dataset was practically convenient due to the size of this dataset in terms of the number of proteins assayed and the sample size, furthermore this data was readily available to me as the results were generated by colleagues analysing samples from participants in the INTERVAL study. Finally, I performed colocalisation with risk of cardiovascular and autoimmune disorders as these disease outcomes are those which are known to be influenced by blood cell function.

My analysis has annotated genetic determinants of Sysmex parameters to a range of autoimmune disorders such as atopic dermatitis, multiple sclerosis, and celiac disease. My results are informative for drug design and target selection, demonstrated by two examples: replicating the known mechanism of action of Daclizumab by it's influence of lymphocyte cell properties via *IL2RA* (Section 5.3.2.1) a treatment for Multiple Sclerosis (MS), and evidence for a common genetic determinant influencing *IL-18R1* plasma protein concentration, NE-FSC, and risk for celiac disease (Section 5.3.2.4). I have identified many common genetic determinants between white cell granulation as measured by Sysmex parameters and the blood plasma proteome (Section 5.3.3). Furthermore, I perform MR analysis to assess causal association between white cell parameters and disease, and identify causal relationships between NE-SSC and coronary artery disease (CAD) or lung cancer and EO-FSC and asthma (Section 5.3.4).

Cell Type	CD Marker	Cardiogenics	CEDAR	WTCHG	BP	Total
Monocyte	CD14	758	300	432		1,490
Granulocyte Neutrophil	CD15		300			300
T-Lymphocyte	CD4		300			300
T-Lymphocyte	CD8		300			300
B-lymphocyte	CD19		300			300
Platelet	PLA		268		156	424

Table 5.1: Contribution of studies to the eQTL analysis of blood cell types. Number of individuals for which genotype and expression data was available is presented, all individuals are healthy participants except for 363 individuals in the Cardiogenics dataset whom have a previous history of coronary artery disease, BLUEPRINT (BP).

5.1.1 Quantitative trait loci

Quantitative trait loci (QTL) are genetic loci associated with variation in a quantitative trait. Examples include eQTL or pQTL, representing transcript abundance (expression) and protein concentration respectively. However, we must always enquire about the specific nature of these phenotypes, for example, eQTL analysis can use blood cell type specific RNA or RNA extracted from whole blood. The eQTL and pQTL results utilised in my work are quantitative GWAS studies of transcripts expressed by specific blood cell types and proteins in the blood plasma respectively.

5.1.2 Expression quantitative trait loci

I used eQTL data generated from a number of specific blood cell types separated using immunophenotyping cluster of differentiation (CD) markers [94]. The CD method utilises markers such as cell surface carbohydrates or proteins which can be signalling molecules or cell adhesion proteins to identify cell types [59]. In total, I obtained eQTL data for platelets, monocytes, lymphocytes, and neutrophils [94]. This eQTL analysis was performed by Kreuzhuber *et al.,* who combined individual level genotype and phenotype data collected from a number of studies: Cardiogenics, CEDAR, Wellcome Trust Centre for Human Genetics (WTCHG), and BLUEPRINT [94]. Sample sizes for each cell type ranged from between 300 to 1,490 individuals (Table 5.1). Individuals participating in the Cardiogenics, WTCHG and BLUEPRINT studies were healthy, in CEDAR of 758 participants, 395 were healthy individuals and 363 had a history of coronary artery disease [94].

5.1.3 Protein Quantitative Trait Loci (pQTL)

Sun *et al.*, performed an aptamer based assay (SOMAscan) which allowed the quantification of 3,622 plasma proteins in 3,301 participants from the INTERVAL study, of which 2,994 were studied in a GWAS of 10.4 million imputed variants. This analysis identified 1,927 significant associations ($P < 1.5 \times 10^{11}$) with 1,478 plasma proteins [164]. Protein levels were quantified with binding of aptamers to circulating proteins, in some cases it is possible that associations could result from genetic variations which influence the affinity of aptamer binding rather than changes in protein levels [164]. Sun *et al.* addressed this by comparing the GWAS results for a subset of the proteins studied by SOMAscan with a complementary antibody based Olink assay and found strong concordance between the predicted effect size of associations ($r = 0.83$) [164].

5.1.4 Disease risk GWAS

Many case-control based GWAS of disease outcomes have been performed and many associations with disease risk have been identified. However, a GWAS study in isolation does not provide a full picture on which cell types, cell functions, proteins, and transcripts are causally mediating a particular association. I have colocalised associations of haematological phenotypes with associations with disease risk and identified genetic associations with biological mechanisms involving both blood cell biology and disease aetiology. My dataset of GWAS summary statistics for 28 disease outcomes is largely focused on cardiovascular and immune disorders (Appendix 5.4), due to the known role of blood cells in mediating such disorders [83, 15, 97, 49, 150].

5.2 Methods

5.2.1 Colocalisation

Methods for colocalisation analysis (Section 1.6.2) have been implemented in software packages by a number of authors. In my analysis I utilise two implementations: *coloc* by Giambartolomei *et al.* [70] and *gwas-pw* by Pickrell *et al.* [137].

coloc uses a user-set prior for the chance of association between a SNP and the phenotype and the variance for this effect size. *gwas-pw* estimates prior parameters for association of SNPs to the phenotype using a genome-wide optimisation procedure, and averages over a set of priors for the variance of SNP effect sizes. The *gwas-pw* approach is more computationally burdensome and also requires full genome-wide association summary statistics. This is not available in the case of eQTL colocalisation as variants only 1 MB on either side of each gene were tested for association. The genome-wide approach for estimating priors is also inappropriate for pQTL colocalisation due to the very small number of signals identified genome-wide per protein (1,927 associations for 1,478 proteins tested).

In contrast, the *coloc* approach uses predefined priors, this is my preferred approach

eQTL Celltype	Sysmex Parameter
PLA	H-IPF, P-LCR, PLT-FSC, PLT-FSC-DW, PLT-SFL, PLT-SFL-DW, PLT-SSC, PLT-SSC-DW
CD19	LY-FSC, LY-FSC-DW, LY-SFL, LY-SFL-DW, LY-SSC, LY-SSC-DW, RE-LYMP(L)%, RE-LYMP#, RE-LYMP%
CD15	NE-FSC, NE-FSC-DW, NE-SFL, NE-SFL-DW, NE-SSC, NE-SSC-DW
CD14	MO-FSC, MO-FSC-DW, MO-SFL, MO-SFL-DW, MO-SSC, MO-SSC-DW
CD8	LY-FSC, LY-FSC-DW, LY-SFL, LY-SFL-DW, LY-SSC, LY-SSC-DW, RE-LYMP(L)%, RE-LYMP#, RE-LYMP%
CD4	LY-FSC, LY-FSC-DW, LY-SFL, LY-SFL-DW, LY-SSC, LY-SSC-DW, RE-LYMP(L)%, RE-LYMP#, RE-LYMP%

Table 5.2: Colocalisation between haematological parameters and cell type matched between eQTL.

A table of cell types for which eQTL data was collected by Kreuzhuber [94] and the cell type matched Sysmex parameters for which colocalisation was performed. In total colocalisation was performed for 47 eQTL and Sysmex parameter pairs.

for colocalisation with eQTL and pQTL data. Furthermore, the *coloc* was used by the authors of the blood plasma proteome pQTL study to colocalise their results with a series of eQTL and disease datasets [164]. Alternatively, for disease risk colocalisation I utilise the *gwas-pw* method. Here there are a larger number of signals across the genome, but fewer sets of GWAS summary statistics to be colocalised making the *gwas-pw* method computationally tractable. I further discuss my reasoning for implementation of two colocalisation methods in Section 5.3.1.

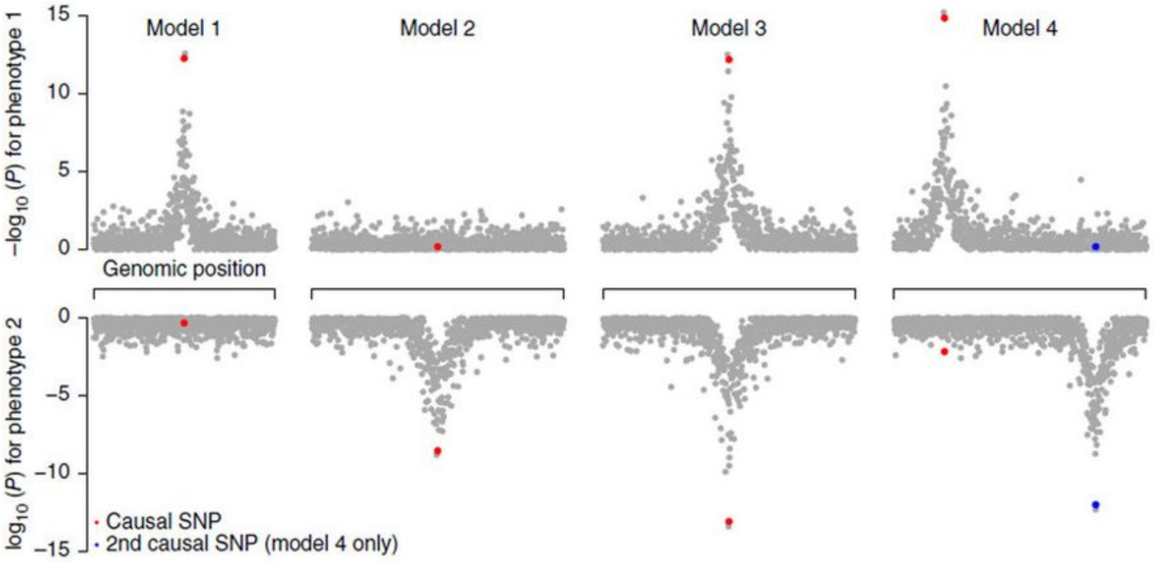

Figure 5.1: Schematic of the models considered by *gwas-pw*.
Posterior probabilities are calculated for each of the described models. The first and second model defines a single causal variant with phenotypes 1 and 2 respectively, the third model describes a single causal variant with both phenotypes, and model 4 shows two distinct causal variants for phenotypes 1 and 2. The null model with no association in either phenotype is not shown (Figure source: [137]).

5.2.1.1 Colocalisation with gwas-pw by Pickrell et al., 2016

Using a Bayesian approach, posterior probabilities are computed for following models which represent the underlying genetic architecture assuming at most one causal signal for each phenotype in the locus. The total posterior probability will sum to 100%, thus Pickrell *et al.* assume that all possible outcomes are captured in the following scenarios (Fig. 5.1):

- Null Hypothesis: No associated genetic variants with either trait.
- Model 1: The locus contains one genetic variant which influences the first phenotype.
- Model 2: The locus contains one genetic variant which influences the second phenotype.
- Model 3: The locus contains one genetic variant which influences both phenotypes.
- Model 4: The locus contains two separate genetic variants which influence the first and second phenotype respectively.

Calculation of posterior probabilities for the models (Fig. 5.1) begins with calculation of Bayes factors which represent the evidence for a variant being associated with the first or second phenotype: $BF^{(p)}$, where $p \in \{1, 2\}$ is an index over the two phenotypes. Bayes factors are the ratio between the likelihood ratio of a null (the variant is not associated) and alternate hypotheses (the variant is associated). Bayes factors for SNPs in a chosen locus are summed to calculate Regional Bayes Factors, RBF_p for each model in the locus, and RBFs are then used to calculate the posterior probability for each of the four models in the locus.

Bayes factors are calculated using the Wakefield approximation, the Wakefield approximation allows computation of Bayes factors from frequentist p-values [174]. A Bayes factor is calculated for the association of each SNP with phenotype 1 or 2 indexed by $p \in \{1, 2\}$:

$$WABF_p = \sqrt{1 - r_p} exp[\frac{Z_p^2}{2} r_p] \tag{5.1}$$

Where $Z_p = \frac{\hat{\beta}_p}{\sqrt{V_p}}$ and $r_p = \frac{W_p}{V_p + W_p}$, $\hat{\beta}_p$ is the estimated effect size for association of the SNP with phenotype p, and $\sqrt{V_p}$ is the standard error of the effect size estimate. r_p is a shrinkage factor computing the ratio between variance of the prior W and total variance. Thus, effect sizes are distributed as follows $\beta_p \sim \mathcal{N}(0, W_p)$ with W_p set to 0.01, 0.1, or 0.5 and the Bayes factors are averaged over those values. Not much reasoning is given by Pickrell *et al.,* to justify this choice for their prior on the variance of the effect size. However, it could be argued that the degree of justification for the assignment of a prior should be proportional to the degree of influence that prior will have on the final estimated outcome. Most priors are assigned on the mean of an estimate (such as in the *coloc* method), here Pickrell *et al.,* assign a very broad set of priors on the variance. Therefore, I do not find their lack of justification for their choice problematic. Since the publication of gwas-pw in 2016 a better alternative has not been proposed (at the time of writing). As mentioned, Bayes factors are averaged over the three prior values and are defined as follows for the first three models:

$$BF^{(1)} = WABF_1 \tag{5.2}$$

$$BF^{(2)} = WABF_2 \tag{5.3}$$

$$BF^{(3)} = WABF_1 WABF_2 \tag{5.4}$$

Broadly speaking, the calculation of the Bayes factors with the Wakefield approximation relies on the asymptotic assumption that the sample size is “large” which is generally satisfied in the context of a GWAS study. For a formal proof of Eqn. 5.1 see the Appendix of Wakefield 2009 [174]. Regional bayes factors are defined across an entire locus to represent the models defined above (Fig. 5.1), Regional Bayes Factors are computed as follows for models (m) 1, 2 and 3, where K is the total number of SNPs in the locus indexed by i:

$$RBF^m = \sum_{i=1}^{K} \pi_i^{(m)} BF_i^{(m)} \tag{5.5}$$

The Regional Bayes Factor for model 4 is defined as follows:

$$RBF^4 = \sum_{i=1}^{K} \sum_{j=1}^{K} \pi_i^{(1)} BF_i^{(1)} \pi_j^{(2)} BF_j^{(2)} I[i \neq j] \tag{5.6}$$

The purpose of the $I[i \neq j]$ term is to restrict the sum to pairs of Bayes factors for which the corresponding causal variants are distinct for the two phenotypes. $\pi_i^{(m)}$ represents the prior probability of association of SNP i with the phenotype. The SNP priors for all models (m) are set as follows: $\pi_i^{(m)} = \frac{1}{K}$, where K is the total number of SNPs in the locus.

Finally, using the regional Bayes factors, the following likelihood function is constructed indexing over all loci in the genome by I. Prior probabilities for each of the four models are identified by optimisation of the following likelihood function:

$$l(\theta|D) = \sum_{k=1}^{I} \log(\Pi_0 + \sum_{m=1}^{4} \Pi_m RBF_k^{(m)}) \tag{5.7}$$

Where Π_0 is the prior probability that a region has no associated genetic variants, and $\Pi_{(m)}$ is the prior probability of each of the four models (m) described above. $RBF_i^{(m)}$ is the Regional Bayes Factor of each of the four models indexed by m. The summation indexed by m is over each of the four models or hypotheses. $l(\theta|D)$ is the definition of posterior probability for all four models where θ represents parameters for all four models and null hypothesis. We identify the prior probabilities (Π) by maximising the likelihood function $l(\theta|D)$, thus identifying the prior probabilities for each model from the data itself. From here, posterior probabilities (PP) for each locus can be constructed using the Regional Bayes Factors RBF_i^m and prior probabilities Π_m, where m is an index over all possible models:

$$PP_i^{(m)} = \frac{RBF_i^{(m)} \Pi_{(m)}}{\sum_{m=0}^{4} RBF_i^{(m)} \Pi_{(m)}} \tag{5.8}$$

The method assumes at most one causal signal per phenotype in the locus of interest. Furthermore, it is not possible to differentiate between models 3 and 4 if the causal variants for each of the traits are in high LD. Note that calculation of the prior for model three (Π_3) is defined as the prior for the proportion of genomic regions containing a common variant that detectably influences both phenotypes. If there is indeed a common causal variant which effects both phenotypes but this association signal is very weak, this locus will not have a high posterior probability (PP) for colocalisation [137].

5.2.1.2 Colocalisation with coloc by Giambartolomei et al., 2014

Giambartolomei *et al.* begin by defining a null hypothesis and four alternative hypotheses as described above, however the definitions of models 3 and 4 are switched when compared to the colocalisation implementation by Pickrell *et al.* The posterior probability for each of the four hypotheses h is defined as follows:

$$P(H_h|D) \propto \sum_{S \in S_h} P(D|S)P(S) \tag{5.9}$$

Where h is one of four hypotheses, and S_h represents a complete set of all possible SNP 'configurations' which are true under each hypothesis. A configuration is a pair of lists, where each list contains a binary element for each variant in the locus (Fig. 5.2). Every configuration has two lists and each list has up to one binary element set to true to represent the variant which is the causal mediator of the association signal. Each hypothesis, for example, the hypothesis for a colocalisation occurring, has multiple possible configurations. This is because it could be any one (or none) of all the variants in the locus (represented by elements in the list) which could be associated with the phenotype. The hypothesis for a colocalisation occurring, has n total possible configurations, where n is the number of variants in the locus. An example of three configurations is presented in Figure 5.2.

coloc makes the assumption that the prior for association is consistent across SNPs and therefore we can simplify to the following:

$$P(H_h|D) \propto \sum_{S \in S_h} P(D|S)P(S) = P(S|S \in S_h) \times \sum_{S \in S_h} P(D|S) \tag{5.10}$$

Where the summation is summing over every SNP in the configuration set for that model represented by S_h. To avoid calculating the proportionality constant which is the Bayesian normalising constant (Eqn 5.9), Giambartolomei *et al.* divide by $P(H_0|D)$ changing the calculation of posterior probability to that of posterior odds:

$$\frac{P(H_h|D)}{P(H_0|D)} = \sum_{S \in S_h} \frac{P(D|S)}{P(D|S_0)} \times \frac{P(S)}{P(S_0)} \tag{5.11}$$

The first term in the equation $\sum_{S \in S_h} \frac{P(D|S)}{P(D|S_0)}$ is summation of Bayes factors for SNPs in each configuration within each hypotheses (indexed by S). This is the definition of Regional Bayes Factors made above (Eqn. 5.5). The calculation of such is performed in the same way from summary statistics using the Wakefield approximation (Eqn. 5.1). The

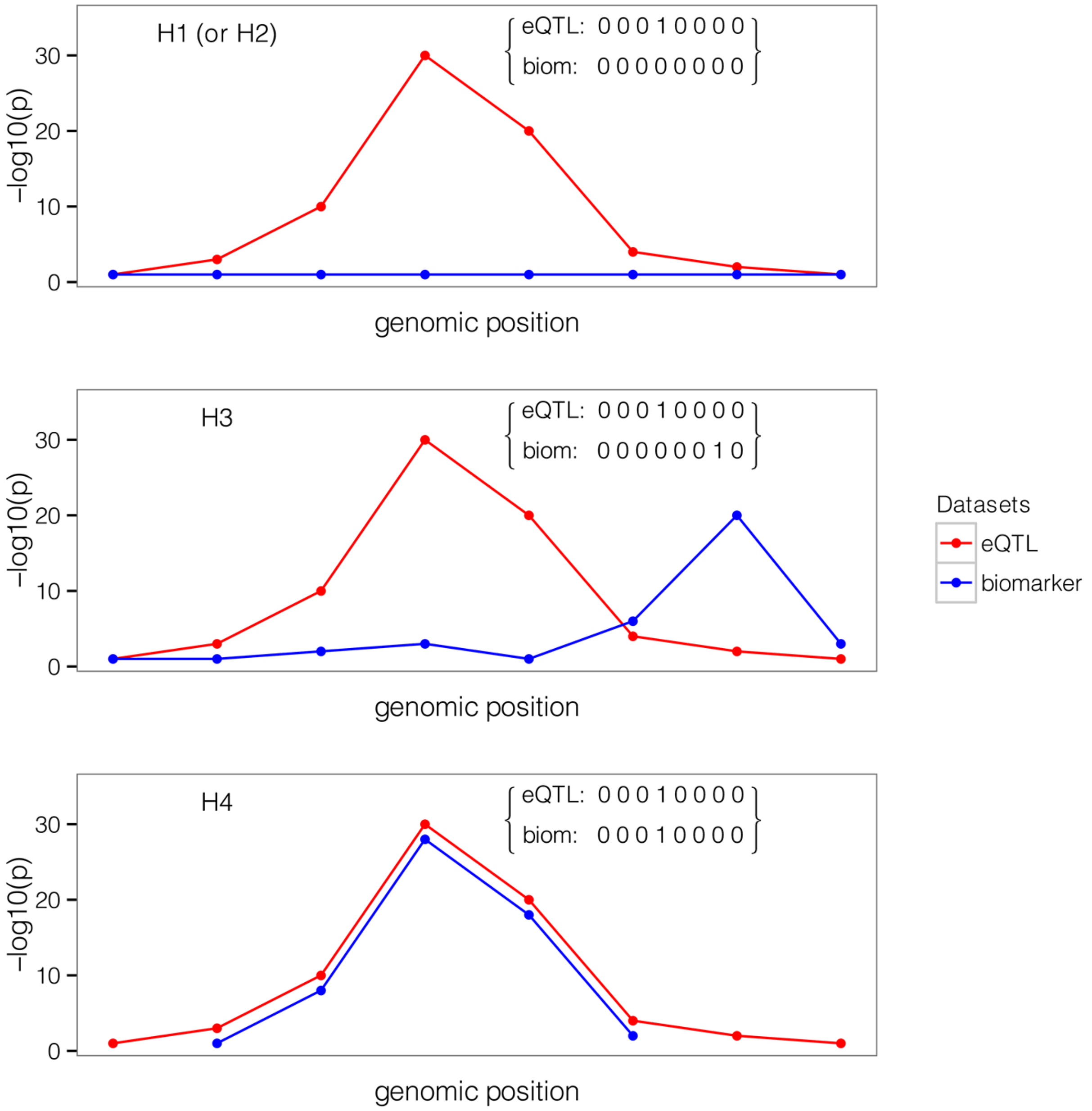

Figure 5.2: Example of one configuration for each of the four hypotheses. Configurations are represented by binary vectors, where each element in the vector is a SNP. A value of 1 indicates the SNP is causally associated with the phenotype, and 0 indicates that the SNP is not associated with the phenotype. One configuration per hypothesis (model) is displayed, each model will have a large number of configurations and only one of which is shown in the figure. The first plot shows the a single causal variant associated with the first or second phenotype, the second plot shows two causal variants associated with phenotypes 1 and 2 respectively, and the final plot shows a single causal variant associated with both phenotypes (Figure source: [70]).

second term $\frac{P(S)}{P(S_0})$ represents the prior odds of the model under consideration and the null hypothesis. Prior probabilities are defined by Giambartolomei as the following, p_1, p_2 representing the prior for association of a SNP with the first or second trait respectively, and p_{12} representing the prior for association of the SNP with both the first and second trait. Given that a SNP must exist in one of the four models defined above (Fig. 5.2), $p_0 + p_1 + p_2 + p_{12} = 1$ where p_0 is the prior for association with no trait.

In my analysis to perform colocalisation between association signals with the Sysmex parameter traits and pQTL or eQTL association signals I used the following prior probabilities:

- The prior for association of the SNP with trait 1 or 2: $p_1 = p_2 = 1 \times 10^{-4}$.
- The prior for the SNP being associated with both traits $p_{12} = 1 \times 10^{-6}$.
- The prior for the SNP being associated with neither trait: $p_0 = 0.999799$.

The choice of priors represents an assumption of my approach, my confidence in this assumption is based on the following factors:

- The successful implementation of *coloc* with the same prior probabilities by peer reviewed publications which colocalise association signals with eQTL [70] and pQTL [164] data, and sensitivity analysis performed by Giambartolomei *et al.* [70].
- The set priors are conservative given the design of my colocalisation experiment. I only ever perform colocalisation in a locus if the following two conditions are met: 1) There is a significant association in that region with both phenotypes. 2) The conditionally significant variant associated with the Sysmex parameter has a significantly associated proxy of LD $r^2 > 0.8$ in the partner phenotype. The significance threshold in the partner phenotype is defined separately based on that specific GWAS study.
- A thorough and manual search through all the purported colocalising loci generated from this prior inspecting colocalising loci and the LD structure within those loci to check the purported colocalisations.

