Using genetic data to determine the effect of routinely measured blood cell traits on disease

Andrei-Emil Constantinescu

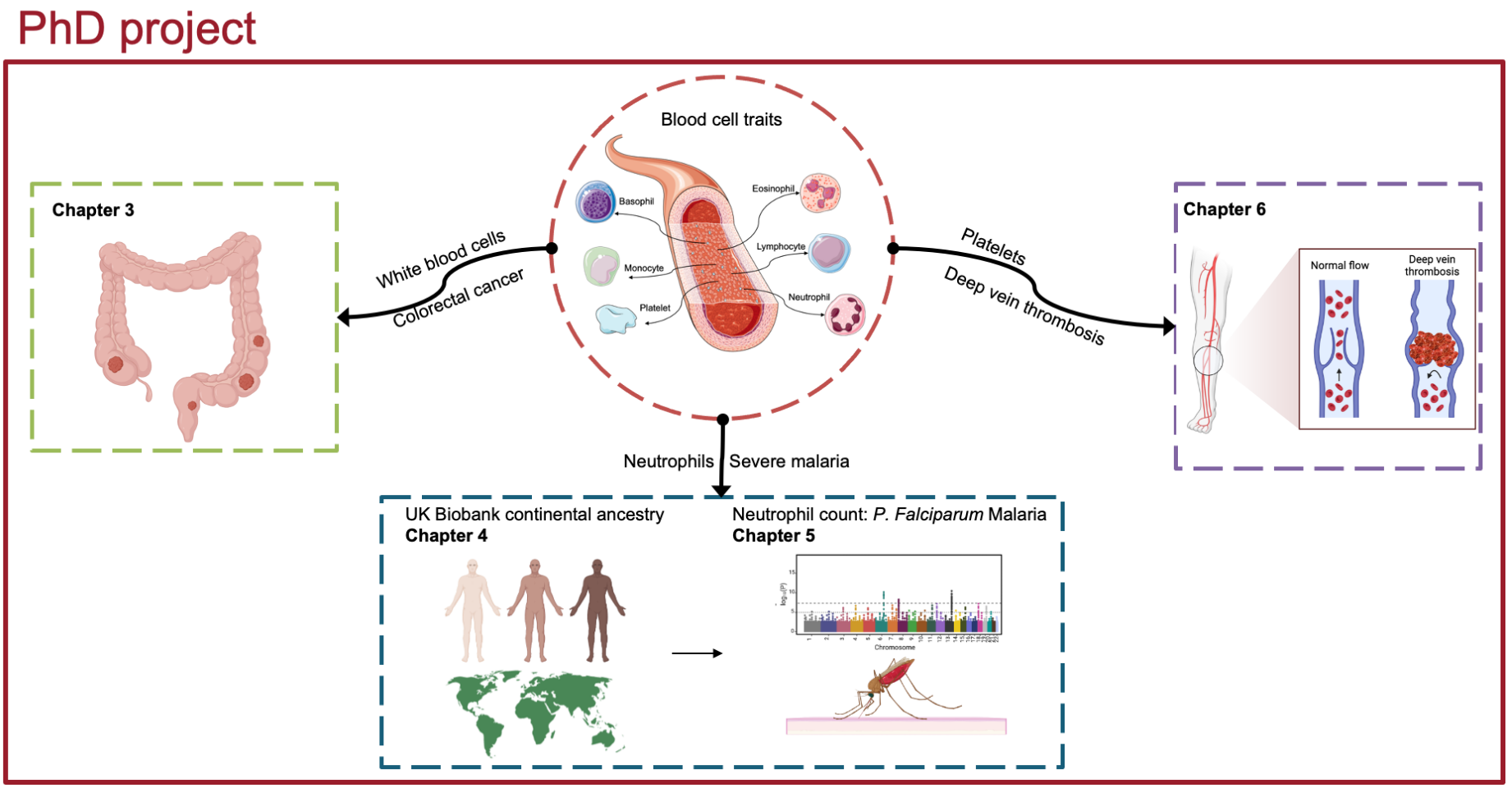
A dissertation submitted to the University of Bristol in accordance with the requirements for award of the degree of
Doctor of Philosophy
in the Faculty of Health Sciences
Bristol Medical School
Apr 2023

Word count: 55,172

ABSTRACT

Blood cell traits (BCTs), including white blood cells (WBCs) and platelets, are commonly measured in a routine blood test or hospital visit. This is because there is a well-established relationship between BCTs and diseases that lead to death and disability. Most studies on diseases associated with BCTs have been observational, and therefore generally prone to confounding and reverse causation. Given the health burden of diseases observationally linked to BCTs, it is desirable to determine whether these relationships are causal. Mendelian randomization (MR) is a method in genetic epidemiology which uses people's genetic data to provide a causal estimate between an exposure and an outcome. Therefore, the overarching aim of my thesis was to use MR to advance the knowledge on diseases associated with BCTs. To investigate this, I focused on three diseases, each having their own methodological challenges: **Chapter 3** – colorectal cancer (CRC); **Chapter 4** & **Chapter 5** – *P. falciparum* malaria; **Chapter 6** – deep vein thrombosis (DVT). In **Chapter 3** I provided evidence that a higher eosinophil and lymphocyte count reduced the risk of CRC, and a follow-up MR analysis revealed a possible protective role for allergic disease in CRC development. In **Chapter 4** I identified a subset of UK Biobank participants that correspond to the African continental ancestry group, allowing me to conduct a genome-wide association study of neutrophil count to *P. falciparum* malaria in **Chapter 5**. Here, the MR analysis showed limited evidence for a causal relationship between neutrophil count and severe malaria. Finally, in **Chapter 6** I conducted a phenome wide MR study to identify novel risk factors for DVT, and a follow-up analysis identified that a protein predominantly present in platelets, plasminogen activation inhibitor 1 (PAI-1), mediates the relationship between adiposity and DVT risk.

VISUAL ABSTRACT



DEDICATION AND ACKNOWLEDGEMENTS

Per aspera ad astra

I never thought I would be here. Each act of selflessness by those with whom I have interacted throughout my time here on this “pale blue dot” has made me the person I am today – for this I am forever grateful.

Special thanks go to my supervisors Caroline, Nic and Emma. Caroline, you’ve guided me through my PhD journey and made me a better researcher. Nic, you’ve been incredibly magnanimous and a great mentor. Emma, you’ve done so much that not only am I unable to put it into words here; if I could, it would take at the minimum an entire book to cover it all.

David, thank you for teaching me so much and I wish you the best going forward. Your positive attitude has inspired me. I have been very lucky to be part of the Integrative Epidemiology Unit; everyone here has been friendly and always eager to help. Bristol has been a great city to live in and, after almost seven years, I am happy to call it my home.

Mamaie, you taught me from a young age to be kind and respectful. Mum, you’ve always believed in me no matter what; without you, I don’t know where I would be now. Same goes to my dad. Andreea, thank you for your love and kindness, and for being by my side while I was going through the toughest period of my PhD.

This work was supported in part by grant MR/N0137941/1 for the GW4 BIOMED MRC DTP, awarded to the Universities of Bath, Bristol, Cardiff and Exeter from the Medical Research Council (MRC)/UKRI. I would like to thank the participants who contributed to the data gathered as part of the consortia used in this thesis, such as UK Biobank, the Avon Longitudinal Study of Parents and Children (ALSPAC), the Genetics and Epidemiology of Colorectal Cancer Consortium (GECCO) and more.

AUTHOR'S DECLARATION

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's Regulations and Code of Practice for Research Degree Programmes and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate's own work. Work done in collaboration with, or with the assistance of, others, is indicated as such. Any views expressed in the dissertation are those of the author.

SIGNED: .....  ..... DATE: .....27/04/2023.....

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

1KG: 1000 Genomes
1SMR: One-sample Mendelian randomization
2SMR: Two-sample Mendelian randomization
ACKR1/DARC: Atypical chemokine receptor 1 / Duffy antigen receptor for chemokines
ACOT12: Acyl-CoA Thioesterase 12
ACRC: Advanced Computing Research Centre
AFR: African
ALSPAC: Avon Longitudinal Study of Parents and Children
AMR: Americas
ANOVA: Analysis of variance
APC: Antigen-presenting cell
BBJ: BioBank Japan
BCT: Blood cell trait
BEB: Bengali in Bangladesh
BEN: Benign ethnic neutropenia
BMI: Body mass index
BMR: Basal metabolic rate
C/EBP-α: CCAAT/enhancer-binding protein alpha
C12orf54: Chromosome 12 Open Reading Frame 54
CA: Correspondence analysis
CAG: Continental ancestry group
CBC: Complete blood count
CDX: Chinese Dai in Xishuangbanna, China
CEU: Utah residents with Northern and Western European ancestry
CHB: Han Chinese in Beijing, China
CHD1L: Chromodomain helicase DNA binding protein 1 like
CHS: Han Chinese South
CI: Confidence interval
CLP: Common lymphoid progenitor
CM: Cerebral malaria
CMP: Common myeloid progenitor
COB: Country of birth
COPD: Chronic obstructive pulmonary disease
CRC: Colorectal cancer
DARS/CXCR4: Aspartyl-TRNA Synthetase 1/C-X-C Motif Chemokine Receptor 4
DC: Dendritic cell
DG: Dense granule
DLL-4: Delta-like ligand 4
DOAC: Direct oral anticoagulants
DVT: Deep vein thrombosis
EAS: East-Asian
ECP: Eosinophilic cationic protein
EDN: Eosinophil-derived neurotoxin
EET: Eosinophil extracellular trap
EPX: Eosinophil peroxidase
eQTL: Expression quantitative trait locus
ESN: Esan in Nigeria
EUR: European
FAP: Familial adenomatous polyposis
FBC: Full blood count
FIN: Finnish in Finland
FSH: Follicle-stimulating hormone

Fst: Fixation index
Galectin-10: Charcot-Leyden crystals
GBR: British in England and Scotland
GC: Genomic control
GIH: Gujarati Indian in Houston, Texas
GM-CSF: Granulocyte-macrophage colony stimulating factor
GMP: Granulocyte-macrophage precursor
GRM: Genetic relationship matrix
GTEx: Genotype-Tissue Expression
GWAS: Genome-wide association study
GWD: Gambian in Western Division, The Gambia - Mandinka
H3Africa: Human Heredity and Health in Africa
Hbs: Haemoglobin S
HCT: Haematocrit
HSC: Haematopoietic stem cell
HWE: Hardy-Weinberg equilibrium
IBS: Iberian populations in Spain
ICD-10: International Classification of Disease
IFN-γ: Interferon-γ
IgA: Immunoglobulin A
IL-3: Interleukin 3
IL5R: IL-5 receptor
INHBC: Inhibin beta C chain
inSIDE: Instrument strength independent of direct effect
IRF8: Interferon regulatory factor-8
ITU: Indian Telugu in the UK
IVW: Inverse-variance weighted
JPT: Japanese in Tokyo, Japan
K-pop / K-cluster / K: K-means cluster
Kb: Kilobases
KHV: Kinh in Ho Chi Minh City, Vietnam
KLF4: Krüppel-like factor 4
LD: Linkage disequilibrium
LMM: Linear mixed model
LWK: Luhya in Webuye, Kenya
MAC: Minor allele count
MAF: Minor allele frequency
MBP: Major basic protein
MCH: Mean corpuscular haemoglobin
MCHC: Mean corpuscular haemoglobin concentration
MCV: Mean corpuscular volume
MHC: Major histocompatibility complex
MIF: Macrophage migration inhibitory factor
MPV: Mean platelet volume
MR-PheWAS: Mendelian randomization phenome-wide association study
MR-PRESSO: Mendelian randomization pleiotropy residual sum and outlier
MR: Mendelian randomization
MSL: Mende in Sierra Leone
MVMR: Multivariable MR
NET: Neutrophil extracellular trap
NF-E2: Nuclear factor-erythroid 2
NHS: National Health Service
NK: Natural killer
NLR: Neutrophil-to-lymphocyte ratio
NOTCH1: Neurogenic locus notch homolog protein 1
OFH: OurFutureHealth

OR: Odds ratio
OTHER: Other severe malaria
PAGE: Population Architecture using Genomics and Epidemiology
PAI-1: Plasminogen activator inhibitor-1
PC: Principal components
PCA: Principal component analysis
PE: Pulmonary embolism
PHESANT: Phenome Scan Analysis Tool
PJL: Punjabi in Lahore, Pakistan
PLT: Platelet count
pQTL: Protein quantitative trait locus
PRS: Polygenic risk score
RBC: Red blood cell
RCT: Randomised controlled trial
ROB: Region of birth
ROS: Reactive oxygen species
RR: Risk ratio
SAS: South-Asian
SD: Standard deviation
SE: Standard error
SEA: South-east Asia
SH3GL2: SH3 Domain Containing GRB2 Like 2, Endophilin A1
SLC22A2: Solute Carrier Family 22 Member 2
SM: Severe malaria
SMA: Severa malaria anaemia
SNP: Single-nucleotide polymorphisms
SOL: Hispanic Community Health Study / Study of Latinos
sQTL: Splicing quantitative-trait locus
STU: Sri Lankan Tamil in the UK
T2D: Type 2 diabetes
TAM: Tumour-associated macrophage
TAN: Tumour-associated neutrophils
TCR: T-cell receptor
TF: Transcription factor
TGF-β: Transforming growth factor beta
Th: T-helper cell
TH: Thyroid hormone
TIL: Tumour-infiltrating lymphocyte
TLR4: Toll-like receptor 4
TME: Tumour microenvironment
Treg: T-regulatory cell
TREM-1: Triggering receptor expressed on myeloid cells-1
TRIB1: Protein kinase Tribbles homolog 1
TRPS1: Transcriptional Repressor GATA Binding 1
TSI: Toscani in Italia
TWAS: Transcriptome-wide association study
TYRP1: Tyrosinase-related protein 1
UKBB: UK Biobank
UVMR: Univariable MR
VEP: Variant effect predictor
VTE: Venous thromboembolism
vWF: von Willebrand factor
WBC: White blood cell
WGS: Whole genome sequencing
WHR: Waist-to-hip ratio
WR: Wald ratio

YRI: Yoruba in Ibadan, Nigeria

PUBLICATIONS ARISING FROM THIS WORK

Below is a list of publications that have arisen from research undertaken throughout this PhD along with the contributions section from each publication.

The following publication is based on Chapter 3

Circulating white blood cell traits and colorectal cancer risk: A Mendelian randomization study. Andrei-Emil Constantinescu, Caroline J Bull, Nicholas Jones, Ruth Mitchell, Kimberley Burrows, Niki Dimou, Stéphane Bézieau, Hermann Brenner, Daniel D Buchanan, Mauro D'Amato, Mark A Jenkins, Victor Moreno, Rish K Pai, Caroline Y Um, Emily White, Neil Murphy, Marc Gunter, Nicholas J Timpson, Jeroen R Huyghe, Emma E Vincent. medRxiv 2023.03.03.23286764; doi: https://doi.org/10.1101/2023.03.03.23286764 (Pre-print).

Author contributions: I led the project and was involved in the study design and planning along with CJB, JRH and EEV. I performed all the data analysis and interpretation of the findings. I wrote the initial manuscript and CJB, JRH and EEV aided with the subsequent revised versions for publication. All authors critically revised the paper for intellectual content and approved the final version of the manuscript.

First author: Andrei-Emil Constantinescu **Date**: 24/04/2023

SIGNED: [redacted]

Last author: Emma Vincent **Date**: 24/04/2023

SIGNED: [redacted]

The following publication is based on Chapter 4

A framework for research into continental ancestry groups of the UK Biobank. Constantinescu, AE., Mitchell, R.E., Zheng, J. *et al.* A framework for research into continental ancestry groups of the UK Biobank. *Hum Genomics* **16**, 3 (2022). https://doi.org/10.1186/s40246-022-00380-5 (Accepted, in print).

Author contributions: I led the project and was involved in the study design and planning along with REM and DH. DH ran the ADMIXTURE analysis while I performed the rest of the analyses and interpretation of the findings. DH provided helpful advice on the principal component analyses. I wrote the initial manuscript and EEV, CJB and DH aided with the subsequent revised versions for publication. All authors read and approved the final manuscript.

First author: Andrei-Emil Constantinescu **Date**: 24/04/2023

SIGNED: 

Last author: David Hughes **Date**: 24/04/2023

SIGNED: 

The following publication is based on Chapter 6

A phenome-wide approach to identify causal risk factors for deep vein thrombosis. Andrei-Emil Constantinescu, Caroline J Bull, Lucy J Goudswaard, Jie Zheng, Benjamin Elsworth, Nicholas J Timpson, Samantha F Moore, Ingeborg Hers, Emma E Vincent. bioRxiv 476135; doi: https://doi.org/10.1101/476135 (Pre-print).

Author contributions: I led the project and was involved in the study design and planning along with IH, CJB and EEV. I performed the data analysis and interpretation of the findings. I wrote the initial manuscript and EEV, CJB and LJG aided with the subsequent revised versions for publication. All authors read and approved the final manuscript.

First author: Andrei-Emil Constantinescu **Date**: 24/04/2023

SIGNED: 

Last author: Emma Vincent **Date**: 24/04/2023

SIGNED: 

Publications *not* associated with the content of this thesis

The effect of interleukin-6 signaling on severe malaria: A Mendelian randomization analysis. Hamilton F, Mitchell RE, Constantinescu A, et al. Int J Infect Dis. 2023;129:251-259. doi: https://doi.org/10.1016/j.ijid.2023.02.008.

Genetically-proxied anti-diabetic drug target perturbation and risk of cancer: a Mendelian randomization analysis. James Yarmolinsky, Emmanouil Bouras, Andrei Constantinescu, Kimberley Burrows, Caroline J Bull, Emma E Vincent, Richard M Martin, Olympia Dimopoulou, Sarah J Lewis, Victor Moreno, Marijana Vujkovic, Kyong-Mi Chang, Benjamin F Voight, Philip S Tsao, Marc J Gunter, Jochen Hampe, Annika Lindblom, Andrew J Pellatt, Paul D P Pharoah, Robert E Schoen, Steven Gallinger, Mark A Jenkins, Rish K Pai, the PRACTICAL consortium, VA Million Veteran Program, Dipender Gill, Kostas K Tsilidis. medRxiv 2022.10.24.22281370; doi: https://doi.org/10.1101/2022.10.24.22281370.

Identifying metabolic features of colorectal cancer liability using Mendelian randomization. Caroline J. Bull, Emma Hazelwood, Joshua A. Bell, Vanessa Y. Tan, Andrei-Emil Constantinescu, Maria Carolina Borges, Danny N. Legge, Kimberly Burrows, Jeroen R. Huyghe, Hermann Brenner, Sergi Castellví-Bel, Andrew T Chan, Sun-Seog Kweon, Loic Le Marchand, Li Li, Iona Cheng, Rish K. Pai, Jane C. Figueiredo, Neil Murphy, Marc J. Gunter, Nicholas J. Timpson, Emma E. Vincent. medRxiv 2023.03.10.23287084; doi: https://doi.org/10.1101/2023.03.10.23287084.

CHAPTER 1. BACKGROUND

1.1. Introduction

The blood as a component of the human circulatory system has been written about and studied for a long time, going as far back as antiquity. The ancient Greek physician Hippocrates is accredited to developing the theory of the four-humours (a.k.a. Humourism) in the 3rd century BC, which stated that the human body was formed from four components: blood, black bile, yellow bile, and phlegm [1]. Galen, another Greek philosopher and follower of this theory, was the first to discover pulmonary circulation in the 2nd century AD [2], along with proposing that imbalances in the proportion of humours would cause disease and mood changes [3]. Analysing the blood has been an essential part of health care ever since.

Major scientific discoveries on the blood and its composition took place predominantly after the Renaissance, which eventually led to the demise of Humourism by the 19th century [3]. In the 1600s, William Harvey was the first to discover that the circulation of blood was a closed system [2,4]. Around the same time, the Dutch scientist Jan Swammerdam used a powerful microscope to view red blood cells (RBCs) for the first time [5], while his colleague, Antoni van Leeuwenhoek, is credited for offering the first description of RBCs [5]. Further studies by William Hewson in the 18th century revealed the presence of white blood cells (WBCs) in circulation, although at that time WBCs were not yet identified as separate subtypes [6]. Finally, almost a hundred years later, Giulio Bizzozero investigated the function of platelets, and described them as the third cell type of the blood, along with erythrocytes (RBCs) and WBCs [7].

Today, we know that the blood is a mix of plasma and blood cells [8]. The former is predominantly water (92%), along with proteins, minerals, sugars and fat, while the latter is formed from three major components: WBCs, platelets, and RBCs [8]. This thesis is focused on the cells of the blood; more specifically, I will show how genetic epidemiology can expand the knowledge on diseases associated with blood cell traits.

1.2. The biology of blood cell traits

Biologically, the blood is part of the connective tissue and develops from the mesoderm, the second germ layer involved in embryonal development [9]. Haematopoietic stem cells (HSC) derive from this layer, being first produced in the aorta-gonad mesonephros region during early embryogenesis and finally settling in the bone marrow in the late stages of development [10]. HSC are multipotent stem cells, with the ability of self-renewal and committing to specific blood cell-lineages to form mature blood cells, a process which is also known as haematopoiesis [11].

Bone marrow is found inside the central and long bones of the human body [12,13] and hosts a mixture of many cell types, including HSCs and stromal cells [14]. The largest proportion of HSCs is within the spongy trabecular bone i.e. metaphysis [12,15,16], a region also referred to as “red” marrow [17]. This is where haematopoiesis largely takes place throughout adulthood [13,17]. A higher proportion of adipocytes is found in the central part of the bone (“yellow” marrow), where haematopoietic activity is limited [18].

Many transcription factors (TFs) and cytokines are involved in the process of haematopoiesis that leads to the eventual transition of a HSC to a mature blood cell [11]. Assuming the classical model of haematopoiesis, HSCs first develop into two cell types: common myeloid progenitor (CMP) and common lymphoid progenitor (CLP) [19]. The former leads to the development of four WBC subtypes (basophils, eosinophils, monocytes and neutrophils), platelets, and RBCs, while the latter develops into the fifth WBC subtype – lymphocytes, encompassing T-cells, B-cells, and natural killer (NK) cells [19] (**Figure 1-1**). Interestingly, recent studies using single-cell RNA-Sequencing suggest evidence for haematopoiesis being a continuous process rather than a series of binary steps [11,20], further complicating the mechanistic landscape of haematopoiesis.

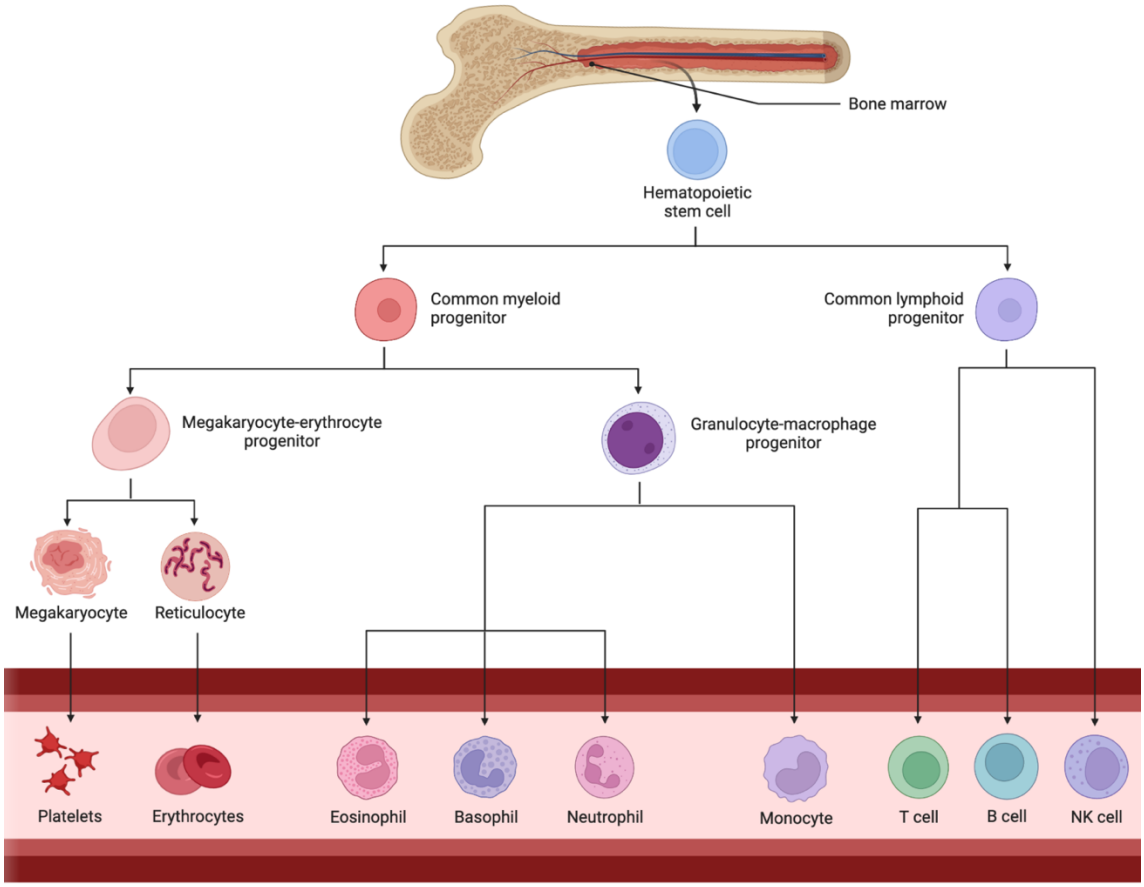


Figure 1-1. *Classical model of haematopoiesis.*
Adapted from "Stem Cell Differentiation from Bone Marrow", by BioRender.com (2023). Retrieved from https://app.biorender.com/biorender-templates.

1.2.1. White blood cells

There are five WBC subtypes (basophils, eosinophils, monocytes, lymphocytes and neutrophils):

Basophils. Granulocytes (i.e. contain intracellular granules) which represent less than 1% of the total WBCs in the blood [21], making their study difficult and were often confused with mast cells, which have common functions and reside in the tissues [22]. Consequently, they were not deemed to be relevant from a biological standpoint, which had resulted in basophils still being an understudied WBC subtype by the early 2000s [23]. However, basophils are now known to come from a different lineage compared to mast cells [24,25], and their biology and function have been studied increasingly in the past two decades [22].

Basophils typically have a diameter of 5-10µm [26], a short half-life of about 1-2 days, and are characterised by their polymorphonuclear cell composition surrounded by many

granules [21]. Basophils typically have a diameter of 5-10μm [26], a short half-life of about 1-2 days, and are characterised by their polymorphonuclear cell composition surrounded by many granules [21]. They develop from CMPs, which then mature into granulocyte-macrophage precursors (GMPs) and then finally into basophil precursors before becoming mature basophils [26,27]. TF GATA-binding factor 2 [28] along with stimulation by the cytokine interleukin 3 (IL-3) are key in the production of basophils in the bone marrow [22]. TF GATA-binding factor 2 [28] along with stimulation by the cytokine interleukin 3 (IL-3) are key in the production of basophils in the bone marrow [22]. However, the exact intermediary processes of haematopoiesis that lead to the development of HSCs into basophils are still under investigation [27].

In terms of their biological function, basophils are commonly associated with the activity of the innate immune system [27]. The basophilic granules contain molecules such as histamine, leukotriene C4, heparin, prostaglandins [29] and basogranulin, the latter being unique to this cell type [30]. When activated, they cause many effects, such skin itchiness and allergic symptoms [31,32], vascular permeability and chemotactic activity that can recruit other immune cells [26]. On the cell surface, basophils contain the receptor FcεRI, which has a high affinity for IgE, one of the five subsets of antibodies i.e. immunoglobulins (Ig, along with IgA, IgG, IgM, IgD) [33]. Cross-linkage of IgE antibodies triggers an intracellular signalling cascade [33], leading to degranulation and release of its contents, along with production and secretion of cytokines IL-4 and IL-13 [27].

However, basophils can also be activated in an IgE-independent manner, outlining two modes of action by basophils. Here, cytokines IL-3, IL-33, complement component C5a and lipopolysaccharide (binding to toll-like receptor 4 TLR4), IgG and IgD activate basophils [27]. In contrast with IgE-dependent activation, basophils do not undergo degranulation, but rather experience a prolonged state of activation during which they secrete cytokines [27]. Further studies have shown that basophils also aid in adaptive immunity, serving as antigen-presenting cells (APCs) for T-cell differentiation into T-helper 2 cells (Th2) [34], and cytokine secretion for B-cell production [29]. All the effector functions of basophils outlined here aid in the clearance of pathogens and resolution of inflammation through type 2 immunity, such as in the case of helminths and ticks, where basophils play key roles for resolution of infection or for acquired immunity [31,35,36].

Eosinophils. Like basophils, eosinophils are polymorphonuclear cells and are commonly associated with the innate immune system [37]. They make up around 1-6% of the total WBCs in circulation [37,38] and initially their inactivation or removal was not

associated with detrimental physiological effects [39]; therefore their biology and functions have also been understudied up until recently [40].

In terms of their development, eosinophils are also matured from the myeloid cell lineage, the CMP, which then mature into GMPs [41]. These then differentiate in the bone marrow into CD34+ eosinophil-committed progenitors [42] that present the IL-5 receptor (IL5R) on their cell surface, making the IL-5 cytokine critical in the maturation and recruitment of mature eosinophils [43]. TFs C/EBP-α and GATA-1, along with the cytokine IL-33, are also vital components that aid the differentiation of early myeloid precursors into eosinophils [41], while the TF protein kinase Tribbles homolog 1 (TRIB1) aids in the final stages of eosinophil-committed progenitors commitment to eosinophil maturation [44].

Eosinophils are also implicated in IgE immunity, although this differs from basophils, and much is still to be discovered. For example, eosinophils only present the FcεRI receptor when under stress [45], as opposed to basophils, where it is present in homeostatic conditions as well [33,46]. The low-affinity IgE receptor FcεRII is also detected in eosinophils, but not always [47], suggesting again that there might be IgE-dependent immunity at play under certain immunological scenarios. As mentioned, IL-5 is one of the key cytokines for eosinophil activation and recruitment [43]. However, other cytokines such as eotaxins CCL11, CCL24 and CCL26 [47], as well as granulocyte-macrophage colony stimulating factor (GM-CSF) and IL-3 [43] are also involved in eosinophil recruitment and degranulation.

Just like basophils, eosinophils can secrete histamine, LTs and PGs [48]. Additionally, studies have also outlined several effector proteins specific to eosinophil degranulation: eosinophilic cationic protein (ECP), major basic protein (MBP), eosinophil-derived neurotoxin (EDN), eosinophil peroxidase (EPX) and Charcot-Leyden crystals (galectin-10) [44]. The first three are all cationic proteins capable of direct cytotoxic effects on pathogens, with ECP and EDN also having RNase capabilities [49]. Moreover, these three proteins can also recruit other immune cells, activate dendritic cells (DCs) and lead to secretion of inflammatory cytokines [50,51]. Meanwhile, EPX produces reactive oxygen species (ROS) that directly kills pathogens [52] and galectin-10 can induce eosinophil extracellular trap release (EETs) and has been found to induce a Th2 response in murine models [53]. Eosinophils are also capable of phagocytosis, cytokine/chemokine release, as well as antibody mediated immunity [47].

These tools allow eosinophils to fulfil many roles in immunity. For example, they are associated with a Th1/Th17 response and release of MBP, ECP, EDN and EETs when

dealing with viral, bacterial or fungal infections [47,48]. Meanwhile, parasites such as helminths and larvae usually trigger a Th2 response [47] and lead to secretion of EPX, MBP and ECP that aid in the killing of the pathogens [53]. Moreover, eosinophils can act as APCs, T-regulatory cells (Tregs) [43], and aid in the regeneration of muscle and liver tissue [53].