Finally, it must be stressed that no statistical procedure can prove an outcome with full certainty. In the colocalisation approach a posterior probability for each of the four aforementioned models including that of a colocalising loci is generated. When discussing results I always resolve to communicate to the reader the posterior probability for the discussed colocalisation being ‘true’ rather than presenting a binary true / false outcome which simplifies the inherent uncertainty in the statistical procedure.

5.2.2 Mendelian randomisation

A MR analysis uses genetic variants as instrumental variables to assess a causal relationship between an exposure and outcome (Fig. 1.21). MR assesses causality utilising 'instrumental variables' derived from the results of a GWAS study (Section 1.6.3). Instrumental variables are independent genetic variants associated with the exposure of interest. A causal association between the exposure and outcome of interest can be tested using the inverse variance weighted (IVW) MR model. The following three assumptions must hold for a genetic variant to be a valid instrumental variable [30]:

1. The variant must be predictive of the exposure, thus have a significant association with the exposure.

2. The variant must be independent of any measured or unmeasured confounding factors which influence both the exposure and outcome.

3. The variant must not influence the outcome through any pathway other than the chosen exposure, often termed the 'exclusion restriction criterion'.

Assumption 1) can be tested with a standard GWAS analysis which determines significance of association between a genetic variant and phenotype. However, assumptions 2) and 3) are more difficult to test as they depend on factors which may not be measured. For example, if a genetic variant influences an alternative unknown factor which also effects the outcome, assumption 2) will be broken. Similarly, if the genetic variant is associated with changes in an unmeasured confounding factor which influences both the outcome and exposure, this could induce a seemingly causal relationship between the exposure and outcome. As we are modelling complex biological systems, assumption 3) is rarely ever true, variants are generally pleiotropic, meaning they influence multiple traits and phenotypes. In the context of IVW analysis assumption 3) is relaxed to assume 'balanced pleiotropy' between all instrumental variables. Balanced pleiotropy suggests that the overall sum of pleiotropy across all instrumental variables should sum to zero. This can be tested qualitatively with a funnel plot, or quantitatively using MR-Egger an extension of the IVW model which allows the intercept of the regression line to vary. The assumptions for valid instrumental variables rarely holds true for complex phenotypes such as the Sysmex parameters studied in my analysis. Therefore, I utilise a number of alternative MR models which relax the assumptions listed above and applied these methods as sensitivity analyses to determine if estimated causal effects are consistent across the multiple MR models with differing assumptions, this is discussed further in Section 5.2.2.2.

5.2.2.1 Mendelian Randomisation Software Analysis Protocol

My MR analysis began with a set of conditionally independent variants which are associated with the exposure and were intended to be instrumental variables for this exposure of interest. Following this, my mendelian randomisation protocol proceeded with the following steps which were implemented in a custom *R* pipeline utilising the *TwoSampleMR* package by Hemani *et al* [80]:

1. The instrumental variables must be independent with each other, not only being conditionally independent but also filtered to ensure none are pairwise LD higher than 0.6 r^2.

2. I collected the univariate summary statistics (estimated effect size and standard error of this estimate) for association of the instrumental variables with the exposure.

3. I collected the univariate summary statistics for association of each instrumental variable with the outcome.

4. In the case where the instrumental variable in question does not exist in the GWAS of the outcome, a close proxy with LD greater than 0.8 r^2 is used instead.

5. The effect size directionality between the exposure and outcome are 'harmonised', as in many cases the definition of reference and alternative allele for a variant differs between GWAS studies. At this stage, insertion or deletion variants are removed due to potential inconsistencies in the way the alleles of such variants can be coded and assigned to a specific base-pair in the genome.

6. I performed, not only the standard IVW and egger MR analysis, but also nine other MR tests to test for robustness of a purported causal association. These sensitivity tests are in Table 5.2.2.2 and discussed further in Section 5.2.2.2.

My usage of an LD threshold of 0.8 r^2 is based on similar analysis by other authors [80] [15], a correlation of 0.8 r^2 (given a maximum possible r^2 of 1.0) between two variants is strong evidence that those variants are largely tagging the same underlying genetic changes in most individuals. In step 4) of my protocol I removed insertion or deletion variants simplifying the analysis, but also removing potentially informative instrumental variables. This simplifying step also allowed me to utilise the *TwoSampleMR R* package and *MRBase* platform by Hemani *et al.,* to collect and harmonise instrumental variables [80].

5.2.2.2 Mendelian randomisation models

In total my MR analysis utilised 12 MR models, each of which varies slightly in the estimation of causal effect between exposure and outcome. The 12 methods can be assigned to the following categories: IVW, median based, and MR egger regression. MR methods may be further modified by: weighting, penalisation, robust regression, and both penalisation and robust regression together (Table 5.2.2.2).

Inverse variance weighted

IVW MR begins by calculating a ratio of association estimates between exposure and outcome for each instrumental variable:

$$\hat{\theta}_j = \frac{\hat{\beta}_{Yj}}{\hat{\beta}_{Xj}} \tag{5.12}$$

Where j is an index over all instrumental variables, $\hat{\beta}_{Yj}$ is the estimated effect size of variant j on the outcome, and $\hat{\beta}_{Xj}$ is the estimated effect size of that variant on the exposure. The estimation of variant effect sizes on a phenotype (exposure or outcome) is performed as part of a GWAS analysis and this was presented in Section 1.5. These ratio of association estimates are combined [36] to estimate the causal association ($\hat{\theta}_{IVW}$) as follows:

$$\hat{\theta}_{IVW} = \frac{\sum_j \hat{\beta}_{Xj}^2 se(\hat{\beta}_{Yj})^{-2} \hat{\theta}_j}{\sum_j \hat{\beta}_{Xj}^2 se(\hat{\beta}_{Yj})^{-2}} \tag{5.13}$$

The IVW estimator has the effect of weighting each instrumental variable by $\hat{\beta}_{Xj}^2 se(\hat{\beta}_{Yj})^{-2}$, or conceptually the ratio between the influence of the instrumental variable on the exposure and uncertainty in the estimated effect of the variant on the outcome. A variant with a small effect on the exposure and highly uncertain influence on the outcome has a down-weighted influence on the final causal estimate $\hat{\theta}_{IVW}$.

Median based

The IVW method for estimating the causal effect becomes biased if even a single instrumental variable breaks the aforementioned assumptions (Section 5.2.2). This was described by Bowden *et al.,* as the IVW method having a 0% ‘breakdown level’ [30]. Bowden *et al.,* proposed the median based MR approach where the causal estimate is simply calculated as the median of all the calculated ratios between estimated effect on outcome and exposure (Eqn. 5.12) [30]. This median based method has a 50% ‘breakdown level’, meaning that up to 50% of the instrumental variables can be invalid without resulting in a biased causal estimate [30].

However, this median based approach is inefficient, particularly when the estimate of effect sizes of the instrumental variables is uncertain. Furthermore, the certainty of the causal estimate will not improve with an increasing number of instrumental variables. Therefore, Bowden *et al.,* also proposed the weighted median estimator. Here each instrumental variable j is given a weight w_j where weights are computed similar to the IVW approach above: $\hat{\beta}^2_{Xj} se(\hat{\beta}_{Yj})^{-2}$. Weights are then standardised in order to sum to 1 the weighting of each variant then informs identification of the 'median' point. Given an ordered list of effect size ratios $\hat{\theta}_j$, a cumulative weight $s_j = \sum_{k=1}^{j} w_k$ is computed and a distribution is created which has an estimate $\hat{\theta}_j$ at it's $p_j = 100(s_j - \frac{w_j}{2})$ percentile [30]. We can see from this equation that the median (50% percentile) point will be shifted to be 'earlier' compared to the classical paradigm if the instrumental variants with lower estimated ratios are up-weighted, and the opposite in the converse scenario [30].

MR Egger regression

The IVW approach assumes that instrumental variants with no effect on the exposure also have no effect on the outcome. This assumption mandates that the instrumental variable should not effect the outcome except through the exposure of interest. However, this assumption is often broken due to the pleiotropic nature of genetic variants which often effect multiple phenotypes. The MR egger approach replaces the aforementioned assumption with a weaker assumption which states that pleiotropic effects of the variants on the outcome may exist, but must be independent and should not correlate with the magnitude of association with the outcome [29].

MR egger introduces an intercept term which represents the overall pleiotropic effect of the instrumental variables on the outcome [29]. As before, $\hat{\beta}_{Xj}$ is the estimated effect on the exposure, and $\hat{\beta}_{Yj}$ estimated effect on the outcome:

$$\hat{\beta}_{Yj} = \gamma_0 + \gamma_E \hat{\beta}_{Xj} \tag{5.14}$$

Here γ_E is be computed as the estimated causal effect of the exposure on the outcome and offers a more flexible approach compared to the IVW method. However the additional degree of freedom introduced by estimating the intercept parameter γ_0 will decrease power to detect causal relationships compared to the IVW approach [29].

It should be noted that before an MR egger analysis, instrumental variables must be re-orientated so that all genetic variants are in the 'positive' quadrant [29]. This means that for all variants where $\hat{\beta}_{Xj} < 0$, both estimated effect sizes on the exposure and outcome are multiplied by -1 to ensure that estimated effect size on the exposure is not less than zero. Orientation of genetic variants is arbitrary and depends on which allele is considered to be the 'effect' allele. Consistent orientation of instrumental variables allows proper estimation of the intercept term (Eqn. 5.14).

Penalised regression

One implicit assumption made by the IVW approach is that the influence of the exposure on the outcome is consistent regardless of the 'pathway' by which the exposure is influenced. More specifically, sub-sets of genetic variants may influence the exposure by differing biological pathways that then have a causal effect on the outcome of a different magnitude. This assumption is unlikely to be true in practice especially when working with complex intermediate traits such as the Sysmex parameters in question. Both 'penalised' or 'robust' regression techniques address this heterogenity by reducing the impact of outlying instrumental variables (or variants) on the causal estimate.

Penalised regression will down-weight instrumental variables with a heterogeneous (outlying) $\hat{\theta}_j$ ratio (Eqn. 5.12) [30], where heterogenity calculated by the Cochrans Q statistic [30]. The Q statistic is calculated as follows:

$$Q_j = \hat{\beta}_{Xj}^2 se(\hat{\beta}_{Yj})^{-2}(\hat{\beta}_j - \hat{\theta}) \tag{5.15}$$

Here $\hat{\theta}$ is the causal estimate of the IVW or egger regression, depending on which method the penalised modifier is applied to. Following this, the weights of the variants are calculated as follows:

$$w_j^* = \hat{\beta}_{Xj}^2 se(\hat{\beta}_{Yj})^{-2}\text{min}(1, 20q_j) \tag{5.16}$$

Where q_j is the one sided upper P-value of Q statistic Q_j on a chi-squared distribution of degree freedom 1 [30]. The effect of this approach is that most variants are not influenced by the penalisation, but the influence of outlying variants on the final causal estimate will be severely down-weighted.

Robust regression

Robust regression is designed to allow greater tolerance to outlying instrumental variables by down-weighting the influence of these data-points on the final estimate [142]. In this example the MM-regression procedure is used [9]. In essence, Tukey's bisquare objective function is used to down-weight outlying instrumental variables in the estimation procedure:

$$w(r_j) = \begin{cases} [1 - \frac{r_j}{c}^2]^2 & |r_j| < c \\ 0 & |r_j| \geq c \end{cases} \tag{5.17}$$

Where r_j is the residual of data-point or instrumental variable j, $w(r_j)$ is the calculated weight, and c is a tuning parameter. In the MM-estimation robust regression protocol the tuning parameter c is initially set to 1.548 then 4.685 in a two step procedure [9] [142].

Model	Description	Assumptions	Ref
IVW	Weights of variants set by variance of association to outcome.	As listed in Section 5.2.2	[29]
Penalised IVW	Wald ratios weighted by heterogenity calculated by Cochran's Q statistic.	"	[184]
Robust IVW	Regression minimises absolute residual (rather than squared residual) to reduce effect of heterogenity.	"	[184]
Penalised Robust IVW	Combination of the above.	"	[184]
Simple Median	Calculation of causal estimate using median of wald ratios.	At least 50% of the IVs must meet IVW assumptions	[30]
Weighted Median	Wald ratios weighted by Cochran's Q statistic.	"	[30]
Penalised Weighted Median	Combination of the above.	"	[30]
MR-Egger	Model directional pleiotropy by modelling intercept of the regression line.	Assumption of balanced pleiotropy replaced with InSIDE. InSIDE: Assume magnitude of IV pleiotropy is not correlated with effect size of association with exposure.	[29]
Penalised MR-Egger	Penalised regression as described above.	"	[184]
Robust MR-Egger	Robust regression as described above.	"	[184]
Penalised Robust MR-Egger	Penalised Robust regression as described above.	"	[184]

Table 5.3: Mendelian randomisation methods utilised for sensitivity analysis.
I utilised 11 MR methods for sensitivity analysis to identify exposure and outcome causal associations which are robust to the assumptions made by various MR models. Assumptions include the three assumptions for instrumental variables described in Section 5.2.2 and additional or relaxation of those assumptions as described in the table.

5.3 Results

In this section, I present my results of colocalisation and MR analysis of the genetic determinants of Sysmex parameters identified in Chapter 4. The purpose of MR is to identify a causal association between tested Sysmex parameters (purported risk factors) and disease outcomes (Section 1.6.3). Colocalisation can determine if two phenotypes share a common genetic determinant in a locus of association. I discussed my motivations for colocalisation regarding biological and aetiological inference in Section 1.6.2.

5.3.1 Colocalisation of genetic determinants of Sysmex parameters with disease risk

The total set of GWAS summary statistics utilised for disease colocalisation is listed in Tables 5.4, 5.5, and 5.6, and summary statistics utilised for eQTL colocalisation are listed in Table 5.1. Furthermore, Sun *et al.,* performed GWAS analysis for 1,478 plasma proteins as measured by SOMAscan (Section 5.1.3) which I also studied in my colocalisation analysis, a full list of plasma proteins can be seen in the relevant publication [164]. As previously discussed in Section 5.2.1, the *gwas-pw* procedure utilised for disease colocalisation assigns prior probabilities based on a genome-wide optimisation procedure. This genome-wide calculation is not possible in the case of eQTL colocalisation as the association study was limited to a 1 MB range around each gene, and inappropriate in the case of pQTL colocalisation due to the small number of signals identified per protein phenotype (1,927 associations identified across 1,478 proteins tested). Therefore, for eQTL and pQTL colocalisation I utilised the *coloc* approach (Section 5.2.1.2).

My colocalisation analysis identified 134, 164, and 74 variant-trait associations which colocalise with atleast one eQTL, pQTL, and disease risk GWAS respectively. Furthermore, there are 6, 15, and 5 variant-trait associations colocalising with atleast one eQTL and pQTL, pQTL and disease risk, and eQTL and disease risk association signal. There are no variant-trait associations with colocalise with a eQTL, pQTL, and disease risk association signal (Fig. 5.4).

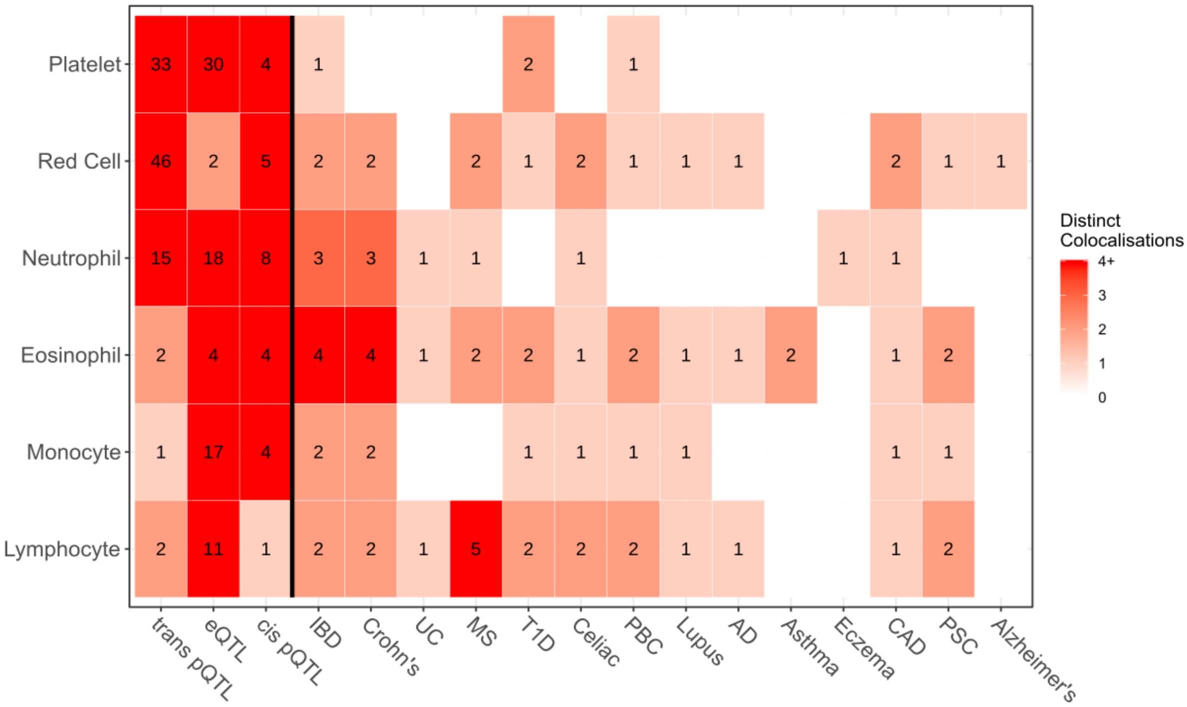

Figure 5.3: Overview of LD sets which colocalise with loci from different GWAS analyses.

Heatmap showing colocalisations between LD sets associated with all Sysmex parameters and pQTL, eQTL, and disease GWAS datasets. Irritable Bowel Disease (IBD), Ulcerative Cholitis (UC), Multiple Sclerosis (MS), Celiac Disease (Celiac), Systemic Lupus Erythematosus (Lupus), Atopic Dermatitis (AD), Coronary Artery Disease (CAD), Primary Sclerosing Cholangitis (PSC), Primary Biliary Cirrhosis (PBC). trans pQTL are associations with the plasma concentration of a protein encoded by a gene which is further than 1 MB from the association signal and cis pQTL are associations located within 1 MB of the appropriate gene.

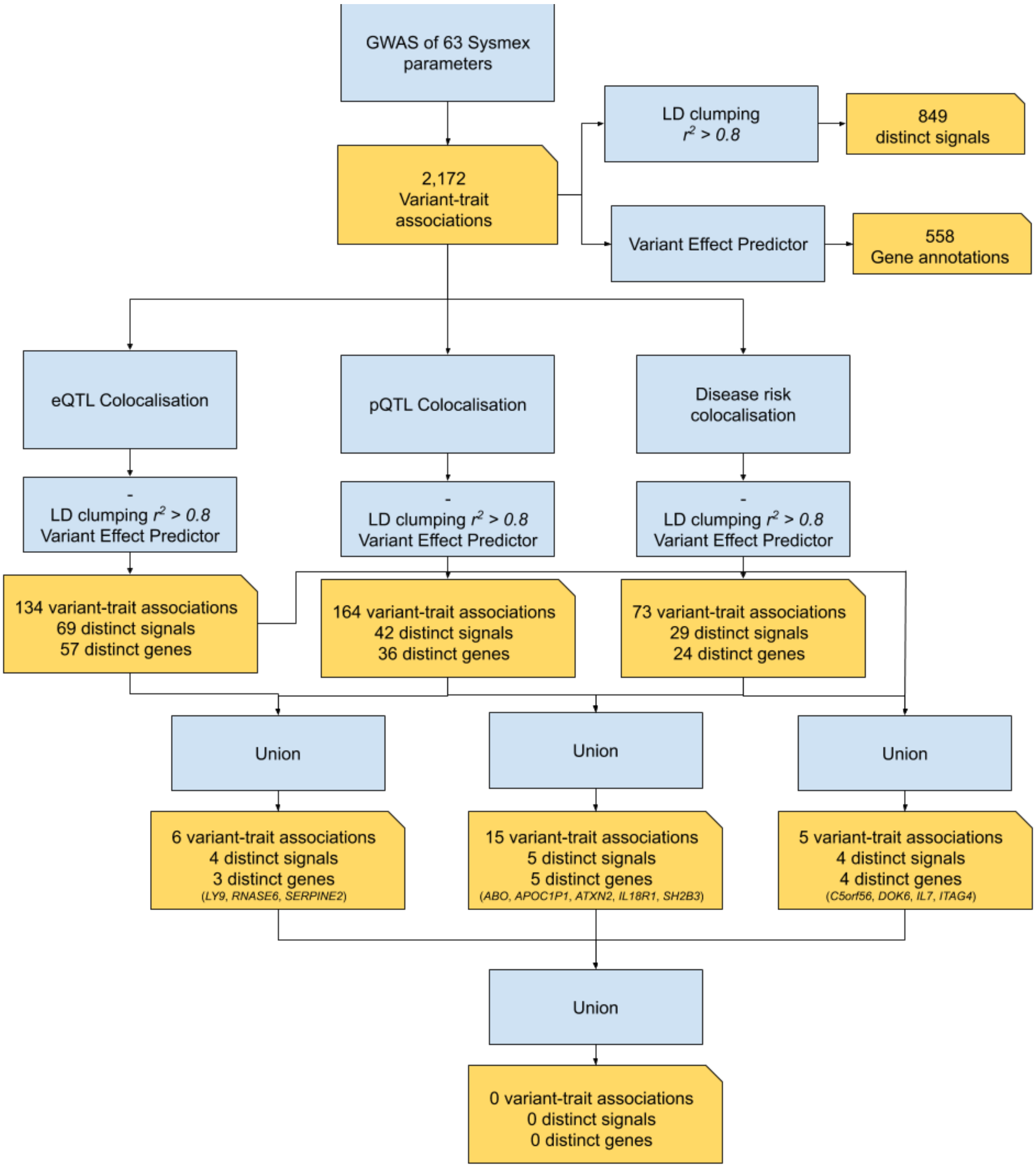
GWAS of 63 Sysmex parameters
2,172 Variant-trait associations
LD clumping $r^2 > 0.8$
849 distinct signals
Variant Effect Predictor
558 Gene annotations
eQTL Colocalisation
pQTL Colocalisation
Disease risk colocalisation
- LD clumping $r^2 > 0.8$ Variant Effect Predictor
- LD clumping $r^2 > 0.8$ Variant Effect Predictor
- LD clumping $r^2 > 0.8$ Variant Effect Predictor
134 variant-trait associations 69 distinct signals 57 distinct genes
164 variant-trait associations 42 distinct signals 36 distinct genes
73 variant-trait associations 29 distinct signals 24 distinct genes
Union
Union
Union
6 variant-trait associations 4 distinct signals 3 distinct genes (LY9, RNASE6, SERPINE2)
15 variant-trait associations 5 distinct signals 5 distinct genes (ABO, APOC1P1, ATXN2, IL18R1, SH2B3)
5 variant-trait associations 4 distinct signals 4 distinct genes (C5orf56, DOK6, IL7, ITAG4)
Union
0 variant-trait associations 0 distinct signals 0 distinct genes

Figure 5.4: Flow chart showing systematic reduction of candidate genes from initial association signals through colocalisation.
Colocalisation allows systematic reduction of association signals (or candidate genes) to smaller sets for which there is evidence for the association signal causally modulating biological factors (eQTL, pQTL, disease risk).