Lymphocytes. These represent a cell group that have traditionally been associated with adaptive (acquired or long-lasting) immunity [19] and up until the 1950s-60s were thought to constitute one cell type [54]. Three major cell types belong to the lymphoid lineage: T-cells (thymic), B-cells (Bursal or bone marrow), and NK cells [19,55–57]. The half-life of a lymphocyte depends on its cell type and activation status i.e. naïve or activated [58–61]. For example, the half-life of NK cells is around 3-4 weeks [61], while for naïve T-cells it can be up to eight years [59].

In terms of their development, lymphoid progenitors mature from HSCs, after which they turn into CLPs [19]. Here, the CLPs can commit further to the T-cell, B-cell or NK lineage [19]. T-cell progenitors migrate from the bone marrow to the thymus, becoming thymocytes [62]. Here, they undergo a process of generating T-cell receptors (TCRs) that can bind to previously unencountered pathogens through immunoglobulin-like gene recombination [63]. This is followed by positive and negative selection, which keep those T-cells with the maximal response towards a pathogen while avoiding damaging the host [62]. B-cells mature from progenitor and precursor B-cells in the bone marrow [64]. Here, cytokines such as CXCL12, IL-7 and SCF, along with TFs like Early B-Cell Factor 1 and Pax5 contribute to B-cell development and lineage commitment [64]. Like T-cells, B-cells also undergo a process of immunoglobulin gene recombination, allowing B-cells to ultimately produce antibodies that can bind to almost every possible antigen [65].

The lymphatic system is key to the functioning of the adaptive immune system [66]. It contains lymph, which is derived from interstitial fluid [67], and acts as a network through which WBCs can activate lymphocytes in the lymph nodes by antigen or engulfed pathogen presentation [68]. Moreover, another component of adaptive immunity is the major histocompatibility complex (MHC) proteins that present parts of a pathogen to T-cells [69,70]. These can be either class I, activating CD8+ T-cells and present on any nucleated cell, or class II, activating CD4+ T-cells and presented by APCs [69,70].

T-cells can differentiate into separate T-cell subtypes: cytotoxic (CD8+) and helper (CD4+) [62]. Cytotoxic T lymphocytes are involved in the killing of cells that might affect the normal functioning of the host, such as cells infected with viruses and bacteria, either

through direct cell-cell interactions or through the secretion of specific cytokines, such as interferon-γ (IFN-γ) [71]. Helper T-cells also further differentiate into subtypes depending on a combination of specific cytokine and TF action: Th1 (IL-12, T-bet), Th2 (IL-4, GATA-3), Th17 (IL-6, IL-23, RORγt), Th9 (TGF-β, IL-4), Th22 [IL-6, tumour necrosis factor (TNF)], T follicular (fh, IL-21, BCL6) and T regulatory (reg, TGFβ, Foxp3) [72–75]. The first five listed T helper cells are involved in different aspects of immunity, and each secrete their specific cytokines which can shift the balance of the immune response [72,73,76]. Tfh cells are present only in the lymphoid tissues and aid in B-cell differentiation and antigen affinity maturation [74]. Tregs, as their name implies, are important in regulating the immune response of T-cells and their differentiation [73].

Just like T-cells, B-cells serve many functions that ensure the functioning of the immune system. They are involved in the production and secretion of antibodies, serving also as memory cells that aid in efficient clearance of reinfection from a pathogen [77]. NK cells are another lymphocyte subset, although they are considered to be part of the innate immune system [78,79]. Similarly to CD8+ T-cells, they are involved in immune-mediated cytotoxicity, and are particularly involved in controlling and clearing viral infections [78,79].

Monocytes. These mononuclear leukocytes are the largest cells in circulation (~20µm) [80] and have a half-life of around 1-2 days [81]. Traditionally, monocytes are known for their capacity to enter tissues and become tissue-resident macrophages [82]. Therefore, despite representing a larger proportion of total WBCs (5-10%) than basophils and eosinophils [80,83], monocytes have predominantly been thought of as blood intermediaries of macrophages[80].

During haematopoiesis, HSCs commit to the CMP lineage in the bone marrow [84]. Studies suggest that here the pathways bifurcate into GMPs and monocyte-DC progenitors, and monocytes can arise from both these precursors [85]. There are several factors involved in the commitment of these progenitors to differentiating into mature monocytes. General TFs such as PU.1 and C/EBP-α combined with the action of monocyte TFs interferon regulatory factor 8 (IRF8) and Krüppel-like factor 4 (KLF4) are key in the commitment of progenitors to the monocyte lineage [84,85].

Biologically, studies looking at gene expression have outlined the presence of classical and non-classical monocytes, as well as a possible third intermediary monocyte type that has features from both subsets [86]. These three monocyte subpopulations are commonly identified by the presence of lipopolysaccharide (CD14) and low-affinity Fcγ (CD16) receptors on their cell surface [87]. Classical monocytes ($CD14^{++}$ $CD16^{-}$) [88] represent over

80% of monocytes [87] and are continuously differentiating, with cytokines such as IFN-γ or TNF-α acting as further enhancers of their differentiation process [82]. Their activation is associated with a Th1 response, wound healing, phagocytosis, and the release of cytokines IL-6, IL-10 and CCL2 [83,86]. The role of intermediate monocytes ($CD14^+$ $CD16^+$) seems to be predominantly to act as APCs [89]. Meanwhile, non-classical monocytes ($CD14^-$ $CD16^+$) [88] are sometimes termed "sentinel" monocytes due to their role in surveillance of the endothelium and production of CD4+ T-cells [83,90]. These cells do not usually become tissue-resident macrophages [90].

Given these functions, monocytes are cells capable of performing multiple roles in immunity. For example, classical monocytes aid in anti-microbial activity [88], and both classical and non-classical monocytes have been found to combat fungal infections [83]. Additionally, non-classical monocytes are involved in anti-viral activity [88], and intermediate monocytes were found to mediate the immune response in the presence of bacterial Staphylococcal enterotoxin [83].

Neutrophils. These polymorphonuclear granulocytes measure around 10μm in size and represent 50-70% of the total WBCs in the blood, making them the most populous WBCs in circulation [91]. Neutrophils typically have a half-life of 8 hours, although it has been reported that they can live up to 5 days in the blood [92]. Like basophils and eosinophils, their appearance is characterised by their multinuclear core surrounded by granules that contain various proteins involved in immunity [93].

In terms of neutrophil development, CMPs differentiate into GMPs [91], these then differentiate into myeloblasts, where they can either commit to the monocyte or granulocyte lineage [91]. For the progenitor cells to become neutrophils, several factors work concomitantly, such as TFs C/EBPα, IRF8, and PU.1, along with the cytokines SCF, G-CSF and GM-CSF [94]. Interestingly, studies have found that apart from their presence in the blood stream, neutrophils are also present in even higher numbers in the bone marrow, where they act as a pool of cells ready to be recruited in stress situations, such as infections [94]. During stress, cytokines IL-1, IL-6 and GM-CSF have been found to prolong the lifespan of neutrophils [95].

Biologically, neutrophils have three main functions: phagocytosis, granule secretion and neutrophil extracellular trap (NET) production [96]. During phagocytosis, the neutrophil engulfs the microbe inside the cell in a phagosome, which can fuse to neutrophilic granules to release their contents on the pathogen and lead to its elimination [97]. The granules are divided into four types: azurophilic granules, specific granules, gelatinase

granules and secretory vesicles [98]. Azurophilic granules contain microbicidal proteins, such as ROS-producing myeloperoxidases [98]. Specific granules contain lactoferrin and lipocalin [98], which have many roles, such as regulation of NETosis (NET production) and lymphocyte maturation, serving as chemokines or mediating cytokine release [99]. Meanwhile, gelatinase granules have matrix metalloproteinase 9, which can lead to chemokine release by neutrophils [98]. Secretory vesicles contain cytokine receptors and other plasma proteins, including components for the nicotinamide adenine dinucleotide phosphate oxidase enzyme, predominantly responsible for microbicidal ROS production in neutrophils [98,100].

Discovered in 2004, neutrophil extracellular traps (NETs), as the name implies, can trap microbes [101]. These extracellular webs are formed from DNA and neutrophilic granule proteins after the neutrophil undergoes programmed cell death [102]. The three main neutrophil processes outlined here do not work independently, but in tandem. For example, if the size of the microbe is too large to be phagocytosed, neutrophil granules are migrated towards the nucleus to start the process of NETosis [103]. Recent studies have found heterogeneous neutrophil populations, where different transcriptional profiles are present depending on the type of infection [104,105].

Neutrophils are vital to the functioning of the innate immune system, as they are the first cells to arrive at the site of infection [98]. They use the tools outlined above to deal with a vast array of pathogens, such as viruses, bacteria, fungi, which activate neutrophils through pattern recognition receptor that recognise the molecules associated with these pathogens [103]. Additionally, neutrophils are also involved in the resolution of inflammation, where they can clear debris through phagocytosis and by releasing C-C chemokine receptor type 5 or NETs to soak up pro-inflammatory cytokines [106].

1.2.2. Platelets

Platelets, also known as thrombocytes, are the smallest cells in circulation, measuring around 2-4µm [107], and have a half-life of around 7-10 days [108].

In terms of development, HSCs differentiate into CMPs, which then commit to the megakaryocyte-erythroid lineage [109]. Here, general TFs GATA-1 and GATA-2 are required, along with nuclear factor-erythroid 2, which is the most important TF in megakaryopoiesis i.e. megakaryocyte maturation [109]. Progenitor cells in the bone marrow transform into megakaryocytes, which are large mononuclear cells [110] whose production is regulated by thrombopoietin [108]. Next, early megakaryocytes undergo a

process called endomitosis, where the cell replicates without finishing mitosis many times until the DNA content reaches up to 128 times the genetic material found in a normal cell [111]. Afterwards, megakaryocytes undergo cytoskeletal remodelling and form protrusions called pro-platelets [108]. These cytoplasmic structures bud off and become platelets, leaving the megakaryocyte predominantly with only its nucleus left, platelets are therefore anuclear cells [112].

Although small in size compared to other blood cells, platelets have many components that help them accomplish their functions, including three granule types: dense granules, α-granules, and lysosomal granules [113]. The dense granules (DGs, or δ-granules) contain ions (e.g. Ca^{2+} or Mg^{2+}), ADP, ATP, serotonin, and phosphate molecules [114,115]. α-granules contain an large number and diversity of proteins, such as clotting factors von Willebrand factor (vWF) and fibrinogen, membrane protein P-selectin, both pro- and anti-angiogenic factors, as well as multiple cytokine and chemokine types [116]. Lysosomal granules, as the name implies, contain lysosomes, which are enzymes that can break down other cells or microbes [117].

Functionally, platelets have traditionally been associated with blood haemostasis i.e. maintaining blood flow and stopping vessel leakage [118]. In brief, when the blood vessel becomes injured, the endothelium releases vWF which anchors platelets to the endothelial wall and triggering the release of the intracellular granules [113,119]. These then bind to other platelets and RBCs in circulation through release of thrombin and collagen, leading to their activation and eventually the formation of a thrombus that prevents bleeding and maintains haemostasis [120].

Additionally, platelets are now known to be involved in other processes, such as immunity. For example, P-selectin is present on the surface of activated platelets and aids in the recruitment of neutrophils and monocytes, as well as in the production of pro-inflammatory cytokines [119]. Moreover, platelets have also been found to play a role in adaptive immunity by acting as APCs and regulating Treg production [118].

1.2.3. Red blood cells

RBCs, also known as erythrocytes, are the predominant cells found in circulation [8]. As mentioned in **Chapter 1**, they were the first cells seen under a microscope [5]. They measure 8µm in size and have the easily recognisable biconcave disk shape [121]. RBCs have a much longer half-life than most blood cells, averaging around 120 days in circulation [122].

In terms of erythropoiesis, HSCs mature into CMPs, which then commit to the megakaryocyte-erythrocyte lineage through regulation by TFs PU.1 and GATA-1 [94]. Afterwards, megakaryocyte-erythrocyte progenitors commit to the erythroid lineage, where the erythropoietin cytokine plays an important role in further maturation of progenitor cells [123]. These cells then undergo further steps before becoming mature RBCs, such as at the orthochromatic erythroblast stage, where their nucleus gets ejected, leading to their maturation into reticulocytes and finally into RBCs [123,124].

RBCs are best known for their role in transporting oxygen from the lungs to tissues [125]. This process is facilitated by haemoglobin, a protein which contains iron (Fe^{2+}) ions [126]. These iron ions have an affinity for both oxygen (O_2) and carbon dioxide (CO_2) molecules, which allows for the delivery of oxygen, while taking CO_2 back to the lungs, where it is ultimately exhaled [125]. The biconcave structure of erythrocytes maximises the surface area for gas exchange while ensuring their plasticity when travelling through small capillaries, which can be several times smaller than their size [122].

At the same time, RBCs are known to have receptors that bind to cytokines or chemokines and regulate immune activity, such as the receptor Atypical chemokine receptor 1 / Duffy antigen receptor for chemokines (ACKR1/DARC) [127].

1.3. The full blood count test

Blood cells are now commonly measured as part of a routine blood test [128]. However, measuring the constituent cells of the blood was not always a straightforward and affordable process.

1.3.1. A history of blood cell counting

The first recorded blood counting method was invented by the German physiologist Karl von Vierordt in the early 19th century [129]. Other scientists gained interest in Vierordt's research and built upon his discoveries, such as the Dutch researcher Antonj Cramer, who designed a chamber made from two spaced glass slides, allowing a diluted blood sample to be spread evenly between them for more efficient and replicable blood counting [130]. Other scientists aimed to improve upon the inventions of these two scientists, one of them being the French histologist Louis-Charles Malassez, who is credited to have been the inventor of the modern haemocytometer in 1874 [130,131]. Further

improvements were made by Karl Bürker, who built a chamber on the same principles of Cramer and the French physician Georges Hayem [130].

At the same time, scientists were able to use blood cell indices to aid in medical practice. Maxwell Wintrobe was an Austrian-born doctor who is considered to be one of the fathers of haematology and a pioneer in the field of blood counting [132]. One of his most well-known inventions was a technique to study the haematocrit i.e. the percentage that RBCs represent from the total blood [132]. Using this method, he was able to diagnose patients with different diseases just by studying percentage differences in the blood constituents after centrifugation [129]. For example, a lower haematocrit indicated anaemia, while a pale white segment above the dark red mature RBC sedimentation indicated leukaemia [129]. Further studies allowed for the development of the Wintrobe indices, which are nowadays known as RBCs indices: mean corpuscular volume (MCV), mean corpuscular haemoglobin (MCH), and mean corpuscular haemoglobin concentration (MCHC) [132].

However, RBCs were not the only measured blood cells. The German pathologist Richard Thoma observed that counting WBCs would not be possible at the dilution used to measure RBCs, and the presence of RBCs made the counting process difficult [130]. Therefore, he used a custom dilution factor along with an acetic acid solution to lyse the RBCs, allowing him to reliably measure WBCs [133]. Scientists tried to perform the counting of platelets as well, although such attempts were not reliable until the invention of the phase contrast microscope in the 1950s [134].

1.3.2. The Coulter method

As mentioned, manual counting used to be the only way to get blood cell indices, which was time-consuming and required a skilled professional [135]. While there are many methods for cell cytometry used today, such as specialised flow cytometry devices, the most well used and reliable method to analyse BCTs (indices associated with the cells of the blood) during a clinic or hospital visit is a Coulter counter [131].

The Coulter method was developed by the American scientist Wallace H. Coulter and his brother through the aid of a government project in 1953 [136]. The first Coulter counter was called “Model A” and its rudimentary appearance was strikingly different to the counters used today [136], and some BCTs like platelet indices were only available to measure automatically in the late 1970s [134].

The Coulter principle employs a concept in electrical engineering known as electrical impedance (i.e. resistance) [135]. The blood sample is diluted with a solution and is pumped into a beaker that has a tube with a small aperture through which each cell can travel through [136]. Two electrodes are also present in the beaker, one on the inside of the tube, and one on the outside [136]. A vacuum is created inside the tube which pulls each cell in the solution through the aperture. Each cell passing through the slit produces an electrical resistance between the two electrodes which is directly proportional to the cell's volume, making it possible for the machine to detect the type of blood cell [136] (**Figure 1-2**).

Modern Coulter counters have evolved greatly in the last 60 years, and models such as the Beckman Coulter LH750 use additional methods to analyse each BCT, such as radiofrequency or light scattering similar to those used in flow cytometers, being able to compute 20 or more BCT indices [131,137] (**Table 1-1**).

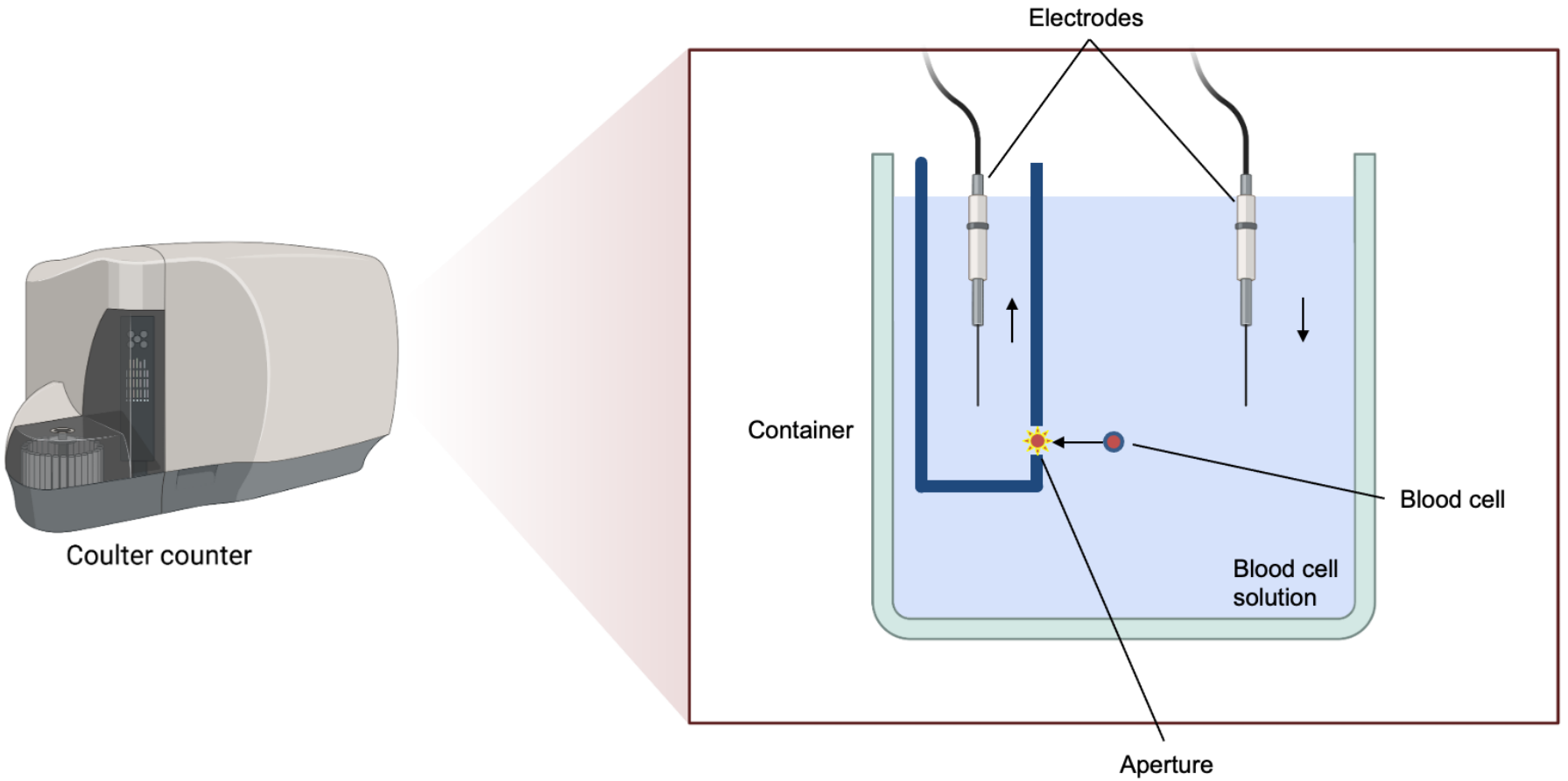


Figure 1-2. *Coulter method.*
The blood cell passes through the aperture, creating electric resistance between the two electrodes, which gives an estimate of the cell's volume. Along with augmentative devices introduced in modern counters, this can be used to estimate the blood cell type. Created with BioRender.com.

Table 1-1. *Blood cell traits as measured in a common full blood count (FBC) test.*

Blood cell trait	Abbreviation	Description	Units
White blood cell count	WBC	The number of white blood cells in the blood	10^9 cells/Litre
Basophil count	BAS #	The number of basophils in the blood	10^9 cells/Litre
Basophil percentage	BAS %	The proportion of basophils of the total WBC count	Percent (%)
Eosinophil count	EOS #	The number of eosinophils in the blood	10^9 cells/Litre
Eosinophil percentage	EOS %	The proportion of eosinophils of the total WBC count	Percent (%)
Lymphocyte count	LYM #	The number of lymphocytes in the blood	10^9 cells/Litre
Lymphocyte percentage	LYM %	The proportion of lymphocytes of the total WBC count	Percent (%)
Monocyte count	MON #	The number of monocytes in the blood	10^9 cells/Litre
Monocyte percentage	MON %	The proportion of monocytes of the total WBC count	Percent (%)
Neutrophil count	NEU #	The number of neutrophils in the blood	10^9 cells/Litre
Neutrophil percentage	NEU %	The proportion of neutrophils of the total WBC count	Percent (%)
Platelet count	PLT #	The number of platelets in the blood	10^9 cells/Litre
Mean platelet volume	MPV	Average platelet volume	Femtolitres (10^{-15} Litres)
Red blood cell count	RBC #	The number of red blood cells in the blood	10^{12} cells/Litre
Haemoglobin concentration	HGB	Haemoglobin mass per unit volume of blood	grams/Decilitre

Blood cell trait	Abbreviation	Description	Units
Haematocrit	HCT	Proportion of red blood cells from the total blood volume	Percent (%)
Mean corpuscular volume	MCV	Average RBC volume	Femtolitres (10^{-15} Litres)
Mean corpuscular haemoglobin	MCH	Mass of haemoglobin per average RBC	Picograms
Mean corpuscular haemoglobin concentration	MCHC	Average haemoglobin mass per relative volume of RBCs in the whole blood sample	grams/Decilitre
Red blood cell distribution width	RDW	Measures the variation in size and volume of the red blood cells	Percent (%)

The introduction of Coulter counters in clinics and hospitals made the blood an easily accessible medium, and blood cell counters are now present in hospitals worldwide [138]. This in turn made it attractive for scientists to investigate the relationship between BCT measurements and different traits, including disease [134].

As part of this thesis, I will study the blood cells measured in a routine blood test, where they are analysed as homogeneous entities. From this point forward, whenever I will use the acronym BCTs, this will refer to the traits measured in a common blood test as presented in **Table 1-1**, unless specified otherwise in the text**.**

1.4. Blood cell traits and disease

As evidenced by the work of Max Wintrobe [132], BCTs proved to be valuable in the diagnosis of disease. Many more studies since then have focused on how BCTs can be indicative of pathological conditions.

1.4.1. Blood as an accessible sample of diagnostics

Due to the low costs and cell counters present in most hospitals, BCTs are now used in the diagnosis of many conditions [128]. For example, a high neutrophil count can indicate

a bacterial infection, while a low platelet count can be used to diagnose bleeding disorders [128]. The introduction of the MCV and RDW in the 80s allowed for a more accurate classification of anaemias in patients [139]. Trans-ethnic studies identified that those of African descent were more likely to have a lower WBC count, which was termed and diagnosed as benign ethnic neutropenia (BEN) and its cause was unknown at the time [140].

1.4.2. Blood cells in disease

Much research has been done in the last couple of decades to investigate the role of blood cells in disease. For example, neutrophils have been extensively studied for their role in cancer, where they were discovered to have dual capacity for cytotoxic killing of tumour cells, as well as promoting metastasis and immune suppression [141]. Eosinophils were identified as factors that can worsen the symptoms of asthma through secretion of eosinophilic granules, such as MBP [142]. CD8 T-cells are known to be key factors in anti-tumoural immunity against cancers such as metastatic melanoma [143]. Meanwhile, platelets were identified as cells capable of producing thrombi both directly and indirectly [144,145], and are well-known for their role in deep vein thrombosis (DVT) [146].

1.4.3. The FBC test for establishing risk factors for disease

By leveraging the BCT indices given by the modern Coulter counter, epidemiologists have been able to assess how changes in BCTs affect the risk and prognosis of disease. In this thesis I will focus on the relationship between BCTs and disease risk, namely - colorectal cancer (CRC), *Plasmodium falciparum* (*P. falciparum*) malaria and DVT. More detail on the relationship between BCTs and each disease is covered in Chapters 4, 6 and 7, respectively.

1.4.4. BCTs as flags of a biological mechanism

As a point of clarification, we should not think of BCTs themselves as factors that can directly affect disease prognosis or development. Unlike direct examples like the impact of a caustic agent on the skin or a carcinogen on cells, BCTs are more akin to flags of a particular biological mechanism, where an increase or decrease in the value of a BCT is associated with a biological effect. This is similar to how body mass index (BMI) is studied in diseases such as type 2 diabetes (T2D), where adiposity itself is not directly increasing the risk of e.g. diabetes, but rather increased adiposity is thought to cause changes in

the metabolic profile of cells, leading to increased risk of developing T2D [147]. Therefore, understanding the biological functions of each blood cell type is important in the interpretation of findings, and hence the motive behind the extended biological background presented earlier in this chapter.

1.5. The genetics evolution

BCTs have been studied genetically through single base germline variation i.e. single-nucleotide polymorphisms (SNP) that could be associated with a trait of interest [148,149]. Up until the 1990s, SNPs were studied using methods that were costly and focused on a small section of the genome, and therefore only a small number of SNPs were associated with biological outcomes [148]. However, major events took place in the early 2000s that would radically transform the field of genetics as it was hitherto known. The completion of The Human Genome project in 2003 marked the start of many further initiatives to explore the genetic architecture of humans [150]. Projects such as HapMap [151] and the 1000 Genomes Project [152] were further initiatives that aimed to map genetic variation across global populations.

The study of hundreds of thousands, or millions of SNPs in a comprehensive manner and how they associate with a trait is called a genome-wide association study (GWAS) [153]. In cohort studies where people aim to understand the genetic architecture of a particular trait, patients undergo both genotyping and a recording of baseline measurements (**Figure 1-3**) [154]. GWAS have been invaluable in pushing the boundaries of what is known in relation to BCTs and disease. SNPs associated with BCTs could be mapped to novel genes and investigated through existing databanks or new methods in relation to auto-immune disease [155], T2D, cancers or blood disorders [156].

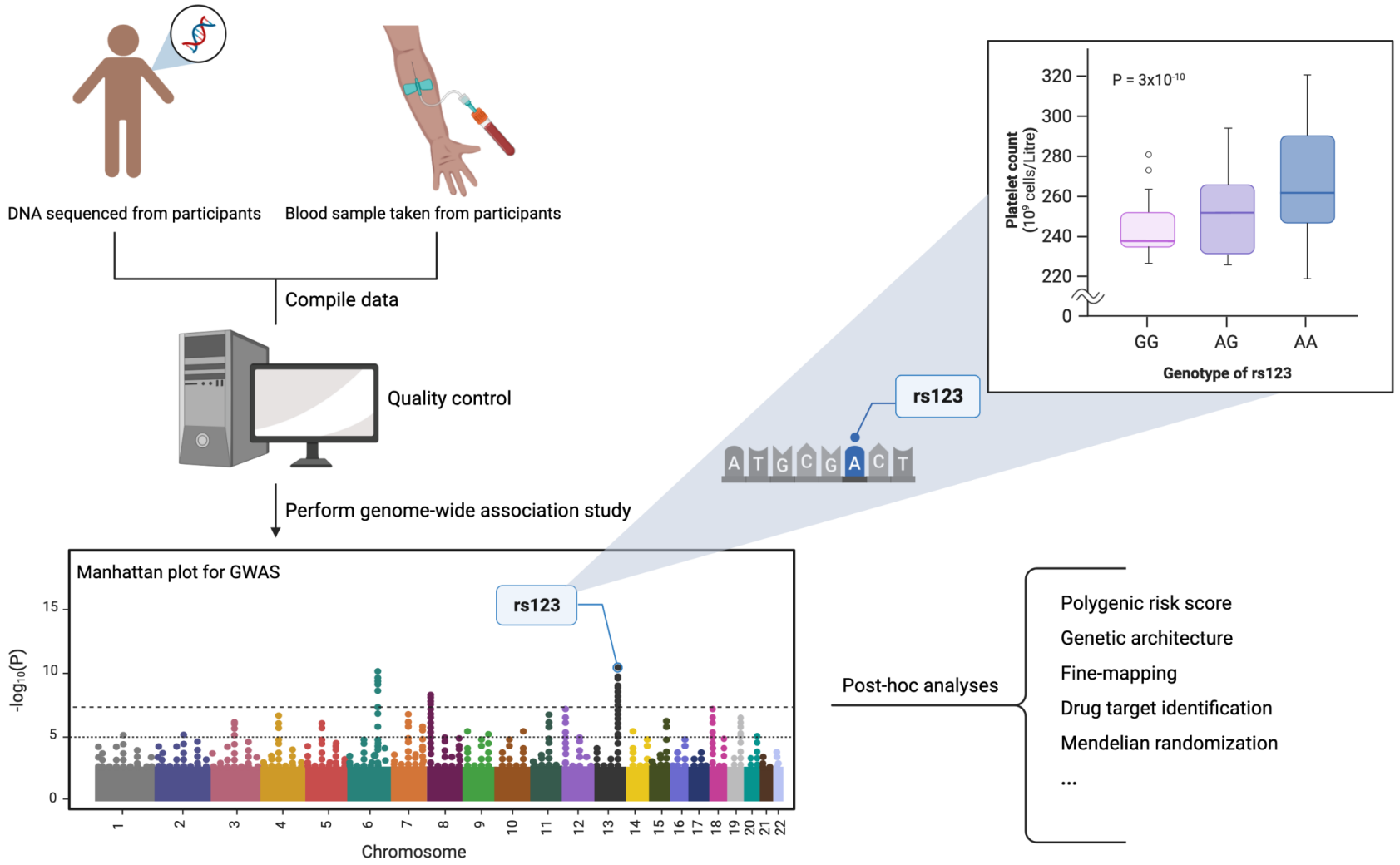


Figure 1-3. *The concept of a genome-wide association study. Made with Biorender.com.*

The first GWAS was performed in 2005 on age-related macular degeneration using only 86 cases and 50 controls [157]. It did not take long until scientists took this concept further and applied it to BCTs. Here however, the sample-sizes were already considerably larger than in the first GWAS [158]. By the late 2000s, GWAS in up to 25,000 thousand participants had been done to investigate the genetic architecture of BCTs [158], outlining loci associated with traits such as MCV, WBC count and MPV [158]. Further smaller scale studies focusing on SNPs associated with gene expression, a.k.a. expression quantitative trait loci (eQTLs), were useful in describing the transcriptional profile of WBCs during pathogenic conditions, establishing mechanistic pathways of immune system activation [159].

However, GWAS could also provide insights into ancestry-specific effects that were previously unknown. One notable example is the GWAS of David Reich and colleagues in 2009, where they identified that the "Duffy" SNP rs2814778 was predominantly responsible for the BEN seen in people of African ancestry [160], outlining that the genetic architecture between populations could differ. This showed that performing GWAS in diverse populations is beneficial, as GWAS done in one population might not be translatable to other populations [153].

1.5.1. Genetics and large-scale biobank studies

The 2010s were marked by large scale initiatives that aimed to combine phenotypic and genotypic data. Such examples include the UK Biobank (UKBB) study in the UK [161] and the Million Veterans Program in the US [162]. However, conducting large scale studies was not the only way to improve sample-sizes for novel loci discovery, as smaller consortia could pool their data together. One such example is the GWAS of CRC risk by Huyghe et al., where over 60 studies pooled their data to generate a final meta-analysed dataset of over 100K CRC cases and controls [163].

Currently, most large consortia have adopted the concept of directly genotyping a selected number of SNPs (e.g. 800K for UKBB [164]), and imputing the remaining millions of genetic variants using a reference panel with whole genome sequenced (WGS) data, such as the 1000 Genomes project [154]. This allows for conducting studies in tens or hundreds of thousands of people in a cost-efficient manner [154].

This increase in power due to large sample-sizes proved to be incredibly helpful in mapping the genetic architecture of traits that were not monogenic i.e. only variation in one SNP or gene is responsible for changes in a phenotype [153]. One of these are polygenic traits, where multiple SNPs/genes contribute to changes in a trait [165]. At the end of this spectrum are

complex traits, where many SNPs contribute to small changes in a trait along with the environment and interactions arising from genes and the environment [153].

1.5.2. Blood cells as complex traits

BCTs also belong to this category of traits, as even early GWAS identified tens of loci that were responsible for BCT variation [158]. Further analyses with higher sample-sizes like the analysis of Astle et al. in UK Biobank using over 300K people of European ancestry identified hundreds of loci associated with BCTs [149]. More recent studies like that of Chen and colleagues took this further, meta-analysing UKBB with other consortia to find additional novel loci affecting BCTs [166]. These initiatives showed that genetics could greatly contribute to the biological knowledge of blood cells, even after decades of laboratory-based research on the topic. As mentioned at the start of this section, this allowed for the identification of how BCTs are involved in disease.

1.6. Causal inference in genetic epidemiology

The release of publicly available summary statistics for GWAS conducted using data from these large consortia made it accessible for genetic epidemiologists to perform analyses, especially with the release of centralised databases, such as the GWAS Catalog [167]. Genetic epidemiology is a branch of epidemiology that looks at the relationship between genetics and disease and it aims to leverage people's genetic data to find a causal relationship between an exposure and an outcome [168].

1.6.1. Mendelian randomization

One important method in genetic epidemiology is Mendelian randomization (MR), which is named after Gregor Mendel [169], an Austrian priest and scientist from the 19th century and now considered to be a major figure in the field of genetics [170]. MR is most known from the popular article published by George Davey-Smith and Shah Ebrahim back in 2003 [169]. The method makes use of the random assignment of theoretically independent alleles at conception, under Mendel's ascribed second law of genetics [169], to give causal estimates between an exposure and an outcome. This has been said to be analogous to a randomised controlled trial (RCT, discussed further in **Chapter 2**) in so much as the genetic variation employed in these studies should be effectively independent of other alleles and the environment/confounding factors [171,172]. Since then, MR has been extensively used to estimate causal relationships between exposures and outcome [173].

Given a set of assumptions, MR can be used to estimate a causal effect of an exposure (e.g. BMI) on an outcome (e.g. T2D) [171]. This has advantages over traditional epidemiological approaches in the context of BCTs as risk factors for disease, which will be further explored in **Chapter 2**. However, the point of MR is not to replace existing traditional epidemiology or lab-based approaches, but rather to serve as a bi-directional avenue which can inform and be informed by these traditional approaches (**Figure 1-4**).

MR has been valuable in addressing the difficult questions by such tradtional approaches. One such example is the relationship between C reactive protein (CRP), a protein produced by the liver and a marker of inflammation, and risk of developing coronary heart disease (CHD) [174]. Observational methods had shown that higher CRP levels were associated with an increase in CHD development, leading researchers to believe that CRP has a causal effect on CHD risk [175]. Scientists from the C Reactive Protein Coronary Heart Disease Genetics Collaboration performed a MR analysis of CRP on CHD risk, where they did not find evidence of a causal effect of CRP on CHD risk [175].

Another example came from a study done by Voight et al [176]. Observationally, a higher high-density cholesterol (HDL) was associated with a decrease in myocardial infarction risk [176]. However, in their MR study of HDL to CHD risk, Voight et al. identified in their MR analysis that there was no evidence of a causal relationship between higher HDL levels and myocardial infarction risk [176].

Furthermore, observational studies had previously found a negative association between selenium intake and prostate cancer risk [177]. In a follow-up RCT by the Selenium and Vitamin E Cancer Prevention Trial (SELECT), selenium had shown no association with risk of developing prostate cancer, contrary to the initial observational studies [178]. In their follow-up analysis of this trial, Yarmolinsky et al. conducted a MR analysis between selenium intake and prostate cancer risk, where they did not find any evidence of a causal effect [177], outlining the value of MR in testing observational studies in a more cost-effective and timely manner than a RCT.

These examples, along with many more over the past decade, have shown that MR is a valuable tool in assessing causality.

1.6.2. Genetically proxied BCTs and Mendelian randomization

The capacity for MR to detect risk factors for disease is useful when a RCT cannot be performed, either due to cost, practicability or ethics [172] – BCTs are included in this category, as for example one might not expect to give BCT altering drugs to patients. Studies using MR to explore the relationship between BCTs and disease have been very few, and given the gaps in knowledge concerning blood cell indices and disease risk, applying this method in the context of BCTs as risk factors is desirable for establishing an avenue for further studies.

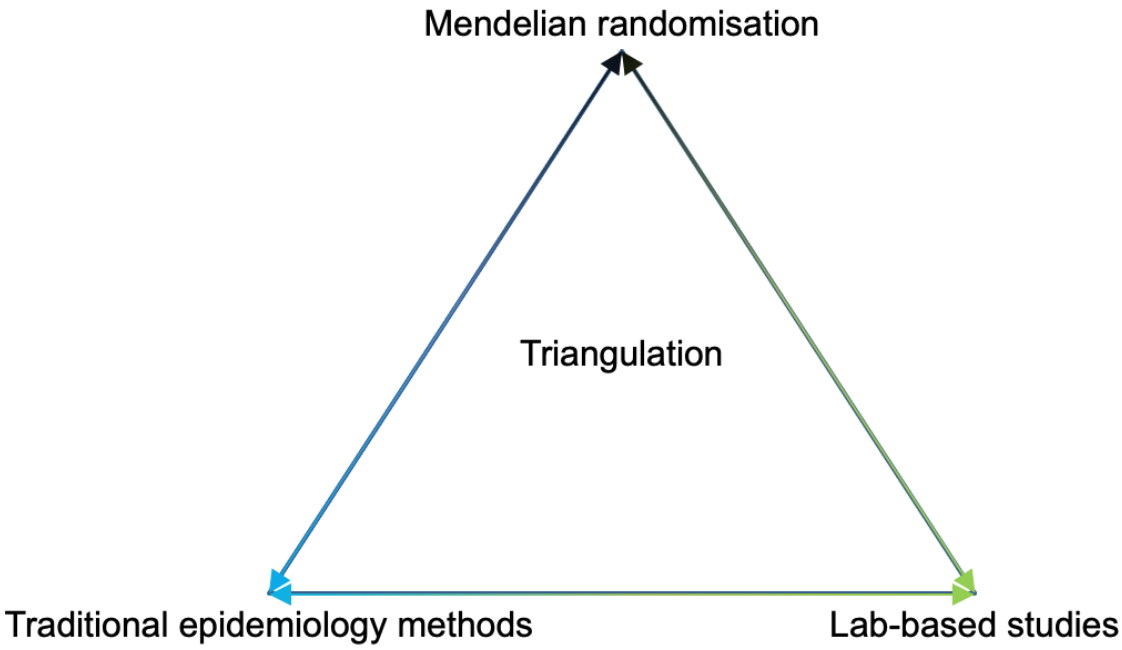


Figure 1-4. *MR in the context of empirical methods used to advance the knowledge in biomedical research.*

1.7. Overarching objective of thesis and aims

The overarching objective of my thesis is to show how studying the relationship between genetically proxied traits that affect BCT driven diseases can improve our understanding of BCTs and disease aetiology. To do this, I will use Mendelian randomization (MR).

I have chosen to focus on three diseases, which are also the foci of the research questions of my thesis, each structured in a results chapter and coming with their own methodological challenges:

i. Do WBC subtype counts affect the risk of developing CRC? (**Chapter 3**)
ii. Does a higher neutrophil count increase the risk of severe P. falciparum malaria? (**Chapter 4**, **Chapter 5**)
iii. What are the mechanisms through which platelet BCTs could affect the risk of DVT? (**Chapter 6**)

I will delve deeper into the biology of each disease and their relationship with the chosen BCTs in the respective results chapters.

CHAPTER 2. METHODS AND DATA SOURCES

Chapter summary

The aim of this chapter is to give a broad overview of the different types of epidemiological studies and to describe in depth the advanced methods in genetic epidemiology that I have used to address the aims of my thesis, e.g., Mendelian randomization (MR) [169]. I will then enumerate, in detail, the data sources that I have used to conduct the analyses in my thesis. Chapter-specific methods have been described in the respective results chapters.

2.1. Traditional methods in epidemiology

There are many types of study designs employed in epidemiological research, most of which use data available at the population level to assess the relationship between an exposure and an outcome with the aim of establishing a causal link between the two [179]. These are typically ranked from the point of view of their potential to provide evidence for causality [180]. For example, a case-report involves the study of only one case and can be useful in investigating a rare disorder or generating new hypotheses, but it carries little weight in establishing a causal relationship between an exposure and an outcome [181]. At the opposite side of the spectrum are randomised controlled trials (RCTs), which are considered to be the "gold standard" in establishing causality (**Table 2-1**) [161,180–183].

Table 2-1. *Traditional study designs in biomedical research.*
Each study design is ranked top-down by its ability to allude to causality. Note: this is a non-exhaustive list (many subtypes exist for each study design) and none are perfect, which is why triangulation (discussed in the previous chapter) is important.

Study design	Description	Evidence description	Type
Randomised controlled trial	Participants randomly allocated into treatment (exposure) and control/placebo groups. Very expensive and time-consuming. Strict selection criteria.	"Gold standard" in establishing causality.	Intervention

Study design	Description	Evidence description	Type
	Might not be practical/ethical e.g. cannot force participants to smoke.		
Cohort study	Participants recruited first, after which they are followed and can eventually develop the outcome. Can be used to study associations of exposures with diseases. Expensive and time-consuming. Needs large sample-size to study rare diseases e.g. UK Biobank.	Less susceptible to bias than case-control study, especially if prospective.	Analytic
Case-control study	Cases recruited first, after which controls are selected to be similar in all traits to cases apart from the outcome. Cannot establish temporality between exposure and outcome. Can be used to study associations with (usually) rare disease. More expensive and lengthy than cross-sectional studies.	Only shows an association, does not estimate a causal effect. Can be used as an initial study to give evidence for possible causal factors. Susceptible to bias such as recall and selection bias.	Analytic

Study design	Description	Evidence description	Type
Cross-sectional study	Snapshot of a study sample at a point in time. Studies individuals. Can assess the association between an exposure and an outcome Fast and cheap.	Cannot establish a causal relationship due to the exposure and outcome being studied at the same time.	Descriptive
Correlational/Ecological study	Studies groups rather than individuals. Provides the correlation between two variables of interest.	Could provide evidence of causality, althoguh unable to distinguish if other sources contribute to the association.	Descriptive
Case-series	Similar to a Case-report, where multiple cases are studied.	Slightly better than Case-reports.	Descriptive
Case-report	A single case is studied. Patient usually has a certain disease manifestation or a syndrome. No controls. Does not test a hypothesis. Fast and very affordable. Can generate new hypotheses.	Lowest potential in establishing causality.	Descriptive

Observational studies have been valuable in establishing causality in well-known cases such as smoking and lung cancer, hepatitis B and liver cancer [184]. However, these types

of studies are susceptible to biases such as confounding that can either provide a false-positive or mask a true causal effect [185].

Confounding is when a variable affects both the exposure and the outcome, which then leads to a non-causal association to be identified between the two traits (**Figure 2-1A**) [179]. The confounding variable can be added as a covariate (i.e. controlled for) in the statistical model, although knowing all potential variables that one should adjust for is an almost impossible task [186]. A specific type of confounding is reverse causation, where an association is seen between an exposure and an outcome, but this is due to the outcome affecting the exposure rather that the other way around (**Figure 2-1B**) [187].

Other cases that affect association statistics exist. One is collider bias, where controlling for a variable that is affected by both the exposure and the outcome can introduce a spurious association between the two [188] (**Figure 2-1C**). Sampling bias is also an issue, where participants recruited in a study are not representative of the whole population and limits the generalizability of the results [188].

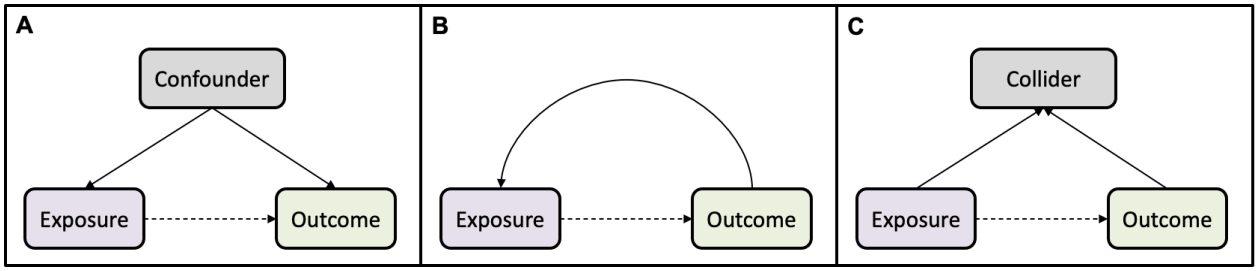


Figure 2-1. *Common causes of bias in observational studies.*
Confounder introducing a non-causal association between the exposure and outcome (A); reverse causation, where the association is due to the outcome affecting the exposure (B); collider bias, where both the exposure and outcome act on the collider variable, introducing bias if controlled for (C).

As they have the greatest capacity to overcome these limitations of observational studies, RCTs are considered to be the "gold standard" in establishing causality [189]. Nevertheless, they also come with some caveats (**Table 2-1**). More specifically, in the case of BCTs and disease, a RCT would most likely be an unrealistic proposition. For example, it would not be practical, let alone ethical, to give patients drugs that alter BCTs, as it would likely have detrimental health consequences. Moreover, it would be costly to investigate the relationship between changes in BCTs and e.g. cancer, as the study would have to run for many years for cancer to develop.

2.2. Mendelian randomization

If one aims to study the relationship between BCTs and disease with the aim of establishing a causal relationship, employing novel methods is required. MR is a one such method in genetic epidemiology that makes use of Mendel's 2nd ascribed law of genetics [169], which refers to the random allocation of alleles that takes place during meiosis [190]. As mentioned in the previous chapter, a GWAS analysis studies the association between SNP allele variation and unit changes in a trait of interest [154]. MR uses these SNPs associated with a trait to proxy for an exposure (i.e. the GWAS trait) to estimate a causal effect between an exposure and an outcome [169]. As the alleles for the SNPs instrumenting an exposure are randomly assigned at conception, this makes MR work akin to a RCT [191] (**Figure 2-2**).

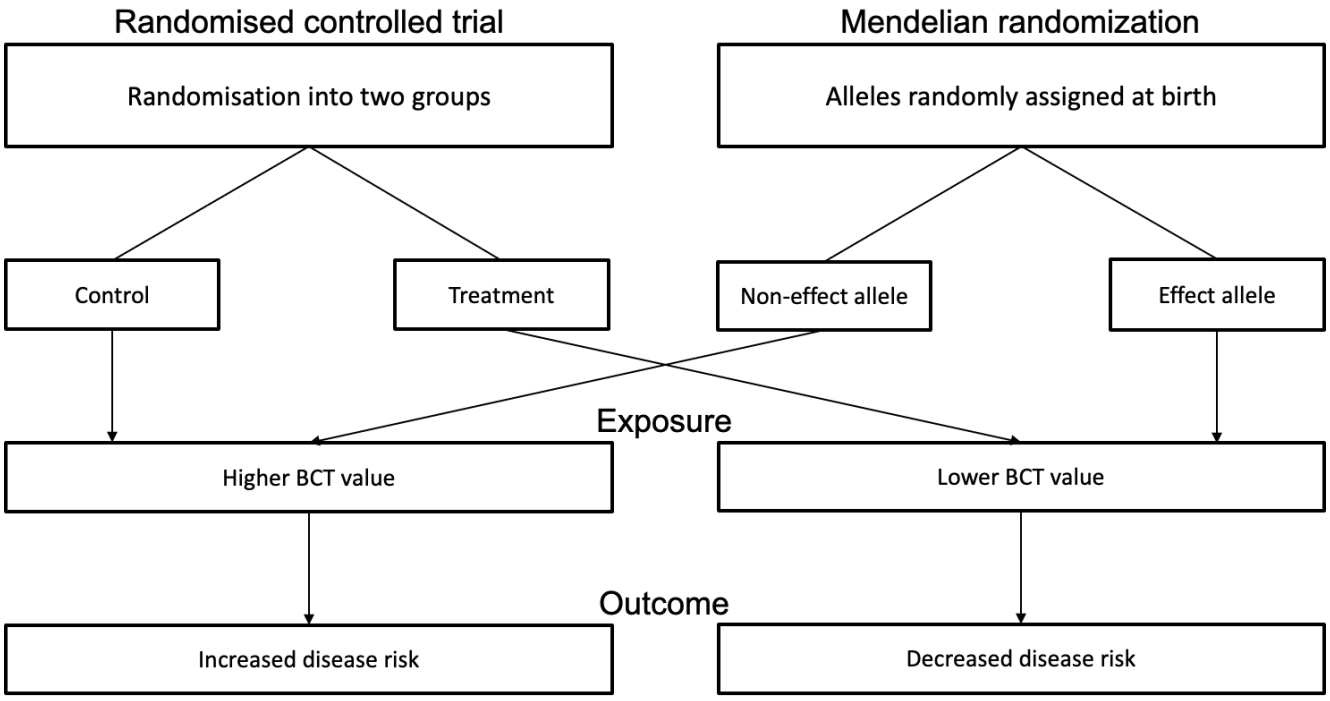


Figure 2-2. *MR compared with a RCT.*
In MR, the RCT equivalent of the control group is the non-effect allele, while the equivalent of the treatment (exposure) group is the effect allele.

In the context of existing study designs in epidemiology, MR is a relatively new method that under certain assumptions [192] can be used in the discovery of novel risk factors or for testing previously known associations [193,194]. The capacity for MR to establish a causal relationship makes it a powerful method that can be useful when conducting a RCT is not possible, either due to cost, ethics, or practicability [189]. As described previously, this is the case for blood cell traits (BCTs), making MR the next best method for causal inference in BCTs and disease (**Figure 2-3**) [171,180–183].

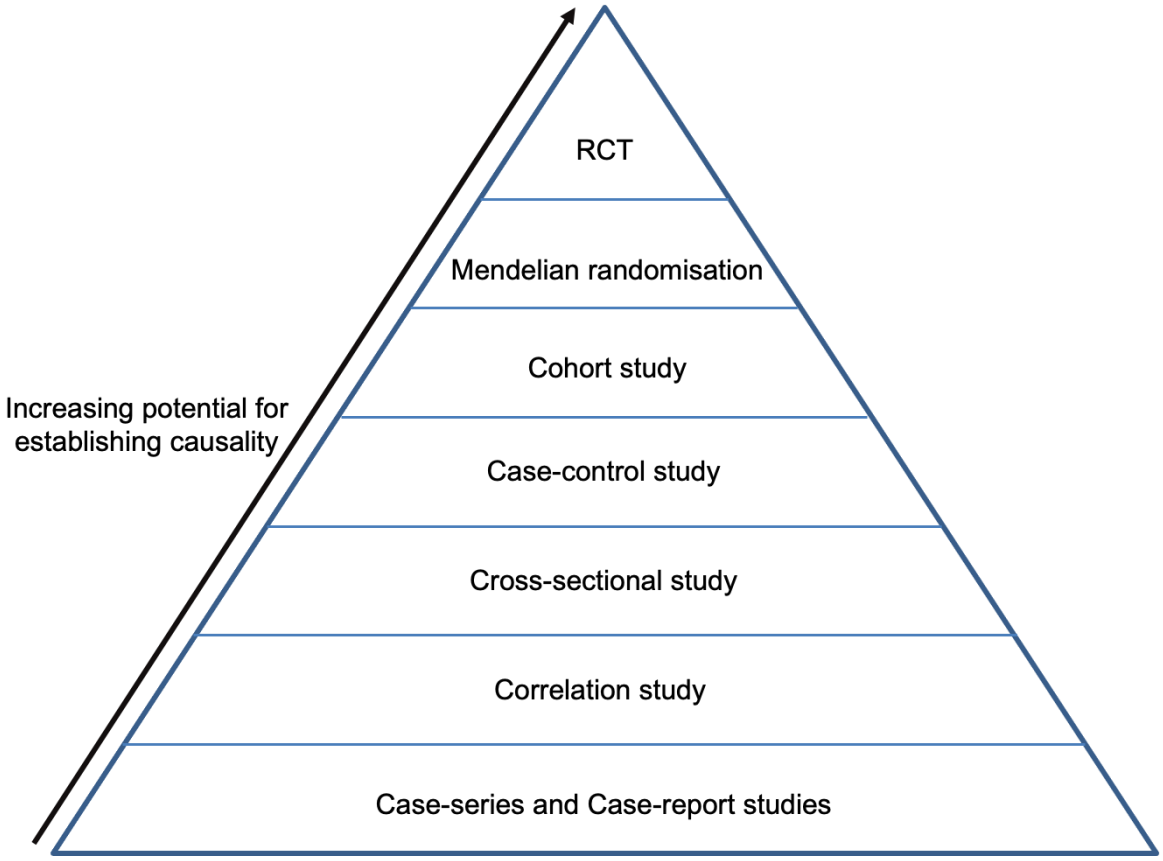


Figure 2-3. *MR in the context of biomedical study designs.*
Note that this hierarchical pyramid assumes that each study is done appropriately – there can be poor RCTs that would rank lower than e.g. a cohort study.

MR can provide a causal estimate if and only if the following conditions are met [169,192] (**Figure 2-4**):

1) The genetic variant (G) is a valid instrument, in that it is reliably associated with the exposure. By valid instruments I refer to genotypes to act as instruments which offer properties close to those necessary for running a MR analysis. A general rule is a SNP with an association P-value with the exposure of < 5e-8 [195].
2) There is no independent association with the outcome, except through the exposure.
3) The instrument is independent of any measured or unmeasured confounding factors.

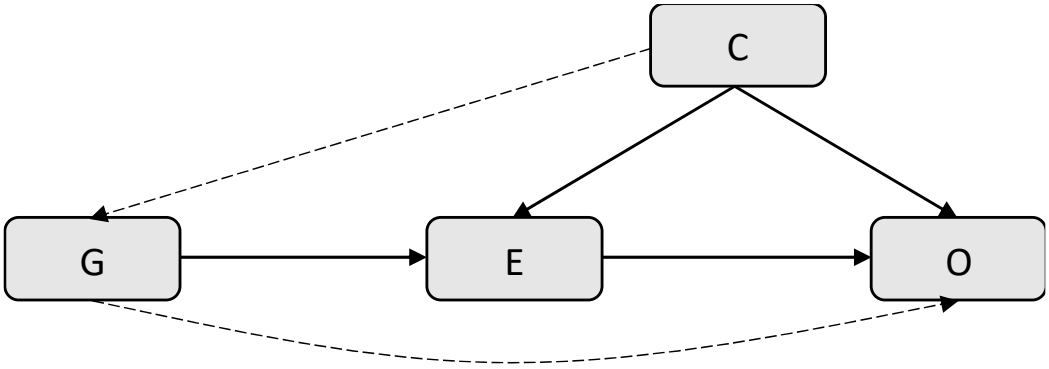


Figure 2-4. *MR assumptions.*
MR works in a similar way to a RCT, exploiting the essentially random allocation of alleles at conception and the independent assortment of parental variants at meiosis. MR uses genetic variants (G) as proxies (instruments) to investigate whether an exposure (E), is causally associated with an outcome (O).

2.2.1. One-sample MR

The first MR studies were conducted in a one-sample (1SMR) manner i.e. both the exposure and outcome data come from the same cohort, and individual-level data is used [196,197]. In current 1SMR studies, a polygenic risk score (PRS) for each individual in the dataset is created by summing-up all the risk alleles (0, 1 or 2) for all *k* SNPs used as instruments for the exposure [198]. Afterwards, a two-stage least-square regression (adjusting for other variables) is conducted between the exposure PRS and the outcome, estimating an effect [191].

2.2.2. Two-sample MR

With the rise in publicly available GWAS, a further development to MR was made – the two-sample MR (2SMR) method, where the exposure data and outcome data come from separate GWAS [199]. Here, an additional requirement is that the participants from the exposure and outcome GWAS are from the same population (e.g. both datasets are from individuals of European ancestry) and there should not be any overlap between the two studies [200].

The sample overlap requirement is more nuanced, however. If the exposure and outcome effects are estimated using the same participants, the resulting bias aligns with the confounded observational association [201]. In contrast, when the datasets for exposure and outcome are entirely distinct, the bias tends towards the null, while for datasets with partial overlap, the bias lies somewhere between these two extremes [201].

The particularities of 2SMR give it certain advantages over 1SMR. For example, it is usually more accessible, statistically powerful and cost-efficient due to summary-level data for the exposure and outcome coming from two separate studies [197]. While these benefits make 1SMR a more expedient method, they don't necessarily imply that 2SMR is less biased than 1SMR [200]. In the next section I will explore the most common methods used in 2SMR studies.

Due to its popularity, many methods have been designed for 2SMR analyses to either test the assumptions of MR or act as extensions to the method [200]. In the next section I will explore the most common methods used in 2SMR studies.

2.3. Common Mendelian randomization methods

2.3.1. Wald ratio

In 2SMR, the effect of one single-nucleotide polymorphism (SNP) is estimated using the Wald ratio (WR) method, which is the two-stage least squares 1SMR estimate when only one SNP is available to use as an instrument for the exposure [189]. The WR is calculated through the formula:

$$\hat{\beta}_{IV} = \frac{\hat{\beta}_{GY}}{\hat{\beta}_{GX}}\text{ , where:}$$

$\hat{\beta}_{GY}\ is\ the\ effect-size\ of\ the\ SNP-outcome\ association$ or Γ

$\hat{\beta}_{GX}\ is\ the\ effect-size\ of\ the\ SNP-exposure\ association$ or γ

2.3.2. Inverse-variance weighted

The most common method used in MR when more than one SNP is available as an instrument is the fixed-effects inverse-variance weighted (IVW) method. It represents the meta-analysis of all ratio estimates (WRs) and works akin to the meta-analysis employed in traditional epidemiological approaches [202]. The MR IVW weights each SNP by the inverse of the variance of the SNP-outcome association [202]. This weighting is done under the no measurement error assumption of the SNP-exposure association [203].

Given *k* number of SNPs used to instrument for an exposure [204]:

$$\Gamma_k = \beta\gamma_k$$

$$W_k = \frac{1}{\sigma_\Gamma^2}\ ,where:$$

$$W_k \text{ is the weight for SNP } k \text{ and}$$
$$\sigma_{\Gamma}^2 \text{ is the the variance of the SNP - outcome association for SNP } k$$

Given these, the formula for the fixed-effects IVW method is [202]:

$$\hat{\beta}_{\mathrm{IVW}} = \frac{\sum_{k=1}^{K} \hat{\beta}_k W_k}{\sum_{k=1}^{K} W_k} \quad ,$$

and a visual representation of the equation is showed below in **Figure 2-5**.

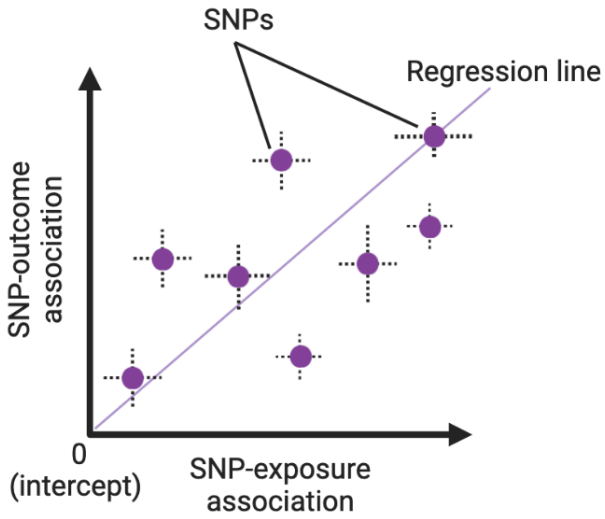


Figure 2-5. *IVW estimate example.*
Each purple dot on the scatterplot represents a SNP used to instrument for the exposure, along with the zero-constrained regression line. The dotted lines represent the effect size and 95% confidence intervals (CIs) for both the SNP-exposure (x-axis) and SNP-outcome (y-axis) association. Made with BioRender.com.

The resulting beta ($\hat{\beta}_{IVW}$) can be exponentiated to generate an odds ratio (OR), and combined with the standard error (SE), can then provide the CIs, aiding in the interpretability of the MR results.

The IVW method has the most power to detect an effect between an exposure and an outcome, as it assumes that all the SNPs used in the MR analysis are valid [205]. This comes with its disadvantages, such as increased susceptibility of the MR analysis to bias [194]. There are some possible ways in which MR estimates can be biased due to invalid instruments [195]. One common type of bias is weak-instrument bias and happens when the SNPs explain too little of the variance in the exposure [204,206]. The conditional F-statistic is calculated for each SNP and a rule of thumb is that a F < 10 is indicative of

potential weak-instrument bias [206]. Another common type of bias can arise due to the presence of horizontal pleiotropic SNPs [207].

2.3.3. Vertical and horizontal pleiotropy

The two common forms of pleiotropy discussed in MR studies are vertical and horizontal pleiotropy [208]. In the case of vertical pleiotropy, the SNP used to instrument for the exposure is acting downstream of the exposure through a biological intermediate to affect the outcome [194,208]. This does not invalidate MR assumptions, as the SNP is still acting through the exposure, and is in actuality necessary for MR [197]. The presence of vertical pleiotropy is indicated by Cochran's Q statistic for heterogeneity applied to the MR method [197].

On the other hand, horizontal pleiotropy is when a SNP does not act on the outcome through the exposure, but rather through a different biological pathway [194]. The inclusion of one or more horizontal pleiotropic SNPs can bias the IVW MR estimate if the horizontal pleiotropic effect is unbalanced i.e. the sum of the pleiotropic effects don't cancel-out in both directions [195]. The issue of detecting and dealing with horizontal pleiotropy led to the development of methods to detect and correct for it.