Data	Disease	Short name	Year	PMID	Study
allergic_disease_EUR_ferreira_2017	Allergic Disease	AD	2017	29083406	GWAS
alzheimers_lambert_2013	Alzheimers Disese	Alzheimer's	2013	24162737	GWAS
asthma_EUR_tagc_2018	Asthma	Asthma	2018	29273806	GWAS
cad_nikpay_2015	Coronary Artery Disease	CAD	2015	26343387	GWAS
celiac_disease_dubois_2010	Celiac disease	Celiac	2010	20190752	GWAS
celiac_disease_IC_trynka_2011	Celiac disease	Celiac	2011	22057235	Immunochip
eczema_eagle_2015	Eczema	Eczema	2015	26482879	GWAS
hayfever_or_rhinitis	Hayfever or Rhinitis	AD	2018	NA	GWAS
IBS_CD_delange_2017	Crohn's disease	Crohn's	2017	28067908	GWAS
IBS_CD_liu_2015	Crohn's disease	Crohn's	2015	26192919	GWAS
IBS_CD_IC_liu_2015	Crohn's disease	Crohn's	2015	26192919	Immunochip
IBS_delange_2017	Inflammatory bowel disease	IBD	2017	28067908	GWAS
IBS_LIU_2015	Inflammatory bowel disease	IBD	2015	26192919	GWAS
IBS_IC_liu_2015	Inflammatory bowel disease	IBD	2015	26192919	Immunochip
multiple_sclerosis_sawcer_2011	Multiple sclerosis	MS	2011	21833088	GWAS
multiple_sclerosis_patsopoulos_2017	Multiple sclerosis	MS	2018	BioRxiv	GWAS
multiple_sclerosis_IC_beecham_2013	Multiple sclerosis	MS	2013	24076602	Immunochip
primary_biliary_cirrhosis_cordell_2015	Primary biliary cirrhosis	PBC	2015	26394269	GWAS
primary_biliary_cirrhosis_IC_liu_2012	Primary biliary cirrhosis	PBC	2012	22961000	Immunochip
systemic_lupus_erythematosus_bentham_2015	Systemic lupus erythematosus	Lupus	2015	26502338	GWAS
type_1_diabetes_IC_gumuscu_2015	Type_1 diabetes	T1A	2015	25751624	Immunochip
type_1_diabetes_meta_IC_gumuscu_2015	Type_1 diabetes	T1A	2015	25751624	Immunochip
IBS_UC_delange_2017	Ulcerative cholitis	UC	2017	28067908	GWAS
IBS_UC_liu_2015	Ulcerative colitis	UC	2015	26192919	GWAS
IBS_UC_IC_liu_2015	Ulcerative colitis	UC	2015	26192919	Immunochip
ulcerative_cholitis_anderson_2011	Ulcerative colitis	UC	2011	21297633	GWAS
primary_sclerosing_cholangitis_ji_2016	Primary Sclerosing Cholangitis	PSC	2016	27992413	GWAS

Table 5.4: Disease outcomes studied in colocalisation analysis

Data from GWAS studies for a number of cardiovascular and immune related disease outcomes were collected and studied with colocalisation analysis to identify common genetic determinants with phenotypes from the INTERVAL and UK Biobank study cohorts.

Data	No. of cases	No. of controls	Total_sample	Polulation
allergic_disease_EUR_ferreira_2017	180,129	180,709	360,838	European
alzheimers_lambert_2013	28,640	48,466	77,106	European
asthma_EUR_tagc_2018	23,948	118,538	142,486	European
cad_nikpay_2015	60,801	123,504	184,305	European
celiac_disease_dubois_2010	4,533	10,750	15,283	European
celiac_disease_IC_trynka_2011	12,041	12,228	24,269	European
eczema_eagle_2015	21,399	95,464	116,863	European + Non European
hayfever_or_rhinitis	20,904	91,787	112,691	European
IBS_CD_delange_2017	12,194	28,072	40,266	European
IBS_CD_liu_2015	22,575	46,693	69,268	European + Non European
IBS_CD_IC_liu_2015	22,575	46,693	69,268	European + Non European
IBS_delange_2017	25,042	34,915	59,957	European
IBS_LIU_2015	42,950	53,536	96,486	European + Non European
IBS_IC_liu_2015	42,950	53,536	96,486	European + Non European
multiple_sclerosis_sawcer_2011	9,772	17,376	27,148	European
multiple_sclerosis_patsopoulos_2017	47,351	68,284	115,635	European
multiple_sclerosis_IC_beecham_2013	14,498	24,091	38,589	European
primary_biliary_cirrhosis_cordell_2015	2,764	10,475	13,239	European
primary_biliary_cirrhosis_IC_liu_2012	2,861	8,514	11,375	European
systemic_lupus_erythematosus_bentham_2015	7,219	15,991	23,210	European
type_1_diabetes_IC_gumuscu_2015	6,683	12,173	18,856	European
type_1_diabetes_meta_IC_gumuscu_2015	6,683	12,173	18,856	European
IBS_UC_delange_2017	12,366	33,609	45,975	European
IBS_UC_liu_2015	20,417	52,230	72,647	European + Non European
IBS_UC_IC_liu_2015	20,417	52,230	72,647	European + Non European
ulcerative_cholitis_anderson_2011	6,687	19,718	26,405	European
primary_sclerosing_cholangitis_ji_2016	4,796	19,955	24,751	European

Table 5.5: Study size and population cohort for GWAS data used for disease colocalisation
Populations studied in disease risk GWAS were largely European with some studies containing a limited number of non-European samples.

Data	Data download source	Source link
allergic_disease_EUR_ferreira_2017	Paper Methods	www.genepi.qimr.edu.au/ staff/manuelf/gwas_results/main.html
alzheimers_lambert_2013	Paper Methods	web.pasteur-lille.fr/ en/recherche/u744/igap/igap_download.php
asthma_EUR_tagc_2018	Phenoscanner	www.phenoscanner.medschl.cam.ac.uk/login/?next=/data/
cad_nikpay_2015	Paper Methods	www.cardiogramplusc4d.org/ data-downloads/
celiac_disease_dubois_2010	ImmunoBase	www.immunobase.org/downloads/protected_data/GWAS_Data/
celiac_disease_IC_trynka_2011	ImmunoBase	www.immunobase.org/downloads/protected_data/iChip_Data/
eczema_eagle_2015	Direct Link	data.bris.ac.uk/data/dataset/28uchsdpmub118uex26ylacqm
hayfever_or_rhinitis	Direct Link	docs.google.com/spreadsheets/d/ 1kvPoupSzsSFBNSztMzl04xMoSC3Kcx3CrjVf4yBmESU/ edit?ts=5b5f17db#gid=178908679
IBS_CD_delange_2017	GWAS catalog	ftp://ftp.ebi.ac.uk/pub/databases/ gwas/summary_statistics/deLangeKM_28067908_GCST004132/
IBS_CD_liu_2015	IBD_genetics	www.ibdgenetics.org/downloads.html
IBS_CD_IC_liu_2015	IBD_genetics	www.ibdgenetics.org/downloads.html
IBS_delange_2017	GWAS catalog	ftp://ftp.ebi.ac.uk/pub/databases/gwas/summary_statistics/ deLangeKM_28067908_GCST004131/
IBS_LIU_2015	IBD_genetics	www.ibdgenetics.org/downloads.html
IBS_IC_liu_2015	IBD_genetics	www.ibdgenetics.org/downloads.html
multiple_sclerosis_sawcer_2011	ImmunoBase	www.immunobase.org/downloads/protected_data/GWAS_Data/
multiple_sclerosis_patsopoulos_2017	From colaborator	From colaborator
multiple_sclerosis_IC_beecham_2013	ImmunoBase	www.immunobase.org/downloads/protected_data/iChip_Data/
primary_biliary_cirrhosis_cordell_2015	ImmunoBase	www.immunobase.org/downloads/protected_data/GWAS_Data/
primary_biliary_cirrhosis_IC_liu_2012	ImmunoBase	www.immunobase.org/downloads/protected_data/iChip_Data/
systemic_lupus_erythematosus_bentham_2015	ImmunoBase	www.immunobase.org/downloads/protected_data/GWAS_Data/
type_1_diabetes_IC_gumuscu_2015	ImmunoBase	www.immunobase.org/downloads/protected_data/iChip_Data/
type_1_diabetes_meta_IC_gumuscu_2015	ImmunoBase	www.immunobase.org/downloads/protected_data/iChip_Data/
IBS_UC_delange_2017	GWAS catalog	ftp://ftp.ebi.ac.uk/pub/databases/gwas/summary_statistics/ deLangeKM_28067908_GCST004133/
IBS_UC_liu_2015	IBD_genetics	www.ibdgenetics.org/downloads.html
IBS_UC_IC_liu_2015	IBD_genetics	www.ibdgenetics.org/downloads.html
ulcerative_cholitis_anderson_2011	ImmunoBase	www.immunobase.org/downloads/protected _data/GWAS_Data/primary_sclerosing_cholangitis_ji_2016
primary_sclerosing_cholangitis_ji_2016	Paper Methods	www.ipscsg.org/

Table 5.6: Source of summary statistics for GWAS data used for disease colocalisation
Summary statistics were obtained from a range of sources including Immunobase, GWAS catalog, with a direct request to the author, or from a hyperlink listed in the methods section of the publication.

5.3.2 Disease aetiology and drug target validation

I have identified 73 associations with Sysmex parameters which colocalise with GWAS risk for cardiovascular inflammation or immune related disease (PP $> 80\%$) (Table A.1) (Fig. 5.3). My analysis annotates signals associated with risk for disease and functionally informative blood cell properties as measured by Sysmex which are discussed below. In Section 1.6.5 I discussed the challenges of drug development, in particular:

- In the clinical pipeline of AstraZeneca between 2005 - 2010, 88% of drugs failing at Phase IIb did so due to lack of efficacy. In 40% of these cases the reason for failure due to efficacy was cited to be target linkage to disease not established or no validated models available.
- Usually, $1 billion dollars of development costs are incurred prior to Phase II, the first real chance to test efficacy of a drug compound in man.

During drug development a billion dollars and many years of research may go by before the first chance to test a purported efficacious drug target in humans. Furthermore, after a huge commitment of time and resources many candidates are failing because they simply are not efficacious! It is this context which motivates my thesis and the work of other statistical geneticists. I do not aim to definitively prove a biological mechanism or provide overwhelming evidence for a new drug candidate. Instead, my work shows it is possible to find evidence for a purported efficacious drug target from genetic data derived from a human cohort living in 'wild-type' conditions. This analysis can occur early in the drug development pipeline, before resources are committed to a candidate pathway or molecule and can prioritise candidates.

5.3.2.1 Lymphocyte traits and multiple sclerosis

Five signals associated with lymphocyte parameters colocalised with genetic risk of multiple sclerosis (MS) (Fig. 5.3). The colocalising associations were located in the transcription factor encoding gene BACH2, and in the genes encoding receptors for Interleukin(IL)-2 (*IL2RA*) and IL-7 (*IL7R*) and in IL-7 itself. The conditionally independent variants representing these associations are: rs142376788 located in the 5' untranslated region (UTR) of *IL7R*, rs11567705 located in an intron of *IL7R*, rs72928038 located in an intron of *BACH2*, rs10957897 located in an intron of *IL7R*, and rs3118471 located in an intron of *IL2RA*.

5.3.2.2 Interleukin 2 receptor alpha

IL-2RA is a transmembrane protein present on nearly all activated T cells, but not on resting T cells [172]. IL-2RA is the subject of therapeutic antibody Daclizumab found effective for treatment of MS by blocking T-lymphocyte IL-2RA receptors. This results in significant expansion of natural killer (NK) cell population and gradual reduction in numbers of activated T-lymphocyte cells [24].

It could be expected that a lymphocyte pool with higher numbers of activated cells or NK cells which are granular will have a higher LY-SSC measurement [32]. Therefore, we could hypothesise that Daclizumab would raise LY-SSC measurement in patients due to it's effect on increasing NK population.

A variant in *IL2RA* (rs3118471, -Log$_{10}$P: 8.60, MAF: 29.9%, VEP: intronic *IL2RA*) is associated with a reduction in LY-SSC, an increase in LYMPH# and colocalises with increased risk of MS (PP: 99.9%, Fig. 5.5) [152]. Supporting my aforementioned hypothesis: at this locus an association annotated to the mechanistic target for Daclizumab is increasing the risk of MS and decreasing LY-SSC. It is interesting that the variant annotated to *IL2-RA* is associated with an increased risk of MS and decreased LY-SSC, because the variant is acting consistently in the opposite direction to Daclizumab. Daclizumab acts 'against' MS aetiology, and increases NK population, a granular subset of lymphocytes [32] - thus the opposite of this genetic association. This result shows that colocalisation of Sysmex parameters with disease risk GWAS can identify association signals annotated to genes which are already validated therapeutic drug targets.

Furthermore, from these results, perhaps it could also be argued that MS patients with lower LY-SSC are those most likely to benefit from Daclizumab? This would be supported by the genetic evidence, an association increasing the risk of MS decreases LY-SSC. However, there doesn't seem to be a convincing causal relationship between genome-wide instrumental variables for LY-SSC and MS as assessed by MR (MR-Egger, P-value: 0.37, causal estimate: 1.60), a full set of MR results are available in Supplementary file A.3. Separately, a counter-argument could be made that the increase in LY-SSC is purely as a function of the increase in lymphocyte count. This is somewhat unlikely due to the anti-correlation between LY-SSC and LYMPH#, with a Pearson correlation between phenotypes of -0.183 and a genetic correlation of -0.248 (P-value: $5.13x10^{-5}$). Furthermore, Shirley *et al.,* reported that the clinical effect of Daclizumab is not thought to result from 'broad immunodepletion' of overall lymphocyte cell counts, but instead through immunomodulation of lymphocyte cell subtypes [156].

However, an interpretation of the LY-SSC parameter, or any lymphocyte scatter parameter is difficult to make compared to other white cell Sysmex parameters, because of the heterogeneous nature of the lymphocyte population (Section 6.1.1). It is possible that a comparison of SSC as an index of granulation between NK cells and other lymphocytes is confounded by other structural differences between lymphocyte cell types (Section 1.1.3). Therefore, confident interpretations of lymphocyte parameters will require better understanding of the cytometry properties of different lymphocyte subtypes in isolation. For example, a SSC comparison between purified NK, T, and B lymphocyte populations.

5.3.2.3 Interleukin 7 receptor

I identified four conditionally significant variants annotated to *IL7R*, assigned to two statistically distinct signals (Signal ID: 236 and 237 in Table A.1), which are associated with seven lymphocyte traits including LYMPH#, LY-SFL, LY-SSC, RE-LYMP%, and RE-LYMP# (Table A.1). LY-SFL and LY-SSC indicate lymphocyte populations with higher nucleic acid content and cell granulation respectively. RE-LYMP# is the count of lymphocyte cells with high SFL values, a reactive and activated lymphocyte sub-population (Section 3.2.2). The locus of association at *IL7R* decreased the value of the aforementioned lymphocyte parameters which include proxies for properties related to cell activation. This association also reduces the genetic risk for hay fever and rhinitis (PP: 99.1%). This result between *IL7R* and hayfever or rhinitis is consistent with a clinical study showing the expression of *IL7R* increasing 14% following allergen immunotherapy, and also a decrease expression following ragweed season, a time of year with high allergen concentrations [19]. Variants in *IL7R* also colocalised with a range of immune disorders including MS (Fig. 5.6) [152], and primary biliary cirrhosis (PP: 95.7% and 85.6%) [48]. The targeting of IL-7R using antibodies in preclinical mouse models of MS shows dramatic therapeutic effects [96]. Our association and colocalisation adds statistical genetic evidence from human cohorts for the therapeutic effect of this drug target. My finding adds proof of efficacy for this drug target by leveraging genetic studies of lymphocyte parameters, and risk of multiple immune disorders (hayfever and rhinitis, MS and primary biliary cirrhosis.

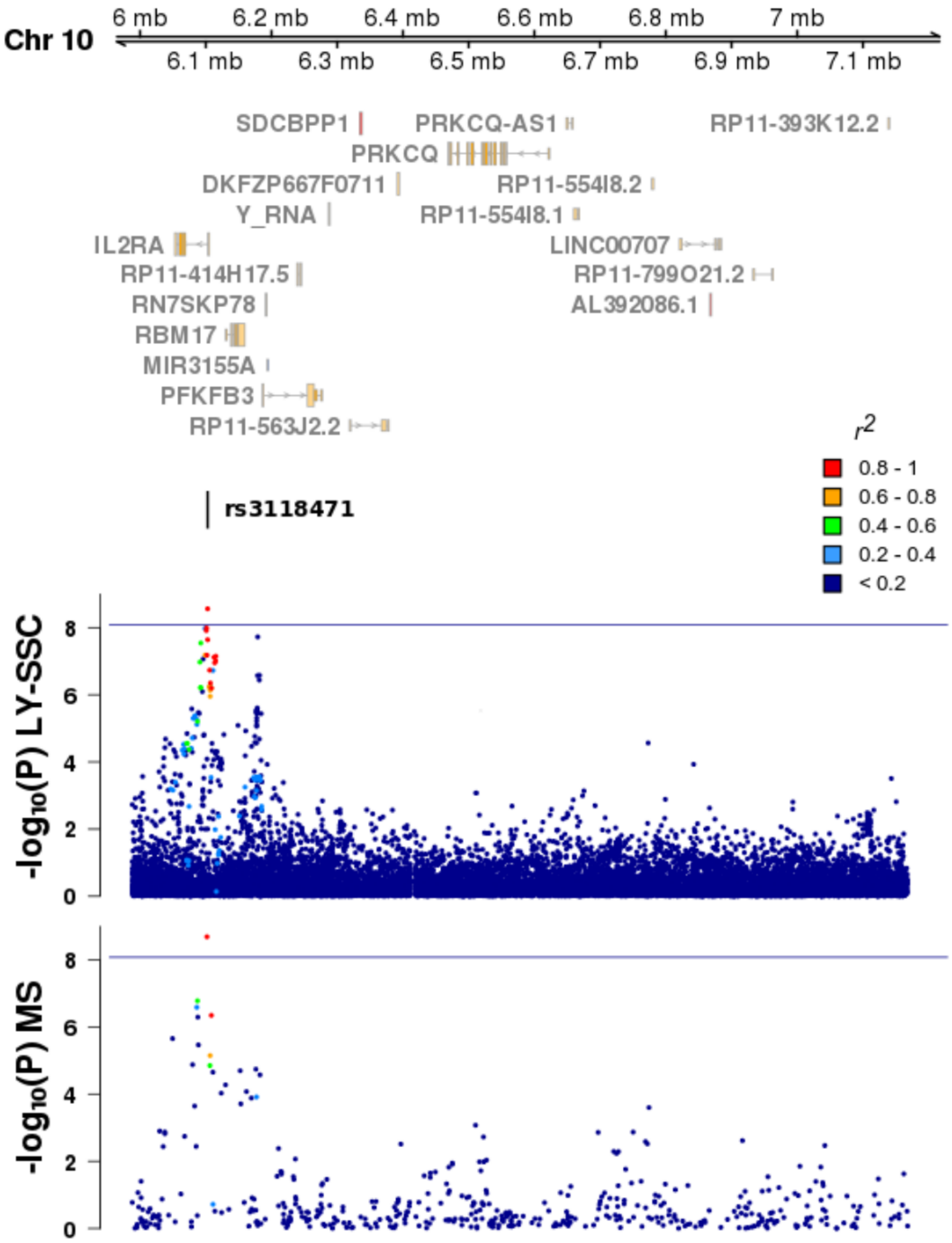

Figure 5.5: Plot showing colocalisation between disease risk for multiple sclerosis and lymphocyte side scatter.
Each data point represents a genetic variant and the position of that data point in the y is the -Log$_{10}P$ for association with the phenotype. Variants are coloured according to their LD with the conditionally significant variant in this locus. Colocalisation between the two association signals occurs with posterior probability of 99.9%, LD between the conditionally significant variant (rs3118471) and sentinel (rs3118470) in disease risk GWAS is 0.96 r^2. The GWAS of MS is performed by Sawcer *et al.*, 2011 [152].

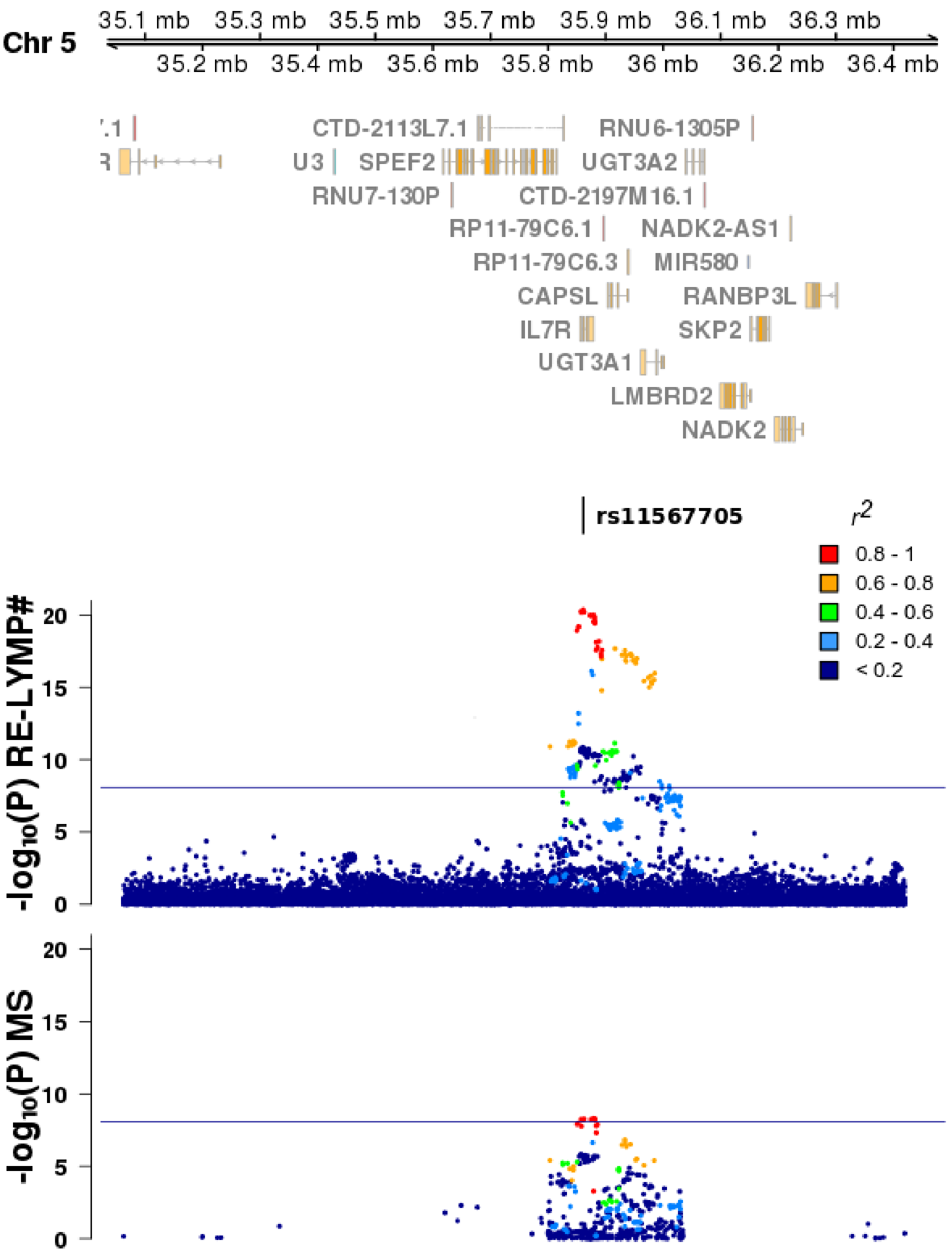

Figure 5.6: Plot showing colocalisation between disease risk for multiple sclerosis and reactive lymphocyte count.

Each data point represents a genetic variant and the position of that data point in the y axis is the -Log$_{10}P$ for association with the phenotype. Variants are coloured according to their LD with the conditionally significant variant in this locus. Colocalisation between the two association signals occurs with posterior probability of 95.7%. LD between the conditionally significant variant (rs11567705) and sentinel (rs6881706) in disease risk GWAS is 1.00 r^2. The GWAS of MS was performed by Beecham *et al.*, 2013 [2]. The sparsity in genetic variants in the disease GWAS figure (bottom) is caused by the difference in genotyping panel between the two studies.

5.3.2.4 NE-FSC, IL-18R1 and atopic dermatitis

My GWAS analysis of NE-FSC identifies 70 independent conditionally significant variants, including an association annotated to the *IL-18R1* gene (rs1035127, MAF: 22.3%, -Log$_{10}$P: 10.9, VEP: *downstream IL18-R1*). The ligand of *IL-18R1*, *IL-18*, is a neutrophil activator and blood samples with higher levels of interleukins such as *IL-18* contain neutrophil populations with higher NE-FSC values [88]. *IL-18* expression has been shown to contribute to aetiology of celiac disease [100], atopic dermatitis [95], and psoriasis [67].

rs1035127 annotated to *IL-18R1* associated with a decrease in NE-FSC, which also colocalises with a pQTL signal for decrease in plasma *IL-18R1* (PP: 98.0%), and increased risk of celiac disease (PP: 93.4%), but decreased risk of eczema and IBD (PP: 93.0% and 84.1% respectively). This is a confusing result, as celiac disease is considered to be a differential diagnosis of IBD and shares much of the same disease aetiology [133]. Therefore, this locus is already raising interesting questions about the potential differential role of *IL-18R1* regarding celiac disease and IBD. I found explaining this result difficult, I hope that by publication of this thesis I may attract colleagues to address this finding.