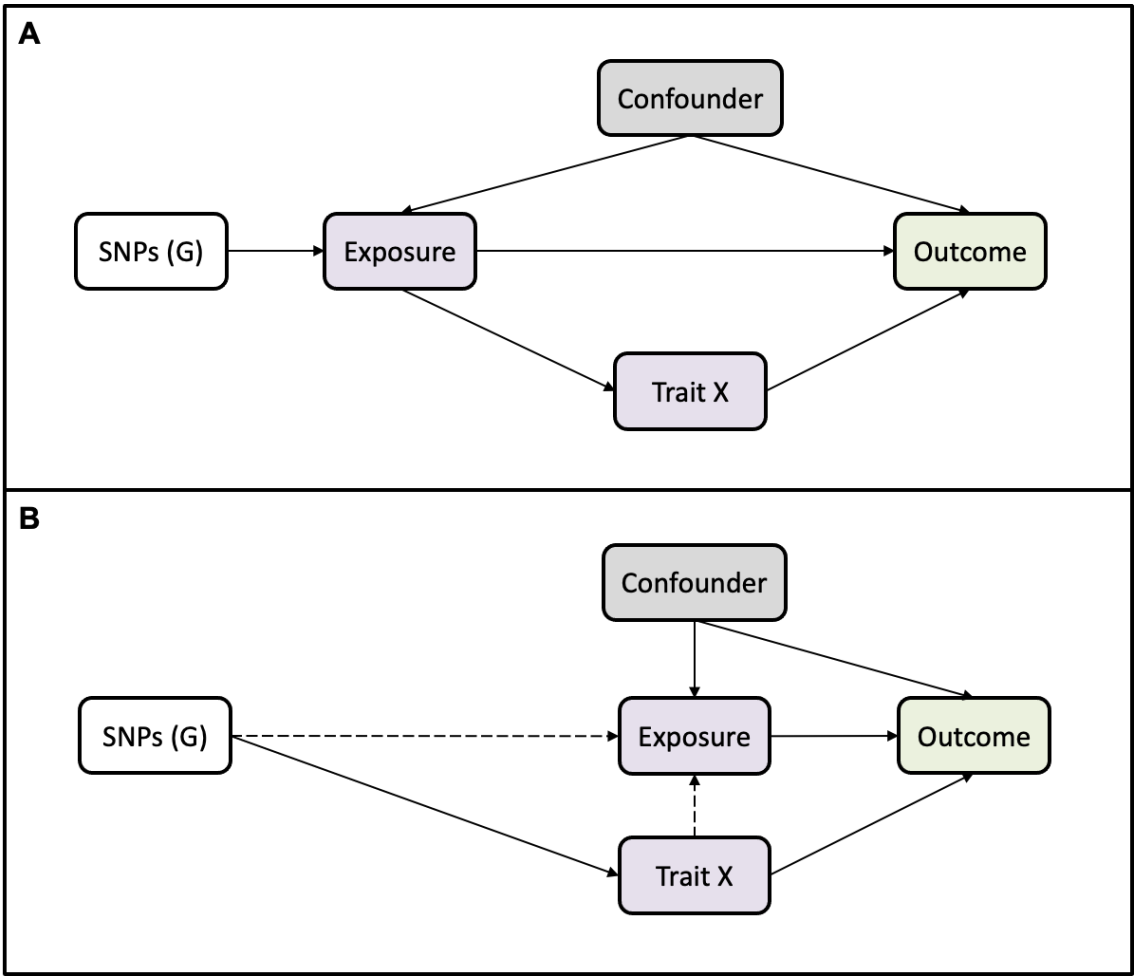


Figure 2-6. *Vertical and horizontal pleiotropy in MR.*
*Vertical pleiotropy, where genetic proxies (SNPs – G) act only through the exposure, and the Exposure either affects the Outcome directly or indirectly downstream through another trait 'Trait X'; this does not invalidate MR assumptions (**A**) . Horizontal pleiotropy, where genetic proxies act fully or in part through trait 'Trait X', which then affects the outcome directly or indirectly downstream through the Exposure; this does invalidate MR assumptions (**B**) .*

2.3.4. MR-Egger

One such method is MR-Egger, first described by Bowden et al. [205]. As was the case with the IVW method, MR-Egger actually has its roots in the Egger method used to identify small-study bias in meta-analyses described by Egger et al. [209]. The MR-Egger assumptions *2* and *3* are the same as those in **Figure 2-4**, while the *1st* assumption is replaced with a more relaxed instrument strength independent of direct effect (inSIDE) assumption [205]. The latter states that if horizontal pleiotropic SNPs are present, these are equally distributed across separate biological pathways [205].

Unlike the IVW method, the regression intercept is not constrained to zero, but rather estimated as part of the analysis [210]. If the intercept does equal zero, then the MR-Egger estimate is equivalent to that of the IVW method and suggests that horizontal pleiotropy is either not present or is balanced, while the opposite applies when the slope is non-zero [210].

While MR-Egger regression is valuable in assessing the presence of horizontal pleiotropy, the method has some limitations. For example, the power to detect an effect is much lower compared to the IVW method, and therefore many SNPs are required to proxy for the exposure for a reliable estimate [211]. Moreover, the inSIDE assumption can be invalidated if all pleiotropic SNPs act through the same pathway [210]. Therefore, additional sensitivity MR methods were further developed to be used in parallel with the IVW and MR-Egger analyses.

2.3.5. Weighted median

Another sensitivity method developed to investigate horizontal pleiotropy is the weighted median [204]. In brief, a distribution is generated from the ratio estimates, and each ratio estimate contributes to this distribution based on their W_k weight [204]. The 50th percentile (median) of this distribution represents the weighted median estimate [204]. Unlike MR-Egger, the weighted median assumes that at least 50% of the weight comes from valid instruments [204], giving it more power to detect an effect while also being less susceptible to pleiotropic outliers than the IVW method [211].

2.3.6. Weighted mode

The weighted mode approach is another sensitivity MR method developed by Hartwig et al. [212]. The same empirical distribution method as in the weighted median is applied, although in this case the estimator is the mode (most frequent) ratio estimate [212]. Unlike the median-based estimator, the weighted mode can provide an estimate for an effect even when only a plurality (<50%, can vary) of the SNPs are valid instruments, as only the modal SNPs have to be valid instruments [212]. While this decreases the influence of pleiotropic outliers on the MR result, it also means that the mode-based estimate has lower power compared to the IVW and weighted-based estimates [212].

2.3.7. MR-PRESSO

Another sensitivity analysis is the MR Pleiotropy RESidual Sum and Outlier (MR-PRESSO) method [207]. The first test detects if there is evidence of horizontal pleiotropy and identifies the possible outlier pleiotropic SNPs through a residual sum of squares approach [207]. After detection of these outliers, an IVW analysis is run again on the remaining SNPs, and the result is compared with the initial uncorrected MR to assess if these outliers influence the MR results to a large degree, which can inform if horizontal pleiotropy is responsible for the unadjusted effect [207]. Like MR-Egger, MR-PRESSO has the inSIDE assumption, and that at least 50% of the SNPs are valid instruments for the exposure [207].

2.3.8. MR-Steiger

Another sensitivity 2SMR method is MR-Steiger, which is an extension of Steiger's Z-test of correlated correlations [213,214]. It is suitable when conducting MR using large-scale GWAS summary statistics [213,214]. In essence, the method assumes that the variance explained by the SNPs used in the MR is higher for the exposure than for the outcome [213,214]. Based on this, it can suggest if one or more SNPs used in the MR are reverse causal i.e. act on the outcome to affect the exposure, and eliminates those from the main MR analysis if that is the case [213,214].

2.3.9. Multivariable MR

Sometimes in a MR analysis, the exposure of interest is genetically correlated with other traits [215]. For example, 30% of the SNPs used to proxy for exposure A might also be used to proxy for exposure B, making it more difficult to untangle the biological mechanism that affects an outcome. Therefore, an extension of MR is multivariable MR (MVMR), developed to account for the shared SNPs between two or more exposures of interest and to estimate their direct effect on the outcome [215,216] (**Figure 2-7**). This is useful in the study of the direct effects of e.g. WBC subtype counts on disease, as these are known to be genetically correlated [166].

Moreover, MVMR can also be used to correct for the presence of confounders that might bias effect estimates. Instead of trying to estimate the direct effect of two correlated traits, the exposure's estimate can be corrected by accounting for the shared genetic instruments with the confounder [216].

MVMR shares the same assumptions and limitations of univariable MR [216]. There are, however, additional limitations and assumptions that MVMR can have in addition to univariable MR. For example, weak instrument bias can also take place if the proportion of variance explained by SNPs instrumenting for trait A also explain a high degree of the variance in trait B, as the remaining trait B-specific SNPs used in the MVMR might not be enough to give a reliable MR result [217]. Additional limitations apply to each added exposure in the MVMR analysis, including the assumption of linear and homogeneous effects of the exposure on the outcome [216].

In their recent publication, Sanderson et al. provide an R-package that can perform sensitivity analyses for MVMR as well, such as heterogeneity and conditional F-statistic values, which are useful in assessing the validity of the MVMR estimates [217]. However, this requires a covariance matrix of the effect of each SNP on each exposure or a correlation matrix between the phenotypes, which can only be calculated from individual-level data [217]. Otherwise, the covariance is assumed to be 0 [217].

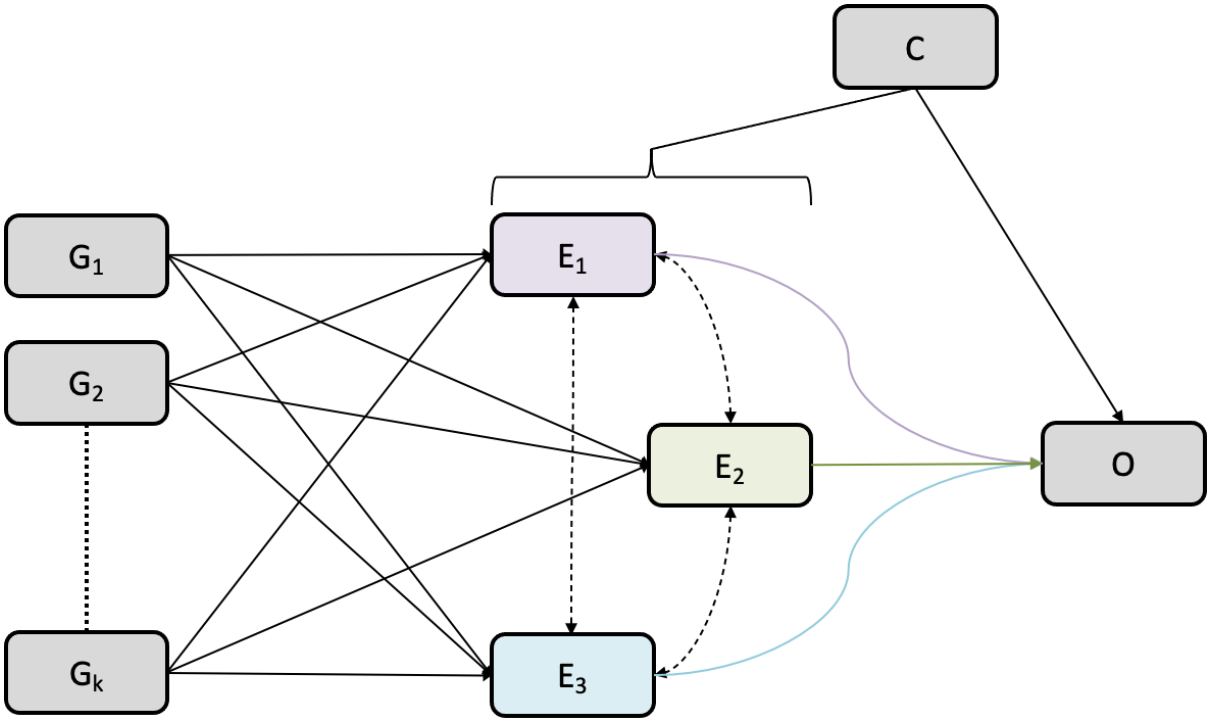


Figure 2-7. *Multivariable MR schematic.*
The concept is similar to that of univariable MR. Given k number of SNPs (G) proxying for exposures E1, E2, and E3, MVMR can estimate the direct effect of each exposure on the outcome (O) of interest.

2.3.10. MR for mediation analyses

Mediation analysis is a method now integrated in MR that was traditionally used in observational epidemiology to investigate intermediary factors that lie on the pathway between an exposure and an outcome, and is useful in understanding the aetiology of e.g. disease [218].

There are two ways a mediation MR can be performed. The first one is also known as two-step MR, where the exposure and the mediator each use separate SNPs as proxies [218]. The second one is through MVMR, where the exposure and the mediator share a combined genetic instrument repertoire, and the direct effect of the exposure on the outcome is estimated [218]. The mediation MR analysis assumes no interaction between the exposure and the mediator, and that there are no other pathways through which proxy SNPs act apart from the exposure and mediator [218].

The two-step MR and MVMR approaches allow for the estimation of the indirect effect of the exposure on the outcome, which can be calculated using the product of coefficient method or the difference method [218]. The product of coefficients is essentially the effect of the exposure-mediator MR multiplied by the meditator-outcome MR, while the difference is given by subtracting the direct effect of the exposure-outcome MR from the direct effect of exposure-outcome MVMR result [218]. The indirect effect can then be used to calculate the proportion mediated by the mediator in the exposure-outcome relationship. This is only possible when the indirect and total effect are in the same direction and is given by the following formula: indirect effect / total effect * 100 (**Figure 2-8**) [218].

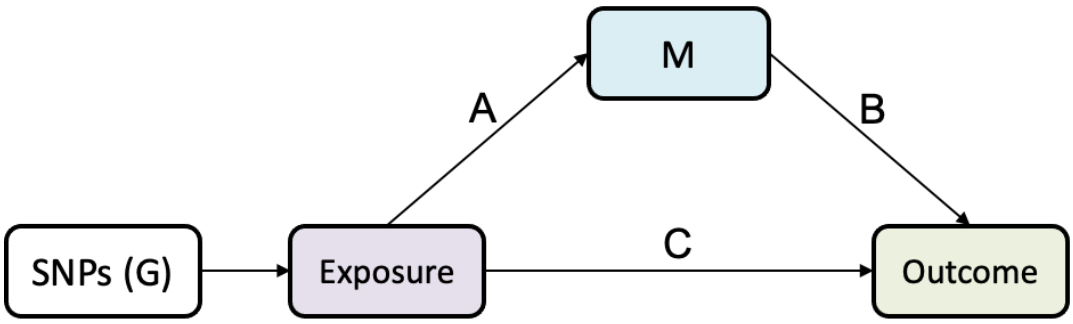


***Figure 2-8**. Two-step mediation MR schematic.*
The exposure of interest is proxied through SNPs (G), while the mediator is proxied through its own instruments. The effect of the exposure on the mediator is estimated (A). Afterwards, a MR analysis is done between the mediator and the outcome (B). Next, the total effect of the exposure on the outcome is estimated (C). This then allows for the

calculation of the indirect effect and estimation of proportion mediated by the mediator in the exposure-outcome causal relationship. In MVMR, the direct effect C' is estimated, after which the difference method is applied to calculate the indirect effect.

2.4. GWAS summary statistics

To run a 2SMR analysis, one needs to have access to the summary statistics of both the exposure and outcome GWAS [219]. In the previous chapter, I briefly described the concept of GWAS and how they are conducted. By "summary statistics", one refers to a dataset which contains the minimum amount of information (non-identifiable) that can help the reader understand which, in which direction, and to what degree a SNP is associated with a particular trait [220]. Most GWAS report summary statistics as a file with at least the following columns: the ID of the tested SNP (e.g. rs123 or 1:1231231_A_T if no rsID present), the chromosome number and base-pair position to which the SNP maps to, the effect allele (studied allele in relation to the trait), non-effect allele, effect allele frequency, beta coefficient of the regression, standard error, P-value, and sample-size [167].

2.5. Non-MR analytical approaches

2.5.1. Population genetics tools

ADMIXTURE. Software designed to map individuals from a sample to a *k* number of populations given two or more reference datasets where the populations are already known [221,222]. This allows for the estimation of the percentage ancestry compared to the reference datasets for each individual in the study sample [221,222].

Principal component analysis. Is a statistical method designed to reduce the number of dimensions to a set of uncorrelated variables called principal components [223]. Interestingly, PCs generated from genetic data have been found to correspond to geographical locations on a map, indicating the usefulness of PCs as measures of population structure [224,225]. Software such as EIGENSOFT are able to take as an input people's genetic data to generate PCs that are then loaded and displayed with the R and Python programming languages, as well as being added as covariates in GWAS to reduce bias due to population stratification [226–229].

K-means clustering. K-means clustering is an unsupervised classification algorithm used in machine learning that partitions *n* observations into *k* clusters in such a way that

the data points within a cluster are more similar than those (outside) in another cluster [230]. This method was used in **Chapter 4**.

Fixation index. Estimations of fixation index (Fst), which range from 0 to 1, provide a measure of population differentiation among populations, which describe the proportion of total variation at a SNP that is explained by variation between populations [231]. For any SNP, a value of 0 would indicate that minimal variation is attributable to variation between populations, while a value of 1 would indicate a fixed difference i.e., the two populations are both invariable but for alternative alleles [231]. This method was used in **Chapter 4**.

2.5.2. GWAS software

SNPTEST. Is a software used to run linear models for GWAS [232]. In essence, it represents a linear regression run between the dosages of the effect allele for each SNP included in the model (0, 1, or 2) and a phenotype. Additional variables e.g. genetic sex, age, can be added as covariates in the model.

META/METAL. META [232] and METAL [233] are software in statistical genetics that can take multiple GWAS conducted on the same trait and meta-analyse them into a single data table. These can either run the fixed-effects or random-effects IVW method of meta-analysis.

BOLT-LMM. Is a software designed for running GWAS through a linear mixed model approach. It builds a genetic relationship matrix to adjust for population relatedness that can bias traditional linear model GWAS analyses [234]. A GRM is essentially a matrix with *n* rows and *y* columns, where n = number of individuals in sample and y = number of SNPs [235]. The *ny* matrix contains the minor allele counts for each SNPs of each individual [235234]. Additionally, it was designed designed to be very fast when running on datasets with hunderds of thousands of inidivduals, such as UKBB. Like SNPTEST, it can adjust for covariates such as genetic sex, age and other traits of interest. All software described in this subsection were used in **Chapter 5**.

2.6. Data sources

2.6.1. UK Biobank

The UK Biobank (UKBB) is an ongoing prospective cohort study conducted in the United Kingdom [161,164]. The aim behind UKBB was the recruitment of hundreds of thousands of individuals to provide a centralized data source for scientists to perform biomedical

research that would have enough power to detect even disease risk factors with a small effect-size on disease development [161]. Participants were recruited between the years 2006-2010 and were 40–69 years old, as the aim of the study was to select participants prior to developing diseases that occur later in life, such as cancer [164].

An extensive amount of data was gathered from each study participant attending UKBB recruitment centres throughout the UK [161,164]. Here, participants answered a long list of questions, ranging from dietary patterns, to personality traits, to self-report disease status, and linkage with external datasets such as medical records was also done [164]. One of the steps of the data collection procedures was the collection blood samples, which were analysed to produce information on BCTs [164]. Participant blood samples were analysed using four Beckman Coulter LH750 instruments designed for high throughput screening [236], which employ the Coulter method for blood cell measurement [137] discussed in the previous chapter. Total white blood cell (WBC) count and WBC subtype percentages (%) were measured through the sampling of the blood, with absolute WBC subtype count derived as "WBC subtype % / 100 x total WBC" and expressed as 10^9 cells/Litre [236].

Apart from gathering of phenotypic data, the genomes of the participants were sequenced using two arrays specifically made for the study: the UKBB Axiom Array and UK BiLEVE array, which were used to directly genotype around 800K SNPs [164]. Further analyses were done, such as PC generation and multiple imputation of millions of genetic variants, allowing researchers to run GWAS [164]. UKBB data was used across all chapters of my thesis. All participants provided written informed consent, and each study was approved by the relevant research ethics committee or institutional review board. UK Biobank received ethical approval from the NHS National Research Ethics Service North West (11/NW/0382; 16/NW/0274) and was conducted in accordance with the Declaration of Helsinki. Individual-level UKBB data from application ID "81499" were used for **Chapter 3**; for **Chapter 4** and **Chapter 5**, these were provided as part of application ID "15825". I am a co-investigator in both applications, which cover all the UKBB data I used in this thesis for each respective chapter.

2.6.2. White blood cell count data

Summary statistics for WBC count and subtypes were obtained from a recent study by Chen et al. as part of the "Blood Cell Consortium" (BCX) [166]. These were generated through a meta-analysis of GWAS previously conducted in people of European, African, East Asian, South Asian and Hispanic-American ancestries, with a total sample-size of

746,667 participants [166]. Genetic sex, age, age^2, study-specific variables and principal components (PCs) 1 to 10 were used as covariates. Effect estimates were meta-analysed by ancestry, and in a trans-ancestry setting [166]. A brief description of each meta-analysed study is available in **Appendix 1**. Specific details on QC steps and association testing are available in the source manuscripts. This dataset was used in **Chapter 3** and **Chapter 5** [237,238]. All participants provided written informed consent, and each study included in the meta-analysis (see **Appendix 1**) was approved by the relevant research ethics committee or institutional review board [166]. Summary statistics for WBC counts were downloaded from the following website: http://www.mhi-humangenetics.org/en/resources/.

2.6.3. Colorectal cancer data

The GWAS summary statistics for CRC and its anatomical subtypes come from the most comprehensive meta-analysis of CRC risk to date [163,239]. In the first stage, 45 studies part of the Genetics and Epidemiology of Colorectal Cancer Consortium (GECCO), Colorectal Cancer Transdisciplinary Study (CORECT), and Colon Cancer Family Registry (CCFR) consortia were meta-analysed (34,869 cases and 26,783 controls) [240–242]. In the second stage, an additional 24 studies were included [163]. The final sample was predominantly of European ancestry, with ~5% representing East Asians due to their similar CRC genetic architecture [243]. Genetic sex, age, study-specific variables, and PCs were used as covariates. An overview of all consortia included in the CRC meta-analysis is available in **Appendix 2**. The summary-level GWAS data for CRC used in this study were made available following an application to the GECCO. This resource was used in **Chapter 3** [237]. All participants provided written informed consent, and each study included in the meta-analysis (see **Appendix 2**) was approved by the relevant research ethics committee or institutional review board [163,239]. The summary-level GWAS data on outcomes used in this study were made available following an application to the Genetics and Epidemiology of Colorectal Cancer Consortium (GECCO): https://www.fredhutch.org/en/research/divisions/public-health-sciences-division/research/cancer-prevention/genetics-epidemiology-colorectal-cancer-consortium-gecco.html.

2.6.4. Allergic disease data

Summary statistics for allergic disease were obtained from a meta-analysis of 13 GWAS done in people of European ancestry by Ferreira et al. [244]. Cases were defined as those who self-reported suffering from at least one of the following: asthma, hay fever, or

eczema – while controls were defined as those who did not report any of these afflictions. This resulted in a final sample-size of 360,838 (cases = 180,129, controls = 180,709). Included covariates and specific numbers for each meta-analysed study is available in **Appendix 3**. This dataset was used in **Chapter 3** [237]. All participants provided written informed consent, and each study included in the meta-analysis (see **Appendix 3**) was approved by the relevant research ethics committee or institutional review board [244]. Exposure summary statistics for allergic disease can be downloaded on the manuscript's journal web page: https://doi.org/10.1038/ng.3985.

2.6.5. Severe malaria data

GWAS summary statistics for *P. falciparum* severe malaria were downloaded from a case-control study that spanned nine African and two Asian countries [245]. In brief, controls samples were gathered from cord blood, and in some cases, from the general population. Cases were diagnosed according to WHO definitions of severe malaria [246]. Summary statistics were made available on the following link: https://www.malariagen.net/sppl25/.

2.6.6. MR-Base

MR-Base is an online platform which automates the process of conducting MR analyses. It is also available as a standalone R package "TwoSampleMR". The platform/package connects to the OpenGWAS database (see next subsection). There are two main steps undertaken prior to an MR analysis that the "TwoSampleMR" package automates.

Genetic confounding may bias MR estimates by double counting SNPs that are in LD [187]. Therefore, the TwoSampleMR package runs the PLINK genetic software [247] to clump the SNPs used to proxy for the exposure. Clumping is a process in which SNPs in LD (based on the correlation coefficient threshold r^2) over a genomic window of X kilobases (kb) are removed and only the SNP with the lowest GWA P-value is kept [247]. The default clumping parameters used in MR studies are stringent (radius = 10,000 kb; r^2 = 0.001; P-value = 5e-8) [219] and most use the 1000 Genomes (1KG) reference panel [248].

Another pre-MR analysis step is harmonisation [249]. The majority of GWAS present the effects of a SNP on a trait in relation to the allele on the forward strand [250]. However, the allele present on the forward strand can change as reference panels get updated [250]. This requires correction (harmonisation) so that both exposure and outcome data reference the same strand [250]. For exposure and outcome data harmonisation, incorrect

but unambiguous alleles are corrected, while ambiguous alleles are removed. In the case of palindromic SNPs (A/T or C/G), allele frequencies are used to solve ambiguities. These resources developed at the integrative epidemiology unit (IEU) Bristol were used across all chapters of this thesis.

2.6.7. OpenGWAS

OpenGWAS is an online platform which allows the user to access over 14,000 harmonized GWAS summary statistics through a programming language (R and Python) interface [251]. It is linked with other packages and platforms, such as the online MR-Base database or the TwoSampleMR R package [219], allowing for efficient and standardised running of MR analyses. For example, in the case of **Chapter 6**, the outcome deep vein thrombosis (DVT) was presented in OpenGWAS as as "Non-cancer illness code self-reported: deep venous thrombosis (dvt)". These summary results describe a GWAS of Europeans (6,767 cases and 330,392 controls) conducted using UK Biobank data by Benjamin Neale and colleagues (http://www.nealelab.is/uk-biobank).

2.6.8. The 1000 Genomes Project

Another dataset I employed was the 1000 Genomes Project (1KG), which I also mentioned in the previous chapter [152]. The latest iteration of the project contains whole-genome sequence (WGS) data for 2,504 individuals from five continents, each divided into subpopulations by the location of sampling within the continent e.g. Kenya and Nigeria for the African continent [248]. The 1KG has been extensively used as a reference panel due to its valuable diverse WGS data, such as in the imputation of UKBB SNPs [248] or removing SNPs in linkage disequilibrium (LD, i.e. correlated) prior to MR analysis to avoid double-counting [170]. The 1KG data was used across all chapters of my thesis – either directly, as in the case of **Chapter 4**, or indirectly, in Chapters **Chapter 3**,**Chapter 5** and **Chapter 6**, where it was used as a reference panel when clumping SNPs prior to conducting the MR analyses.

In this chapter I described the methods and data sources that were part of my thesis. The next four chapters encompass the results that have been generated from my work. The first one is **Chapter 3**, where I studied the relationship between circulating white blood cells and colorectal cancer using Mendelian randomization [169].

CHAPTER 3. CIRCULATING WHITE BLOOD CELL TRAITS AND COLORECTAL CANCER RISK: A MENDELIAN RANDOMIZATION STUDY

Chapter summary

The aim of this chapter was to assess the relationship between levels of circulating white blood cells (WBCs) and the risk of colorectal cancer (CRC) [237] (**Figure 3-1**). As introduced in the Background (**Chapter 1**), studies using genetic epidemiology to study the causal relationship between WBC count and CRC development have not been explored, and most research on the topic has used traditional observational approaches which suffer from known limitations [191,197]. Therefore, to advance our current understanding of how immune cells affect the development of CRC, I used univariable (UV) [252] and multivariable (MV) [216] two-sample Mendelian randomization (2SMR) [253]. I will start with an introduction to CRC, its health burden and risk factors. Afterwards, the role of WBC count in CRC will be explored, where I will give an overview of what its currently known about each WBC subtype in relation to CRC, and what is lacking. I will discuss the use of MR to address this knowledge gap by outlining its advantages over observational epidemiology. Finally, I will present this study's objective and break it down into smaller aims that I accomplished. The rest of the chapter follows the common paper format of methods, results and discussion, the latter in which I will relate my findings back to the literature.

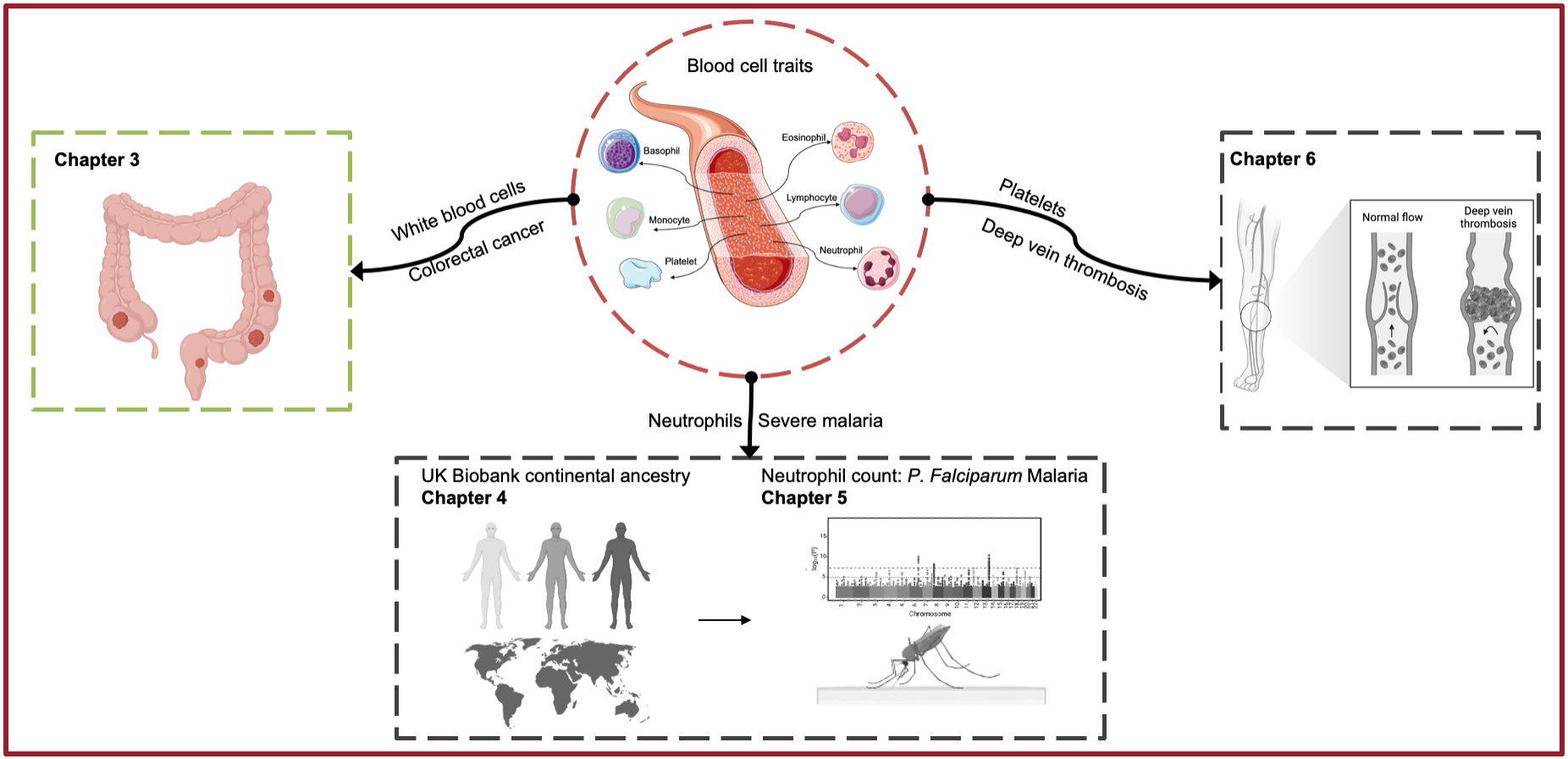


Figure 3-1. *PhD project and current chapter (3 - coloured).*
Created with BioRender.com.

3.1. Introduction

Colorectal cancer (CRC) is the combined term for colon and rectal cancer [254]. It accounts for over 10% of cancer cases worldwide and is the second leading cause of cancer-related deaths globally [255–257]. Overall, the number of CRC cases is rising [258], and in 2020 alone, 1.93 million people were diagnosed with CRC and over 940,000 died as a result of the disease [258]. As living standards and lifestyle patterns in developing countries continue to become more Westernised it is estimated that the number of CRC cases will continue to grow by over 60% through 2035 [259]. Given current challenges, and the estimation that 50% of CRC cases may be preventable [260], focus on identifying novel risk factors, and subsequent prophylactic and treatment options is warranted to limit the future healthcare burden of this disease.

Biologically, the colorectum is divided into anatomical subsites: the cecum, ascending and transverse colon, which form the proximal colon subsite, the descending and sigmoid colon, representing the distal colon subsite, and the rectum [261,262] (**Figure 3-2**). In most cases, CRC develops in the proximal (right side) colon [263] and comes with an increased mortality risk compared to left-sided CRCs [264].

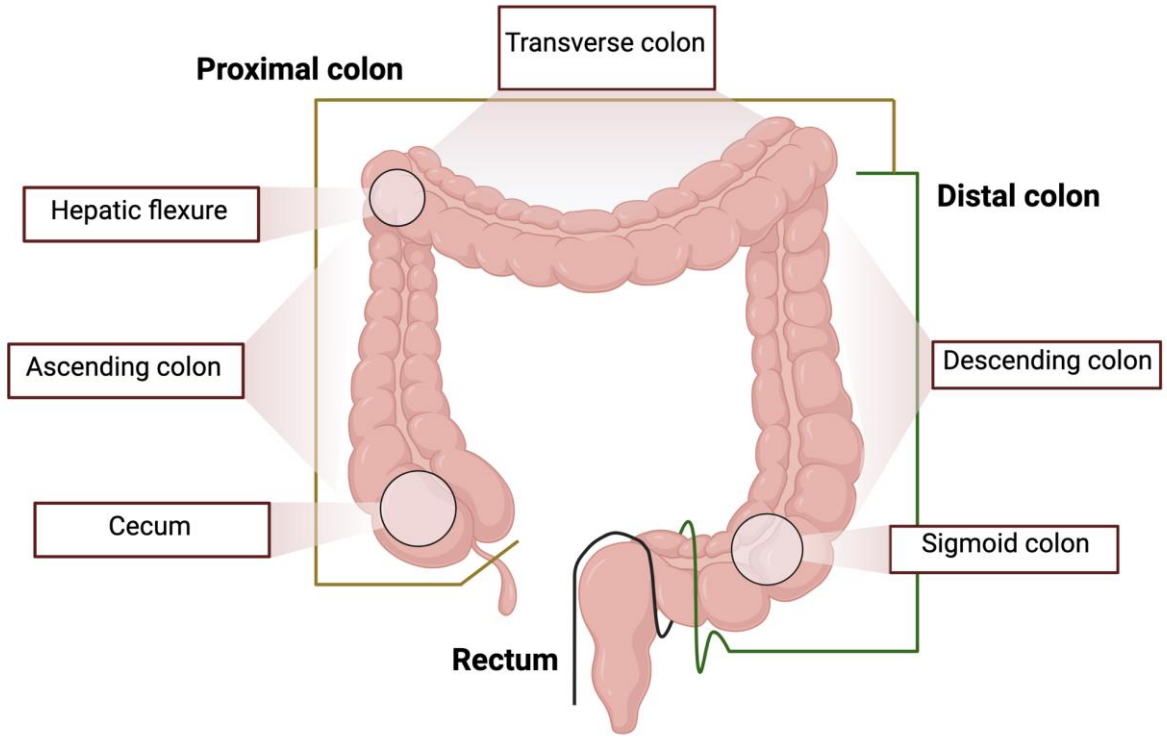


Figure 3-2. *Anatomy of the colon and rectum.*
Adapted from "Colon Callout (Layout)", by BioRender.com (2022). Retrieved from https://app.biorender.com/biorender-templates.

3.1.1. CRC aetiology

CRC is multifactorial in nature, meaning a combination of environmental and genetic factors contribute to disease development [265–267]. Notable modifiable risk factors stated in the latest World Cancer Research Fund (WCRF) report were high body mass index (BMI), red meat and alcohol consumption [268]. Complex traits, such as type 2 diabetes (T2D) and chronic inflammation, have also been cited as established as risk factors for CRC [259,269]. Along with established hereditary conditions, such as familial adenomatous polyposis (FAP) [270] and Lynch syndrome [271], common genetic variants associated with

CRC have also been identified by genome-wide association studies (GWAS) [272]. Early efforts by Tomlinson et al. in 2007 led to the first known GWAS of CRC, finding one single-nucleotide polymorphism (SNP) associated with disease risk: rs6983267, mapped to the 8q24.21 locus and the *CCAT2* gene, a SNP which had been previously associated with prostate and general cancer risk [273]. The need for increased statistical power to detect genetic associations led to the creation of consortia with larger sample-sizes, designed either specifically for the study of CRC, such as the Colon Cancer Family Registry (CCFR) [274], or for general healthcare research, such as UK Biobank (UKBB) [161,164]. Eleven years after the first CRC GWAS, the latest iteration of these combined efforts was a comprehensive analysis by Huyghe et al. in over 120,000 individuals, which identified 40 novel independent single nucleotide polymorphisms (SNPs) associated with CRC risk and brought the total number of variants associated with CRC risk to 150 [163].

3.1.2. CRC and immunity

Inflammatory activity is a risk factor for CRC and has predominantly been linked to WBCs [275]. Indeed, immune cells are implicated in tumour surveillance [275] and are one of the constituents of the CRC tumour microenvironment (TME) [276]. The role of blood cells, such as WBCs, in CRC biology has been studied extensively over the past decades, the result of which has been to identify inflammation markers for both CRC- and CRC-associated cells, as well as establishing prognostic factors, such as neutrophil-to-lymphocyte ratio (NLR) [277].

Historically, WBCs have been studied for their role in the functioning of the innate and adaptive immune systems [278]. WBCs are now commonly measured in routine blood tests and are divided into five subtypes: basophils, eosinophils, lymphocytes, monocytes and neutrophils [278]. Moreover, they have been found to play a role in disease risk, severity, and progression. For example, an increase in eosinophil count has been associated with lower disease odds and increased neutrophil count with higher odds of general disease [279–284]. Importantly, observational studies have found a relationship between WBC count variation and CRC risk and mortality [285–292]. While their counts are correlated to a degree both phenotypically and genetically [149,166], WBC subtypes have different effector functions [293], making it desirable for them to be studied separately to establish the different biological pathways through which they might act to affect CRC development.

3.1.3. WBC count and CRC

While in this project I aimed to assess the relationship between WBC count and the risk of CRC, the number of studies looking at WBC count and CRC risk have been limited in comparison to those looking at WBC count and CRC prognosis. Therefore, I have also included what is known about WBC count and CRC mortality, as it might give some indication on the possible ways that WBC count could be associated with a biological mechanism that prevents or causes CRC. Studies analysing WBC count trends over time prior to CRC diagnosis have been included here as well, as they may explain the findings of studies exploring risk, which have predominantly been done a year or less prior to CRC diagnosis. More general information about WBCs and their biological relevance is available in the background chapter (**Chapter 1**).

Basophils:

Understudied for a long time due to their low numbers in circulation, basophils are now known to be involved in IgE-mediated immunity and in regulating the function of the adaptive immune system [294]. They differentiate in the bone marrow from granulocyte-macrophage progenitors through action of IL-3, the principal cytokine involved in this WBC's production [295].

The role of basophil count in CRC has been studied in relation to both CRC prognosis and risk. On CRC risk, Goshen et al. looked at WBC count quintiles in 56,485 individuals (1,755 cases, 54,730 controls) from an Israeli population between 30 and 180 days prior to CRC diagnosis [291]. In this cohort study, they used the bottom quantile as reference and compared it to the top quintile, finding a 40% increased odds of CRC diagnosis for men and 19% for women in the top quintile for basophil count [291]. Using data from the UK health improvement database (THIN, cases = 4,929, controls = 11,311), the analysis of Boursi et al. pointed towards, but did not show association for, a positive relationship between basophil count and CRC diagnosed between 6 to 12 months prior to blood sampling [292].

With regards to CRC prognosis, a study following CRC patients over a ten year period (N=569) found that a low basophil count was associated with a lower CRC overall and disease-free survival [283]. Similarly, in their study looking at preoperative CRC patients, Liu et al. (N=1,029) showed that a low basophil count was associated with an increase in CRC mortality [294]. A higher basophil count was associated with increased overall survival in Wu et al.'s study done on 153 CRC patients from China [286], showing that increased basophil count was also associated with reduced mortality.

From a biological perspective, basophils have been associated with the TME [32,296], and were shown to aid in recruiting cytotoxic CD8+ T-cells to the tumour site to aid in the clearance of cancer cells [297]. Therefore, basophil count could act both as a flag for improved adaptive system activity and work concomitantly with other immune cells for CRC tumour clearance.

Overall, the studies presented here show an apparent contrast between basophil count in relation to CRC risk vs. survival. One explanation might be that under homeostatic conditions, basophils act through a biological pathway that promotes carcinogenesis, while after the growth of the tumour, basophils switch to a predominantly anti-tumourigenic role. For example, cytokines IL-3 and IL-13 are associated with activity of blood basophils, known to induce a Th2 pro-tumourigenic response [298,299]. At the same time, current studies looking at CRC risk have used a short time window between blood sampling and CRC diagnosis, making it possible that the associations observed are due to production of basophils in response to the presence of cancer rather than a causal effect by basophils on CRC development. However, while these findings point to a possible detrimental effect by basophils on CRC risk, the limited number and scope of current studies have not established a clear role of basophil count in CRC aetiology.

Eosinophils:

Similarly to basophils, eosinophils are best known for their role in IgE-mediated immunity and allergy severity, such as hay fever and asthma [300], a key cytokine in the process of differentiation is IL-5 [32]. Eosinophils also have eosinophil-specific molecules that are used to perform effector functions, such as eosinophil cationic protein (ECP) and eosinophil-derived neurotoxin (EDN) [301].

With regards to CRC risk, Goshen et al. showed a 62% and 103% percent increase in relative risk for CRC when comparing the bottom to top quintile in eosinophil count in men and women, respectively [291]. When studied 6 to 12 months prior to diagnosis, eosinophil count displayed no association with CRC odds in Boursi et al.'s study [292]. However, these two studies employed a short time window (1-12 months) between blood sampling and diagnosis, and the results could indicate increased eosinophil production in response to cancer. A more comprehensive study aimed specifically at eosinophil count was conducted by Prizment et al. using data from the atherosclerosis risk in communities (ARIC) cohort [300]. Here, 15,792 people were followed between the years 1987 to 2006, a timeframe during which 242 incident CRC cases developed [300].

Eosinophil count was split into tertiles and the trend was studied in relation to CRC risk, which showed a negative association between eosinophil count and CRC risk for colon cancer, but not for rectal cancer [300]. This trend was also apparent when accounting for education, body mass index, smoking status, pack-years of smoking, alcohol use, diabetes, fibrinogen, and total WBC count [300]. Importantly, analysing only those who developed CRC at least 5 to 10 years after blood sampling showed the same negative association [300].

Looking at count trends in blood samples from 16,668 Israeli individuals, Rosman et al. showed a linear increase in eosinophil count in the years (7 years to 3 months) preceding a CRC diagnosis [289]. Virdee et al. conducted a study in 939,949 people living in the UK and compared the trends in WBC count between those who ended up developing CRC (cases = 17,408) and those who did not (controls = 922,541). Here, eosinophil count was lower at baseline in those aged 50 who ended up developing CRC versus those who did not, and increased slowly over time until diagnosis [302].

In terms of prognosis of CRC, Wei et al. showed that a low eosinophil count was associated with decreased overall and disease-free CRC survival [283], while Wu et al. did not find any association [286]. Using data from 381 Austrian patients diagnosed with CRC, Harbaum et al. found a positive association between peri-tumoural eosinophil count and disease-free survival [303]. Similarly, Väyrynen et al. used a machine learning method to scan stained sections of CRC samples from 934 US patients and showed a positive association between eosinophil density and cancer-specific survival [304]. When comparing patients with colorectal polyps and controls in a sample of 1,799 individuals from China, Feng et al. found a higher eosinophil count present in those with polyps [305].

A number of laboratory studies have investigated the role of eosinophils in CRC that could explain these results. For example, Legrand et al. studied the effect of eosinophils on intestinal carcinoma cells (Colo-205), and identified EDN, ECP, and granzyme A mediated killing of CRC cells [306]. In two subsequent studies, eosinophils were found to display antitumour effects with the aid of cytokines IL-18 (Colo-205) and IL-33 (CT26) [307,308]. Moreover, eosinophils have been found to act directly in eliminating tumour cells through the aid of IFN-γ, even without action by CD8+ T-cells [309]. While the current literature indicates that eosinophil count might be protective against CRC development, the causal relationship between eosinophil count and CRC remains to be established.

Lymphocytes:

Lymphocytes are represented by both B and T-cells and are critical for the functioning of the adaptive immune system [310]. In terms of CRC risk, the analysis by Wu et al. in 426 individuals who had their blood sampled at CRC diagnosis (CRC = 162, colorectal polyp = 132, controls = 108) outlined a lower lymphocyte count in CRC vs. control patients, but not in pre-cancerous polyps vs. controls [311]. Similarly, Goshen et al. found lower odds of CRC diagnosis when comparing the top vs. bottom quintile for lymphocyte count when blood sample was taken between 30 to 180 days prior to CRC diagnosis [291]. Boursi et al. analysed the relationship between lymphocyte count [adjusted for haematocrit (HCT), mean corpuscular volume (MCV) and NLR] and CRC odds 6 to 12 months prior to diagnosis [292]. Here however, they showed that lymphocyte count was associated with increased CRC odds [292]. Interestingly, in the same study by Virdee et al. outlined above, lymphocyte count was lower 9 years prior to CRC diagnosis in those aged 50 compared to healthy controls [302].

While NLR is one of the best prognostic markers for CRC [277], absolute lymphocyte count has also been associated with CRC prognosis. In their study of 95 Chinese patients with CRC, Yang et al. found that a low lymphocyte count was associated with decreased progression-free and overall survival [312]. Similarly, Tanio et al. showed a negative association between low lymphocyte count and overall survival in their analysis of 361 pre-operative Japanese patients with CRC [313].

Biologically, tumour-infiltrating lymphocytes (TILs) represent a component of the TME [275], and their numbers have been associated with improved CRC survival [284]. Indeed, natural killer (NK) and CD8+ T-cells are known to lead to the destruction of cancer cells through cytotoxic activity [314,315]. Similarly, B-cells are hypothesised to be involved in supporting a favourable outcome in CRC due to release of anti-tumourigenic cytokines, or through acting as antigen presenting cells (APCs) [316]. At the same time, when looking at 125 CRC samples from a French hospital, Th1 helper cells were associated with improved CRC survival, and Th17 with worse prognosis [317].

Given the current knowledge on the relationship between lymphocytes and cancer, lymphocyte count may display a protective effect against CRC development. However, most studies looking at lymphocyte count and CRC have only analysed this relationship up to a year prior to diagnosis, and a causal relationship has not been established.

Monocytes:

Monocytes are mononuclear phagocytes which play many roles in immunity, as well as being able to differentiate into tissue-resident macrophages [318]. Monocytes have been studied in relation to CRC risk. When sampling participants at diagnosis, Wu et al. identified a higher monocyte count in those with CRC compared to controls, but not in those with benign polyps compared to controls [286]. The analysis by Goshen et al. showed higher odds of CRC diagnosis when comparing the top vs. bottom quintile for monocyte count (30 to 180 days prior to diagnosis) [291]. Similarly, Boursi et al.'s study found increased odds of CRC diagnosis per 1-SD increase in monocyte count (6-12 months prior to diagnosis) [292]. Looking at WBC trends, Virdee et al. reported a higher monocyte count in men 9 years prior to CRC diagnosis compared to healthy controls, but not in women [302].

Monocytes have also been linked to CRC survival. Tanio et al. reported higher mortality in those CRC patients with high monocyte count [313]. A more comprehensive analysis on the relationship between monocyte count and CRC mortality was assessed in a systematic review and meta-analysis by Shu et al., which included the aforementioned study [319]. Here, the pooled data showed a positive relationship between higher monocyte count and worse CRC overall and progression-free survival [319].

From a biological perspective, monocytes can differentiate into tumour-associated macrophages (TAMs) [320], a process that takes place with the help of cytokines such as CCL2 [321]. Shibutani et al. conducted a study in 168 pre-operative CRC patients, where they investigated the relationship between monocyte count and TAM density using blood and immunohistochemical samples [322]. Here, monocyte count was positively associated with the density of TAMs in the CRC TME, and the density of TAMs was associated with worse CRC prognosis [322]. More generally, monocytes have been associated with a number of factors detrimental to CRC development, such as their inefficient killing of cancer cells, reduction of CD8+ T-cell numbers in the TME, and release of angiogenetic factors [323].

In contrast to lymphocyte count, current epidemiological and laboratory studies indicate that monocyte count could have a detrimental effect on CRC development.

Neutrophils:

Neutrophils are the first cells to arrive at the site of infection and form a critical part of the innate immunity [324]. Neutrophil count has been studied in terms of CRC risk. In their sampling at diagnosis study, Wu et al. found a higher neutrophil count mean in those

with CRC vs. controls, and in pre-cancerous polyps vs. controls [286]. Goshen et al. reported increased odds of CRC diagnosis when comparing the top vs. bottom quintile for neutrophil count [291]. When patients were sampled 6 to 12 months prior to diagnosis, neutrophil count was associated with a 24% increase in CRC odds in Boursi et al.'s analysis [292].

Looking at trends, Virdee et al. report a lower neutrophil count in patients 9 years prior to CRC across all age groups, which rose sharply 2-3 years prior to CRC diagnosis [302]. Neutrophil count has been studied is relation to CRC survival. The review by Yamamoto et al. assessing studies between 2008 to 2021 found that in most cases, a higher NLR was associated with worse CRC survival [325]. In terms of absolute counts, a study done by Wyatt et al. in 508 patients undergoing CRC elective resection showed a negative association between circulating neutrophils and CRC prognosis [287].

There have been studies looking at the biological significance of neutrophils in CRC that could explain agreement between neutrophil count in the context of CRC risk and mortality. As an example, in a CRC murine model, Saurer et al. show that neutrophils carrying the triggering receptor expressed on myeloid cells-1 (TREM-1) could be implicated in CRC development [326]. Similarly, tumour-associated neutrophils (TANs) have generally been described as detrimental in cancer through multiple mechanisms [327], one example being neutrophil extracellular trap (NET) shielding of cancer cells against CD8+ lymphocytes [328]. On the other hand, neutrophils have been found to act concomitantly with CD8+ T-cells to improve survival in CRC patients, although this effect was linked predominantly to action of CD8+ T-cells alone [329].

As in the case of monocyte count, the evidence for neutrophil count seems suggest a pro-tumourigenic role in CRC development.

3.1.4. Risk factor heterogeneity in CRC

Overall, the combined evidence shows that variation in WBC count could affect the risk of CRC, and that the possible effect could be different depending on the WBC subtype. Additionally, risk factors for, and the degree to which risk factors are associated with CRC, have been found to vary by the anatomical subsite of the disease [254,330,331]. For example, age, type 2 diabetes, ancestry, height and body mass index (BMI) are examples of risk factors which present heterogenous associations with CRC risk [254,330]. These kinds of differences are also evident with sex, where males are more likely to develop left-sided cancers than females [254,332,333].

This apparent heterogeneity in CRC risk has not only been presented in observational settings, as recent GWAS have found that the genetic architecture of CRC development differs by its anatomical subsite, indicating that distinct genes are involved depending on the location of the primary tumour [163,239,242]. For example, the *MLH1* gene involved in DNA mismatch repair has been linked to proximal colon cancer risk, while the *KLF14* gene involved in TGFβ signalling was found to be linked to distal cancer risk [239]. This heterogeneity was also apparent in GWAS of CRC mortality, where a number of SNPs were associated with proximal colon and distal colon cancer survival, but not with rectal cancer survival [334].

Given these findings, one might be inclined to study how a trait, such as WBC count, affects the risk of CRC stratified by genetic sex and by the anatomical location of where the primary tumour formed inside the colorectum.

3.1.5. Current limitations on WBC count and the risk of CRC

As evidenced from the research referenced above, observational studies have accounted for the majority of epidemiological studies on CRC, which suffer from particular limitations (**Chapter 2**). The large increase in the number of GWAS on CRC risk has been simultaneous with advances in the field of genetic epidemiology [163,335], making it more accessible for research to investigate risk factors for CRC in a MR framework. As mentioned in **Chapter 2**, in two-sample MR (2SMR), the summary statistics for the exposure and the outcome come from different samples [191,197,208,335]. Univariable (UV) MR is used to estimate the total effect of a single exposure on an outcome [218], whereas multivariable (MV) MR is able to estimate the direct effect of multiple exposures by adjusting for their shared proxy instruments [218].

3.1.6. What can MR add in the context of WBC count and CRC risk?

MR has several advantages over observational studies in the context of WBC count and CRC:

- Unmeasured confounding – confounders can be a source of bias in statistical analyses [253], and unmeasured confounding is a particular weak feature of most observational studies [194]. For example, people with higher neutrophil count could be affected by an unknown variable that also increases the risk of CRC, providing

evidence for an association with, but not a causal relationship to CRC. MR could overcome this issue due to its similarity to randomized control trials (RCTs), as alleles are randomly allocated at conception [172].

- Reverse causation – as evidenced previously, most studies investigating WBC count and CRC risk have had a short time window between blood sampling and CRC diagnosis. This makes it hard to determine whether e.g. a rise in neutrophil count is associated with increased odds of CRC diagnosis, or if the presence of CRC led to this increase in WBC count. MR could overcome this, as the genetic proxies used to instrument for WBC count are assigned at conception, well before the development of a colorectal tumour [172].
- Practicability – a RCT would be the best method in traditional epidemiology to give evidence of causality between WBC count and CRC risk. However, in this context, it would be practically impossible to run a RCT due to several intertwined limitations: ethical (can we give WBC count altering drugs to participants?), practical (are there WBC altering drugs available for healthy people?), time and cost related. Summary statistics from GWAS done on WBC count (ethical-exposure) and CRC risk (ethical-outcome) are readily available (time) and can be analysed in a 2SMR framework with ease on a computer (cost), having the capacity to overcome the issues of running a RCT.
- Causality – while observational studies can be done when RCTs are not feasible, these only establish the presence of an association between an exposure and an outcome [336]. As the aim here is to establish a causal relationship between WBC count and CRC risk, a MR analysis would be the best suited method to do so.
- Additionally, MVMR has an advantage over UVMR – it can account for the shared genetic instruments of two or more exposures to estimate the direct effect of each exposure on the outcome [216]. In the context of WBC count, this can be useful, as there is overlap in the genetic architecture of the five WBC subtype counts [149,166]. For example, some of the SNPs used to instrument for lymphocyte count are likely to also instrument for neutrophil count, and in a UVMR context this poses challenges in untangling the direct biological effect of each subtype. Therefore, adding all five WBC counts into a MVMR analysis would allow me to contrast with the UVMR analysis to establish a specific biological mechanism through which WBC count affects CRC development.

3.1.7. Main study objective

Given the advantages of MR and its potential to fill a gap in our knowledge, my overarching aim was to explore the role of circulating levels of WBCs in CRC

development using state of the art methods in genetic epidemiology, namely UV and MV MR, exemplifying how genetically predicted full blood count measures can aid in the understanding of disease aetiology.

3.1.8. Study aims

I have divided this chapter's main objective into four aims that I plan to address:

1) Investigate if variation in WBC count affects the risk of developing CRC, and which WBC subtypes contribute to this link
2) Assess if there is a WBC subtype-specific effect on CRC risk, independent of the counts of the other WBC subtypes
3) Interpret the MR results in comparison with an observational analysis looking at WBC count and CRC risk
4) Triangulate these findings to discuss possible biological mechanisms and assess how these could be taken forward in a follow-up analysis

3.2. Methods

3.2.1. Study design

I assessed the relationship between circulating WBCs and CRC odds using both genetic epidemiologic and observational methods. First, a UVMR analysis was undertaken to estimate the effect of WBC subtype counts on CRC (**Aim 1**), followed by a MVMR analysis where the direct effect of each WBC subtype count was estimated by adding all five WBC subtypes into the model (**Aim 2**) (**Figure 3-3A**). STROBE-MR guidelines were followed (**Appendix 4**) [337]. Afterwards, I ran the largest cohort study between WBCs and CRC to date to compare with the MR estimates (**Figure 3-3B, Aim 3**). Here, subtype specific WBC counts were first studied individually, and then by adding them together into the model. Cohort STROBE guidelines were followed (**Appendix 5**). For these analyses, units were interpreted as odds ratio (ORs) for CRC per a normalized standard deviation (1-SD) increase in WBC count. Finally, a MR analysis between allergic disease and CRC was done after results analysis (**Figure 3-3C**, **Aim 4**). Here, analyses were interpreted as a OR increase in CRC per 1-log OR increase in allergic disease.

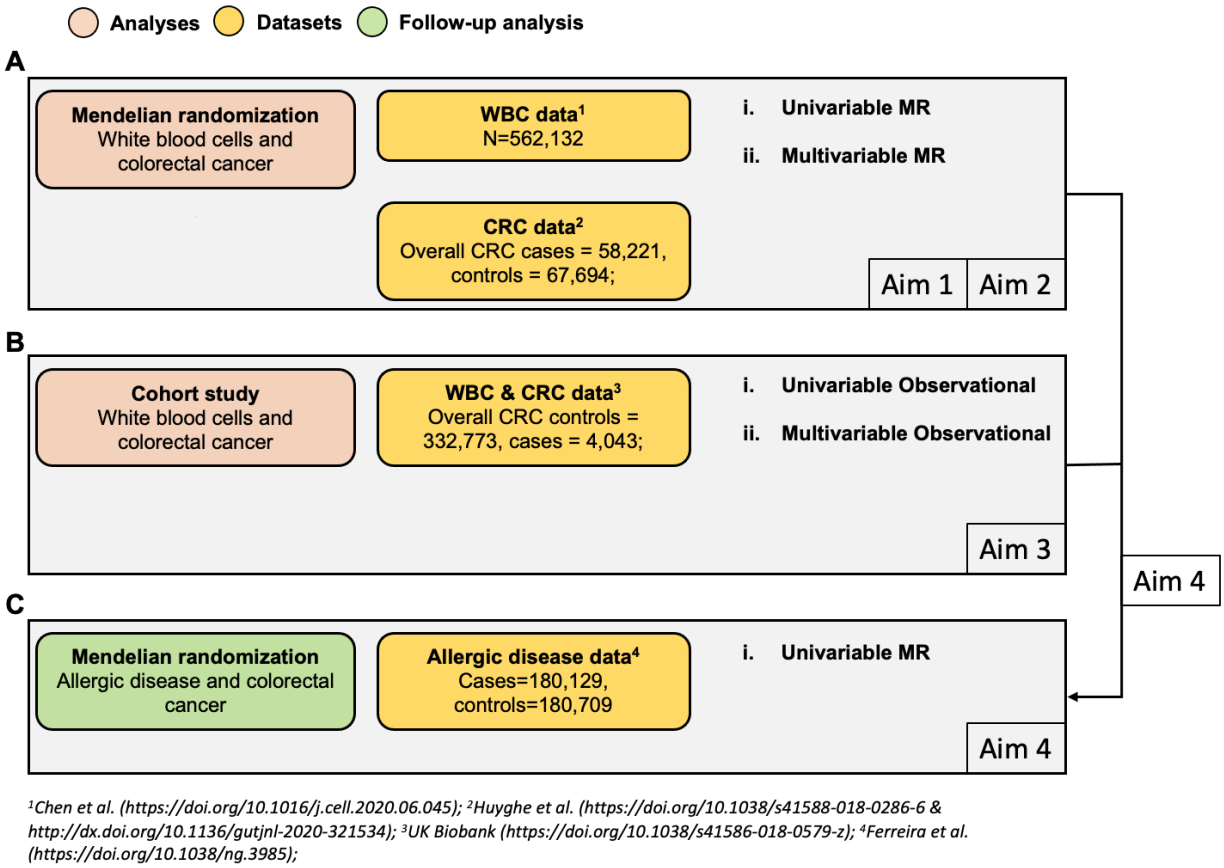


Figure 3-3. *Study design of the project.*

3.2.2. WBC count GWAS data

As mentioned in **Chapter 2**, GWAS summary statistics contain the minimum amount of information that can help the reader understand which, in which direction, and to what degree a SNP is associated with a particular trait [220].

Summary statistics for WBC count and each subtypes were obtained from a recent study by Chen et al. as part of the "Blood Cell Consortium" (BCX) [166] (**Chapter 2**). For the purposes of this project, I used the meta-analysed data for those people of European ancestry, which were predominantly from UKBB (N=~562,243) [166]. This was done so that the exposure and outcome samples were of similar ancestry composition in order to prevent bias of MR estimates arising from residual population structure [338,339]. These instruments were available as sex-combined only. A brief description of each meta-analysed study is available in **Appendix 1**.

3.2.3. CRC GWAS data

The GWAS summary statistics for CRC and its anatomical subtypes come from the most comprehensive meta-analysis of CRC risk to date [163] (**Chapter 2**). For this project, to avoid bias due to sample overlap with the exposure in two-sample MR [196], UKBB participants were excluded, resulting in a final sample-size of 98,815 (52,775 cases and 45,940 controls). The final sample was predominantly of European ancestry, with 5.36% representing East Asians, these were included due to their similar CRC genetic architecture [243]. An overview of all consortia included in the CRC meta-analysis is available in **Appendix 2,** and a breakdown of the sample-size for each CRC summary statistics is presented in **Table 3-1**. The summary-level GWAS data for CRC used in this study were made available following an application to the Genetics and Epidemiology of Colorectal Cancer Consortium (GECCO). These instruments were available as sex-combined for the CRC anatomical subsites data. For overall CRC, instruments were available as sex-combined and sex-specific (male/female).

Table 3-1. *Description of CRC cases and controls by anatomical subsite*

Colorectal cancer type	N Cases	N Controls
Overall CRC	52775	45940
Overall CRC, female	24594	23936
Overall CRC, male	28271	22351
Colon cancer	28736	43099
Proximal colon cancer	14416	43099
Distal colon cancer	12879	43099
Rectal cancer	14150	43099

3.2.4. Allergic disease GWAS data

Summary statistics for allergic disease were obtained from a meta-analysis of 13 GWAS done in people of European ancestry, with a final sample-size of 360,838 (cases = 180,129, controls = 180,709) [244] (**Chapter 2**). Included covariates and specific numbers for each meta-analysed study is available in **Appendix 3**. These instruments were available as sex-combined only.

3.2.5. Genetic data processing

To select for valid MR instruments, I processed the summary statistics for the exposures using the "TwoSampleMR" R package [213,219] prior to running the analyses. The exposure SNPs were linkage disequilibrium (LD) clumped (r^2=0.001, window=10Mb, P-value threshold=5e-8) to avoid can violate the MR SNP independence assumption and lead to double counting and instrument strength overestimation [187] (**Chapter 2**). This was done with the integrated PLINK v1.9 function "clump_data()" [247] using the 1000 Genomes European dataset [152,248] as a reference panel. Following this step, the exposure and outcome datasets were "harmonised" i.e. had their effect alleles placed on the same reference strand [250] (**Chapter 2**).

3.2.6. Univariable MR analysis of WBC counts on CRC risk

I undertook the primary UVMR analysis using the inverse-variance weighted (IVW) method, which is the fixed-effects meta-analysis of the estimated effect of all exposure SNPs on the CRC outcome [202]. Conditional F-statistics were calculated to detect for weak instrument bias [206] for each exposure SNP using previously described methodology [340]. Several sensitivity analyses were undertaken to compare with the main IVW estimates. The presence of vertical pleiotropy i.e. when a trait is downstream of the genetic variant but on the same biological pathway as the exposure [195], was measured using Cochran's Q heterogeneity test [341]. Horizontal pleiotropy, when a SNP or more act through a different pathway to the exposure [195], can violate one of the main MR assumptions. A number of sensitivity MR analyses were undertaken to suggest horizontal pleiotropy: MR-Egger (where the regression intercept is not constrained to zero) [205], Weighted median (the median of all SNP ratio estimates, where each ratio is weighted by the inverse of the variance) [204], Weighted mode (assumes that the most frequent estimate in a set of instruments is zero) [212] and MR-PRESSO (detects individual SNPs that might contribute to horizontal pleiotropy and generates a new IVW estimate where those SNP outliers are removed) [207]. The direction of the causal relationship between WBC traits and CRC was tested using the MR Steiger method, which uses Steiger's test to test the difference between the Pearson correlations of genetic variants with both the exposure and outcome [213]. More information on these methods is available in **Chapter 2**.