I previously noted that rs1035127 is associated with a decrease in NE-FSC and increase in plasma *IL-18R1*, furthermore, *IL-18R1* is expressed with Log$_2$(FPKM) 4.6 in neutrophil cells [43]. This shows a common genetic determinant between NE-FSC, *IL-18R1* expression in the blood plasma, and the aetiology of immune disorders. From these results it might be tempting to suggest that that neutrophil cells with lower FSC result in decreased IL-18R1 in blood plasma and changes in risk of immune disorders. This is one possible explanation which would explain the observed results. However, colocalisation analysis cannot not show a causal relationship between these factors, only the high probability of a common genetic determinant, and even this must be interpreted with an understanding of the limitations of colocalisation (Section 6.1.3).

Mendelian randomisation analysis of NE-FSC as an exposure did not show a significant causal relationship between NE-FSC and autoimmune disorders: celiac disease (P-value: 0.88) and eczema (P-value: 0.58). This MR included instrumental variables from across the genome, not just *IL-18R1.* However, a MR of IL-18R1 blood plasma concentration by Sun *et al.* suggested that IL-18R1 may have a causal relationship with atopic dermatitis (P-value: 1.5×10^{-28}) [164]. Atopic dermatitis is a diagnosis under the umbrella of eczema which includes dermatitis syndromes generally [179]. It is unfortunate my colocalisation analysis was performed with GWAS results of eczema and the MR by Sun *et al.,* was performed using GWAS results of atopic dermatitis. Further work to explore this signal of association could begin with repeating the analysis with summary statistics from the respective outcome.

5.3.3 Genetic characterisation of white cell granulation

Degranulation of white cells has long been established as an important mechanism of immune response. My analysis identifies associations annotated to granule proteins by VEP such as *DEF*, *CTSH*, *CTSC*, *ELANE*, *ARSB*, *LYZ*, *RNASE2*, *RNASE3*, and *RNASE6* (Appendix A.1). In particular, my analysis is the first GWAS of blood phenotypes to annotate associations to *DEF*, *CTSH*, *CTSC*, *ELANE*. This is despite extensive study showing the importance of such granule proteins in white cell immune function [28]. Many proteins known to be present in granules of white cells have also been identified as circulating blood plasma proteins [164]. My analysis shows there is often a common genetic architecture underlying blood plasma protein concentrations and blood cell Sysmex parameters, in particular SSC an index of cell granulation.

Azurophilic granules are present in a range of white blood cells and most prominent in neutrophils, where they are loaded with a range of antimicrobial proteins, and play a critical role in neutrophil immune response [144]. Neutrophil scatter measurements SFL, SSC, and FSC have been proven to be indicative of neutrophil action and immune response [103] [146] [188]. GWAS of neutrophil indices identified variants annotated to *MPO*, *PRTN3*, *ELANE*, *BPI*, *ARSB*, *CTSC*, *CTSH*, *LYZ*, and *RNASE2*, microcidal proteins which are also known to localise in azurophil granules. Furthermore, signals in these genes also colocalise with pQTL signals for plasma proteomics of the same granule proteins (Appendix A.1). An association signal located in *ARSB* containing conditionally significant associations with EO-SSC, NE-SSC, NE-SFL, and NE-FSC colocalise (PP: 93%, 98%, 98%, 99%) with a pQTL signal for *ARSB* in the plasma proteome. Monocyte side fluorescence is associated with four conditionally independent signals in the *RNASE* region on chromosome 14. Associations include rs1045922 (-Log$_{10}$P: 44.9, MAF: 23.8%, VEP: missense) located in *RNASE6*, rs6571511 (-Log$_{10}$P: 58.9, MAF: 7.7%, VEP: upstream) located in the *RNASE3* gene, rs151169198 (-log10P: 10.7, MAF: 0.80%, VEP: missense) and rs2771358 (-Log$_{10}$P: 152.2, MAF: 25.4%, VEP: upstream) both of which are annotated to *RNASE2* by VEP.

RNASE6 has been shown to be localised to the granules of leukocytes and granulocyte cells. Exocytosis of granules secretes *RNASE* and other proteins which conduct antimicrobial activity [18]. Variant rs1045922 is a conditionally significant association for monocyte side fluorescence colocalising with eQTL analysis of *RNASE6* transcripts in CD14 cells (PP: 99.4%). Variant rs1045922 also has a pairwise LD of 1.00 r^2 and colocalises (PP: 98%) with rs11622942 a conditionally significant pQTL variant for the *RNASE6* protein in the plasma proteome [164].

My analysis of novel blood indices has identified granule proteins which are known to be crucial in immune function, but remained unidentified by GWAS of traditional blood cell indices. Furthermore, my integration of data from pQTL analysis by colocalisation has shown common genetic architecture modulating white cell granularity indices and many blood plasma proteins. A likely explanation for these colocalisations is that granule proteins in plasma are originating from granule proteins in blood cells, this hypothesis and has been suggested by others [164, 5, 101].

5.3.3.1 ANCA-associated vasculitis

ANCA-associated vasculitis (AAV) is an autoimmune syndrome characterised by vascular inflammation and autoantibodies against neutrophil granule proteins *MPO* or proteinase-3 (PR3, encoded by the *PRTN3* gene) [107]. It has been shown that genetic variation which affects PR3 abundance in circulation influences risk of vasculitis with anti-PR3 antibodies [164]. I identified a locus in the *PRTN3* region associated with an increase in neutrophil SSC, SFL, FSC (indices of granule content and of nucleic acid content, membrane composition, and cell size respectively). This genetic signal also colocalises with an increase for eQTL in whole blood (GTEX PP: 98.9%) and the pQTL in PR3 plasma concentration (PP: 99.7%). Together, these results suggest that genetic effects on *PRTN3* transcription are reflected in changes in neutrophil granule content and influence abundance of PR3 in the circulation and thereby disease risk. PR3 is also expressed in other myeloid cells including eosinophils (Supp. Table 2). Notably this same genetic signal also colocalises with EO-SFL (a marker of nucleic acid content and membrane composition), suggesting that the genetic effects on PR3 may also be acting through eosinophils. Unfortunately, due to the pleiotropic nature of this genetic variant, it is not possible to identify from these results whether the association in question is acting through eosinophils, neutrophils, or a combination of both. Intriguingly, a subset of AAV patients have eosinophilia, although this is more common in the context of antibodies to MPO rather than to PR3.

5.3.4 Causal association with Mendelian randomisation

Mendelian randomisation can identify casual associations between risk factors and outcomes of interest using genetic variants as instrumental variables. In the context of MR, Sysmex parameters are treated as the exposure and disease risk as the outcome. This is testing for Sysmex parameters as causal mediators of disease risk.

I selected 15 white cell Sysmex parameters to assess for causal relationships with risk for 23 complex diseases (Table 5.7), summary statistics for disease outcomes were collected using the MR Base package [80] (Table 5.8). The parameters selected are SSC, SFL, and FSC measurements of white cells and RE-LYMP#. Genetic variants associated with each phenotype are selected as 'instrumental variables' (IVs) which are used to determine causality (Section 5.2.2). I began by performing a robust penalised MR-Egger regression, the intercept of the MR-Egger models directional pleiotropy of the IVs. If the Wald statistic of the intercept was less than one, I interpreted the IVs as having balanced pleiotropy, thus having satisfied the assumptions of inverse variance weighted (IVW) regression. In such cases I assessed causality with a robust penalised IVW model which offers greater power than the MR-Egger method [29]. P-values are Bonferroni corrected for the 15 traits and 23 diseases which are tested for causal association. To assess the robustness of results I tested causal association with 11 types of Mendelian randomisation models (Table 5.2.2.2) including IVW models which assume balanced pleiotropy across the IVs, MR-Egger models which allow for unbalanced pleiotropy across IVs, median based methods which assume at least 50% of IVs are valid and do not suffer from pleiotropy, multivariable models to test for association accounting for effects of the instrumental variables on other Sysmex parameters and traditional blood cell phenotypes. Weighted, penalised, and robust methods of linear regression were used to account for standard error of IVs, and heterogeneity of IVs respectively. Three pairs showed statistically significant causal associations as assessed by MR between Sysmex parameters and complex disease: NE-SSC and the risk for lung cancer and CAD and EO-FSC and the risk of asthma.

I integrated and visualised results and integrated all MR tests in an interactive report which includes LD between instrumental variables in the model, integration of LD set data to annotate instrumental variables, and leave one out analysis (LOO) analysis where the IVW estimate is recalculated with each of the instrumental variables excluded (Supp. A.3). Reports are generated in simple hypertext markup language (HTML) format which can be opened with any electronic device with an internet browser without the need to install or prepare any additional software. A full set of reports is available in HTML format in the supplementary (Supp. A.3).

Sysmex Parameters	Disease Outcomes
BASO-FSC, BASO-SFL, EO-FSC, EO-SFL, EO-SSC, LY-FSC, LY-SFL, LY-SSC, MO-FSC, MO-SFL, MO-SSC, NE-FSC, NE-SFL, NE-SSC, RE-LYMP#,	Abdominal Aortic Aneurysm, Amyotrophic Lateral Sclerosis, Asthma, Autism, Cardioembolic Stroke, Celiac Disease, Chronic Kidney Disease, Coronary Artery Disease, Crohns, Eczema, Gout, Iga Neuropathy, Inflammatory Bowel Disease, Ischemic Stroke, Lung Adenocarcinoma, Lung Cancer, Major Depressive Disorder, Multiple Sclerosis, Pagets Disease, Pancreatic Cancer, Rheumatoid Arthritis, Type II Diabetes, Ulcerative Colitis

Table 5.7: Sysmex parameters and diseases tested for pairwise causal association.
A set of 15 Sysmex parameters and 23 disease outcomes were tested for causal associations. Sysmex parameters and disease outcomes were selected based on those deemed to be most interesting candidates for follow up analysis due to the higher number of conditionally independent variants.

Disease	Author	Year	Number of Cases	Number of Controls	Reference (PMID)
Abdominal Aortic Aneurysm	Jones	2017	4,972	99,858	27899403
Amyotrophic Lateral Sclerosis	van Rheenen	2016	12,577	23,475	27455348
Asthma	Moffatt	2007	10,365	16,110	20860503
Autism	Smoller	2013	14,525	14,890	23453885
Cardioembolic Stroke	Malik	2016	1,859	19,326	26935894
Celiac Disease	Dubois	2010	4,533	10,750	20190752
Chronic Kidney Disease	Kottgen	2010	5,807	56,430	20383146
Coronary Artery Disease	Nikpay	2015	60,801	123,504	26343387
Crohns	Liu	2015	5,956	14,927	26192919
Eczema	Paternoster	2014	10,788	30,047	26482879
Gout	Kottgen	2013	2,115	67,259	23263486
IGA Neuropathy	Feehally	2010	977	4,980	20595679
Inflammatory Bowel Disease	Liu	2015	12,882	21,770	26192919
Ischemic Stroke	Malik	2016	10,307	19,326	26935894
Lung Adenocarcinoma	Wang	2014	3,442	14,894	24880342
Lung Cancer	Wang	2014	15,861	27,209	24880342
Major Depressive Disorder	Sullivan	2013	9,240	9,519	22472876
Paget's Disease	Albagha	2011	741	2,699	21623375
Multiple Sclerosis	Beecham	2013	14,498	24,091	24076602
Pancreatic Cancer	Amundadottir	2009	1,896	1,939	19648918
Rheumatoid Arthritis	Okada	2014	14,361	43,923	24390342
Type II Diabetes	Wood	2016	4,040	116,246	26961502
Ulcerative Colitis	Liu	2015	6,968	20,464	26192919

Table 5.8: Source of summary statistics for GWAS data used in mendelian randomisation (MR)

Summary statistics were obtained from a range of studies across 23 immune and cardiovascular related disorders. Studies often overlap with those analysed in disease colocalisation analysis.

5.3.4.1 Heterogenity and pleiotropy in instrumental variables

The results indicated a high degree of heterogenity and pleiotropy in the instrumental variables suggesting multiple pathways contributing to the exposure of interest with differential association with the disease risk. An example is a test for causality between NE-SSC and CAD (Fig. 5.7, 5.8). The MR test for causality with an IVW model resulted in no significance with a P-value of 0.01, compared to a much stronger significance with Penalised Robust IVW MR model (P-value: 7.26×10^{-6}). This could be caused by heterogeneity in the data, as the Penalisation and Robust methods both reduce the influence of outlying data points on the regression (Chapter 5.2.2). Cochran's Q statistic suggests heterogenity for NE-SSC and CAD ($Q : 82.2, df : 53, \text{P-value} : 6.23 \times 10^{-3}$) (Section 5.3.4.3), but no heterogenity for NE-SSC and lung cancer ($Q : 44.4, df : 53, \text{P-value} : 0.773$) (Section 5.3.4.4), and EO-FSC with asthma ($Q : 7.22, df : 10, \text{P-value} : 0.704$) (Section 5.3.4.5).

5.3.4.2 Multivariable Mendelian randomisation

I performed a multivariable MR analysis in order to assess whether causal estimates are consistent even when considering effects of instrumental variables on other Sysmex and FBC parameters. I identified Sysmex parameters for which instrumental variables may be acting through pleiotropy for each of the three identified associations (NE-SSC & CAD or lung cancer, and EO-FSC with asthma). This was done using the LD clumping approach (Chapter 2.2.9). If an instrumental variable is in the same clump as associations with other parameters that parameter is considered to be a potential pleiotropic factor. Furthermore, I also considered the cell type appropriate FBC count and percentage (of all white blood cells) measurements for multivariable analysis.

My analysis suggests that NE-SSC and CAD or lung cancer causal associations are robust for possible pleiotropy with other Sysmex parameters (Fig. 5.9, 5.10). However, in the case of EO-FSC and asthma the multivariable MR analysis is complicated by high correlation between the estimated effect size of the instrumental variables on EO-FSC and other parameters studied in the multivariable analysis such as EO-SSC, EO-SSC-DW, EO-SFL with r^2 correlations of 0.88, 0.79, and 0.71 respectively (Fig. 5.11). It is generally expected that estimates of covariate effect size in a linear model become unreliable with collinearities of above 0.8 r^2. However, causal estimates for EO-FSC with asthma are robust to inclusion of eosinophil count and percentage FBC parameters. This suggests that there is additional information regarding the causal role of eosinophils in asthma beyond that already proposed by MR of eosinophil count and risk of asthma [15]. More detailed analysis is required to assess whether EO-FSC is indeed causally mediating risk of asthma or if there is a potential pleiotropic effect via other eosinophil parameters, or a combination of both of these factors.

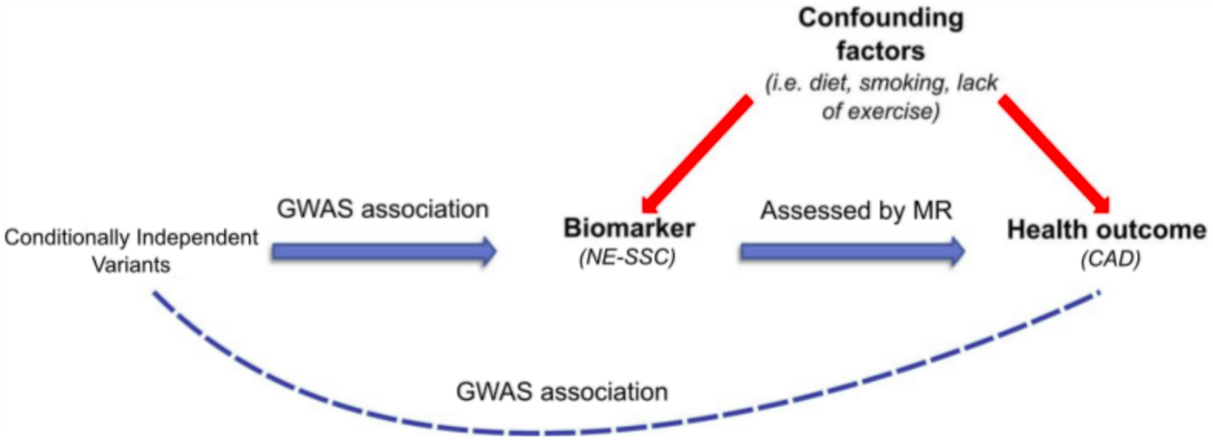

Figure 5.7: Mendelian randomisation to test for causal association between NE-SSC and CAD.

A diagram representing the framework of a MR analysis using conditionally independent variants associated with NE-SSC to test for causal association between NE-SSC and CAD.

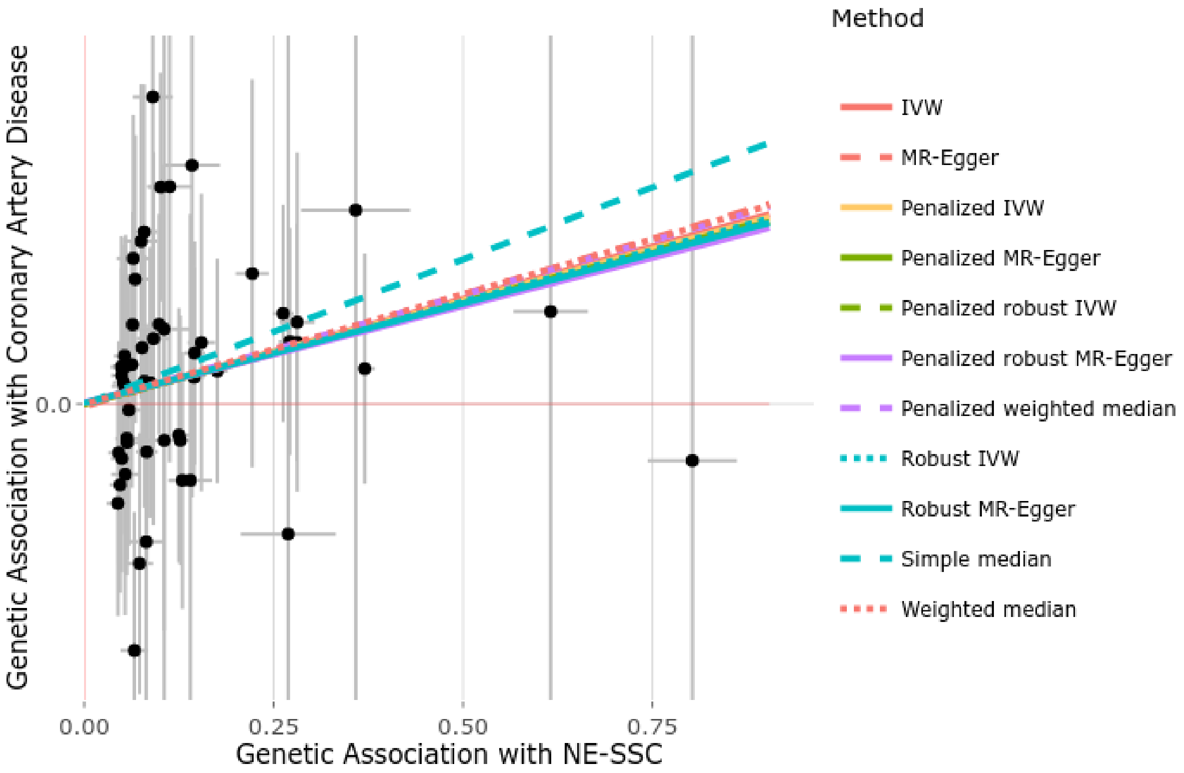

Figure 5.8: Scatterplot of instrumental variables and association with NE-SSC and CAD.

Bars represent 95% confidence intervals for estimate of effect size of instrumental variables with either the exposure or outcome. Results show good concordance between causal estimates of sensitivity analyses. The Penalised Robust IVW model predicts causal association between NE-SSC and CAD ($\beta : 0.0355, \text{P-value} : 7.26 \times 10^{-6}$).

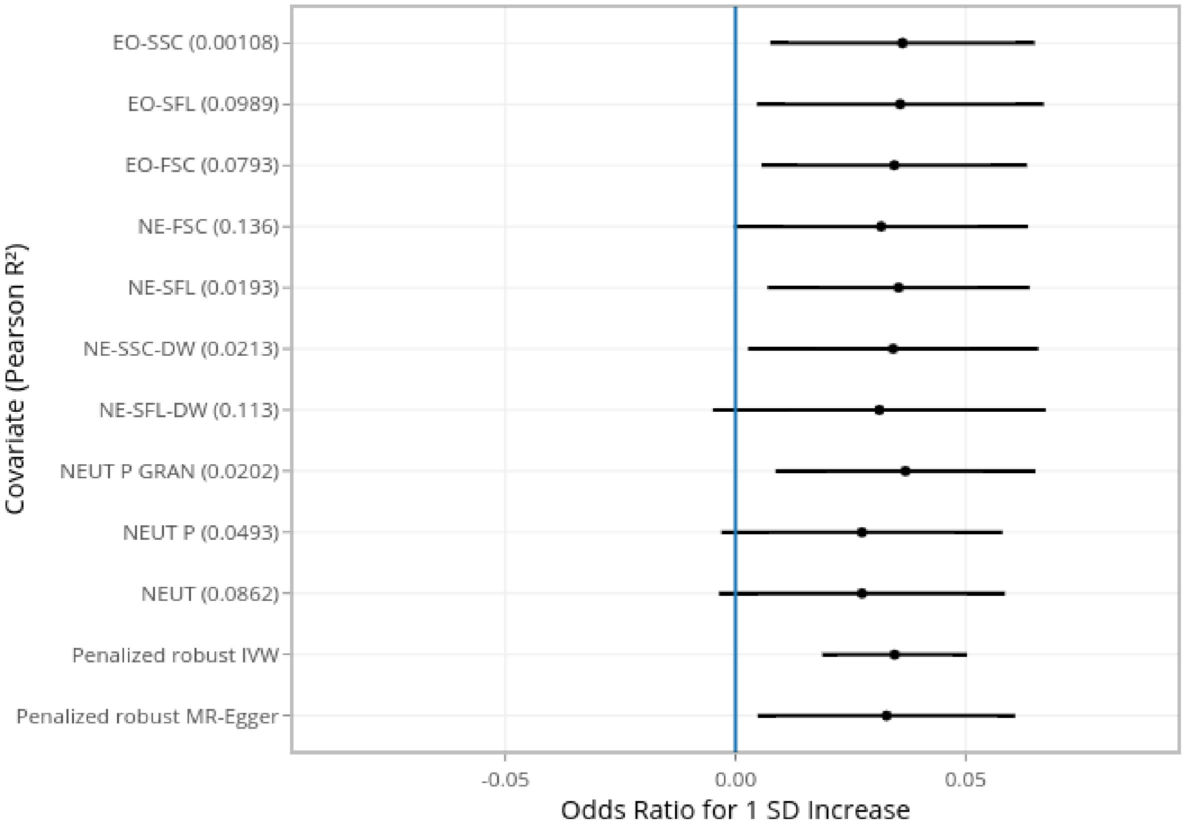

Figure 5.9: Pairwise multivariable IVW MR models for CAD containing NE-SSC and a covariate.

Each covariate is listed on the y axis with the Pearson correlation between the phenotype of that covariate and NE-SSC. The x axis is the calculated change in odds ratio for disease risk given 1 SD increase in NE-SSC in a multivariable MR model containing the corresponding covariate. This is compared to the estimates for causal association from Penalised robust IVW and Penalised robust MR-Egger containing NE-SSC only. The Pearson correlation (r^2) between instrumental variable effect sizes for each of the covariates is in brackets.