3.2.7. Multivariable MR analysis of WBC counts on CRC risk

As done in the UVMR analysis, I used the IVW method for the MVMR approach. First, a pair-wise analysis between all five WBC subtype counts was done, where the proportion of variance explained for SNPs used to instrument a WBC trait was estimated in the other four WBC subtypes using previously described methodology [340]. The direct effect of each WBC subtype was estimated by adding in all five WBC subtypes into the MVMR model. P-values were adjusted using a False Discovery Rate method for 35 independent tests i.e. 5 WBC subtypes x 7 CRC outcomes [342]. Bias arising from weak instruments was also assessed here. This was done using methodology described by Sanderson et al., where a generalized version of Cochran's Q was employed to evaluate instrument strength [216]. Standard Cochran's Q statistic [341] was calculated to detect the presence of heterogeneity. For those traits with an F-statistic <10, a follow-up MVMR analysis was done accounting for the presence of weak instruments. All functions are available as part of the "MVMR" R package (https://wspiller.github.io/MVMR/).

3.2.8. UK Biobank phenotypic data

Specific details about the UK Biobank study are available in **Chapter 2**. The blood sampling date variable was split into year, month, day, and minutes (passed since the start of the day of the appointment visit). Additional variables were gathered: recruitment centre, sampling device ID, age, genetic sex, principal components 1 to 10 (geographical structure [226]), BMI, Townsend deprivation index (socioeconomic status [343]), smoking and alcohol drinker status (self-report questionnaire – UKBB codes 20116 and 20117). CRC cases were identified through hospital inpatient records coded to the 10th version of the International Classification of Disease (ICD-10).

3.2.9. Filtering and selection criteria

The UK Biobank dataset underwent a series of steps prior to further analyses. Withdrawn participants and those of non-European ancestry were excluded. Viable controls and incident CRC cases were defined using methodology previously described by Burrows et al. [344] (**Table 3-2**). Here however, incident CRC cases were defined as those diagnosed at least one year after blood sampling. Participants with no WBC measurement data and/or sampling date were removed, as were those known to be pregnant, with chronic conditions (e.g. HIV, blood cancers, thalassaemia), or undergoing erythropoietin treatment (as done in Astle et al. [149] and Chen et al. [166]), given the effects of these traits on WBC measurements. Those with acute conditions (e.g. upper

respiratory infections) diagnosed less than 3 months prior to blood sampling were also excluded. Finally, missing values in “Townsend Deprivation Index”, “Body mass index”, “Smoking status” and “Alcohol drinker status” variables were removed.

Table 3-2. *Selection criteria for cases and controls in the cohort analysis.*

Cases	Selection	ICD10 - UKBB code 41270	ICD10 C18.0-C18.9; C19; C20
	Inclusions	Behaviour of tumour - UKBB code 40012	Malignant, primary site Malignant, microinvasive Malignant, metastatic site Malignant, uncertain whether primary or metastatic site
		White blood cell acquisition date - UKBB code 30002	CRC diagnosed at least 12 months after blood sample taken
	Exclusions	ICD10 - UKBB code 41270	All D codes
		Behaviour of tumour - UKBB code 40012	Benign Uncertain whether benign or malignant Carcinoma in situ
Controls	Inclusions	All eligible participants not defined as cases	
	Exclusions	ICD10 - UKBB code 41270	Any C and D code
		Self-report cancer - UKBB code 20001	Any cancer Any site-specific cancer

Data items available at https://biobank.ndph.ox.ac.uk/showcase/

3.2.10. Descriptive analysis of phenotypic data

A descriptive analysis was undertaken prior to the observational analysis. First, a phenotypic correlation matrix between each WBC subtype was generated using a Spearman's rank test. Univariable, analysis of variance (ANOVA) type I and ANOVA type II models were then used to calculate the variance explained in WBC count by the following covariates: UKBB assessment centre, blood sampling device ID, sample year, sample month, sample day, minutes passed in sample day, sex, age, principal components 1 to 10, BMI, Townsend deprivation index, smoking status and alcohol drinker status.

3.2.11. Observational analysis between WBC count and CRC

Following the descriptive analyses, an observational analysis was undertaken between circulating WBCs and incident CRC. WBC count values were log-transformed, after which they were adjusted for the following covariates: sex, age, age^2, PCs 1 to 10, as by Chen et al [166]. The resulting residuals were rank-inverse normal transformed and then used in a logistic regression on CRC incidence. This main observational analysis was termed "Model 1", which was the minimally adjusted model. A separate analysis was also done, the fully adjusted "Model 2", where BMI, Townsend DI, smoker status and alcohol drinker status were added as additional covariates, as these were shown to explain some of the variation in WBC count. Following this, another pair of analyses were run, where all five WBC subtype counts were added together into the model. Analyses where each WBC trait was studied individually were termed as "univariable", while those where they were added together were termed as "multivariable". For all analyses, units were interpreted as odds ratio (ORs) for CRC per normalized standard deviation (1-SD) increase in WBC count.

3.2.12. Working environment

All analyses were performed with R version 4.1.2 (Bird Hippie) [345] in a Linux environment supported by the University of Bristol's Advanced Computing Research Centre (ACRC). Genetic data preparation, as well as the UVMR analyses, were done with the "TwoSampleMR" R package [213,219]. The MVMR analyses were done with "MVMR" R package [217].

3.3. Results

3.3.1. Univariable MR between WBC count and CRC

Prior to running the UVMR analysis, the average F-statistic for each WBC trait was calculated. This is used to detect the presence of weak instrument bias, which can give unreliable MR estimates and is indicated by an average F-statistic < 10 [206]. For overall CRC, the average F-statistic was 64.48 (basophil count), 124.72 (eosinophil count), 105.85 (lymphocyte count), 147.44 (monocyte count) and 98.84 (neutrophil count), indicating strong instruments for MR analyses (**Table 3-3**).

Table 3-3*. F-statistics for the overall CRC outcome.*

WBC subtype count	CRC type	Avg. F-stat	No. SNPs[1]
Basophil	Overall	64.48	171
Eosinophil	Overall	124.72	396
Lymphocyte	Overall	105.85	444
Monocyte	Overall	147.44	477
Neutrophil	Overall	98.84	387

[1]Number of SNPs instrumenting for the trait after PLINK clumping

After computing the average F-statistic, I conducted the UVMR analysis. Here, the main MR analysis was done using the IVW method (**Chapter 2**), which has the most power to detect the presence of an effect, but is also the most liable to being biased by horizontal pleiotropy [204]. Therefore, the IVW analysis was followed by four sensitivity analyses which have less statistical power, but are commonly used to detect horizontal pleiotropy in MR [204,205,207,212]. In MR it is more important to identify consistency in the direction of the effect between the IVW method and the sensitivity methods rather than comparing their P-values [346]. Agreement in direction and effect estimates between one or more MR methods provides evidence that there is a causal effect between the exposure and outcome.

The IVW method showed evidence of a protective effect for basophil count (OR: 0.88, 95% CI: 0.78-0.99, P-value: 0.037) on overall CRC odds. This was also true for the MR-PRESSO method (OR: 0.9, 95% CI: 0.81-0.99, P-value: 0.039) for overall CRC, showing a similar result to the IVW method when eliminating SNPs that might display horizontal pleiotropy, strengthening the result. Similarly, the weighted median method pointed to a protective effect in male CRC (OR: 0.81, 95% CI: 0.67-0.99, P-value: 0.037) (**Figure 3-4**, **Appendix 6**). For eosinophil count, the IVW method showed evidence of a protective effect for overall (OR: 0.93, 95% CI: 0.88-0.98, P-value: 0.012) and female (OR: 0.91,

95% CI: 0.85-0.99, P-value: 0.021) CRC odds. These results were supported by the MR-Egger, weighted median and MR-PRESSO methods for overall CRC, by MR-Egger and MR-PRESSO methods for female CRC, and by MR-Egger for male CRC (**Figure 3-4**, **Appendix 6,** further description of sensitivity analyses results below).

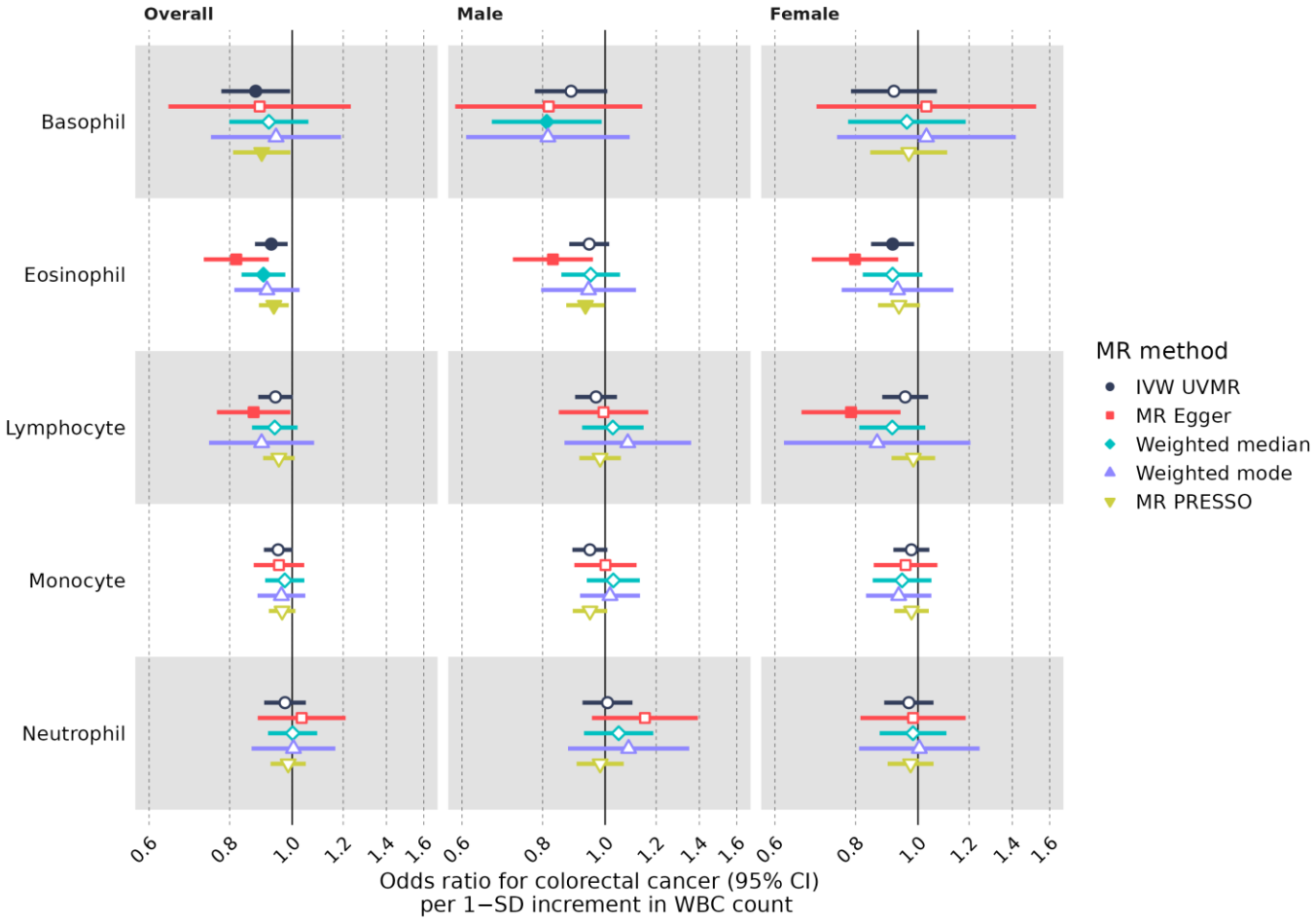


Figure 3-4. *Univariable MR analysis of WBC count on overall CRC, and stratified by genetic sex.*

WBC traits are separated into rows on the X-axis. Each column is an analysis looking at overall, male-specific, and female-specific CRC. The estimated effect given by each method is presented on the Y-axis. Point estimates are filled where the P < 0.05. Results are interpreted as ORs (95% CI) for CRC per 1-SD normalized increment in WBC count.

Because risk factors for CRC can vary depending on the anatomical subsite of tumour development, I next considered whether altered levels of circulating WBCs impact risk of CRC differently by subsite. The univariable IVW method showed evidence for a protective effect of basophil count on colon (OR: 0.85, 95% CI: 0.74-0.98, P-value: 0.022) and distal colon (OR: 0.82, 95% CI: 0.70-0.97, P-value: 0.019) cancers (**Figure 3-5**, **Appendix 6**). For eosinophil count, the main IVW analysis gave evidence of a protective effect for colon (OR: 0.90, 95% CI: 0.84-0.96, P-value: 0.001), proximal colon (OR: 0.89, 95% CI: 0.82-0.96, P-value: 0.003) and distal colon (OR: 0.89, 95% CI: 0.82-0.97, P-value: 0.007) cancers (**Figure 3-5**, **Appendix 6**). These results were generally supported by sensitivity analyses (further information on sensitivity analyses below).

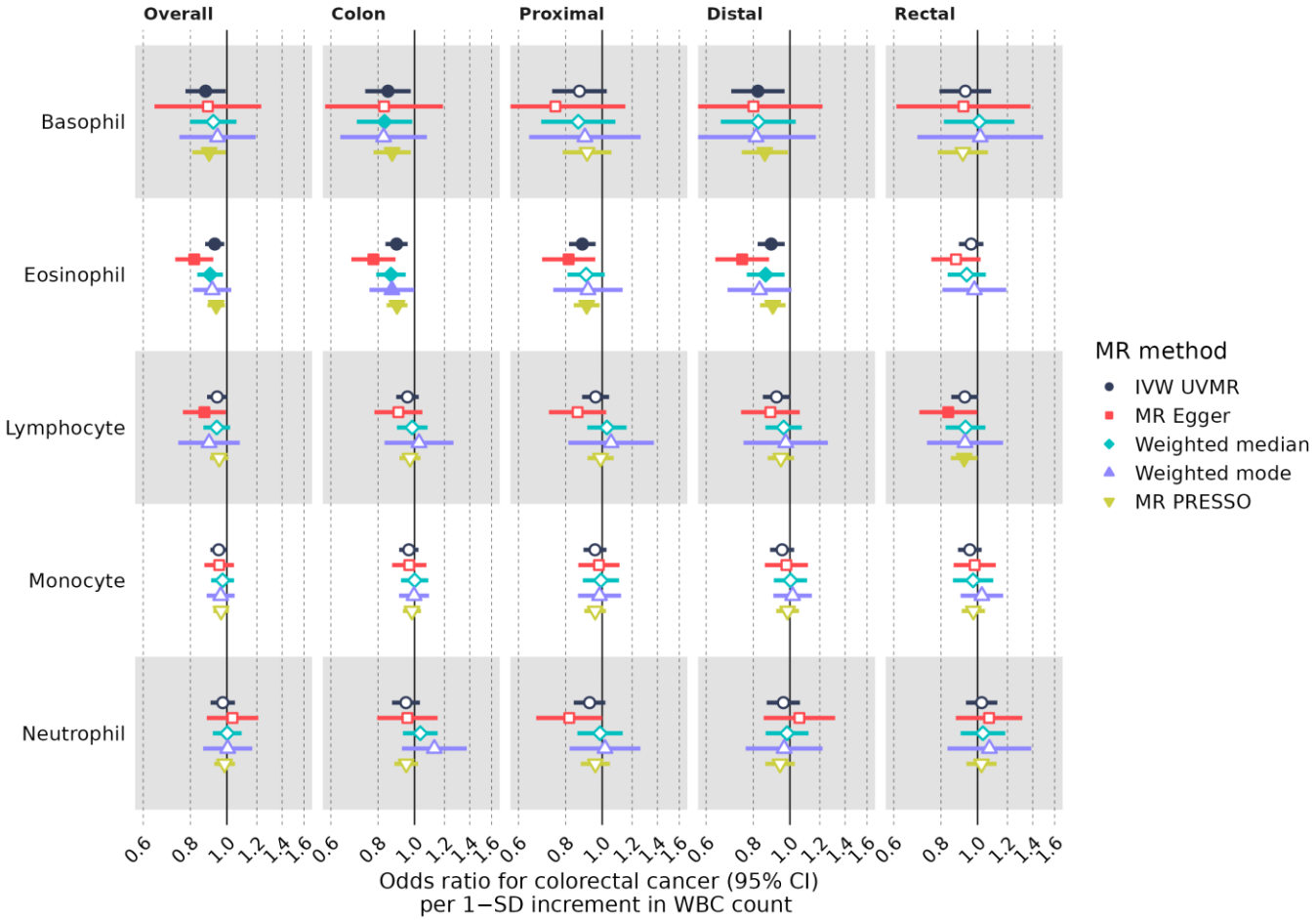


Figure 3-5. *Univariable MR analysis of WBC count on overall CRC, and stratified by CRC subsite.*

WBC traits are separated into rows on the X-axis. Each column is an analysis looking at overall, colon-, proximal-, distal-, and rectal-specific CRC. The estimated effect given by each method is presented on the Y-axis. Point estimates are filled where the P < 0.05. Results are interpreted as ORs (95% CI) for CRC per 1-SD normalized increment in WBC count.

Additional analyses were performed to study the presence of vertical and horizontal pleiotropy in the UVMR analysis. Cochran's heterogeneity test provided evidence for the presence of vertical pleiotropy in all but one (basophil count-male CRC, P_{HET}=0.104) WBC trait-CRC pair (**Table 3-4**). This suggests that some of the SNPs instrumenting for the count of a WBC subtype might also act on other traits downstream on the same biological pathway to affect CRC risk, which is expected for complex traits such as WBC count and does not invalidate the MR results [347]. Following this, the MR-Egger intercept test for horizontal pleiotropy was undertaken. Here, evidence for this type of pleiotropy was suggested for eosinophil count and colon (P_{PLT}=0.018), distal colon (P_{PLT}=0.015) cancers, and female (P_{PLT}=0.049), male (P_{PLT}=0.041) and overall (P_{PLT}=0.015) CRCs, indicating that the main IVW method might be unreliable due to violation of MR assumptions. However, the sensitivity analyses generally gave estimates that were in agreement with the main IVW method (**Figure 3-4, Appendix 6**), which would not be the case if horizontal pleiotropy was the main driver behind the effect seen in the IVW analysis. Therefore, I wanted to investigate whether horizontal pleiotropic SNPs were responsible for the effect seen in the IVW analysis. To do this, I ran an MR-PRESSO analysis, which can identify the presence of SNPs responsible for horizontal pleiotropy (outliers) and re-run an IVW MR analysis without them, to compare with the unadjusted (main) IVW MR method [207]. Outlier SNPs were identified in all but one WBC trait-CRC pair (basophil count to male CRC, P_{PRESSO}=0.13) (**Table 3-5**). However, there was no evidence that the removal of these outliers contributed to a notable shift in the point estimates (**Table 3-5**), providing evidence that the estimates of the main IVW MR analysis for eosinophil count are reliable.

Table 3-4. *UVMR Cochran's Q and MR-Egger intercept sensitivity analyses.*

Exposure	Outcome	Het P	Ple intercept	Ple P	Steiger Correct direction	Steiger P
Basophil	Colon	1.37E-10	0.00052582	0.88721	TRUE	< 2.22e-16
Basophil	Distal	2.15E-05	0.00060567	0.88963	TRUE	< 2.22e-16
Basophil	Female	5.27E-07	< 2.22e-16	0.53026	TRUE	< 2.22e-16
Basophil	Male	0.1043086	0.00174581	0.61188	TRUE	< 2.22e-16
Basophil	Overall	3.65E-13	< 2.22e-16	0.92721	TRUE	< 2.22e-16
Basophil	Proximal	8.54E-08	0.0032254	0.46441	TRUE	< 2.22e-16
Basophil	Rectal	0.0002336	0.00025644	0.95194	TRUE	< 2.22e-16
Eosinophil	Colon	1.39E-15	0.00448397	0.01819	TRUE	< 2.22e-16
Eosinophil	Distal	1.77E-08	0.00558396	0.01513	TRUE	< 2.22e-16

Eosinophil	Female	2.93E-10	0.00420644	0.04975	TRUE	< 2.22e-16
Eosinophil	Male	8.01E-06	0.00409192	0.04108	TRUE	< 2.22e-16
Eosinophil	Overall	< 2.22e-16	0.00395443	0.01536	TRUE	< 2.22e-16
Eosinophil	Proximal	1.34E-10	0.00266324	0.24554	TRUE	< 2.22e-16
Eosinophil	Rectal	0.0103057	0.00292038	0.16374	TRUE	< 2.22e-16
Lymphocyte	Colon	3.36E-14	0.00147868	0.40159	TRUE	< 2.22e-16
Lymphocyte	Distal	2.76E-07	0.00099121	0.6442	TRUE	< 2.22e-16
Lymphocyte	Female	7.71E-11	0.00501015	0.01609	TRUE	< 2.22e-16
Lymphocyte	Male	1.21E-06	< 2.22e-16	0.70971	TRUE	< 2.22e-16
Lymphocyte	Overall	< 2.22e-16	0.00206385	0.18727	TRUE	< 2.22e-16
Lymphocyte	Proximal	1.50E-09	0.00295324	0.16271	TRUE	< 2.22e-16
Lymphocyte	Rectal	1.99E-05	0.00265733	0.20589	TRUE	< 2.22e-16
Monocyte	Colon	< 2.22e-16	< 2.22e-16	0.95013	TRUE	< 2.22e-16
Monocyte	Distal	2.19E-11	< 2.22e-16	0.64084	TRUE	< 2.22e-16
Monocyte	Female	3.06E-08	0.00069274	0.66554	TRUE	< 2.22e-16
Monocyte	Male	2.14E-06	< 2.22e-16	0.23622	TRUE	< 2.22e-16
Monocyte	Overall	< 2.22e-16	< 2.22e-16	0.93759	TRUE	< 2.22e-16
Monocyte	Proximal	6.77E-11	< 2.22e-16	0.65344	TRUE	< 2.22e-16
Monocyte	Rectal	6.91E-10	< 2.22e-16	0.58219	TRUE	< 2.22e-16
Neutrophil	Colon	< 2.22e-16	< 2.22e-16	0.91815	TRUE	< 2.22e-16
Neutrophil	Distal	< 2.22e-16	< 2.22e-16	0.3202	TRUE	< 2.22e-16
Neutrophil	Female	4.21E-11	< 2.22e-16	0.86017	TRUE	< 2.22e-16
Neutrophil	Male	1.35E-12	< 2.22e-16	0.11737	TRUE	< 2.22e-16
Neutrophil	Overall	< 2.22e-16	< 2.22e-16	0.39846	TRUE	< 2.22e-16
Neutrophil	Proximal	1.84E-12	0.003245	0.17042	TRUE	< 2.22e-16
Neutrophil	Rectal	9.61E-10	< 2.22e-16	0.61914	TRUE	< 2.22e-16

Table 3-5. *MR-PRESSO summary.*

Global P-value indicates the presence of horizontal-pleiotropic SNPs, while the Distortion P-value shows whether there is evidence for a difference between the outlier-adjusted IVW MR and the unadjusted (main) IVW MR.

Exposure	Outcome	Beta	SE	P-value	Global P-value	Distortion P-value	No SNPs	N outliers
Basophil	Colon	-0.14	0.06	0.02	<3.33e-4	0.57	173	4
Basophil	Distal	-0.15	0.07	0.03	<3.33e-4	0.41	174	3
Basophil	Female	-0.03	0.07	0.64	<3.33e-4	0.2	174	4

Basophil	Male				0.13		173	0
Basophil	Overall	-0.11	0.05	0.04	<3.33e-4	0.59	171	4
Basophil	Proximal	-0.09	0.08	0.23	<3.33e-4	0.28	174	3
Basophil	Rectal	-0.09	0.08	0.25	<3.33e-4	0.87	176	1
Eosinophil	Colon	-0.11	0.03	0	<3.33e-4	0.93	397	5
Eosinophil	Distal	-0.1	0.04	0.01	<3.33e-4	0.8	397	5
Eosinophil	Female	-0.07	0.04	0.08	<3.33e-4	0.42	393	3
Eosinophil	Male	-0.07	0.03	0.05	<3.33e-4	0.75	398	3
Eosinophil	Overall	-0.07	0.03	0.02	<3.33e-4	0.69	396	7
Eosinophil	Proximal	-0.09	0.04	0.02	<3.33e-4	0.41	392	3
Eosinophil	Rectal				0.01		393	1
Lymphocyte	Colon	-0.03	0.03	0.39	<3.33e-4	0.39	453	5
Lymphocyte	Distal	-0.06	0.04	0.18	<3.33e-4	0.31	449	4
Lymphocyte	Female	-0.02	0.04	0.67	<3.33e-4	0.23	443	5
Lymphocyte	Male	-0.02	0.04	0.64	<3.33e-4	0.34	444	2
Lymphocyte	Overall	-0.05	0.03	0.1	<3.33e-4	0.59	444	8
Lymphocyte	Proximal	-0.01	0.04	0.81	<3.33e-4	0.15	455	3
Lymphocyte	Rectal	-0.08	0.04	0.05	<3.33e-4	0.94	452	1
Monocyte	Colon	-0.02	0.03	0.57	<3.33e-4	0.24	484	5
Monocyte	Distal	-0.01	0.04	0.69	<3.33e-4	0.17	479	4
Monocyte	Female	-0.02	0.03	0.46	<3.33e-4	0.99	477	4
Monocyte	Male	-0.05	0.03	0.09	<3.33e-4	0.99	480	2
Monocyte	Overall	-0.04	0.02	0.14	<3.33e-4	0.32	477	6
Monocyte	Proximal	-0.04	0.03	0.2	<3.33e-4	0.98	487	5
Monocyte	Rectal	-0.03	0.04	0.48	<3.33e-4	0.27	484	2
Neutrophil	Colon	-0.05	0.04	0.17	<3.33e-4	0.96	390	8
Neutrophil	Distal	-0.06	0.05	0.19	<3.33e-4	0.75	398	6
Neutrophil	Female	-0.03	0.04	0.52	<3.33e-4	0.87	390	5
Neutrophil	Male	-0.02	0.04	0.68	<3.33e-4	0.17	391	5
Neutrophil	Overall	-0.01	0.03	0.66	<3.33e-4	0.37	387	10
Neutrophil	Proximal	-0.04	0.05	0.35	<3.33e-4	0.21	382	5
Neutrophil	Rectal	0.02	0.05	0.61	<3.33e-4	0.98	396	4

3.3.2. Multivariable MR between WBC count and CRC

After the UVMR analysis, I conducted a MVMR analysis. As mentioned in the introduction, the counts of each WBC are correlated to a degree with each other, genetically and phenotypically [149,166]. In a UVMR setting, this makes it more difficult to establish what is driving an identified effect i.e. is eosinophil count alone reducing the risk of CRC? Therefore, the MVMR analysis was done to identify if the count of a specific WBC subtype is influencing the risk of CRC, allowing me to pinpoint a more specific biological pathway.

Prior to running the MVMR analysis, I assessed whether adjusting between the instruments of all five WBC subtype counts would be feasible i.e. not lead to weak instrument bias [217], making the MVMR results unreliable. For example, if the proportion of variance explained by SNPs instrumenting for trait A explain a similar amount of variance in trait B, adding both A and B in a MVMR analysis might lead to weak instrument bias for the trait A estimate, as the analysis would adjust for almost all SNPs used to instrument for trait A.

Therefore, I calculated the proportion of variance explained for the SNPs used to instrument for basophil count in the other four WBC subtype counts. This was done in a pair-wise manner for the other four WBC subtypes. Here, only the SNPs used to instrument for basophil count explained a similar proportion of variance to another WBC count (2.44% vs. 2.39% when instrumenting neutrophil count) (**Table 3-6**). The overall results indicated that statistical power is not detrimentally diminished by adding all five WBC subtype counts in the MVMR analysis.

Table 3-6. *Pair-wise analysis of estimated proportion of variance explained for each WBC subtype count.*

	Basophil	**Eosinophil**	**Lymphocyte**	**Monocyte**	**Neutrophil**
Basophil	2.44	1.44	2.26	2.03	2.39
Eosinophil	0.52	10.43	1.49	3.47	1.59
Lymphocyte	0.75	1.73	9.04	2.73	1.97
Monocyte	0.85	2.52	3.00	13.57	2.99
Neutrophil	0.80	2.15	2.24	4.65	7.42

Blue cells are the proportion of variance explained by the MR SNPs instrumenting for the trait in column 1.

	Basophil	Eosinophil	Lymphocyte	Monocyte	Neutrophil

The other cells represent the proportion of variance explained by the SNPs instrumenting for the WBC trait in column 1 inside each WBC trait from row 1.

Therefore, the MVMR analysis was done by adding all five WBC traits into the model. Here, the MVMR IVW method estimated a protective effect of eosinophil count on overall (OR: 0.88, 95% CI: 0.80-0.97, P-value: 0.011) and female CRC (OR: 0.83, 95% CI: 0.73-0.94, P-value: 0.004) (**Figure 3-6**). This effect was more pronounced than in the UVMR analysis (OR: 0.88 vs. 0.93), suggesting that eosinophil count specifically reduces the risk of CRC. Similarly, lymphocyte count was estimated to have a protective effect on overall (OR: 0.84, 95% CI: 0.76-0.93, P-value: 0.0007) and female CRC (OR: 0.76, 95% CI: 0.67-0.86, P-value: 6.46E-05) (**Figure 3-6**). While the UVMR analysis was pointing towards a protective effect of lymphocyte count on CRC risk, the MVMR analysis showed a strong protective effect (OR: 0.84 vs. 0.94), suggesting a lymphocyte-specific action that protects against CRC development.

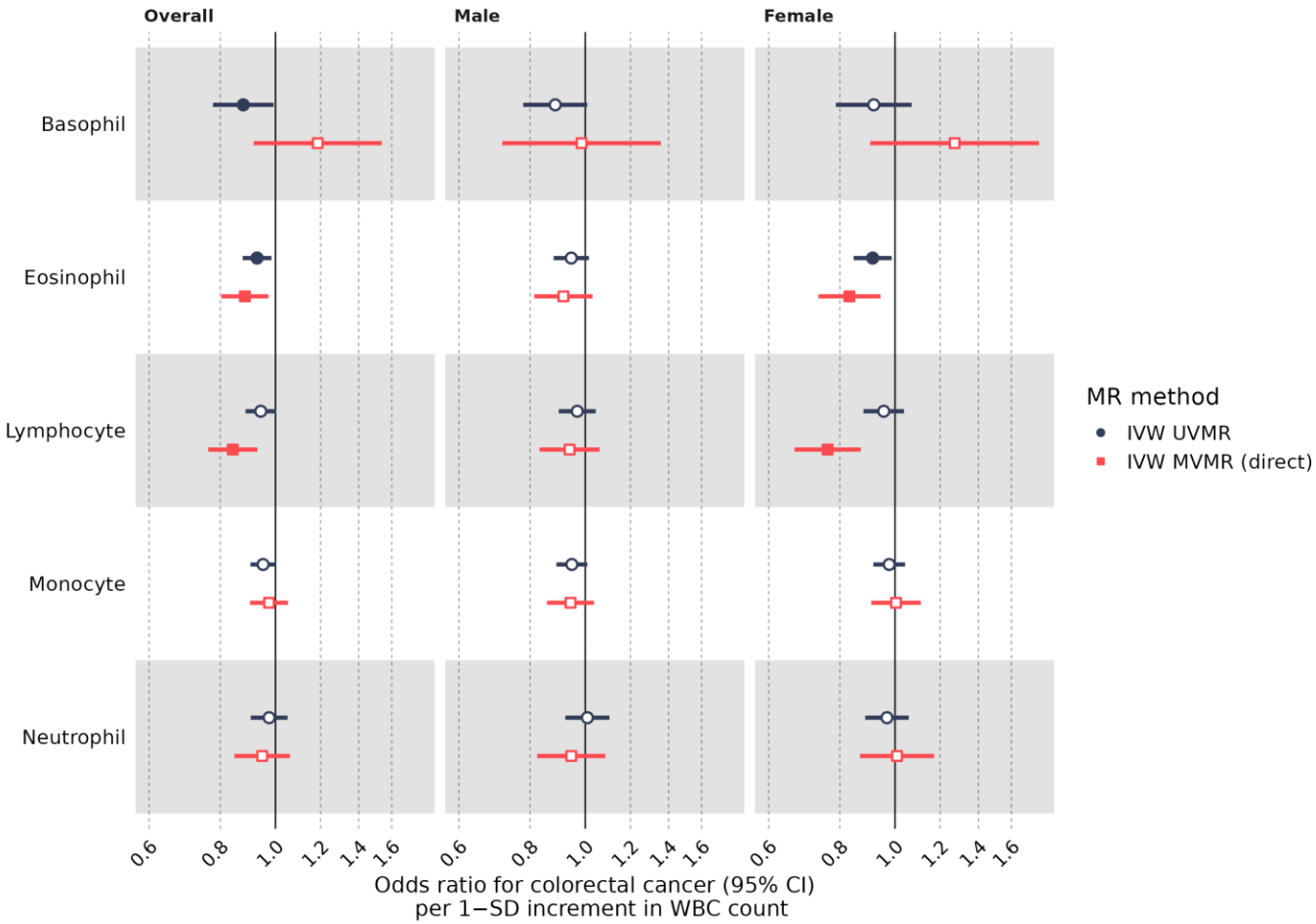


Figure 3-6. *Multivariable MR analysis of WBC count on overall CRC, and stratified by genetic sex.*

WBC subtypes are separated into rows on the X-axis. Each column is an analysis looking at overall, male-specific, and female-specific CRC. The estimated effect given by each method is presented on the Y-axis. Point estimates are filled where the P < 0.05. Results are interpreted as ORs (95% CI) for CRC per 1-SD normalized increment in WBC count.

Eosinophil count was estimated to have a direct protective effect on colon cancer (OR: 0.84, 95% CI: 0.75-0.94, P-value: 0.002), proximal colon cancer (OR: 0.92, 95% CI: 0.85-1.00, P-value: 0.042) and distal colon cancer (OR: 0.88, 95% CI: 0.74-1.00, P-value: 0.049) anatomical subsites (**Figure 3-7**, **Appendix 7**). Similar to the overall and sex-specific analyses, the MVMR analysis pointed towards an increased protective effect by eosinophil count on colon cancer compared to the UVMR analysis. Like the UVMR analysis, there was no evidence for an effect on rectal cancer by eosinophil count, suggesting a CRC subsite-specific effect. Lymphocyte count was estimated to have a direct protective effect on colon (OR: 0.85, 95% CI: 0.76-0.96, P-value: 0.007), distal colon (OR: 0.77, 95% CI: 0.67-0.88, P-value: 0.0001) and rectal (OR: 0.86, 95% CI: 0.75-0.98, P-value: 0.022) cancer anatomical subsites (**Figure 3-7**, **Appendix 7**). Unlike eosinophil count, lymphocyte count was shown to be protective across all CRC anatomical subsites, suggesting a more systemic role in reducing the risk of CRC.

Following the MVMR analysis, I applied a Benjamini-Hochberg (a.k.a. False Discovery Rate) [342] multiple hypotheses testing correction to the results. This was not done in the case of the UVMR analysis, as the genetic correlation between the WBC traits might have masked an effect which would have been detected in the MVMR analysis when adjusting between the WBC traits e.g. lymphocyte count protecting against CRC. Following multiple testing correction adjusting for 35 indepdendent tests in the MVMR analysis, only eosinophil count on distal CRC and lymphocyte count on rectal CRC had P-values >0.05 compared to the uncorrected results, adding evidence for the robustness of the MVMR results (**Appendix 7**).

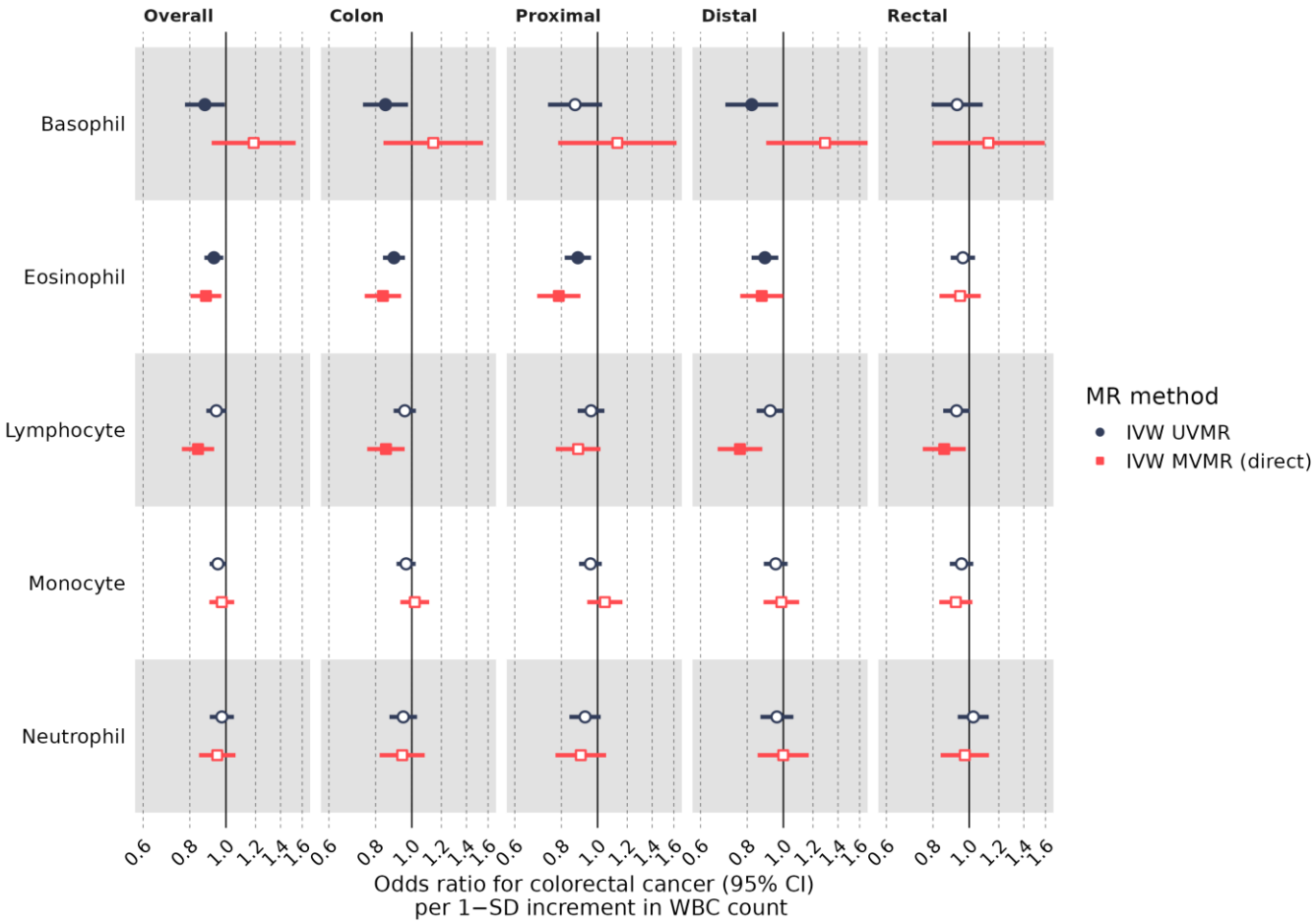


Figure 3-7. *Multivariable MR analysis of WBC count on overall, and by subsite CRC.*

WBC subtypes are separated into rows on the X-axis. Each column is an analysis looking at overall, colon-, proximal-, distal-, and rectal-specific cancer. The estimated effect given by each method is presented on the Y-axis. Point estimates are filled where the P < 0.05. Results are interpreted as ORs (95% CI) for CRC per 1-SD normalized increment in WBC count.

I undertook several sensitivity analyses to assess heterogeneity and the presence of weak instruments in the MVMR analysis. There was evidence of heterogeneity (vertical pleiotropy) in all WBC trait-CRC pairs (**Table 3-7**), as in the UVMR analysis, showing that the SNPs proxying specifically for a WBC subtype count also act on other traits downstream on the same biological pathway to affect CRC risk, which does not invalidate MR assumptions. The instrument strength analysis given by the conditional F-statistic showed evidence of weak instruments (F<10) for basophil count (**Table 3-7**). Based on these results, an additional MVMR analysis was run adjusting for weak instruments for basophil count (overall CRC OR_{Weak}: 1.3; male CRC OR_{Weak}: 1.2; female CRC OR_{Weak}: 1.7; colon cancer OR_{Weak}: 1.5; proximal colon cancer OR_{Weak}: 1.5; distal colon cancer OR_{Weak}: 1.5; rectal cancer OR_{Weak}: 1.6) on CRC (**Table 3-7**). This pointed towards a detrimental effect of basophil count on CRC risk, although these effect estimates did not have confidence intervals.

Table 3-7. *MVMR sensitivity analyses summary.*

Exposure	Outcome	Fstat	Weak	Het P-value	OR weak	OR IVW
Basophil	Colon	4.7	Yes	3.37E-14	1.5	1.1
Basophil	Distal	4.8	Yes	5.88E-07	1.5	1.3
Basophil	Female	4.7	Yes	2.72E-07	1.7	1.3
Basophil	Male	4.8	Yes	1.55E-05	1.2	0.99
Basophil	Overall	4.8	Yes	< 2.22e-16	1.3	1.2
Basophil	Proximal	4.8	Yes	8.48E-09	1.5	1.1
Basophil	Rectal	4.8	Yes	0.0011561	1.6	1.1
Eosinophil	Colon	20	No	3.37E-14		0.84
Eosinophil	Distal	20	No	5.88E-07		0.88
Eosinophil	Female	20	No	2.72E-07		0.83
Eosinophil	Male	19	No	1.55E-05		0.92
Eosinophil	Overall	19	No	< 2.22e-16		0.88
Eosinophil	Proximal	19	No	8.48E-09		0.79
Eosinophil	Rectal	19	No	0.0011561		0.95
Lymphocyte	Colon	18	No	3.37E-14		0.85
Lymphocyte	Distal	18	No	5.88E-07		0.77
Lymphocyte	Female	17	No	2.72E-07		0.76
Lymphocyte	Male	17	No	1.55E-05		0.94
Lymphocyte	Overall	17	No	< 2.22e-16		0.84
Lymphocyte	Proximal	18	No	8.48E-09		0.89
Lymphocyte	Rectal	17	No	0.0011561		0.86

Exposure	Outcome	Fstat	Weak	Het P-value	OR weak	OR IVW
Monocyte	Colon	27	No	3.37E-14		1
Monocyte	Distal	26	No	5.88E-07		0.99
Monocyte	Female	24	No	2.72E-07		1
Monocyte	Male	27	No	1.55E-05		0.94
Monocyte	Overall	29	No	< 2.22e-16		0.97
Monocyte	Proximal	24	No	8.48E-09		1
Monocyte	Rectal	30	No	0.0011561		0.92
Neutrophil	Colon	19	No	3.37E-14		0.94
Neutrophil	Distal	22	No	5.88E-07		1
Neutrophil	Female	20	No	2.72E-07		1
Neutrophil	Male	22	No	1.55E-05		0.94
Neutrophil	Overall	20	No	< 2.22e-16		0.95
Neutrophil	Proximal	23	No	8.48E-09		0.9
Neutrophil	Rectal	21	No	0.0011561		0.97

3.3.3. Phenotypic data preparation for observational analysis

After performing the MR analyses, I aimed to investigate the relationship between WBC count and CRC risk using traditional epidemiology and individual-level data from UKBB. This was done to compare and contrast with the MR analysis and to illustrate the advantages of MR over observational analyses in relation to the overarching study objective.

The first step prior to running an observational analysis is the filtering of participants, which also involves designating them as cases or controls. This is done either because of missing data, or due to particular health traits that might affect the results of the statistical analysis. For example, as mentioned in the methods section, some participants were removed if they did not have WBC count data, or if they had an illness known to affect WBC count. The number of missing values for BMI, alcohol drinker status, smoking status and Townsend DI was very low. The largest number was for the participants who preferred not to answer in the smoking status self-report questionnaire (N=1,145; 0.34% of total sample) (**Table 3-8**). Prior to analysis, those with either “missing” or “prefer not to answer” values in these variables were removed.

Table 3-8. *Participants with missing data.*

Variable	Level	Count	Percent
Body mass index	Not missing	338,841	99.6923
	Missing	1,046	0.307749
Alcohol drinker status	Missing	209	0.061491
	Not Missing	339,335	99.8376
	Prefer not to answer	343	0.100916
Smoking status	Missing	209	0.061491
	Not Missing	338,533	99.6016
	Prefer not to answer	1,145	0.336877
Townsend Depravation Index	Not missing	339,464	99.8755
	Missing	423	0.124453

336,816 participants remained after passing the filtering and selection criteria (**Figure 3-8**). I then aimed to provide descriptive statistics about the sample that I would run the observational analysis on. This was done to compare with previous observational studies of WBC count in UKBB and to help with the interpretability of findings. Moreover, it can show if there are notable differences in traits (e.g. BMI, smoking status) between cases and controls, which can then be further investigated in relation to how they might affect variation in WBC count, and if their inclusion as covariates in an observational model is warranted.

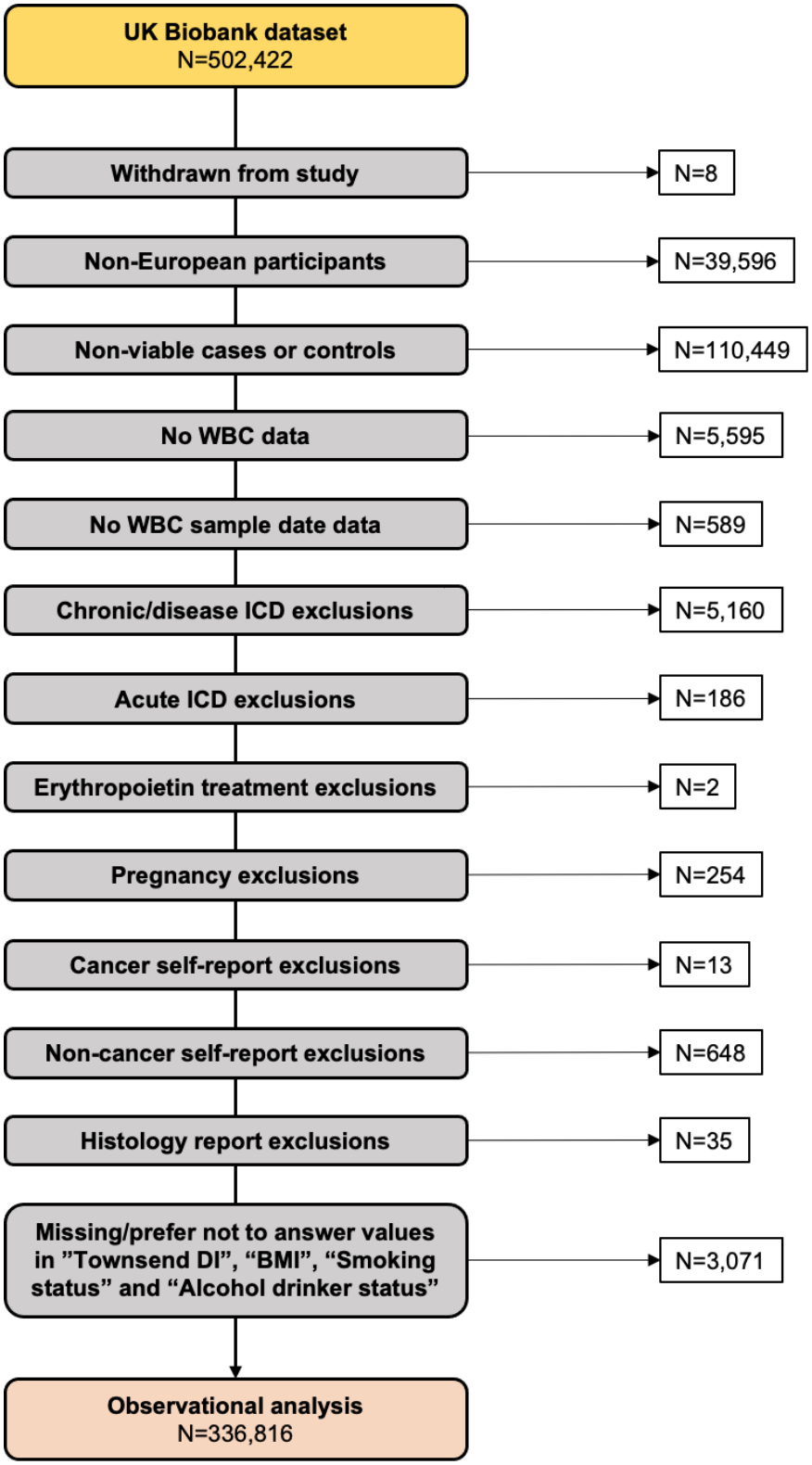


Figure 3-8. *Flowchart describing the filtering and selection criteria for the observational analysis.*

There were 332,773 controls and 4,043 incident CRC cases in the study sample. When split by genetic sex, there were 46% male controls and 57% male cases. Those with CRC were more likely to be male (57% vs. 46%), had a higher average age (60.7 vs.

55.8 years), slightly higher BMI (28.0 vs. 27.4 kg/m^2) and were more likely to have been cigarette smokers in the past (46% vs. 55% never smokers and 44% vs. 34% pervious smokers) (**Table 3-9**).

Table 3-9. *Descriptive statistics of UK Biobank study sample.*

Characteristic	Control, N = 332,773[1]	Case, N = 4,043[1]	P-valu
Sex			<0.00
Female	178,144 / 332,773 (54%)	1,727 / 4,043 (43%)	
Male	154,629 / 332,773 (46%)	2,316 / 4,043 (57%)	
Age (years)	55.794 (8.062)	60.663 (6.593)	<0.00
Body mass index	27.366 (4.750)	27.954 (4.660)	<0.00
Smoking status			<0.00
Never	183,987 / 332,773 (55%)	1,861 / 4,043 (46%)	
Previous	114,349 / 332,773 (34%)	1,790 / 4,043 (44%)	
Current	34,437 / 332,773 (10%)	392 / 4,043 (9.7%)	
Alcohol drinker status			0.7
Never	10,655 / 332,773 (3.2%)	122 / 4,043 (3.0%)	
Previous	10,982 / 332,773 (3.3%)	140 / 4,043 (3.5%)	
Current	311,136 / 332,773 (93%)	3,781 / 4,043 (94%)	
Townsend deprivation index	-1.440 (3.002)	-1.560 (2.982)	0.005
Basophil count	0.034 (0.051)	0.034 (0.056)	0.8
Eosinophil count	0.174 (0.137)	0.176 (0.131)	0.5
Lymphocyte count	1.939 (0.625)	1.933 (0.613)	0.4
Monocyte count	0.475 (0.200)	0.498 (0.177)	<0.00
Neutrophil count	4.224 (1.385)	4.350 (1.424)	<0.00
Overall WBC count	6.852 (1.745)	6.998 (1.772)	<0.00

[1]n / N (%); Mean (SD)

[2]Pearson's Chi-squared test; Wilcoxon rank sum test

Afterwards, I studied the descriptive statistics of WBC count. This was done to show how the WBC counts in the study sample compare with the general population, and if there are deviations from normal parameters e.g. if basophil count percentage of total WBC count is larger than 1% in the study sample. As a percentage of the total WBC count based on the median values, basophils accounted for 0.3%, eosinophils for 2.11%, lymphocytes for 28.16%, monocytes for 6.78%, and neutrophils for 60.39% (**Table 3-10**).

Table 3-10. *Descriptive statistics of WBC count. Units are 10^9 cells/Litre*

Trait	Mean	SD	Median	MAD[1]	Min	Max	Range	Skew
Basophil count	0.03	0.05	0.02	0.03	0	2.6	2.6	10.01

Trait	Mean	SD	Median	MAD[1]	Min	Max	Range	Skew
Eosinophil count	0.17	0.14	0.14	0.09	0	9.6	9.6	5.62
Lymphocyte count	1.94	0.62	1.87	0.55	0	79.99	79.99	8.55
Monocyte count	0.48	0.2	0.45	0.15	0	12.26	12.26	8.89
Neutrophil count	4.23	1.39	4.01	1.2	0	25.1	25.1	1.15
Overall WBC count	6.85	1.75	6.64	1.57	0	101.9	101.9	1.45

[1]Median absolute deviation

Furthermore, I aimed to study the correlation indices between each WBC subtype count, to compare with previous studies and assess and notable differences which might affect the interpretability of the results in relation to the MR analyses. Therefore, I generated a pair-wise correlation matrix between each WBC subtype. Here, low correlation (R = 0.1 to 0.3) was observed (**Figure 3-9**), suggesting that all five WBC subtypes could be added in a multivariable observational model, although this would not be as effective as the MVMR method in outlining which WBCs drive the associations with CRC risk.

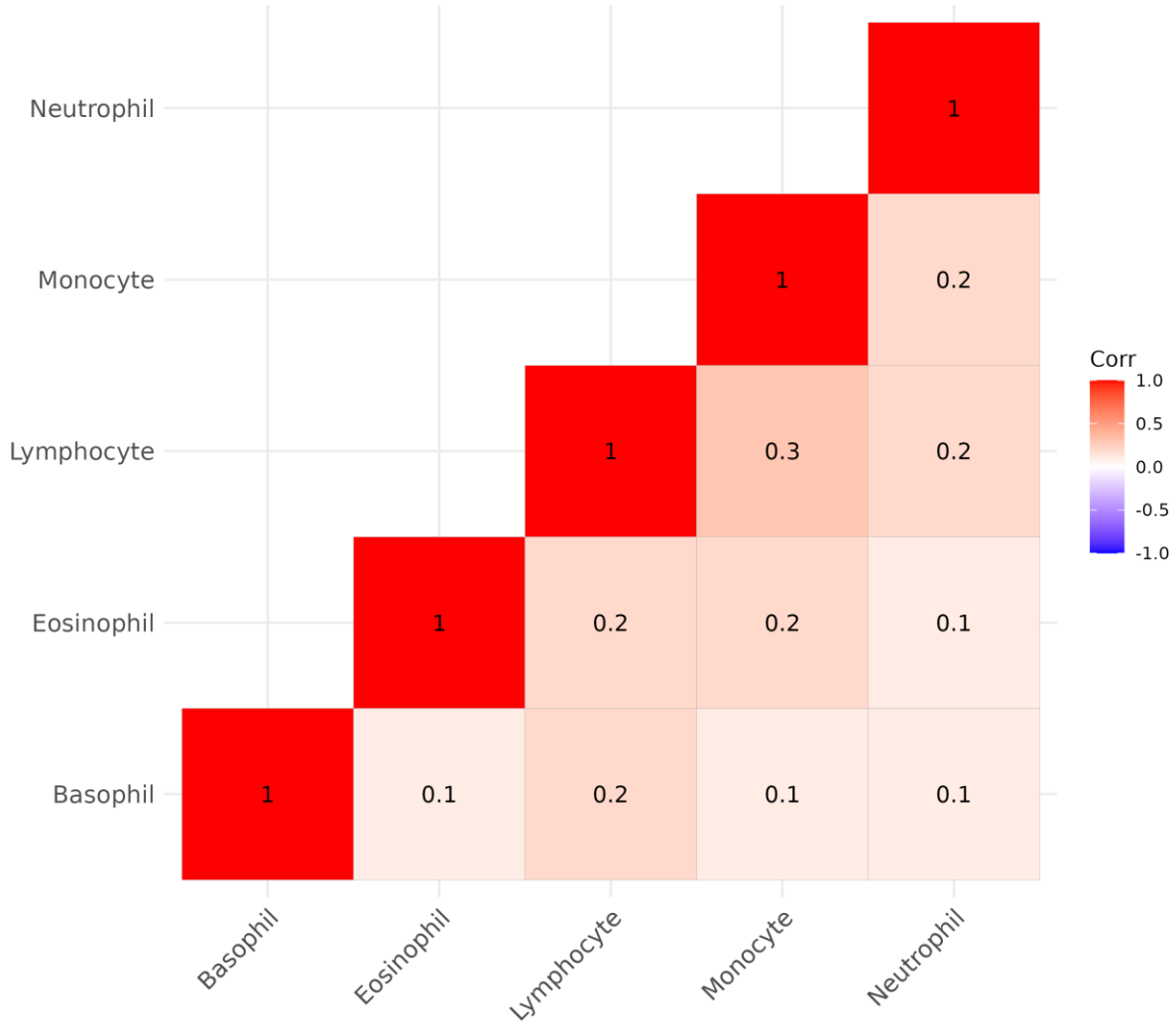


Figure 3-9. *Pair-wise correlation matrix.*
The total WBC count and WBC subtype counts were analysed between each other using Spearman correlation. WBC traits are on the X and Y axes. The number inside each square represents the correlation coefficient between the studied traits for each WBC trait pair.

I explored the variance explained by several variables on WBC count with the aim of adding them as covariates in the fully adjusted “Model 2”. This was done through an ANOVA type II approach, where I investigated the variance explained by each variable on WBC count after adjusting for all other variables.

In the main ANOVA type II analysis, geographical structure (PCs 1 to 10) had little effect on WBC count variation. Batch variables (e.g. blood sample device and sampling date), Townsend DI, and alcohol drinker status explained some of the variance in WBC count

(0% to 0.66%). Depending on the WBC subtype, genetic sex explained between 0.23% to 2.93% of the variance. This was also true for BMI, which explained between 0.14% to 3.75% of the variance, and smoking status, with values between 0.44% and 5.56% (**Figure 3-10**).

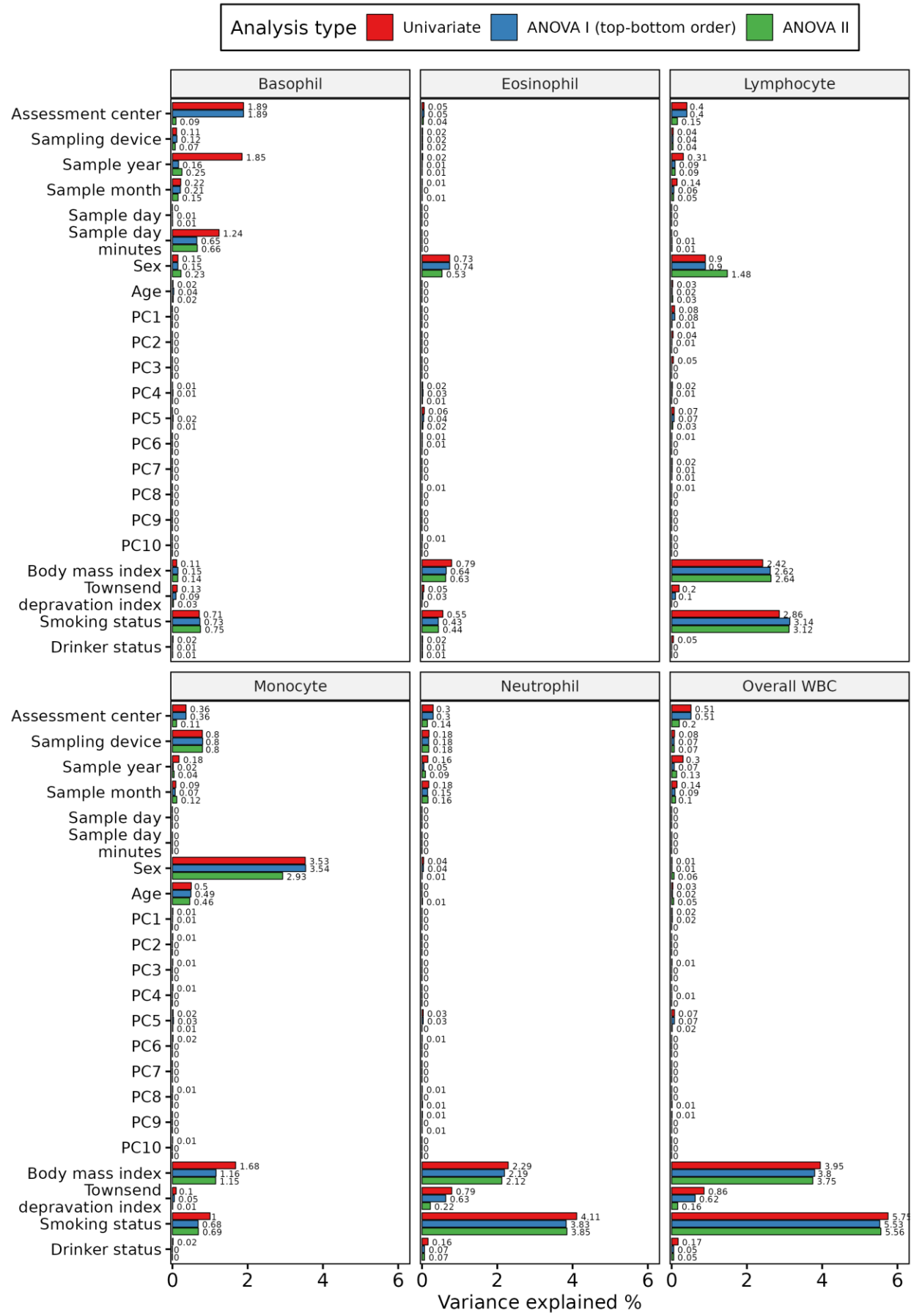

Analysis type
Univariate
ANOVA I (top-bottom order)
ANOVA II
Basophil
Eosinophil
Lymphocyte
Monocyte
Neutrophil
Overall WBC
Assessment center
Sampling device
Sample year
Sample month
Sample day
Sample day minutes
Sex
Age
PC1
PC2
PC3
PC4
PC5
PC6
PC7
PC8
PC9
PC10
Body mass index
Townsend depravation index
Smoking status
Drinker status
0
2
4
6
Variance explained %

Figure 3-10. *Variance explained by variables on WBC count.*
Univariable, type I ANOVA and type II ANOVA were run to determine the variance explained on overall WBC count, as well as each WBC subtype count, which are represented as separate facets in the figure. Studied variables were ordered top-down on the Y-axis, which is the way the type I ANOVA was run. The X-axis represents the variance explained (%) on WBC count for the corresponding variable for each analysis type.

3.3.4. Observational analysis between WBC count and CRC

After filtering participants and investigating the sample data through descriptive statistics, I proceeded to run the observational analysis. Here, the main analysis was represented by "Model 1", which used the same covariates as the GWAS of WBC count that I employed in the MR analysis [166]. This was done because my aim was to compare the observational approach with the MR analysis. However, the analysis of variance conducted above pointed to several variables explaining some of the variance in WBC count, which could affect the observational findings if unadjusted for. Therefore, as a secondary analysis, I added those variables in the fully adjusted "Model 2".