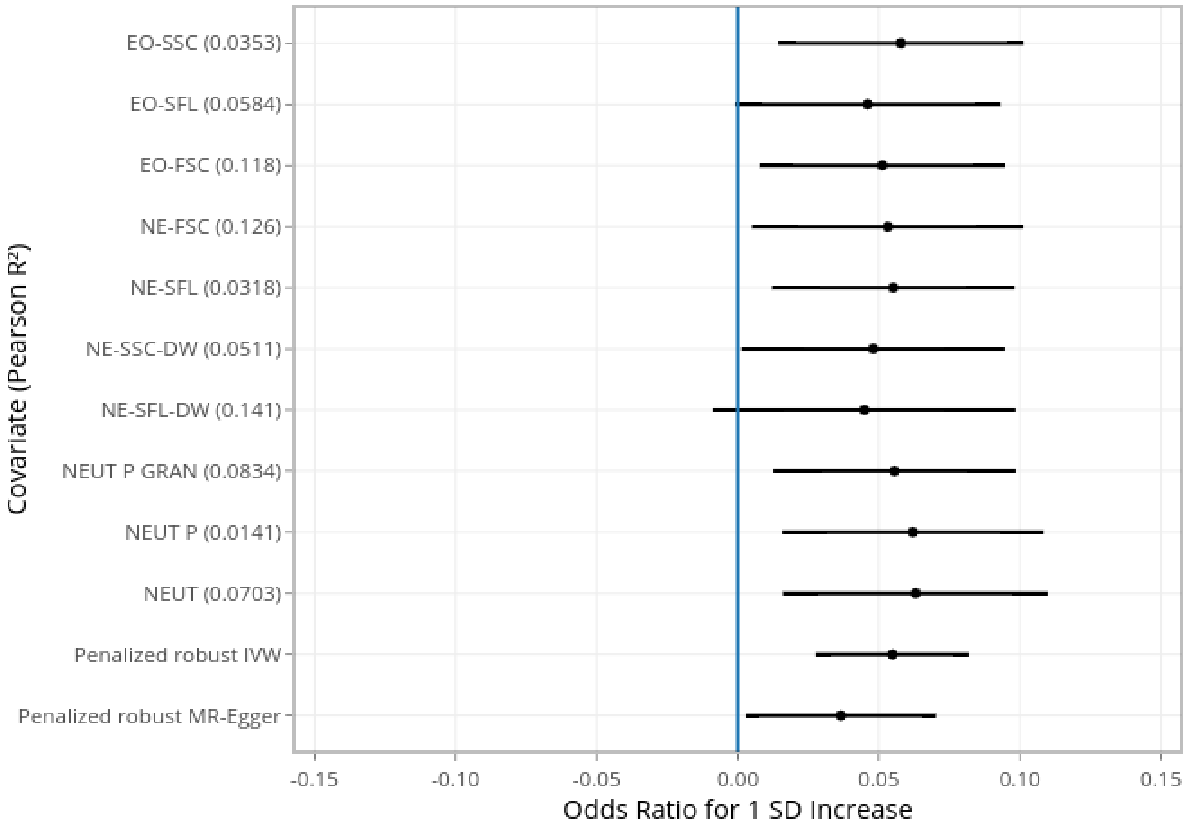

Figure 5.10: Pairwise multivariable IVW MR models for Lung cancer containing NE-SSC and a covariate.

Each covariate is listed on the y axis with the Pearson correlation between the phenotype of that covariate and NE-SSC. The x axis is the calculated change in odds ratio for disease risk given 1 SD increase in NE-SSC in a multivariable MR model containing the corresponding covariate. This is compared to the estimates for causal association from Penalised robust IVW and Penalised robust MR-Egger containing NE-SSC only. The Pearson correlation (r^2) between instrumental variable effect sizes for each of the covariates is in brackets.

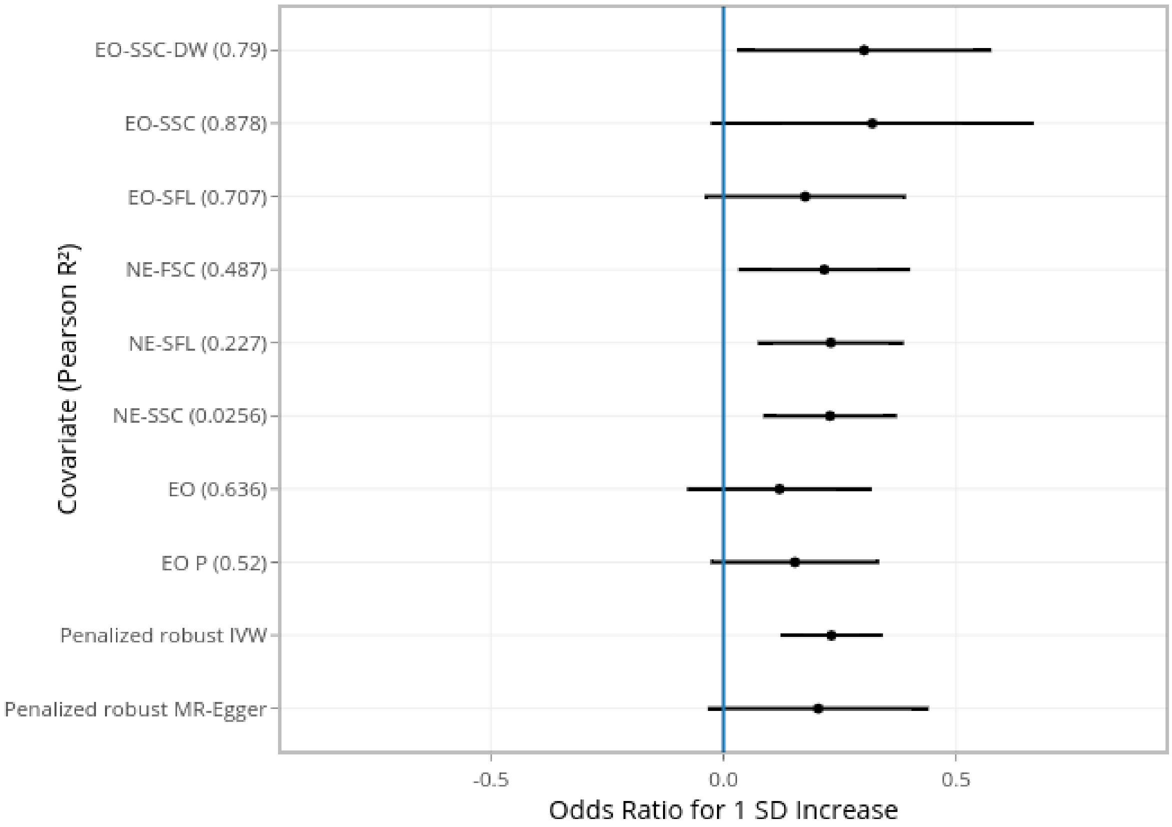

Figure 5.11: Pairwise multivariable IVW MR models containing EO-FSC and a covariate.
Each covariate is listed on the y axis with the Pearson correlation between the phenotype of that covariate and EO-FSC. The x axis is the calculated change in odds ratio for disease risk given 1 SD increase in EO-FSC in a multivariable MR model containing the corresponding covariate. This is compared to the estimates for causal association from Penalised robust IVW and Penalised robust MR-Egger containing EO-FSC only. The Pearson correlation (r^2) between instrumental variable effect sizes for each of the covariates is in brackets.

5.3.4.3 Neutrophil side scatter and coronary artery disease

Neutrophils play a crucial role in thrombosis and acute coronary syndromes. In a mouse study of endothelial damage, neutrophils were shown to be the first cells at the site of damage even preceding platelets [51]. In acute coronary syndromes, neutrophil degranulation damages intact cells, the extracellular matrix,promotes further neutrophil recruitment, and increases infarct size [69]. In particular, neutrophil recruitment of monocytes by release of chemoattractants which includes granule proteins has been noted as a cause for the role of neutrophils in acute coronary syndrome [69]. *ARSB* is a blood marker of neutrophil activation correlates with poor prognosis of heart disease [69] [26]. It is notable that variants in this gene are associated with NE-SSC and colocalise with pQTL of *ARSB* in the blood. Two instrumental variables annotated to *ARSB* inform the aforementioned MR analysis of NE-SSC and CAD. Furthermore, 11 instrumental variables annotated to the α-DEFENSIN are present in the MR analysis. To identify if α-DEFENSIN variants are responsible for the identified causal association I removed all variants within the range chr8:4,000,000-7,000,000 (hg19). This removal results in no significant association between NE-SSC and CAD, but simultaneously does now show a significant change in the causal estimate (Table 5.9). This suggests that instrumental variables in the α-DEFENSIN locus are consistent with the estimated causality, but do not explain the observation in entirety

5.3.4.4 Neutrophil side scatter and lung cancer

Neutrophils have been suggested to play both pro-tumorigenic and anti-tumorigenic roles [168]. Serum α-DEFENSIN protein levels have been noted to be elevated in patients with lung cancer and this has been suggested as a diagnostic tool for lung cancer [12]. IVs located in genes *DEFA9P*, *DEFA3*, *DEFA1B*, *DEFA11P* are located along the axis of causality in the robust penalised IVW MR showing that these signals are along the predicted causal axis. Removal of the 11 IVs located in the α-DEFENSIN locus results in a loss of significance in the causal estimate between NE-SSC and Lung Cancer, but no great change in the estimated causal effect (Table 5.9).

Risk Factor	Outcome	Robust Penalised IVW Estimate (P-value)	Removal DEF IVs Estimate (P-value)
NE-SSC	CAD	0.0346 (1.63×10^{-5})	0.0273 (0.0369)
NE-SSC	Lung Cancer	0.0548 (7.47×10^{-5})	0.0571 (0.0109)

Table 5.9: Comparison of MR estimates following removal of α-DEFENSIN locus instrumental variables.

Removal of instrumental variables in the α-DEFENSIN locus shows that the significant causal association identified between NE-SSC with CAD and lung cancer is dependent on the α-DEFENSIN locus.

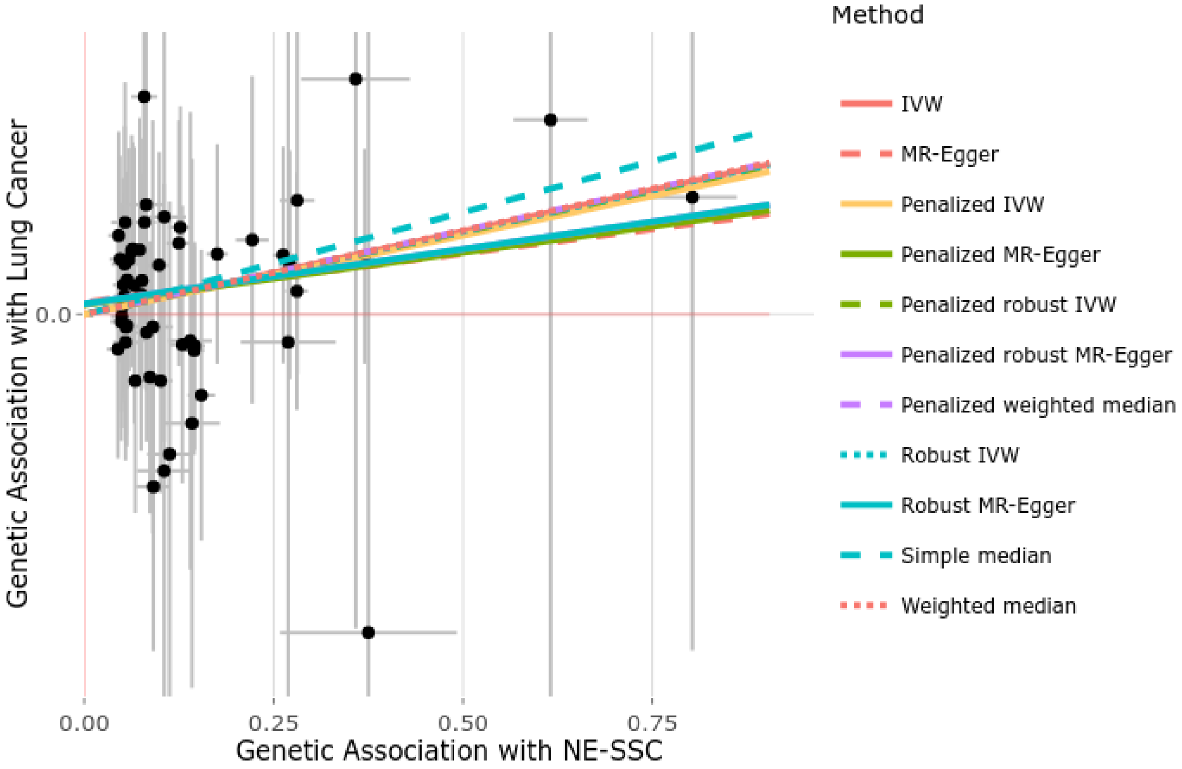

Figure 5.12: Scatterplot of instrumental variables and association with NE-SSC and lung cancer.

Bars represent 95% confidence intervals for estimate of effect size of instrumental variables with either the exposure or outcome. Results show good concordance between causal estimates of sensitivity analyses. The Penalised Robust IVW model predicts causal association between NE-SSC and lung cancer ($\beta : 0.0561, \text{P-value} : 4.55 \times 10^{-5}$).

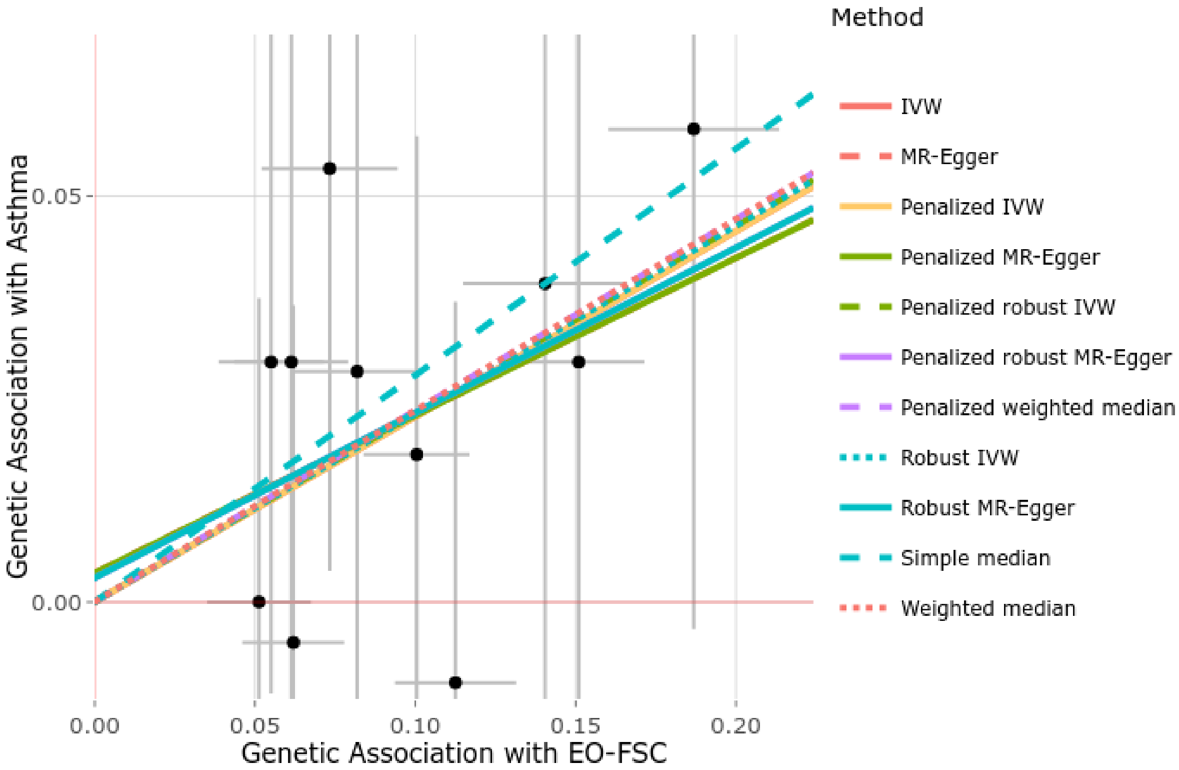

Figure 5.13: Scatterplot of instrumental variables and association with EO-FSC and asthma.
Bars represent 95% confidence intervals for estimate of effect size of instrumental variables with either the exposure or outcome. Results show good concordance between causal estimates of sensitivity analyses. The Penalised Robust IVW model predicts causal association between EO-FSC and Asthma ($\beta : 0.232, \text{P-value} : 3.68 \times 10^{-5}$).

5.3.4.5 Eosinophil forward scatter and asthma

Eosinophil forward scatter shows a positive causal association with risk of asthma according to the univariable MR study and associated sensitivity analyses (Fig. 5.13). However, the associated effect size of instrumental variables for EO-FSC are highly correlated with EO-SSC (pearson r^2 0.93) suggesting pleiotropy. Although, it is interesting to note that this purported causal relationship seems to be largely independent of the known causal association between eosinophil count and asthma [15] Fig. 5.11). This suggests two separate eosinophil properties which causally increase the risk of asthma. More study is required to understand the effect of these instrumental variables which may be acting through another intermediary risk factor, but seem independent to the already shown causal relationship between eosinophil count and asthma [15].

5.4 Discussion

In this chapter I have detailed my downstream analysis of GWAS results to generate hypotheses relating to the study of haematology and disease including cardiovascular disease, immune disorders, and cancer. My colocalisation analysis annotated genetic determinants associated with Sysmex parameters to signals influencing blood cell transcriptome, blood plasma proteome, and disease risk. This analysis identified genetic determinants as causally mediating multiple factors and thus helps with interpretation of the gene and pathways which are mediated by the genetic determinant. My results have identified disease colocalisations with associated signals in genes which are known drug targets such as IL-2RA and Daclizumab and evidence for the role of IL-18R1 in aetiology of celiac disease and IL-7R in a number of autoimmune disorders. Furthermore, I performed a MR analysis to determine causal associations between the risk factors and disease, the risk factors of interest being Sysmex parameters which have been shown to be relevant proxies for cell immune function and activation. Identified associations include NE-SSC and lung cancer or CAD, and EO-FSC and asthma. MR analysis is limited the assumptions of MR models, which I address with a sensitivity analysis to assess robust causal estimates. Interpretation of these results should be made within context of the limitations of genetic studies and the phenotypes of interest which are discussed further in Chapter 6.

Chapter 6

Conclusion

I have performed the first ever GWAS of flow cytometry parameters derived from a Sysmex analyser. Many of these parameters have been shown to be clinically and functionally relevant readouts of the haematological system (Section 3.2.3). My analysis identified hundreds of new genetic loci and association signals, many of which are located in genes known to be relevant for immune cell function and activation (Section 4). Furthermore, I contribute to the largest ever study of classical FBC haematological parameters and identify a large number of distinct signals including those which have not previously been associated with blood cell phenotypes.

In **Chapter 2** I describe my contributions to the BCX consortium and our collective work to perform the largest GWAS of haematological traits. This includes a conditional analysis of the UK Biobank cohort and subsequent joint modelling to identify distinct signals identified by a larger meta-analysis.

In **Chapter 3** I discuss data collection and QC of extended Sysmex parameters and associated genotypes. This includes correction of data for technical and environmental factors. Technical factors included time of day, time since start of study, and time of year. Environmental factors included, smoking history, age, sex, and weight. Adjustment for these factors reduced variation in parameters and increased power to detect genetic associations.

In **Chapter 4** I detail my GWAS and conditional analysis of Sysmex parameters, this includes discussion of multiple testing and population stratification in GWAS studies. I show that my GWAS and conditional analysis of extended Sysmex parameters identifies 2,142 conditionally independent associations and 849 LD sets across 63 phenotypes.

Finally, in **Chapter 5** I outline my downstream analysis of GWAS results, including disease, eQTL, and pQTL colocalisation. This analysis identifies genetic determinants of Sysmex parameters which also influence other biological properties including disease risk. Examples including *IL18R1* associated with NE-FSC and atopic dermatitis, and *IL2RA* associated with lymphocyte parameters and MS.

6.1 Limitations of work

6.1.1 Interpretability of Sysmex parameters

I present Sysmex parameters as 'functionally relevant' measurements of blood cell properties as justified with prior literature (Section 3.2.3) and also by a literature review of genes identified from annotation of GWAS association signals which identified genes relevant to white cell function (Section 4.3.1.2). However, the interpretability of these phenotypes and importantly, the possibility to intervene on these phenotypes as a clinical end point is limited. For example, neutrophil side scatter (NE-SSC) a known index of granulation which has been shown to correlate with incidence of disease and visual assessments of neutrophil granularity. I show evidence from MR analysis that suggests NE-SSC may be a causal mediator of CAD and lung cancer (Section 5.3.4). However, there is no currently known clinical intervention that could reduce NE-SSC in patients. Alternatively, lymphocyte side scatter (LY-SSC) is a measure of the side scatter of lymphocyte cells and GWAS of this trait identifies a number of association signals located in genes important for lymphocyte function, often colocalising in diseases with known lymphocyte involvement such as MS. However, for associations which raise the LY-SSC property, it is not clear by which lymphocyte subsets the effect is being observed. It could be the granular LGL cells such as NK or cytotoxic T cells, or some other structural changes in other lymphocyte subtypes.

In modern GWAS studies, single sets of phenotypes are rarely considered alone because freely available summary statistics allows more detailed analysis including multiple sets of phenotypes. Sysmex parameters can be used a intermediate traits and colocalised to more interpretable outcomes such as disease risk, or eQTL and pQTL measurements. In this way, Sysmex parameters are yet another layer of information which can be used to annotate genetic determinants and shouldn't be used as the sole source of information to support a hypothesis.

6.1.2 Establishing causality

Establishing causal relationships between biological components is fundamental to our understanding of biology. However, the hypothesis generating nature of GWAS analysis presents challenges in proving causality. For example, the identification of an association signal suggests many, often hundreds, of variants which could be mediating the association signal (Section 1.4). GWAS analyses are not measuring the change in a phenotype as a result of an genetic intervention. Therefore it is not easily possible to define which variant is causally mediating the observed signal. It is possible that the true causal variant may not have been genotyped or imputed. Such a scenario could be difficult or impossible to identify without performing an experiment where the genetic variant is induced and any changes in phenotype observed.

Establishing causal relationships between phenotypes or traits is similarly difficult. In many points in my thesis I describe colocalisation analysis which shows a common genetic determinant between two phenotypes (Sections 5.3.2.1, 5.3.2.4). There are multiple mechanisms by which a colocalisation between two traits would arise. Firstly there could be a causal relationship between two traits where mediation of one trait results in concomitant changes in the other, for example, low density lipoprotein (LDL) and CAD risk. Alternatively, the genetic variant could be mediating both traits independently, through different biological pathways. Followup MR studies can determine causality between two traits, and this approach has been successful in the study of blood cell phenotypes before [15]. However, MR analysis are burdened with a number of assumptions many of which are difficult to test. For example, ensuring that instrumental variables are non-pleiotropic and are not influencing the outcome via a different biological mechanism other than the intended exposure (Section 5.2.2). This is difficult to prove definitively for any single instrumental variable.

6.1.3 Limitations of colocalisation

If a locus contains association signals with two phenotypes, colocalisation analysis can determine if the associations are being caused by the same genetic determinant. An introduction to colocalisation analysis and an overview of the mathematical implementation can be found in Sections 1.6.2 and 5.2.1 respectively. Here I focus my discussion on the limitations of colocalisation which must be held in account when interpreting results colocalisation analysis:

1. If two distinct associations are caused by a two distinct underlying variants which are in LD 1.00 r^2, colocalisation cannot distinguish between such associations.

2. Colocalisation assumes only one association signal per locus, if multiple association signals exist they can bias results towards reporting false negatives.

3. It must be stressed that colocalisation simply identifies a common genetic determinant in a particular locus between two phenotypes, not a causal relationship between the phenotypes

Situations leading to point 1) cannot be tested properly with a statistical genetics approach alone, we must always aim to validate with biological experimentation, such as genetic modification of an animal or tissue model. However, in the case of rare variants which are less likely to be in high LD with other variants this limitation is ameliorated by being less likely to occur. Furthermore, given that most genetic variants in the genome have almost no effect on a phenotype in question, it would seem unlikely that within a small set of variants which are in high LD two variants would have distinct effects on separate phenotypes. Point 2) is less concerning as it can only lead to false negative results, furthermore an associated locus is unlikely to ever contain a single phenotype influencing variant. There likely will exist many possible associations nearby which simply do not reach significance, nearby associations only become problematic when their relative significance is close to the significance of the 'main' association we wish to colocalise. As GWAS is performed on a greater number of phenotypes, colocalisation analysis will become more important to allow proper characterisation of the effect of a genetic determinant.

6.1.4 Replication of results

A unique aspect of my work is being the first analysis of functionally relevant white cell phenotypes. However it follows that my analysis is also lacking in a replication set which would allow me to further validate my results. Confounding effects such as population stratification or technical artefacts can lead to spurious associations, validation can help identify these false positives. I hope my thesis will lead to more interest in Sysmex parameters and thus spur further GWAS analyses of these phenotypes.

6.2 Recommendations for future research

6.2.1 Validation and meta-analysis with the COMPARE study

The COMPARE study has recruited a cohort of 31,000 healthy donors to compare three methods for making haemoglobin measurements: blood extraction from the finger, measurement with a spectrometer placed over the skin, Sysmex automated haematology analysis. At the time of writing genotyping of this cohort has not concluded, in the future this data could be utilised to validate my GWAS of Sysmex parameters in INTERVAL. Furthermore, a meta-analysis with both the INTERVAL and COMPARE cohorts would increase the population cohort to 66,000 individuals. Combining the technical and environmental correction procedures between these studies would allow for a more accurate statistical deduction on the effects of covariates on haematological parameters and make a more effective correction procedure.

6.2.2 Utilising the second measurement from the INTERVAL trial

Blood donors in both the INTERVAL and COMPARE cohorts provide a second measurement for haematological analysis, which in the case of INTERVAL is 2 years following their initial donation, and in the COMPARE study following their first donation. I only utilised the second time-point of measurement for participants in the INTERVAL trial if data for the first time point was not available. This was the case for only a small subset of individuals due to a data storage issue during the trial. A statistical correction procedure which utilises both first and second measurement in all individuals where available could significantly increase the effective power to detect genetic associations. This analysis could be done by firstly performing technical and environmental correction on all measurements as before, then simply calculating the mean between both time points for each individuals. This analysis would be complicated by at least two factors:

1. Individuals in INTERVAL would be two years older in their second measurement. In some participants this may mean they would have experienced menopause which is known to have a large impact on haematological measurements.

2. Donors in the INTERVAL cohort would have been assigned to different donation schedules.

These factors could be mitigated by including age and donation schedule as a factor in the environmental adjustment, menopause status was already included in this adjustment.

Separately, comparing second and first measurement would provide a better sense of intra-individual variability of these parameters. It is not known, for example if individuals generally have fairly consistent eosinophil size (EO-FSC) measurements over the course of time. These observations would be of general interest to haematologists and provide better characterisation of these haematological properties.

6.2.3 Raw flow cytometry data and predictive models

Haematological parameters have long been used to make diagnoses and predict disease status. However, there likely exists a greater degree of information in a Sysmex analysis beyond what is reported as a parameter. The three dimensional position of every cell in the scattergram is not a useful representation of information for a clinical doctor as this data would be too difficult for a them to interpret. Instead of relying on parameters which are inherently a simplification of the rich data available from a scattergram, we could use modern statistical techniques to consider the entire dataset in order to make predictions regarding disease status. Machine learning algorithms such as neural networks have the ability to consider a high-dimensional input of data, training this algorithm over time would allow learning of which features in the dataset are the most important with respect to predicting an outcome. A potential outcome would be to use the raw Sysmex cytometry data to train a neural network to predict the age, sex, or menopause status of a participant. This protocol could also be used to predict disease status of participants, although this would not be possible in using INTERVAL or COMPARE datasets as the participants are healthy blood donors. More interesting would be to combine the Sysmex scattergram data with participant age, sex, weight, height, and other information. Neural networks would effectively integrate these input features and learn relationships which are useful with respect to the outcome.

6.3 Closing statement

In the introduction of my thesis I presented statistical genetics as not just as a methodology to better understand human genetics, but also a method by which we can further our understanding of the biology and disease aetiology. Biology exists as a complex set of interconnected elements, connected deferentially by time, space, and biological compartment. I propose that statistical geneticists of the future should not narrow their analysis to a particular set of phenotypes or a disease outcome. Using genetic variants we can reliably combine data from multiple studies done over time in different sets of individuals, thus making the massive phenotypic profiling required to develop a comprehensive understanding of biological systems tractable. I hope that further study in this field continues to advance towards mathematically grounded study of biological systems which allow not only for accurate exchange of information between researchers, but also ability to make empirical and testable predictions.

Bibliography

[1] A history of the human genome project. *Science*, 291(5507):1195–1195, February 2001.

[2] Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis. *Nature Genetics*, 45(11):1353–1360, September 2013.

[3] A global reference for human genetic variation. *Nature*, 526(7571):68–74, September 2015.

[4] Abbexa. Flow cytometry - the flow cytometer. *Abbexa*, ”july” 2019.

[5] Randa I. Abu-Ghazaleh, Sandra L. Dunnette, David A. Loegering, James L. Checked, Hirohito Kita, Larry L. Thomas, and Gerald J. Gleich. Eosinophil granule proteins in peripheral blood granulocytes. *Journal of Leukocyte Biology*, 52(6):611–618, December 1992.

[6] Parsa Akbari. Github repository: Ld score regression and association comparisons. `https://github.com/ParsaAkbari/LDSC-Comparisons`.

[7] Parsa Akbari. Github repository: Phenotype adjustment. `https://github.com/ParsaAkbari/Phenotype-Adjusts`.

[8] Parsa Akbari. Github repository: Uk biobank 500k conditional analysis. `https://github.com/ParsaAkbari/UKBB500K-Conditional-Analysis`.

[9] Ehab Almetwally and Hisham Almongy. Comparison between m-estimation, s-estimation, and mm estimation methods of robust estimation with application and simulation. 11 2018.

[10] Patrick R. Amestoy. *igraph: Network Analysis and Visualization*, 2019. R package version 1.2.4.1.

[11] Emanuele Di Angelantonio, Simon G Thompson, Stephen Kaptoge, Carmel Moore, Matthew Walker, Jane Armitage, Willem H Ouwehand, David J Roberts, and John Danesh. Efficiency and safety of varying the frequency of whole blood donation

(INTERVAL): a randomised trial of 45 000 donors. *The Lancet*, 390(10110):2360–2371, November 2017.

[12] Y. Arimura, J. Ashitani, S. Yanagi, M. Tokojima, K. Abe, H. Mukae, and M. Nakazato. Elevated serum beta-defensins concentrations in patients with lung cancer. *Anticancer Res.*, 24(6):4051–4057, 2004.

[13] Borros M. Arneth, Maximilian Ragaller, Kathleen Hommel, Oliver Tiebel, Mario Menschikowski, and Gabriele Siegert. Novel parameters of extended complete blood cell count under fluorescence flow cytometry in patients with sepsis. *Journal of Clinical Laboratory Analysis*, 28(2):130–135, jan 2014.

[14] William Astle. Github repository: gen2hd5.

[15] William J. Astle, Heather Elding, Tao Jiang, Dave Allen, Dace Ruklisa, Alice L. Mann, Daniel Mead, Heleen Bouman, Fernando Riveros-Mckay, Myrto A. Kostadima, John J. Lambourne, Suthesh Sivapalaratnam, Kate Downes, Kousik Kundu, Lorenzo Bomba, Kim Berentsen, John R. Bradley, Louise C. Daugherty, Olivier Delaneau, Kathleen Freson, Stephen F. Garner, Luigi Grassi, Jose Guerrero, Matthias Haimel, Eva M. Janssen-Megens, Anita Kaan, Mihir Kamat, Bowon Kim, Amit Mandoli, Jonathan Marchini, Joost H.A. Martens, Stuart Meacham, Karyn Megy, Jared O'Connell, Romina Petersen, Nilofar Sharifi, Simon M. Sheard, James R. Staley, Salih Tuna, Martijn van der Ent, Klaudia Walter, Shuang-Yin Wang, Eleanor Wheeler, Steven P. Wilder, Valentina Iotchkova, Carmel Moore, Jennifer Sambrook, Hendrik G. Stunnenberg, Emanuele Di Angelantonio, Stephen Kaptoge, Taco W. Kuijpers, Enrique Carrillo de Santa-Pau, David Juan, Daniel Rico, Alfonso Valencia, Lu Chen, Bing Ge, Louella Vasquez, Tony Kwan, Diego Garrido-Martín, Stephen Watt, Ying Yang, Roderic Guigo, Stephan Beck, Dirk S. Paul, Tomi Pastinen, David Bujold, Guillaume Bourque, Mattia Frontini, John Danesh, David J. Roberts, Willem H. Ouwehand, Adam S. Butterworth, and Nicole Soranzo. The allelic landscape of human blood cell trait variation and links to common complex disease. *Cell*, 167(5):1415–1429.e19, nov 2016.

[16] Gavin Band and Jonathan Marchini. Qctool. `https://www.well.ox.ac.uk/~gav/qctool_v2/#documentation`, 2018.

[17] Marie-Eve Beaulieu, Toni Jauset, Daniel Massó-Vallés, Sandra Martínez-Martín, Peter Rahl, Loka Maltais, Mariano F. Zacarias-Fluck, Sílvia Casacuberta-Serra, Erika Serrano del Pozo, Christopher Fiore, Laia Foradada, Virginia Castillo Cano, Meritxell Sánchez-Hervás, Matthew Guenther, Eduardo Romero Sanz, Marta Oteo, Cynthia Tremblay, Génesis Martín, Danny Letourneau, Martin Montagne, Miguel

Ángel Morcillo Alonso, Jonathan R. Whitfield, Pierre Lavigne, and Laura Soucek. Intrinsic cell-penetrating activity propels omomyc from proof of concept to viable anti-MYC therapy. *Science Translational Medicine*, 11(484):eaar5012, March 2019.

[18] Brian Becknell, Tad E. Eichler, Susana Beceiro, Birong Li, Robert S. Easterling, Ashley R. Carpenter, Cindy L. James, Kirk M. McHugh, David S. Hains, Santiago Partida-Sanchez, and John D. Spencer. Ribonucleases 6 and 7 have antimicrobial function in the human and murine urinary tract. *Kidney International*, 87(1):151–161, January 2015.

[19] V.B. Behncke, G. Alemar, D.A. Kaufman, and F.J. Eidelman. Azelastine nasal spray and fluticasone nasal spray in the treatment of geriatric patients with rhinitis. *Journal of Allergy and Clinical Immunology*, 117(2):S263, February 2006.

[20] Celine Bellenguez, Amy Strange, Colin Freeman, and Peter Donnelly. A robust clustering algorithm for identifying problematic samples in genome-wide association studies. *Bioinformatics*, 28(1), November 2011.

[21] Celine Bellenguez, Amy Strange, Colin Freeman, and Peter Donnelly. A robust clustering algorithm for identifying problematic samples in genome-wide association studies. *Bioinformatics*, 28(1):134, November 2011.

[22] Derrick A Bennett and Michael V Holmes. Mendelian randomisation in cardiovascular research: an introduction for clinicians. *Heart*, 103(18):1400–1407, June 2017.

[23] Beben Benyamin, Manuel A R Ferreira, Gonneke Willemsen, Scott Gordon, Rita P S Middelberg, Brian P McEvoy, Jouke-Jan Hottenga, Anjali K Henders, Megan J Campbell, Leanne Wallace, Ian H Frazer, Andrew C Heath, Eco J C de Geus, Dale R Nyholt, Peter M Visscher, Brenda W Penninx, Dorret I Boomsma, Nicholas G Martin, Grant W Montgomery, and John B Whitfield. Common variants in TMPRSS6 are associated with iron status and erythrocyte volume. *Nature Genetics*, 41(11):1173–1175, October 2009.

[24] B. Bielekova, M. Catalfamo, S. Reichert-Scrivner, A. Packer, M. Cerna, T. A. Waldmann, H. McFarland, P. A. Henkart, and R. Martin. Regulatory CD56bright natural killer cells mediate immunomodulatory effects of IL-2r -targeted therapy (daclizumab) in multiple sclerosis. *Proceedings of the National Academy of Sciences*, 103(15):5941–5946, April 2006.

[25] UK Biobank. 2015. `http://www.ukbiobank.ac.uk/wp-content/uploads/2014/04/imputation_documentation_May2015.pdf`.

[26] Erik Biros, Corey S. Moran, Jane Maguire, Elizabeth Holliday, Christopher Levi, and Jonathan Golledge. Upregulation of arylsulfatase b in carotid atherosclerosis is associated with symptoms of cerebral embolization. *Scientific Reports*, 7(1), June 2017.

[27] I. Björk and U. Lindahl. Mechanism of the anticoagulant action of heparin. *Molecular and Cellular Biochemistry*, 48(3):161–182, Jan 1982.

[28] Niels Borregaard, Ole E. Sørensen, and Kim Theilgaard-Mönch. Neutrophil granules: a library of innate immunity proteins. *Trends in Immunology*, 28(8):340–345, August 2007.

[29] J. Bowden, G. Davey Smith, and S. Burgess. Mendelian randomization with invalid instruments: effect estimation and bias detection through egger regression. *International Journal of Epidemiology*, 44(2):512–525, April 2015.

[30] Jack Bowden, George Davey Smith, Philip C. Haycock, and Stephen Burgess. Consistent estimation in mendelian randomization with some invalid instruments using a weighted median estimator. *Genetic Epidemiology*, 40(4):304–314, April 2016.

[31] Gabriel Brisou, Delphine Manzoni, Stephane Dalle, Pascale Felman, Dominique Morel, Marouane Boubaya, Jean Pierre Magaud, and Lucile Baseggio. Alarms and parameters generated by hematology analyzer: New tools to predict and quantify circulating sezary cells. *Journal of Clinical Laboratory Analysis*, 29(2):153–161, mar 2014.

[32] Yenan T. Bryceson, Michael E. March, Domingo F. Barber, Hans-Gustaf Ljunggren, and Eric O. Long. Cytolytic granule polarization and degranulation controlled by different receptors in resting NK cells. *The Journal of Experimental Medicine*, 202(7):1001–1012, October 2005.

[33] Brendan K Bulik-Sullivan, , Po-Ru Loh, Hilary K Finucane, Stephan Ripke, Jian Yang, Nick Patterson, Mark J Daly, Alkes L Price, and Benjamin M Neale. LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics*, 47(3):291–295, February 2015.

[34] Annalisa Buniello, Jacqueline A L MacArthur, Maria Cerezo, Laura W Harris, James Hayhurst, Cinzia Malangone, Aoife McMahon, Joannella Morales, Edward Mountjoy, Elliot Sollis, Daniel Suveges, Olga Vrousgou, Patricia L Whetzel, Ridwan Amode, Jose A Guillen, Harpreet S Riat, Stephen J Trevanion, Peggy Hall, Heather Junkins, Paul Flicek, Tony Burdett, Lucia A Hindorff, Fiona Cunningham, and Helen

Parkinson. The NHGRI-EBI GWAS catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research*, 47(D1):D1005–D1012, November 2018.

[35] Sabrina Buoro1, Michela Seghezzi, Mauro Vavassori, Paola Dominoni, Sara Apassiti Esposito, Barbara Manenti, Tommaso Mecca, Gianmariano Marchesi, Enrico Castellucci, Giovanna Azzarà, Cosimo Ottomano, and Giuseppe Lippi. Clinical significance of cell population data (CPD) on sysmex XN-9000 in septic patients with our without liver impairment. *Annals of Translational Medicine*, 4(21):418–418, nov 2016.

[36] Stephen Burgess, Verena Zuber, Apostolos Gkatzionis, Jessica M. B. Rees, and Christopher Foley. Improving on a modal-based estimation method: model averaging for consistent and efficient estimation in mendelian randomization when a plurality of candidate instruments are valid. August 2017.

[37] Mauro Buttarello and Mario Plebani. Automated blood cell counts. *American Journal of Clinical Pathology*, 130(1):104–116, July 2008.

[38] Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O'Connell, Adrian Cortes, Samantha Welsh, Gil McVean, Stephen Leslie, Peter Donnelly, and Jonathan Marchini. Genome-wide genetic data on ˜500, 000 UK biobank participants. July 2017.

[39] Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T. Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O'Connell, Adrian Cortes, Samantha Welsh, Alan Young, Mark Effingham, Gil McVean, Stephen Leslie, Naomi Allen, Peter Donnelly, and Jonathan Marchini. The UK biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203–209, October 2018.

[40] Dawn J. Caster, David W. Powell, Irina Miralda, Richard A. Ward, and Kenneth R. McLeish. Re-examining neutrophil participation in GN. *Journal of the American Society of Nephrology*, 28(8):2275–2289, June 2017.

[41] D. M. Chalmers, A. J. Levi, I. Chanarin, W. R. S. North, and T. W. Meade. Mean cell volume in a working population: the effects of age, smoking, alcohol and oral contraception. *British Journal of Haematology*, 43(4):631636, Dec 1979.

[42] John C Chambers, Weihua Zhang, Yun Li, Joban Sehmi, Mark N Wass, Delilah Zabaneh, Clive Hoggart, Henry Bayele, Mark I McCarthy, Leena Peltonen, Nelson B

Freimer, Surjit K Srai, Patrick H Maxwell, Michael J E Sternberg, Aimo Ruokonen, Gonçalo Abecasis, Marjo-Riitta Jarvelin, James Scott, Paul Elliott, and Jaspal S Kooner. Genome-wide association study identifies variants in TMPRSS6 associated with hemoglobin levels. *Nature Genetics*, 41(11):1170–1172, October 2009.

[43] L. Chen, M. Kostadima, J. H. A. Martens, G. Canu, S. P. Garcia, E. Turro, K. Downes, I. C. Macaulay, E. Bielczyk-Maczynska, S. Coe, S. Farrow, P. Poudel, F. Burden, S. B. G. Jansen, W. J. Astle, A. Attwood, T. Bariana, B. de Bono, A. Breschi, J. C. Chambers, F. A. Choudry, L. Clarke, P. Coupland, M. van der Ent, W. N. Erber, J. H. Jansen, R. Favier, M. E. Fenech, N. Foad, K. Freson, C. van Geet, K. Gomez, R. Guigo, D. Hampshire, A. M. Kelly, H. H. D. Kerstens, J. S. Kooner, M. Laffan, C. Lentaigne, C. Labalette, T. Martin, S. Meacham, A. Mumford, S. T. Nurnberg, E. Palumbo, B. A. van der Reijden, D. Richardson, S. J. Sammut, G. Slodkowicz, A. U. Tamuri, L. Vasquez, K. Voss, S. Watt, S. Westbury, P. Flicek, R. Loos, N. Goldman, P. Bertone, R. J. Read, S. Richardson, A. Cvejic, N. Soranzo, W. H. Ouwehand, H. G. Stunnenberg, M. Frontini, and A. Rendon and. Transcriptional diversity during lineage commitment of human blood progenitors. *Science*, 345(6204):1251033–1251033, September 2014.

[44] G. Cheng, S. Swaidani, M. Sharma, M. E. Lauer, V. C. Hascall, and M. A. Aronica. Correlation of hyaluronan deposition with infiltration of eosinophils and lymphocytes in a cockroach-induced murine model of asthma. *Glycobiology*, 23(1):4358, Aug 2012.

[45] Wikimedia Commons. Schematic of neutrophil cell, 2017.

[46] The UK10K Consortium. The UK10k project identifies rare variants in health and disease. *Nature*, 526(7571):82–90, September 2015.

[47] David Cook, Dearg Brown, Robert Alexander, Ruth March, Paul Morgan, Gemma Satterthwaite, and Menelas N. Pangalos. Lessons learned from the fate of AstraZeneca's drug pipeline: a five-dimensional framework. *Nature Reviews Drug Discovery*, 13(6):419–431, May 2014.

[48] Heather J. Cordell, Younghun Han, George F. Mells, Yafang Li, Gideon M. Hirschfield, Casey S. Greene, Gang Xie, Brian D. Juran, Dakai Zhu, David C. Qian, James A. B. Floyd, Katherine I. Morley, Daniele Prati, Ana Lleo, Daniele Cusi, M. Eric Gershwin, Carl A. Anderson, Konstantinos N. Lazaridis, Pietro Invernizzi, Michael F. Seldin, Richard N. Sandford, Christopher I. Amos, Katherine A. Siminovitch, and and. International genome-wide meta-analysis identifies new primary biliary cirrhosis risk loci and targetable pathogenic pathways. *Nature Communications*, 6(1), September 2015.

[49] C.J. Corrigan and A.B. Kay. T cells and eosinophils in the pathogenesis of asthma. *Immunology Today*, 13(12):501–507, January 1992.

[50] Wallace H Coulter. Means for counting particles suspended in a fluid, 1953. `https://patents.google.com/patent/US2656508A/en`.

[51] Roxane Darbousset, Grace M. Thomas, Soraya Mezouar, Corinne Frère, Rénaté Bonier, Nigel Mackman, Thomas Renné, Françoise Dignat-George, Christophe Dubois, and Laurence Panicot-Dubois. Tissue factor–positive neutrophils bind to injured endothelial wall and initiate thrombus formation. *Blood*, 120(10):2133–2143, September 2012.

[52] S. Das, N. Nikolaidis, H. Goto, C. McCallister, J. Li, M. Hirano, and M. D. Cooper. Comparative genomics and evolution of the alpha-defensin multigene family in primates. *Molecular Biology and Evolution*, 27(10):2333–2343, May 2010.

[53] Luke C Davies, Stephen J Jenkins, Judith E Allen, and Philip R Taylor. Tissue-resident macrophages. *Nature Immunology*, 14(10):986–995, September 2013.

[54] Charles DeLisi. Santa fe 1986: Human genome baby-steps. *Nature*, 455(7215):876–877, October 2008.

[55] Foivos I. Diakogiannis, Geraint F. Lewis, and Rodrigo A. Ibata. Resolving the mass–anisotropy degeneracy of the spherically symmetric jeans equation – i. theoretical foundation. *Monthly Notices of the Royal Astronomical Society*, 443(1):598–609, July 2014.

[56] R Dulbecco. A turning point in cancer research: sequencing the human genome. *Science*, 231(4742):1055–1056, March 1986.

[57] K Dunning and AO Safo. The ultimate wright-giemsa stain: 60 years in the making. *Biotechnic & Histochemistry*, 86(2):69–75, March 2011.

[58] R. Durbin. Efficient haplotype matching and storage using the positional burrows-wheeler transform (PBWT). *Bioinformatics*, 30(9):1266–1272, jan 2014.

[59] Pablo Engel, Laurence Boumsell, Robert Balderas, Armand Bensussan, Valter Gattei, Vaclav Horejsi, Bo-Quan Jin, Fabio Malavasi, Frank Mortari, Reinhard Schwartz-Albiez, Hannes Stockinger, Menno C. van Zelm, Heddy Zola, and Georgina Clark. CD nomenclature 2015: Human leukocyte differentiation antigen workshops as a driving force in immunology. *The Journal of Immunology*, 195(10):4555–4563, November 2015.

[60] Mikkel Faurschou, Ole E Sørensen, Anders H Johnsen, Jon Askaa, and Niels Borregaard. Defensin-rich granules of human neutrophils: characterization of secretory properties. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research*, 1591(1-3):29–35, August 2002.

[61] Manuel A.R. Ferreira, Jouke-Jan Hottenga, Nicole M. Warrington, Sarah E. Medland, Gonneke Willemsen, Robert W. Lawrence, Scott Gordon, Eco J.C. de Geus, Anjali K. Henders, Johannes H. Smit, Megan J. Campbell, Leanne Wallace, David M. Evans, Margaret J. Wright, Dale R. Nyholt, Alan L. James, John P. Beilby, Brenda W. Penninx, Lyle J. Palmer, Ian H. Frazer, Grant W. Montgomery, Nicholas G. Martin, and Dorret I. Boomsma. Sequence variants in three loci influence monocyte counts and erythrocyte volume. *The American Journal of Human Genetics*, 85(5):745–749, November 2009.

[62] Lars Feuk, Andrew R. Carson, and Stephen W. Scherer. Structural variation in the human genome. *Nature Reviews Genetics*, 7(2):85–97, February 2006.

[63] M. J. Fulwyler. Electronic separation of biological cells by volume. *Science*, 150(3698):910–911, November 1965.

[64] J. R. Furundarena, M. Araiz, M. Uranga, M. R. Sainz, A. Agirre, M. Trassorras, N. Uresandi, M. C. Montes, and N. Argoitia. The utility of the sysmex XE-2100 analyzer's NEUT-x and NEUT-y parameters for detecting neutrophil dysplasia in myelodysplastic syndromes. *International Journal of Laboratory Hematology*, 32(3):360–366, jun 2010.

[65] Kevin J. Galinsky, Gaurav Bhatia, Po-Ru Loh, Stoyan Georgiev, Sayan Mukherjee, Nick J. Patterson, and Alkes L. Price. Fast principal-component analysis reveals convergent evolution of ADH1b in europe and east asia. *The American Journal of Human Genetics*, 98(3):456–472, March 2016.

[66] Santhi K Ganesh, Neil A Zakai, Frank J A van Rooij, Nicole Soranzo, Albert V Smith, Michael A Nalls, Ming-Huei Chen, Anna Kottgen, Nicole L Glazer, Abbas Dehghan, Brigitte Kuhnel, Thor Aspelund, Qiong Yang, Toshiko Tanaka, Andrew Jaffe, Joshua C M Bis, Germaine C Verwoert, Alexander Teumer, Caroline S Fox, Jack M Guralnik, Georg B Ehret, Kenneth Rice, Janine F Felix, Augusto Rendon, Gudny Eiriksdottir, Daniel Levy, Kushang V Patel, Eric Boerwinkle, Jerome I Rotter, Albert Hofman, Jennifer G Sambrook, Dena G Hernandez, Gang Zheng, Stefania Bandinelli, Andrew B Singleton, Josef Coresh, Thomas Lumley, André G Uitterlinden, Janine M vanGils, Lenore J Launer, L Adrienne Cupples, Ben A Oostra, Jaap-Jan Zwaginga, Willem H Ouwehand, Swee-Lay Thein, Christa Meisinger, Panos Deloukas, Matthias

Nauck, Tim D Spector, Christian Gieger, Vilmundur Gudnason, Cornelia M van Duijn, Bruce M Psaty, Luigi Ferrucci, Aravinda Chakravarti, Andreas Greinacher, Christopher J O'Donnell, Jacqueline C M Witteman, Susan Furth, Mary Cushman, Tamara B Harris, and Jing-Ping Lin. Multiple loci influence erythrocyte phenotypes in the CHARGE consortium. *Nature Genetics*, 41(11):1191–1198, October 2009.

[67] S Gangemi, RA Merendino, F Guarneri, PL Minciullo, G DiLorenzo, M Pacor, and SP Cannavo. Serum levels of interleukin-18 and s-ICAM-1 in patients affected by psoriasis: preliminary considerations. *Journal of the European Academy of Dermatology and Venereology*, 17(1):42–46, January 2003.

[68] Thomas P Garner, Andrea Lopez, Denis E Reyna, Adam Z Spitz, and Evripidis Gavathiotis. Progress in targeting the BCL-2 family of proteins. *Current Opinion in Chemical Biology*, 39:133–142, August 2017.

[69] Daniel S. Gaul, Sokrates Stein, and Christian M. Matter. Neutrophils in cardiovascular disease. *European Heart Journal*, 38(22):1702–1704, June 2017.

[70] Claudia Giambartolomei, Damjan Vukcevic, Eric E. Schadt, Lude Franke, Aroon D. Hingorani, Chris Wallace, and Vincent Plagnol. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genetics*, 10(5):e1004383, May 2014.

[71] Christian Gieger, Aparna Radhakrishnan, Ana Cvejic, Weihong Tang, Eleonora Porcu, Giorgio Pistis, Jovana Serbanovic-Canic, Ulrich Elling, Alison H. Goodall, Yann Labrune, Lorna M. Lopez, Reedik Mägi, Stuart Meacham, Yukinori Okada, Nicola Pirastu, Rossella Sorice, Alexander Teumer, Katrin Voss, Weihua Zhang, Ramiro Ramirez-Solis, Joshua C. Bis, David Ellinghaus, Martin Gögele, Jouke-Jan Hottenga, Claudia Langenberg, Peter Kovacs, Paul F. O'Reilly, So-Youn Shin, Tõnu Esko, Jaana Hartiala, Stavroula Kanoni, Federico Murgia, Afshin Parsa, Jonathan Stephens, Pim van der Harst, C. Ellen van der Schoot, Hooman Allayee, Antony Attwood, Beverley Balkau, François Bastardot, Saonli Basu, Sebastian E. Baumeister, Ginevra Biino, Lorenzo Bomba, Amélie Bonnefond, François Cambien, John C. Chambers, Francesco Cucca, Pio D'Adamo, Gail Davies, Rudolf A. de Boer, Eco J. C. de Geus, Angela Döring, Paul Elliott, Jeanette Erdmann, David M. Evans, Mario Falchi, Wei Feng, Aaron R. Folsom, Ian H. Frazer, Quince D. Gibson, Nicole L. Glazer, Chris Hammond, Anna-Liisa Hartikainen, Susan R. Heckbert, Christian Hengstenberg, Micha Hersch, Thomas Illig, Ruth J. F. Loos, Jennifer Jolley, Kay-Tee Khaw, Brigitte Kühnel, Marie-Christine Kyrtsonis, Vasiliki Lagou, Heather Lloyd-Jones, Thomas Lumley, Massimo Mangino, Andrea Maschio, Irene Mateo Leach, Barbara McKnight, Yasin Memari, Braxton D. Mitchell, Grant W. Montgomery,

Yusuke Nakamura, Matthias Nauck, Gerjan Navis, Ute Nöthlings, Ilja M. Nolte, David J. Porteous, Anneli Pouta, Peter P. Pramstaller, Janne Pullat, Susan M. Ring, Jerome I. Rotter, Daniela Ruggiero, Aimo Ruokonen, Cinzia Sala, Nilesh J. Samani, Jennifer Sambrook, David Schlessinger, Stefan Schreiber, Heribert Schunkert, James Scott, Nicholas L. Smith, Harold Snieder, John M. Starr, Michael Stumvoll, Atsushi Takahashi, W. H. Wilson Tang, Kent Taylor, Albert Tenesa, Swee Lay Thein, Anke Tönjes, Manuela Uda, Sheila Ulivi, Dirk J. van Veldhuisen, Peter M. Visscher, Uwe Völker, H.-Erich Wichmann, Kerri L. Wiggins, Gonneke Willemsen, Tsun-Po Yang, Jing Hua Zhao, Paavo Zitting, John R. Bradley, George V. Dedoussis, Paolo Gasparini, Stanley L. Hazen, Andres Metspalu, Mario Pirastu, Alan R. Shuldiner, L. Joost van Pelt, Jaap-Jan Zwaginga, Dorret I. Boomsma, Ian J. Deary, Andre Franke, Philippe Froguel, Santhi K. Ganesh, Marjo-Riitta Jarvelin, Nicholas G. Martin, Christa Meisinger, Bruce M. Psaty, Timothy D. Spector, Nicholas J. Wareham, Jan-Willem N. Akkerman, Marina Ciullo, Panos Deloukas, Andreas Greinacher, Steve Jupe, Naoyuki Kamatani, Jyoti Khadake, Jaspal S. Kooner, Josef Penninger, Inga Prokopenko, Derek Stemple, Daniela Toniolo, Lorenz Wernisch, Serena Sanna, Andrew A. Hicks, Augusto Rendon, Manuel A. Ferreira, Willem H. Ouwehand, and Nicole Soranzo. New gene functions in megakaryopoiesis and platelet formation. *Nature*, 480(7376):201–208, November 2011.

[72] M GRAHAM. The coulter principle: foundation of an industry. *Journal of the Association for Laboratory Automation*, 8(6):72–81, December 2003.

[73] Ralph Green and Sebastian Wachsmann-Hogiu. Development, history, and future of automated cell counters. *Clinics in Laboratory Medicine*, 35(1):1–10, March 2015.

[74] Katri Haimila. Genetics of t cell co-stimulatory receptors -cd28, ctla4, icos and pdcd1 in immunity and transplantation. 08 2019.

[75] BD Hames. Molecular biology of homo sapiens. *Biochemical Education*, 16(1):51, January 1988.

[76] Trevor Hastie. *Generalized additive models.* Chapman and Hall, London New York, 1990.

[77] Trevor Hastie and Robert Tibshirani. *The elements of statistical learning : data mining, inference, and prediction.* Springer, New York, 2009.

[78] Douglas M. Hawkins. The problem of overfitting. *Journal of Chemical Information and Computer Sciences*, 44(1):1–12, January 2004.

[79] Balthasar A. Heesters, Cees E. van der Poel, Abhishek Das, and Michael C. Carroll. Antigen presentation to b cells. *Trends in Immunology*, 37(12):844–854, December 2016.

[80] Gibran Hemani, Jie Zheng, Benjamin Elsworth, Kaitlin H Wade, Valeriia Haberland, Denis Baird, Charles Laurin, Stephen Burgess, Jack Bowden, Ryan Langdon, Vanessa Y Tan, James Yarmolinsky, Hashem A Shihab, Nicholas J Timpson, David M Evans, Caroline Relton, Richard M Martin, George Davey Smith, Tom R Gaunt, and Philip C Haycock. The MR-base platform supports systematic causal inference across the human phenome. *eLife*, 7, May 2018.

[81] A. Victor Hoffbrand and Paul A. H. Moss. *Hoffbrand's Essential Haematology (Essentials)*. Wiley-Blackwell, 2015.

[82] Edward J Hollox and John AL Armour. Directional and balancing selection in human beta-defensins. *BMC Evolutionary Biology*, 8(1):113, 2008.

[83] Benjamin D. Horne, Jeffrey L. Anderson, Jerry M. John, Aaron Weaver, Tami L. Bair, Kurt R. Jensen, Dale G. Renlund, and Joseph B. Muhlestein. Which white blood cell subtypes predict increased cardiovascular risk? *Journal of the American College of Cardiology*, 45(10):1638–1643, May 2005.

[84] Bryan Howie, Christian Fuchsberger, Matthew Stephens, Jonathan Marchini, and Gonçalo R Abecasis. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nature Genetics*, 44(8):955–959, July 2012.

[85] Bryan Howie, Jonathan Marchini, and Matthew Stephens. Genotype imputation with thousands of genomes. *Genes Genomes and Genetics*, 1(6):457–470, nov 2011.

[86] Shiro Ikegawa. A short history of the genome-wide association study: Where we were and where we are going. *Genomics & Informatics*, 10(4):220, 2012.

[87] Miten Jain, Ian T Fiddes, Karen H Miga, Hugh E Olsen, Benedict Paten, and Mark Akeson. Improved data analysis for the MinION nanopore sequencer. *Nature Methods*, 12(4):351–356, February 2015.

[88] Amit K Dey Aditya A Joshi Raza Yunus Joseph B Lerman Tsion M Aberra Abhishek Chaturvedi Qimin Ng Joanna Silverman Tarek Aridi Martin P Playford Ramesh Mazhari Ju H Kim, Heather Teague and Nehal Mehta. A novel population of neutrophils, low density granulocytes, are upregulated in acute st segment elevation myocardial infarction. *Circulation Abstract*, 134(A14406), November 2016.

[89] Goo Jun, Matthew Flickinger, Kurt N. Hetrick, Jane M. Romm, Kimberly F. Doheny, Gonçalo R. Abecasis, Michael Boehnke, and Hyun Min Kang. Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. *The American Journal of Human Genetics*, 91(5):839–848, nov 2012.

[90] Steffen Jung. Macrophages and monocytes: of tortoises and hares. *Nature Reviews Immunology*, 18(2):85–86, January 2018.

[91] Ish Khanna. Drug discovery in pharmaceutical industry: productivity challenges and trends. *Drug Discovery Today*, 17(19-20):1088–1102, October 2012.

[92] Robert Kieschnick and B D McCullough. Regression analysis of variates observed on (0, 1): percentages, proportions and fractions. *Statistical Modelling: An International Journal*, 3(3):193–213, October 2003.

[93] Anna Kovalszki and Peter F. Weller. Eosinophilia. *Primary Care: Clinics in Office Practice*, 43(4):607–617, December 2016.

[94] Roman Kreuzhuber. *The effect of non-coding variants in gene transcription in human blood cell types.* PhD thesis, University of Cambridge, "2018".

[95] Ji Lee, Dae Cho, and Hyun Park. IL-18 and cutaneous inflammatory diseases. *International Journal of Molecular Sciences*, 16(12):29357–29369, December 2015.

[96] L.-F. Lee, R. Axtell, G. H. Tu, K. Logronio, J. Dilley, J. Yu, M. Rickert, B. Han, W. Evering, M. G. Walker, J. Shi, B. A. de Jong, J. Killestein, C. H. Polman, L. Steinman, and J. C. Lin. IL-7 promotes TH1 development and serum IL-7 predicts clinical response to interferon- in multiple sclerosis. *Science Translational Medicine*, 3(93):93ra68–93ra68, July 2011.

[97] Laurine Legroux and Nathalie Arbour. Multiple sclerosis and t lymphocytes: An entangled story. *Journal of Neuroimmune Pharmacology*, 10(4):528–546, May 2015.

[98] R. I. Lehrer, D. Szklarek, A. Barton, T. Ganz, K. J. Hamann, and G. J. Gleich. Antibacterial properties of eosinophil major basic protein and eosinophil cationic protein. *J. Immunol.*, 142(12):4428–4434, Jun 1989.

[99] Jennifer W. Leiding. Neutrophil evolution and their diseases in humans. *Frontiers in Immunology*, 8, August 2017.

[100] A. J. León, J. A. Garrote, A. Blanco-Quirós, C. Calvo, L. Fernández-Salazar, A. Del Villar, A. Barrera, and E. Arranz. Interleukin 18 maintains a long-standing inflammation in coeliac disease patients. *Clinical and Experimental Immunology*, 146(3):479–485, December 2006.

[101] Ofer Levy. Antimicrobial proteins and peptides of blood: templates for novel antimicrobial agents. *Blood*, 96(8):2664–2672, October 2000.

[102] Xin Lin, Ge Wei, Zhengzheng Shi, Laurence Dryer, Jeffrey D. Esko, Dan E. Wells, and Martin M. Matzuk. Disruption of gastrulation and heparan sulfate biosynthesis in EXT1-deficient mice. *Developmental Biology*, 224(2):299–311, August 2000.

[103] J. Linssen, S. Aderhold, A. Nierhaus, D. Frings, C. Kaltschmidt, and K. Zänker. Automation and validation of a rapid method to assess neutrophil and monocyte activation by routine fluorescence flow cytometry in vitro. *Cytometry Part B: Clinical Cytometry*, 74B(5):295–309, sep 2008.

[104] Pingyu Liu, Yijun Wang, and Xin Li. Targeting the untargetable KRAS in cancer therapy. *Acta Pharmaceutica Sinica B*, March 2019.

[105] Qianying Liu, Dan L. Nicolae, and Lin S. Chen. Marbled inflation from population structure in gene-based association studies with rare variants. *Genetic Epidemiology*, 37(3):286–292, March 2013.

[106] Po-Ru Loh, George Tucker, Brendan K Bulik-Sullivan, Bjarni J Vilhjálmsson, Hilary K Finucane, Rany M Salem, Daniel I Chasman, Paul M Ridker, Benjamin M Neale, Bonnie Berger, Nick Patterson, and Alkes L Price. Efficient bayesian mixed-model analysis increases association power in large cohorts. *Nature Genetics*, 47(3):284–290, feb 2015.

[107] Paul A. Lyons, Tim F. Rayner, Sapna Trivedi, Julia U. Holle, Richard A. Watts, David R.W. Jayne, Bo Baslund, Paul Brenchley, Annette Bruchfeld, Afzal N. Chaudhry, Jan Willem Cohen Tervaert, Panos Deloukas, Conleth Feighery, Wolfgang L. Gross, Loic Guillevin, Iva Gunnarsson, Lorraine Harper, Zdenka Hrušková, Mark A. Little, Davide Martorana, Thomas Neumann, Sophie Ohlsson, Sandosh Padmanabhan, Charles D. Pusey, Alan D. Salama, Jan-Stephan F. Sanders, Caroline O. Savage, Mårten Segelmark, Coen A. Stegeman, Vladimir Tesař, Augusto Vaglio, Stefan Wieczorek, Benjamin Wilde, Jochen Zwerina, Andrew J. Rees, David G. Clayton, and Kenneth G.C. Smith. Genetically distinct subsets within ANCA-associated vasculitis. *New England Journal of Medicine*, 367(3):214–223, July 2012.

[108] Donald W. MacGlashan, Susan Ishmael, Susan M. MacDonald, Jacqueline M. Langdon, Jonathan P. Arm, and David E. Sloane. Induced loss of syk in human basophils by non-IgE-dependent stimuli. *The Journal of Immunology*, 180(6):4208–4217, March 2008.

[109] Ani Manichaikul, Josyf C. Mychaleckyj, Stephen S. Rich, Kathy Daly, Michèle Sale, and Wei-Min Chen. Robust relationship inference in genome-wide association studies. *Bioinformatics*, 26(22):2867–2873, October 2010.

[110] Alice Louise Mann. *Using genetic and genomic approaches to understand haematopoietic cellular biology and dysregulation in disease.* PhD thesis, University of Cambridge, 11 2017.

[111] Bideau L. Dubreuil Y. Marcandier, M. Applications de la photometrie a la numeration des hemities. *CR Soc Biol Paris*, 99:741, August 1928.

[112] Jonathan Marchini and Bryan Howie. Genotype imputation for genome-wide association studies. *Nature Reviews Genetics*, 11(7):499–511, June 2010.

[113] Michael McHeyzer-Williams, Shinji Okitsu, Nathaniel Wang, and Louise McHeyzer-Williams. Molecular programming of b cell memory. *Nature Reviews Immunology*, 12(1):24–34, December 2011.

[114] William McLaren, Laurent Gil, Sarah E. Hunt, Harpreet Singh Riat, Graham R. S. Ritchie, Anja Thormann, Paul Flicek, and Fiona Cunningham. The ensembl variant effect predictor. *Genome Biology*, 17(1), June 2016.

[115] T Meade. Menopausal status and haemostatic variables. *The Lancet*, 321(83148315):2224, Jan 1983.

[116] Stephan Menzel, Chad Garner, Ivo Gut, Fumihiko Matsuda, Masao Yamaguchi, Simon Heath, Mario Foglio, Diana Zelenika, Anne Boland, Helen Rooks, Steve Best, Tim D Spector, Martin Farrall, Mark Lathrop, and Swee Lay Thein. A QTL influencing f cell production maps to a gene encoding a zinc-finger protein on chromosome 2p15. *Nature Genetics*, 39(10):1197–1199, September 2007.

[117] Alan J. Miller. The convergence of efroymson's stepwise regression algorithm. *The American Statistician*, 50(2):180–181, 1996.

[118] Mario Mitt, Mart Kals, Kalle Pärn, Stacey B Gabriel, Eric S Lander, Aarno Palotie, Samuli Ripatti, Andrew P Morris, Andres Metspalu, Tõnu Esko, Reedik Mägi, and Priit Palta. Improved imputation accuracy of rare and low-frequency variants using population-specific high-coverage WGS-based imputation reference panel. *European Journal of Human Genetics*, 25(7):869–876, April 2017.

[119] Lauren E Mokry, Omar Ahmad, Vincenzo Forgetta, George Thanassoulis, and J Brent Richards. Mendelian randomisation applied to drug development in cardiovascular disease: a review. *Journal of Medical Genetics*, 52(2):71–79, December 2014.

[120] A. Moldavan. Photo-electric Technique for the Counting of Microscopical Cells. *Science*, 80(2069):188–189, August 1934.

[121] Carmel Moore, Jennifer Sambrook, Matthew Walker, Zoe Tolkien, Stephen Kaptoge, David Allen, Susan Mehenny, Jonathan Mant, Emanuele Di Angelantonio, Simon G. Thompson, Willem Ouwehand, David J. Roberts, and John Danesh. The interval trial to determine whether intervals between blood donations can be safely and acceptably decreased to optimise blood supply: study protocol for a randomised controlled trial. *Trials*, 15(1):363, Sep 2014.

[122] Gary Moore, Gavin Knight, and Andrew Blann. *Haematology (Fundamentals of Biomedical Science, 2nd Edition).* Oxford University Press, 2016.

[123] R. A. Murav'ev, V. A. Fomina, and V. V. Rogovin. *Biology Bulletin of the Russian Academy of Sciences*, 30(4):317–321, 2003.

[124] Matthew R Nelson, Hannah Tipney, Jeffery L Painter, Judong Shen, Paola Nicoletti, Yufeng Shen, Aris Floratos, Pak Chung Sham, Mulin Jun Li, Junwen Wang, Lon R Cardon, John C Whittaker, and Philippe Sanseau. The support of human genetic evidence for approved drug indications. *Nature Genetics*, 47(8):856–860, June 2015.

[125] Jenny Nesje. *Impacts of Organic Matter Removal Efficiency on the Microbial Carrying Capacity and Stability of Land-Based Recirculating Aquaculture Systems.* PhD thesis, Norwegian University of Science and Technology, 02 2018.

[126] Jared O'Connell, Kevin Sharp, Nick Shrine, Louise Wain, Ian Hall, Martin Tobin, Jean-Francois Zagury, Olivier Delaneau, and Jonathan Marchini. Haplotype estimation for biobank-scale data sets. *Nature Genetics*, 48(7):817–820, June 2016.

[127] University of Utah. Genetic linkage. `https://learn.genetics.utah.edu/content/pigeons/geneticlinkage/images/linkage-3.jpg`.

[128] George Oliver. A contribution to the study of the blood and the circulation. *The Lancet*, 1:1699, 1896.

[129] Kouichi Ozaki, Yozo Ohnishi, Aritoshi Iida, Akihiko Sekine, Ryo Yamada, Tatsuhiko Tsunoda, Hiroshi Sato, Hideyuki Sato, Masatsugu Hori, Yusuke Nakamura, and Toshihiro Tanaka. Functional SNPs in the lymphotoxin-upalpha gene that are associated with susceptibility to myocardial infarction. *Nature Genetics*, 32(4):650–654, nov 2002.

[130] Sanjeev Palta, Richa Saroa, and Anshu Palta. Overview of the coagulation system. *Indian Journal of Anaesthesia*, 58(5):515, 2014.

[131] S. H. Park, C.-J. Park, B.-R. Lee, K.-S. Nam, M.-J. Kim, M.-Y. Han, Y. J. Kim, Y.-U. Cho, and S. Jang. Sepsis affects most routine and cell population data (CPD) obtained using the sysmex XN-2000 blood cell analyzer: neutrophil-related CPD NE-SFL and NE-WY provide useful information for detecting sepsis. *International Journal of Laboratory Hematology*, 37(2):190–198, may 2014.

[132] Hugh Parry, Sheldon Cohen, Janet E. Schlarb, David A. J. Tyrrel, Andrew Fisher, Michael A. H. Russell, and Martin J. Jarvis. Smoking, alcohol consumption, and leukocyte counts. *American Journal of Clinical Pathology*, 107(1):64–67, January 1997.

[133] Virginia Pascual. Inflammatory bowel disease and celiac disease: Overlaps and differences. *World Journal of Gastroenterology*, 20(17):4846, 2014.

[134] Steven M. Paul, Daniel S. Mytelka, Christopher T. Dunwiddie, Charles C. Persinger, Bernard H. Munos, Stacy R. Lindborg, and Aaron L. Schacht. How to improve r&d productivity: the pharmaceutical industry's grand challenge. *Nature Reviews Drug Discovery*, 9(3):203–214, February 2010.

[135] J. M. Paulus. Platelet size in man. *Blood*, 46(3):321–336, Sep "1975".

[136] M S Peters, M Rodriguez, and G J Gleich. Localization of human eosinophil granule major basic protein, eosinophil cationic protein, and eosinophil-derived neurotoxin by immunoelectron microscopy. *Lab. Invest.*, 54(6):656–662, June 1986.

[137] Joseph K Pickrell, Tomaz Berisa, Jimmy Z Liu, Laure Ségurel, Joyce Y Tung, and David A Hinds. Detection and interpretation of shared genetic influences on 42 human traits. *Nature Genetics*, 48(7):709–717, May 2016.

[138] Mikael J. Pittet, Matthias Nahrendorf, and Filip K. Swirski. The journey from stem cell to macrophage. *Annals of the New York Academy of Sciences*, 1319(1):1–18, March 2014.

[139] M. Pop. Genome assembly reborn: recent computational challenges. *Briefings in Bioinformatics*, 10(4):354–366, May 2009.

[140] A Rad. Simplified hematopoiesis, 2009. `https://commons.wikimedia.org/wiki/File:Hematopoiesis_simple.svg`.

[141] Richard Redon, Shumpei Ishikawa, Karen R. Fitch, Lars Feuk, George H. Perry, T. Daniel Andrews, Heike Fiegler, Michael H. Shapero, Andrew R. Carson, Wenwei Chen, Eun Kyung Cho, Stephanie Dallaire, Jennifer L. Freeman, Juan R. González,

Mònica Gratacòs, Jing Huang, Dimitrios Kalaitzopoulos, Daisuke Komura, Jeffrey R. MacDonald, Christian R. Marshall, Rui Mei, Lyndal Montgomery, Kunihiro Nishimura, Kohji Okamura, Fan Shen, Martin J. Somerville, Joelle Tchinda, Armand Valsesia, Cara Woodwark, Fengtang Yang, Junjun Zhang, Tatiana Zerjal, Jane Zhang, Lluis Armengol, Donald F. Conrad, Xavier Estivill, Chris Tyler-Smith, Nigel P. Carter, Hiroyuki Aburatani, Charles Lee, Keith W. Jones, Stephen W. Scherer, and Matthew E. Hurles. Global variation in copy number in the human genome. *Nature*, 444(7118):444–454, November 2006.

[142] Jessica M. B. Rees, Angela M. Wood, Frank Dudbridge, and Stephen Burgess. Robust methods in mendelian randomization via penalization of heterogeneous causal estimates. *PLOS ONE*, 14(9):e0222362, September 2019.

[143] Matthew R. Robinson, Naomi R. Wray, and Peter M. Visscher. Explaining additional genetic variation in complex traits. *Trends in Genetics*, 30(4):124–132, April 2014.

[144] Sara Rørvig, Ole Østergaard, Niels H. H. Heegaard, and Niels Borregaard. Proteome profiling of human neutrophil granule subsets, secretory vesicles, and cell membrane: correlation with transcriptome profiling of neutrophil precursors. *Journal of Leukocyte Biology*, 94(4):711–721, October 2013.

[145] G. Le Roux, A. Vlad, V. Eclache, C. Malanquin, J F Collon, M. Gantier, F. Schillinger, J Y Peltier, B Savin, R Letestu, F Baran-Marszak, P Fenaux, and F. Ajchenbaum-Cymbalista. Routine diagnostic procedures of myelodysplastic syndromes: value of a structural blood cell parameter (NEUT-x) determined by the sysmex XE-2100tm. *International Journal of Laboratory Hematology*, 32(6p1):e237–e243, oct 2010.

[146] Jitka Y. Sagiv, Janna Michaeli, Simaan Assi, Inbal Mishalian, Hen Kisos, Liran Levy, Pazzit Damti, Delphine Lumbroso, Lola Polyansky, Ronit V. Sionov, Amiram Ariel, Avi-Hai Hovav, Erik Henke, Zvi G. Fridlender, and Zvi Granot. Phenotypic diversity and plasticity in circulating neutrophil subpopulations in cancer. *Cell Reports*, 10(4):562–573, February 2015.

[147] F. Sanger, G. M. Air, B. G. Barrell, N. L. Brown, A. R. Coulson, J. C. Fiddes, C. A. Hutchison, P. M. Slocombe, and M. Smith. Nucleotide sequence of bacteriophage x174 DNA. *Nature*, 265(5596):687–695, February 1977.

[148] Vijay G. Sankaran and David G. Nathan. Reversing the hemoglobin switch. *New England Journal of Medicine*, 363(23):2258–2260, December 2010.

[149] Vijay G Sankaran and Stuart H Orkin. Genome-wide association studies of hematologic phenotypes: a window into human hematopoiesis. *Current Opinion in Genetics & Development*, 23(3):339–344, June 2013.

[150] Eva Särndahl, Ida Bergström, Veronika Patcha Brodin, Johnny Nijm, Helen Lundqvist Setterud, and Lena Jonasson. Neutrophil activation status in stable coronary artery disease. *PLoS ONE*, 2(10):e1056, October 2007.

[151] Mari KONO Atsushi WADA Sawako KAWAUCHI, Yuri TAKAGI and Takashi MORIKAWA. Comparison of the Leukocyte differentiation Scattergrams Between the XN-Series and the XE-Series of Hematology Analyzers. *Sysmex Journal International*, 24(1), 2014.

[152] Stephen Sawcer, Garrett Hellenthal, Matti Pirinen, Chris C. A. Spencer, Nikolaos A. Patsopoulos, Loukas Moutsianas, Alexander Dilthey, Zhan Su, Colin Freeman, Sarah E. Hunt, Sarah Edkins, Emma Gray, David R. Booth, Simon C. Potter, An Goris, Gavin Band, Annette Bang Oturai, Amy Strange, Janna Saarela, Céline Bellenguez, Bertrand Fontaine, Matthew Gillman, Bernhard Hemmer, Rhian Gwilliam, Frauke Zipp, Alagurevathi Jayakumar, Roland Martin, Stephen Leslie, Stanley Hawkins, Eleni Giannoulatou, Sandra D'alfonso, Hannah Blackburn, Filippo Martinelli Boneschi, Jennifer Liddle, Hanne F. Harbo, Marc L. Perez, Anne Spurkland, Matthew J. Waller, Marcin P. Mycko, Michelle Ricketts, Manuel Comabella, Naomi Hammond, Ingrid Kockum, Owen T. McCann, Maria Ban, Pamela Whittaker, Anu Kemppinen, Paul Weston, Clive Hawkins, Sara Widaa, John Zajicek, Serge Dronov, Neil Robertson, Suzannah J. Bumpstead, Lisa F. Barcellos, Rathi Ravindrarajah, Roby Abraham, Lars Alfredsson, Kristin Ardlie, Cristin Aubin, Amie Baker, Katharine Baker, Sergio E. Baranzini, Laura Bergamaschi, Roberto Bergamaschi, Allan Bernstein, Achim Berthele, Mike Boggild, Jonathan P. Bradfield, David Brassat, Simon A. Broadley, Dorothea Buck, Helmut Butzkueven, Ruggero Capra, William M. Carroll, Paola Cavalla, Elisabeth G. Celius, Sabine Cepok, Rosetta Chiavacci, Françoise Clerget-Darpoux, Katleen Clysters, Giancarlo Comi, Mark Cossburn, Isabelle Cournu-Rebeix, Mathew B. Cox, Wendy Cozen, Bruce A. C. Cree, Anne H. Cross, Daniele Cusi, Mark J. Daly, Emma Davis, Paul I. W. de Bakker, Marc Debouverie, Marie Beatrice D'hooghe, Katherine Dixon, Rita Dobosi, Bénédicte Dubois, David Ellinghaus, Irina Elovaara, Federica Esposito, Claire Fontenille, Simon Foote, Andre Franke, Daniela Galimberti, Angelo Ghezzi, Joseph Glessner, Refujia Gomez, Olivier Gout, Colin Graham, Struan F. A. Grant, Franca Rosa Guerini, Hakon Hakonarson, Per Hall, Anders Hamsten, Hans-Peter Hartung, Rob N. Heard, Simon Heath, Jeremy Hobart, Muna Hoshi, Carmen Infante-Duarte, Gillian Ingram, Wendy Ingram, Talat Islam, Maja Jagodic, Michael Kabesch, Allan G. Kermode, Trevor J. Kilpatrick, Cecilia Kim, Norman Klopp, Keijo Koivisto, Malin Larsson, Mark Lathrop, Jeannette S. Lechner-Scott, Maurizio A. Leone, Virpi Leppä, Ulrika Liljedahl, Izaura Lima Bomfim, Robin R. Lincoln, Jenny Link, Jianjun Liu,

Åslaug R. Lorentzen, Sara Lupoli, Fabio Macciardi, Thomas Mack, Mark Marriott, Vittorio Martinelli, Deborah Mason, Jacob L. McCauley, Frank Mentch, Inger-Lise Mero, Tania Mihalova, Xavier Montalban, John Mottershead, Kjell-Morten Myhr, Paola Naldi, William Ollier, Alison Page, Aarno Palotie, Jean Pelletier, Laura Piccio, Trevor Pickersgill, Fredrik Piehl, Susan Pobywajlo, Hong L. Quach, Patricia P. Ramsay, Mauri Reunanen, Richard Reynolds, John D. Rioux, Mariaemma Rodegher, Sabine Roesner, Justin P. Rubio, Ina-Maria Rückert, Marco Salvetti, Erika Salvi, Adam Santaniello, Catherine A. Schaefer, Stefan Schreiber, Christian Schulze, Rodney J. Scott, Finn Sellebjerg, Krzysztof W. Selmaj, David Sexton, Ling Shen, Brigid Simms-Acuna, Sheila Skidmore, Patrick M. A. Sleiman, Cathrine Smestad, Per Soelberg Sørensen, Helle Bach Søndergaard, Jim Stankovich, Richard C. Strange, Anna-Maija Sulonen, Emilie Sundqvist, Ann-Christine Syvänen, Francesca Taddeo, Bruce Taylor, Jenefer M. Blackwell, Pentti Tienari, Elvira Bramon, Ayman Tourbah, Matthew A. Brown, Ewa Tronczynska, Juan P. Casas, Niall Tubridy, Aiden Corvin, Jane Vickery, Janusz Jankowski, Pablo Villoslada, Hugh S. Markus, Kai Wang, Christopher G. Mathew, James Wason, Colin N. A. Palmer, H-Erich Wichmann, Robert Plomin, Ernest Willoughby, Anna Rautanen, Juliane Winkelmann, Michael Wittig, Richard C. Trembath, Jacqueline Yaouanq, Ananth C. Viswanathan, Haitao Zhang, Nicholas W. Wood, Rebecca Zuvich, Panos Deloukas, Cordelia Langford, Audrey Duncanson, Jorge R. Oksenberg, Margaret A. Pericak-Vance, Jonathan L. Haines, Tomas Olsson, Jan Hillert, Adrian J. Ivinson, Philip L. De Jager, Leena Peltonen, Graeme J. Stewart, David A. Hafler, Stephen L. Hauser, Gil McVean, Peter Donnelly, and Alastair Compston. Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. *Nature*, 476(7359):214–219, August 2011.

[153] J. L. Schering, J. Munoz, E. Raybon, S. Hegab, A. S. Hanbali, and P. Kuriakose. The diagnostic and prognostic implications of nucleated red blood cells in myelophthisis. *Journal of Clinical Oncology*, 29(15_suppl):6576–6576, May 2011.

[154] Cosma Rohilla Shalizi. *Advanced Data Analysis from an Elementary Point of View*. Cambirdge University Press, 2019.

[155] Pak C. Sham and Shaun M. Purcell. Statistical power and significance testing in large-scale genetic studies. *Nature Reviews Genetics*, 15(5):335–346, may 2014.

[156] Matt Shirley. Daclizumab: A review in relapsing multiple sclerosis. *Drugs*, 77(4):447–458, February 2017.

[157] Katherine M. Siewert and Benjamin F. Voight. Detecting long-term balancing selection using allele frequency correlation. *Molecular Biology and Evolution*, 34(11):2996–3005, July 2017.

[158] Robert L. Sinsheimer. The santa cruz workshop—may 1985. *Genomics*, 5(4):954–956, November 1989.

[159] Katarzyna Smietana, Marcin Siatkowski, and Martin Møller. Trends in clinical success rates. *Nature Reviews Drug Discovery*, 15(6):379–380, May 2016.

[160] Megan R. Smith, Ann-Louise Kinmonth, Robert N. Luben, Sheila Bingham, Nicholas E. Day, Nicholas J. Wareham, Ailsa Welch, and Kay-Tee Khaw. Smoking status and differential white cell count in men and women in the EPIC-norfolk population. *Atherosclerosis*, 169(2):331–337, August 2003.

[161] C L Sokol and R Medzhitov. Emerging functions of basophils in protective and allergic immune responses. *Mucosal Immunology*, 3(2):129–137, January 2010.

[162] Nicole Soranzo, Tim D Spector, Massimo Mangino, Brigitte Kühnel, Augusto Rendon, Alexander Teumer, Christina Willenborg, Benjamin Wright, Li Chen, Mingyao Li, Perttu Salo, Benjamin F Voight, Philippa Burns, Roman A Laskowski, Yali Xue, Stephan Menzel, David Altshuler, John R Bradley, Suzannah Bumpstead, Mary-Susan Burnett, Joseph Devaney, Angela Döring, Roberto Elosua, Stephen E Epstein, Wendy Erber, Mario Falchi, Stephen F Garner, Mohammed J R Ghori, Alison H Goodall, Rhian Gwilliam, Hakon H Hakonarson, Alistair S Hall, Naomi Hammond, Christian Hengstenberg, Thomas Illig, Inke R König, Christopher W Knouff, Ruth McPherson, Olle Melander, Vincent Mooser, Matthias Nauck, Markku S Nieminen, Christopher J O'Donnell, Leena Peltonen, Simon C Potter, Holger Prokisch, Daniel J Rader, Catherine M Rice, Robert Roberts, Veikko Salomaa, Jennifer Sambrook, Stefan Schreiber, Heribert Schunkert, Stephen M Schwartz, Jovana Serbanovic-Canic, Juha Sinisalo, David S Siscovick, Klaus Stark, Ida Surakka, Jonathan Stephens, John R Thompson, Uwe Völker, Henry Völzke, Nicholas A Watkins, George A Wells, H-Erich Wichmann, David A Van Heel, Chris Tyler-Smith, Swee Lay Thein, Sekar Kathiresan, Markus Perola, Muredach P Reilly, Alexandre F R Stewart, Jeanette Erdmann, Nilesh J Samani, Christa Meisinger, Andreas Greinacher, Panos Deloukas, Willem H Ouwehand, and Christian Gieger. A genome-wide meta-analysis identifies 22 loci associated with eight hematological parameters in the HaemGen consortium. *Nature Genetics*, 41(11):1182–1190, October 2009.

[163] Timothy K. Starr, Stephen C. Jameson, and Kristin A. Hogquist. Positive and Negative Selection of T-Cells, journal = Annual Review of Immunology. 21(1):139–176, April 2003.

[164] Benjamin B. Sun, Joseph C. Maranville, James E. Peters, David Stacey, James R. Staley, James Blackshaw, Stephen Burgess, Tao Jiang, Ellie Paige, Praveen Surendran, Clare Oliver-Williams, Mihir A. Kamat, Bram P. Prins, Sheri K. Wilcox,

Erik S. Zimmerman, An Chi, Narinder Bansal, Sarah L. Spain, Angela M. Wood, Nicholas W. Morrell, John R. Bradley, Nebojsa Janjic, David J. Roberts, Willem H. Ouwehand, John A. Todd, Nicole Soranzo, Karsten Suhre, Dirk S. Paul, Caroline S. Fox, Robert M. Plenge, John Danesh, Heiko Runz, and Adam S. Butterworth. Genomic atlas of the human plasma proteome. *Nature*, 558(7708):73–79, June 2018.

[165] Sysmex Europe. *Measurement Technology and Scattergram*, 1 2014. `https://www.sysmex-europ,e.com/academy/knowledge-centre/calendar-2014/measurement-technology-and-scattergram.html`.

[166] Johan Tas and Liesbeth H. M. Geenen. Microspectrophotometric detection of heparin in mast cells and basophilic granulocytes stained metachromatically with toluidine blue o. *The Histochemical Journal*, 7(3):231248, May 1975.

[167] the Haplotype Reference Consortium. A reference panel of 64, 976 haplotypes for genotype imputation. *Nature Genetics*, 48(10):1279–1283, August 2016.

[168] Louise W. Treffers, Ida H. Hiemstra, Taco W. Kuijpers, Timo K. van den Berg, and Hanke L. Matlung. Neutrophils in cancer. *Immunological Reviews*, 273(1):312–328, August 2016.

[169] M. Uda, R. Galanello, S. Sanna, G. Lettre, V. G. Sankaran, W. Chen, G. Usala, F. Busonero, A. Maschio, G. Albai, M. G. Piras, N. Sestu, S. Lai, M. Dei, A. Mulas, L. Crisponi, S. Naitza, I. Asunis, M. Deiana, R. Nagaraja, L. Perseu, S. Satta, M. D. Cipollina, C. Sollaino, P. Moi, J. N. Hirschhorn, S. H. Orkin, G. R. Abecasis, D. Schlessinger, and A. Cao. Genome-wide association study shows BCL11a associated with persistent fetal hemoglobin and amelioration of the phenotype of -thalassemia. *Proceedings of the National Academy of Sciences*, 105(5):1620–1625, February 2008.

[170] L. J. Vasquez, A. L. Mann, L. Chen, and N. Soranzo. From GWAS to function: lessons from blood cells. *ISBT Science Series*, 11(S1):211–219, October 2015.

[171] G. L. Verdine and L. D. Walensky. The challenge of drugging undruggable targets in cancer: Lessons learned from targeting BCL-2 family members. *Clinical Cancer Research*, 13(24):7264–7270, December 2007.

[172] Flavio Vincenti, Robert Kirkman, Susan Light, Ginny Bumgardner, Mark Pescovitz, Philip Halloran, John Neylan, Alan Wilkinson, Henrik Ekberg, Robert Gaston, Lars Backman, and James Burdick. Interleukin-2–receptor blockade with daclizumab to prevent acute rejection in renal transplantation. *New England Journal of Medicine*, 338(3):161–165, January 1998.

[173] Peter M. Visscher, Naomi R. Wray, Qian Zhang, Pamela Sklar, Mark I. McCarthy, Matthew A. Brown, and Jian Yang. 10 years of GWAS discovery: Biology, function, and translation. *The American Journal of Human Genetics*, 101(1):5–22, July 2017.

[174] Jon Wakefield. Bayes factors for genome-wide association studies: comparison withP-values. *Genetic Epidemiology*, 33(1):79–86, January 2009.

[175] Samantha Welsh, Tim Peakman, Simon Sheard, and Rachael Almond. Comparison of DNA quantification methodology used in the DNA extraction protocol for the UK biobank cohort. *BMC Genomics*, 18(1), January 2017.

[176] Julie F. Westerlund and Daniel J. Fairbanks. Gregor mendel's classic paper and the nature of science in genetics courses. *Hereditas*, 147(6):293–303, November 2010.

[177] Wolfgang Goehde Wolfgang Dittrich. Flow-through chamber for photometers to measure and count particles in a dispersion medium, 12 1968.

[178] Helen L. Wright, Robert J. Moots, and Steven W. Edwards. The multifactorial role of neutrophils in rheumatoid arthritis. *Nature Reviews Rheumatology*, 10(10):593–601, June 2014.

[179] B. Wuthrich and P. Schmid-Grendelmeier. The atopic eczema/dermatitis syndrome. Epidemiology, natural course, and immunology of the IgE-associated (”extrinsic”) and the nonallergic (”intrinsic”) AEDS. *J Investig Allergol Clin Immunol*, 13(1):1–5, 2003.

[180] ChangJiang Xu, Ioanna Tachmazidou, Klaudia Walter, Antonio Ciampi, Eleftheria Zeggini, and Celia M. T. Greenwood and. Estimating genome-wide significance for whole-genome sequencing studies. *Genetic Epidemiology*, 38(4):281–290, feb 2014.

[181] Jian Yang, Teresa Ferreira, Andrew P Morris, Sarah E Medland, Pamela A F Madden, Andrew C Heath, Nicholas G Martin, Grant W Montgomery, Michael N Weedon, Ruth J Loos, Timothy M Frayling, Mark I McCarthy, Joel N Hirschhorn, Michael E Goddard, and Peter M Visscher and. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nature Genetics*, 44(4):369–375, March 2012.

[182] Jian Yang, Noah A Zaitlen, Michael E Goddard, Peter M Visscher, and Alkes L Price. Advantages and pitfalls in the application of mixed-model association methods. *Nature Genetics*, 46(2):100–106, feb 2014.

[183] Chloe X Yap, Luke Lloyd-Jones, Alexander Holloway, Peter Smartt, Naomi R Wray, Jacob Gratten, and Joseph E Powell. Trans-eQTLs identified in whole blood have

limited influence on complex disease biology. *European Journal of Human Genetics*, 26(9):1361–1368, June 2018.

[184] Olena Yavorska and James Staley. *MendelianRandomization: Mendelian Randomization Package*, 2019. R package version 0.4.1.

[185] Alexander Zarbock and Klaus Ley. Mechanisms and consequences of neutrophil interaction with the endothelium. *The American Journal of Pathology*, 172(1):1–7, January 2008.

[186] Hou-Feng Zheng, Jing-Jing Rong, Ming Liu, Fang Han, Xing-Wei Zhang, J. Brent Richards, and Li Wang. Performance of genotype imputation for low frequency and rare variants from the 1000 genomes. *PLOS ONE*, 10(1):e0116487, January 2015.

[187] L. Ziegler-Heitbrock, P. Ancuta, S. Crowe, M. Dalod, V. Grau, D. N. Hart, P. J. M. Leenen, Y.-J. Liu, G. MacPherson, G. J. Randolph, J. Scherberich, J. Schmitz, K. Shortman, S. Sozzani, H. Strobl, M. Zembala, J. M. Austyn, and M. B. Lutz. Nomenclature of monocytes and dendritic cells in blood. *Blood*, 116(16):e74–e80, July 2010.

[188] Mathias Zimmermann, Malte Cremer, Christina Hoffmann, Karin Weimann, and Andreas Weimann. Granularity index of the SYSMEX XE-5000 hematology analyzer as a replacement for manual microscopy of toxic granulation neutrophils in patients with inflammatory diseases. *Clinical Chemistry and Laboratory Medicine*, 49(7), jan 2011.

Appendix A

Appendix

A.1 Table of Sysmex parameter conditionally significant variants and eQTL, pQTL, disease colocalisations

https://figshare.com/s/c8775dc6d85be9b3afa4

A table of 2,172 conditionally independent variant-trait associations identified from GWAS of the 63 cytometry parameters listed in Table S1. All coordinates are with respect to GRCh37. Signal ID corresponds to a unique identifier for each signal of association defined by an LD clumping procedure ($r^2 > 0.8$). Each variant is given marginal (univariable) summary statistics for association with the corresponding trait and summary statistics for joint association (MULTI) in a model including all other conditionally independent variants. Fine-mapping of each locus allows assignment of variants to credible sets indicated by the FINEMAP Credible Set ID column, the total number of variants in each credible set are also indicated. Further columns include posterior probabilities for colocalisation with pQTL, eQTL, and disease association signals.

A.2 Table of conditionally significant associations with FBC phenotypes from the UK Biobank cohort

https://figshare.com/s/0fe1d830cab86dbe095d

A table containing information regarding each of the 17,042 associations ordered by chromosome and position (all coordinates are with respect to GRCh37). Locus ID is a unique identifier for each locus. The column Novel vs Astle et al. 2016 indicates if the variant or any variants in LD $r^2 > 0.8$ was already found to be associated with the same traits in the cited previously published meta-analysis including the first release of UK Biobank. The unique variant ID is constructed from the chromosome, position and the reference and alternative alleles according to the human genome reference (build 37 coordinates). Where available, the rsID is also given. GWAS summary statistics for univariate and multivariate (conditional) model are provided, as well as the VEP worst consequence annotation.

A.3 Mendelian randomisation reports interactive HTML format

https://figshare.com/s/207ae098eb2db3172676

A compressed folder containing HTML reports for the MR analyses, the reports can be accessed by opening the link and pressing the 'Download' button, the file must then be decompressed resulting in a folder titled 'mendelian_randomisation_html'. The reports can then be visualised by opening the 'index.html' file in the 'mendelian_randomisation_html' folder using any web-browser.

A.4 Disease Colocalisation Locuszoom Plots

https://figshare.com/s/9f3e9165300d6468db97

A series of plots showing the regions of colocalisation between Sysmex parameters and disease outcomes. The x axis represents genomic location, the y axis negative log transformed P-value of association, and each data-point is the significance of association.

A.5 pQTL Colocalisation Locuszoom Plots

https://figshare.com/s/ce075d0f52aff56c67a8

A series of plots showing the regions of colocalisation between Sysmex parameters and blood plasma protein QTL. The x axis represents genomic location, the y axis negative log transformed P-value of association, and each data-point is the significance of association.

A.6 eQTL Colocalisation Locuszoom Plots

https://figshare.com/s/103308605dee49d10a3c

A series of plots showing the regions of colocalisation between Sysmex parameters and blood cell transcript QTL. The x axis represents genomic location, the y axis negative log transformed P-value of association, and each data-point is the significance of association.

A.7 Table of Sysmex parameters

https://figshare.com/s/c3ecdcfcad7a39dcba3d

Each row corresponds to one of the 63 cytometry traits studied in this analysis including columns indicating the most correlated standard FBC hematological measurement and the number of new loci discovered per cytometry parameter compared to GWAS of standard FBC parameters.