In the main analysis, "Model 1" (the minimally adjusted model), basophil count was positively associated with CRC odds (OR: 1.06, 95% CI: 1.02-1.09, P-value: 0.0005), as was monocyte count (OR: 1.05, 95% CI: 1.02-1.08, P-value: 0.003) and neutrophil count (OR: 1.09, 95% CI: 1.06-1.13, P-value: 1.94E-08) (**Appendix 8, Figure 3-11**). For male CRC, neutrophil count (OR: 1.08, 95% CI: 1.04-1.13, P-value: 0.0002) was associated with increased odds of CRC. By contrast, eosinophil count was associated with a lower odds of male CRC (OR: 0.96, 95% CI: 0.92-1.00, P-value: 0.048). For female CRC, monocyte (OR: 1.06, 95% CI: 1.01-1.11, P-value: 0.012) and neutrophil (OR: 1.07, 95% CI: 1.02-1.12, P-value: 0.006) counts were associated with an increase in CRC odds. The results of "Model 2" largely coincided with those of "Model 1", though here effect sizes for basophil count, monocyte count and neutrophil count shifted slightly towards the null. This was the opposite for eosinophil count, where there was a more pronounced negative association than in the minimally adjusted model (overall CRC OR: 0.96, 95% CI: 0.93-0.99, P-value: 0.022), possibly due to increased predictive ability by the model to give a more accurate result following adjustment for the additional covariates (**Appendix 8, Figure 3-11**). While there was no evidence of an association between lymphocyte count and CRC risk here, the "Model 2" effect sizes slightly shifted from the null towards decreased CRC risk.

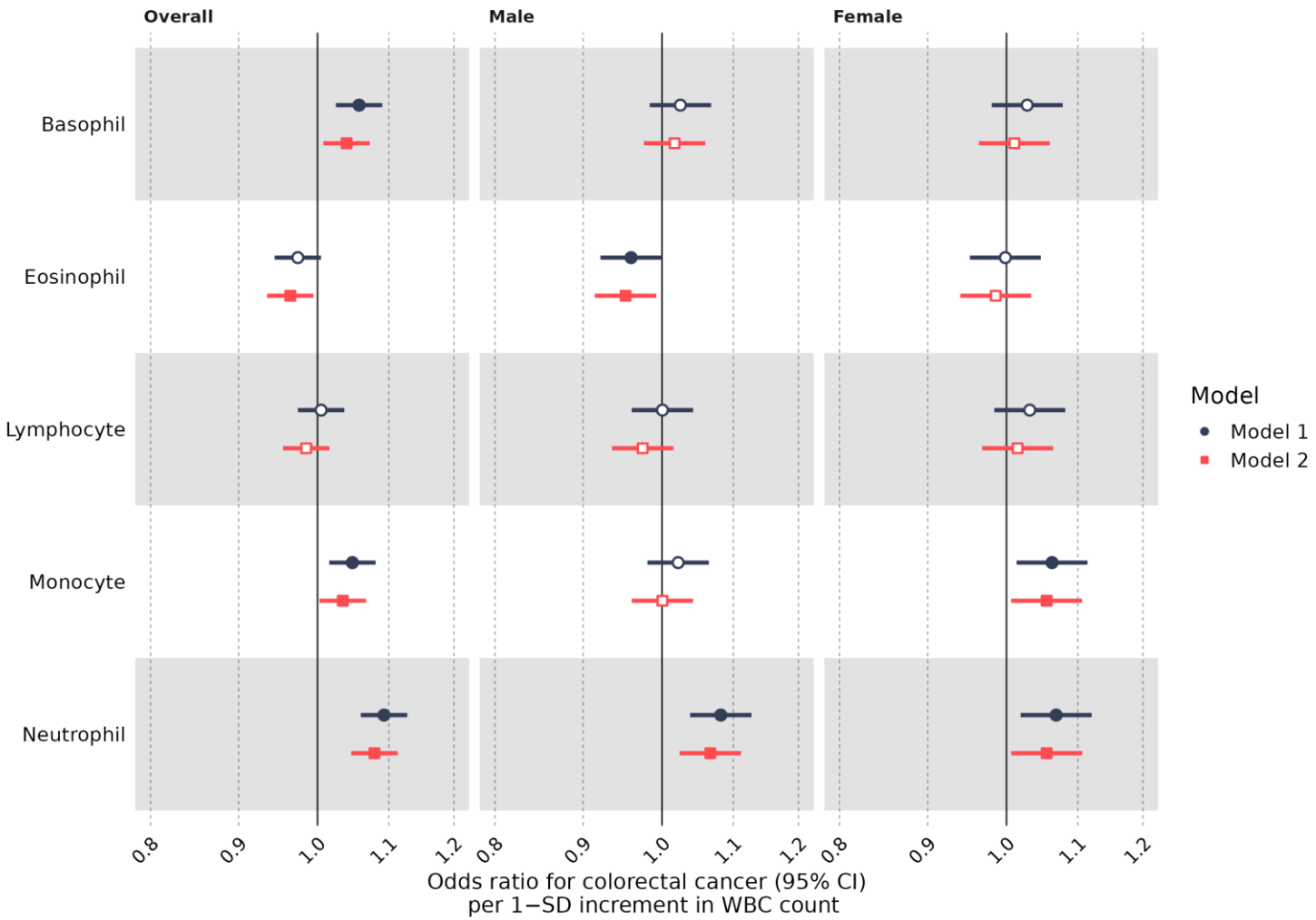


Figure 3-11. *Univariable observational analysis between WBC count and CRC risk.*

WBC subtypes are separated into rows on the X-axis. Each column is an analysis looking at overall, male-specific, and female-specific CRC. The effect sizes from each model are presented on the Y-axis. Point estimates were filled where the P-value was less than 0.05. Results are interpreted as ORs (95% CI) for CRC per 1-SD normalized increment in WBC count.

Following the univariable approach, observational associations were re-computed by adding all five WBC subtype counts together. Once again, the aim was to compare the multivariable observational analysis with the MR approach, and “Model 1” represented the main analysis. I previously studied the correlation indices of WBC count and showed that they are correlated to a small degree, making it possible for me to study the association of e.g. basophil count on CRC risk, adjusting for the counts of the other four WBC subtypes, plus the other model covariates.

In the minimally adjusted “Model 1”, eosinophil count (OR: 0.96, 95% CI: 0.93-0.99, P-value: 0.009) was associated with lower overall CRC odds, while basophil count (OR: 1.04, 95% CI: 1.01-1.08, P-value: 0.008) and neutrophil count (OR: 1.08, 95% CI: 1.05-1.12, P-value: 1.92E-06) were associated with an increase in overall CRC odds (**Figure 3-12**, **Appendix 9**). For male CRC, eosinophil count was associated with lower odds of the disease (OR: 0.96, 95% CI: 0.93-0.99, P-value: 0.009), while neutrophil count was associated with higher disease odds for both male (OR: 1.08, 95% CI: 1.04-1.13, P-value: 0.0003) and female CRC (OR: 1.05, 95% CI: 1.00-1.13, P-value: 0.046). As was the case in the univariable analysis, these results largely coincided with those from the fully adjusted “Model 2” analyses (**Figure 3-12**, **Appendix 9**). Similarly, the effect sizes shifted to the left, towards the estimates of the MR analysis.

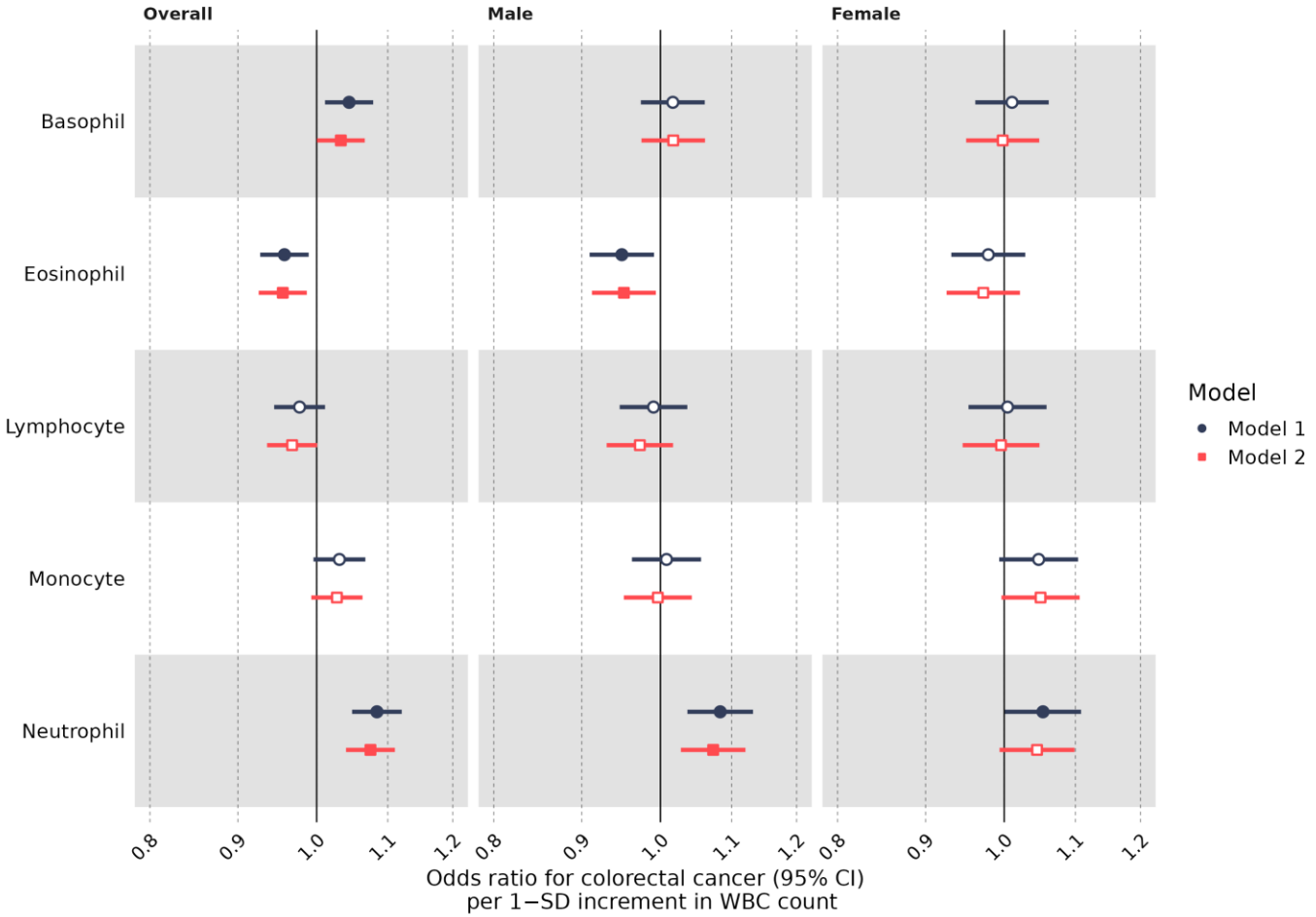


Figure 3-12. *Multivariable observational analysis between WBC count and CRC risk.*

WBC subtypes are separated into rows on the X-axis. Each column is an analysis looking at overall, male-specific, and female-specific CRC. The effect sizes from each model are presented on the Y-axis. Point estimates are filled where the P < 0.05. Results are interpreted as ORs (95% CI) for CRC per 1-SD normalized increment in WBC count.

3.3.5. Univariable MR analysis between allergic disease and CRC

Given the consistency between the UVMR, MVMR, and observational analyses for eosinophil count and given their established role in allergic disease, I conducted a follow-up analysis investigating the effect of allergic disease on CRC. The average conditional F-statistic for the allergic disease instrument was 72.49, indicating strong instruments. In the MR analysis, the IVW method demonstrated evidence for a protective effect of allergic disease on overall CRC (OR: 0.89, 95% CI: 0.82-0.96, P-value: 0.003) (**Figure 3-13**). The sensitivity analyses were largely in agreement with the IVW method. No pleiotropic SNP were suggested by the MR-PRESSO method for the male CRC outcome, and here the MR-Egger had a point estimate in the other direction compared to the other methods, although there was no evidence of an effect (**Figure 3-13**). As was the case with eosinophil count, this protective effect was displayed on colon, proximal colon, and distal colon cancers only (**Figure 3-14, Appendix 10, Table 3-11**).

Given than both eosinophil count and allergic disease had a protective effect on CRC development in the MR analysis, I decided to investigate the proportion mediated by allergic disease using two-step MR [218]. Using the two-step MR products method, allergic disease was estimated to mediate 35.39% of the eosinophil count to CRC risk relationship.

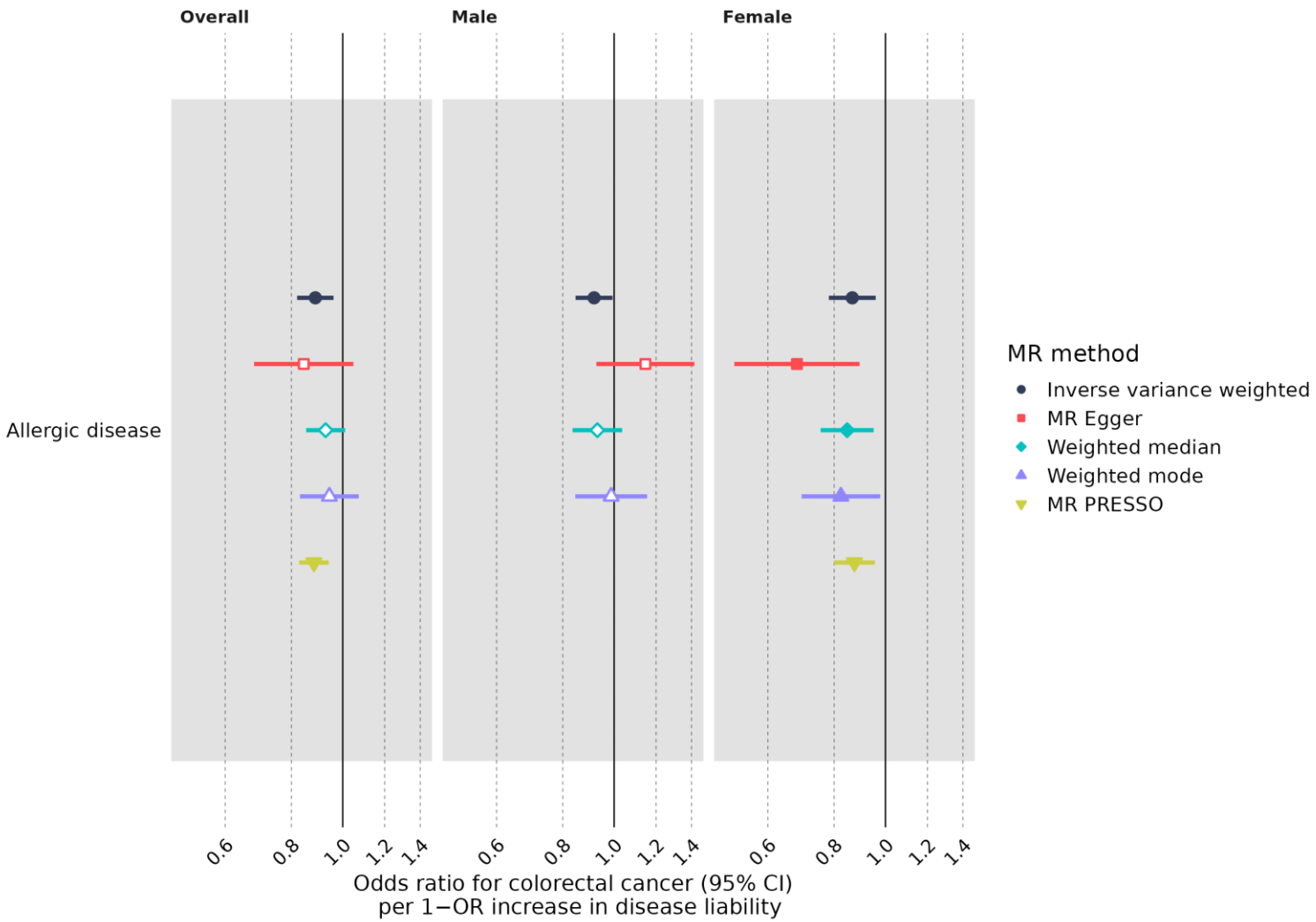


Figure 3-13. *UVMR analysis between allergic disease and CRC risk.*

Allergic disease is presented on the X-axis. The estimated effect is presented on the Y-axis. Point estimates are filled where the P < 0.05. Results are interpreted as ORs (95% CI) for CRC per 1-log(OR) increase in allergic disease.

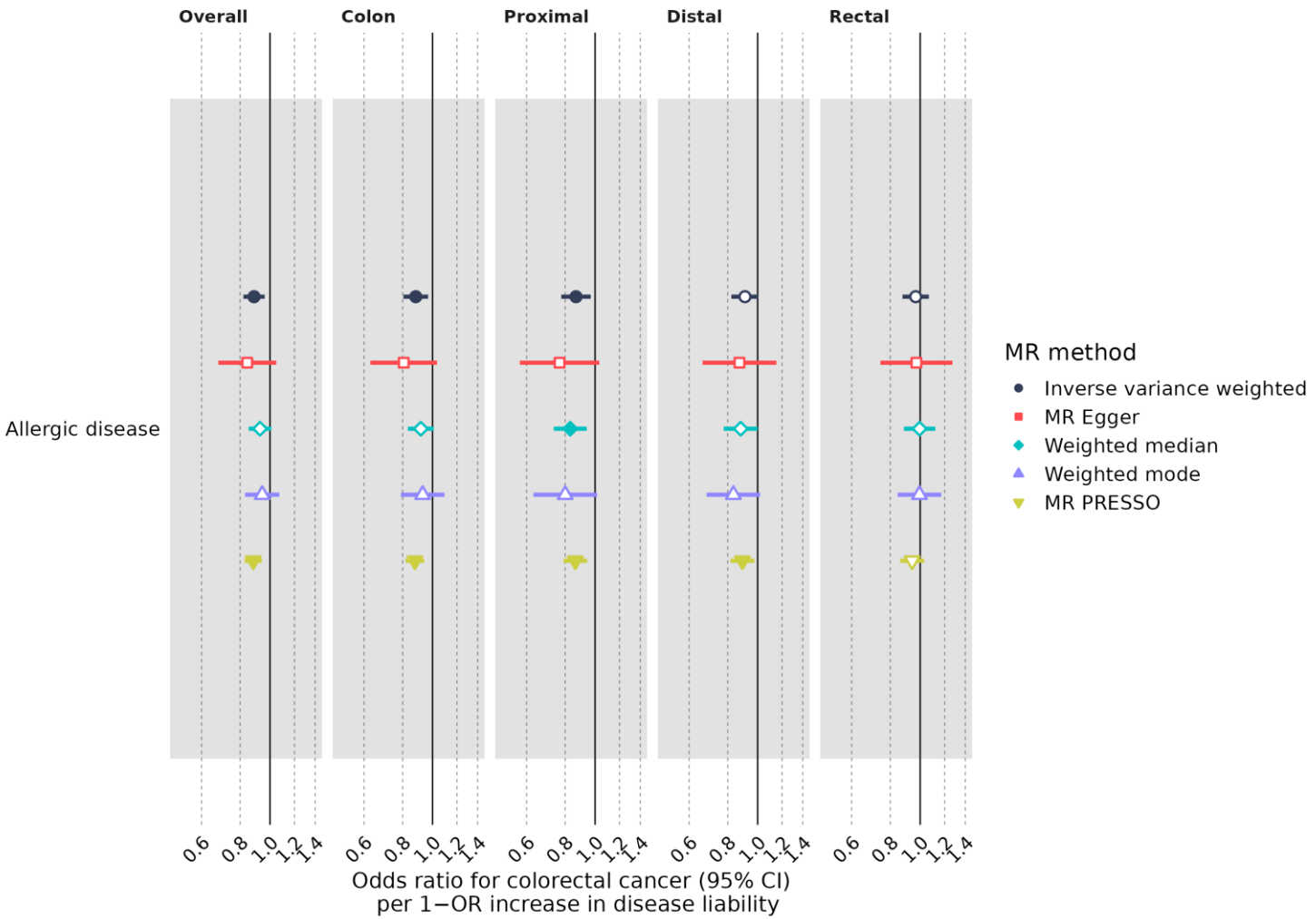


Figure 3-14. *UVMR analysis between allergic disease and CRC risk.*

Allergic disease is presented on the X-axis. The estimated effect is presented on the Y-axis. Each column is an analysis looking at overall, colon-, proximal-, distal-, and rectal-specific cancer. Point estimates are filled where the P < 0.05. Results are interpreted as ORs (95% CI) for CRC per 1-log(OR) increase in allergic disease.

Table 3-11. *Sensitivity analysis for the UVMR analysis between allergic disease and CRC.*

Exposure	Outcome	Het P	Ple intercept	Ple P	Correct direction	Steiger P
Allergic disease	Colon	1.38E-10	0.0048721	0.447312	TRUE	< 2.22e-16
Allergic disease	Distal	0.00012526	0.0022242	0.755872	TRUE	< 2.22e-16
Allergic disease	Female	5.50E-07	0.0132032	0.066381	TRUE	< 2.22e-16
Allergic disease	Male	0.05966146	< 2.22e-16	0.030156	TRUE	< 2.22e-16
Allergic disease	Overall	4.09E-13	0.002755	0.623224	TRUE	< 2.22e-16
Allergic disease	Proximal	6.52E-08	0.0066116	0.385825	TRUE	< 2.22e-16
Allergic disease	Rectal	0.00035543	< 2.22e-16	0.964617	TRUE	< 2.22e-16

3.4. Discussion

In this project I studied the relationship between the five circulating WBC subtypes and CRC risk through a combined genetic epidemiologic and observational framework.

Firstly, I want to address the potential issues with running the MVMR analysis. To address the potential issue of weak instrument bias [217], I estimated the proportion of variance explained by each WBC subtype count in the other four subtypes. Here, only data available on basophil count gave the suggestion that it might be affected by this type of bias, and I prepared for this by running a sensitivity MVMR analysis designed to adjust for it. Moreover, I tried to assess if the MVMR analysis could be compared to the observational analysis. The genetic correlation between WBC counts is greater than their phenotypic correlation [149,166,348], which can be explained due to environmental effects, some of which are unknown [348]. In the analyses taken prior to running the observational methods, I studied the phenotypic correlation between the counts of each WBC subtype. The corelation indices were low, similar to those of Welsh et al., where they studied the link between circulating WBCs and the risk of cardiovascular disease odds and mortality using UK Biobank [282]. Given the low correlations, the MVMR estimates might be better compared to the observational results rather than the UVMR analysis.

The study design was chosen to concomitantly address the four aims I laid out in the introduction which relate back to the overarching objective. First, I used UVMR to assess the relationship between WBC count and CRC risk (Aim 1). The MVMR analysis allowed me to identify if these effects, or lack thereof, were influenced by the known correlations between the WBC subtype counts (Aim 2). Afterwards, an observational analysis was done to augment the MR analysis, which aided in the interpretation of the findings (Aim 3). Therefore, the study design was useful, as it not only identified which WBC subtype counts affect CRC risk, but also which effects were present, or masked, due to the known genetic overlap between WBC counts. Moreover, the observational analysis was helpful in assessing what the MR results mean, and how MR can be advantageous in the study of blood cell traits on human disease. Finally, the combination between these three approaches aided in the discussion of the results in the context of a possible biological mechanism that could explain my findings. Taking this forward, the consistent evidence for eosinophil count prompted me to run a MR analysis between allergic disease and CRC (Aim 4).

The rest of the descriptive statistics I generated prior to the observational analysis are in line with other observational studies that have previously leveraged UK Biobank's phenotypic data. This is helpful, as it can be assumed that the results generated here are not due to an artifact such as human error. For example, the WBC parameters are comparable to those from a study in UK Biobank looking at the association between WBC count and blood pressure [281]. Seasonality has been found to influence WBC variation even when adjusting for factors such as age, sex, BMI, socioeconomic status, alcohol drinker status and smoking status [349], a finding that was confirmed by the analysis of variance I conducted here.

The UVMR analysis suggested a protective effect by basophil count in the main IVW approach, although the sensitivity analyses showed little evidence of an effect. In addition, the MVMR analysis did not show evidence of a direct effect by basophil count on CRC odds. In contrast to the UVMR analysis, the observational approach showed a positive association between basophil count and CRC risk, although this association trended towards the null in the multivariable models. A previous study showed higher basophil count associated increased lung cancer odds [279] (quartile Q4 HR: 1.22 vs. CRC RR: 1.06), supporting the observational findings in my analysis. Similarly, increased basophil count 1-6 and 6-12 months prior to CRC diagnosis has been previously reported, although the latter study found no evidence of an association [291,292]. Compared side-by-side, basophil count shows a diminishing effect size the longer the gap between blood sampling and CRC diagnosis. This trend towards no effect by basophil count is similar to what the MVMR analysis showed, providing evidence that basophil count might not affect CRC development. However, biologically, peripheral blood basophils have the capacity to infiltrate tissues [350] and their activity has been associated with cytokines IL-4 and IL-13, which have been shown to induce a pro-tumorigenic Th2 response [298,299]. Recently, a higher basophil count has also been associated with accumulation of M2 macrophages that promote a pro-tumorigenic environment [351]. Although previous studies looking at pre-treatment CRC have associated a low basophil count with worse CRC severity and prognosis [283,294], this could be a marker of poor overall immune system function rather than a basophil biologically-driven response.

These previous findings contrast with what I found in the UVMR analysis, which showed a protective effect by basophil count. However, the genetic correlation between basophil count and eosinophil count is 0.5 [166], which is far greater than the 0.1 value I found in my phenotypic correlation analysis. This exemplifies the point I have made above, where the MVMR analysis could be more representative of both observational methods. Indeed, the point estimates between the MVMR analysis and the cohort analysis were more

similar than the UVMR estimates. Moreover, in the MVMR analysis, eosinophil count was still estimated to have a protective effect on CRC risk, while basophil count was not. This means that it is likely that the effect seen for basophil count in the UVMR analysis is driven by eosinophil count, due to their high genetic correlation. In essence, MVMR was particularly useful here by overcoming the issues of both UVMR due to correlated instruments, and observational studies due to issues of confounding and reverse causation. Overall, the combined evidence across analyses suggest that basophil count does not influence the development of CRC.

Next, I studied the role of eosinophil count in CRC development. The UVMR analysis displayed a protective effect by the main IVW method, supported by most sensitivity analyses. The MVMR results were more pronounced than the UVMR analysis, further strengthening the evidence that eosinophil count plays a role in CRC development, independent of other WBC types. Similarly, the observational univariable analysis also showed a negative association between eosinophil count and CRC odds, and the strength of this association increased in the multivariable models.

Eosinophil count has been investigated in previous studies which support their protective effect in cancers. A negative association was previously reported between increasing eosinophil count and lung cancer diagnosed >1-year post blood sampling [279]. A similar study in UKBB showed increased prostate cancer odds per 1-SD increase in the trait (OR 0.96 vs. 0.93 for CRC in my analysis) [280]. A MR analysis done in UKBB to assess the effect of WBC traits on endometrial and cervical polyps found a protective effect by eosinophil count, both in the UVMR analysis (OR 0.88 vs. 0.93 for CRC in my analysis), and after adjusting for each WBC subtype count in a pair-wise MVMR setting (ORs between 0.84-0.86 vs. 0.88 for CRC in my analysis) [352]. In support of my findings, Prizment et al. identified a negative association between eosinophil count and colon cancer odds, but no association with rectal cancer odds [300]. This diminishing effect seen further from the start of the colon could be due to eosinophil numbers depending on the colorectal subsites, as eosinophil count and activity is the highest in the cecum and ascending colon and lowest in the rectum [301,353].

As trends, those who developed CRC tended to have lower eosinophil count, which rose up until the time of diagnosis, which has been attributed to increased eosinophil recruitment to the tumour microenvironment [289,354]. As mentioned in the introduction, eosinophil count was higher 1-6 and 6-12 months prior to CRC diagnosis, with no evidence of an association in the latter case [291,292]. This indicates that higher eosinophil count might have resulted from the presence of cancer rather than vice-versa. In the

observational analysis I undertook here, I studied only those participants who developed CRC at least 1-year between blood sampling and diagnosis, aiming not to restrict the number of cases to such a degree that it would greatly diminish statistical power, but at the same time limit the possibility of reverse causation. Given my findings and what is known about eosinophil count risk and trends, the results indicate a trend towards a lower risk of CRC with increasing time gap between blood sampling and CRC diagnosis, consistent with the MR results, which provides evidence for their protective effect in CRC development.

Eosinophils have also been investigated in relation to cancer survival. Increased eosinophil recruitment to the CRC tumour site was associated with better survival, even when adjusting for the effects of $CD8^+$ T-cells [288]. This independent anti-tumoral effect in CRC cells was also observed in vitro [309], and eosinophil-specific granule secretion of granzyme A has been linked with the killing of CRC cells [306]. Moreover, eosinophil density in the tumour microenvironment has been associated with an increase in E-cadherin, a protein that links cancer cells together and hampers their metastatic potential [355].

Eosinophils are well known for their role in allergies, including asthma and allergic rhinitis [356], and previous MR analyses have reported a causal effect of eosinophil count on allergic disease [149,357]. It might therefore be expected for there to be a link between allergies and CRC risk. Indeed, a systematic review looking at the relationship between allergies and cancer suggested a reduced risk of CRC in those with allergic diseases [358]. In a follow-up MR analysis, I discovered that allergic disease is protective against overall CRC development, across all anatomical subsites except the rectum, which was consistent with the eosinophil count results. In their letter to the editor, Yuan et al. showcase an MR analysis between allergic disease and three cancers: oesophageal, gastric, and colorectal [359]. Here, they used the same instruments to proxy for allergic disease that I used in my analysis and found a protective effect by allergy on CRC (overall CRC OR: 0.91 CI: 0.83-0.99 vs. OR: 0.93 CI: 0.88-0.98 in my study) [359]. At the same time, a recent MR study did not find an effect of allergic disease on breast and prostate cancer risk [360], suggesting a specific anti-tumorigenic effect by allergies on CRC development. Overall, the findings generated in my study provide evidence that eosinophil count plays a protective role against CRC.

The findings relating to eosinophil count could be a helpful steppingstone for future studies that could explore a mechanistic effect on CRC. For example, the GTEx database [361] could be used to extract expression quantitative trait loci (eQTLs) that map to genes associated with eosinophil effector proteins (e.g. *RNASE2* encoding EDN) [306].

This could be then used in an MR for a more mechanistic approach to assess how eosinophils could reduce the risk of CRC. Protein quantitate trait loci (pQTL) data [362] could also be used to directly investigate circulating eosinophil proteins on CRC risk, although such datasets are more sparse in comparison. Regarding allergic disease, the role of eosinophils could be explored through a RNA-Seq laboratory analysis [363]. Here, samples from healthy controls and from those with allergic disease could be taken and analysed, outlining differences in gene expression between eosinophils in healthy and affected individuals. Highlighted genes could then be studied further in relation to CRC development.

Returning to the results, the next WBC studied was lymphocyte count. There was limited evidence for a protective effect of lymphocyte count on overall CRC in the UVMR. Interestingly, the MVMR estimates indicated lower ORs for CRC across all anatomical subsites and female CRC. The observational analysis showed no evidence for an association between lymphocyte count and CRC odds in the univariable models, while the multivariable fully adjusted "Model 2" indicated that there may be a negative association with the disease. Employing MVMR was again useful here, as it aided in identifying an effect specific to lymphocytes that was not identified in the main UVMR analysis. The genetic correlation between the counts of lymphocytes with monocytes and neutrophils, both which were not found to have an effect on CRC risk, might be responsible for this pull towards the null in the UVMR analysis.

Tumour-infiltrating lymphocytes (TILs) are lymphocytes which infiltrate the tumour environment and help combat tumour growth through direct action and recruitment of other immune cells [284]. These include $CD8^+$ T-cells and $CD20^+$ B-cells that have been shown to reduce tumour growth and promote cytotoxic effects on tumour cells, both in tandem and independently [364]. High levels of TILs has been associated with better CRC overall and disease-free survival [284,313]. Interestingly, while an increase in $CD3^+$ and $CD8^+$ T-cells was associated with better prognosis in right-sided CRC, an increase in $FoxP3^+$ T-cells was associated with improved prognosis in rectal CRC [365], indicating separate biological mechanisms of lymphocyte action depending on the CRC primary tumour anatomical subsite. The most likely explanation for the MR findings in terms of CRC risk, given the consistency across the results and anatomical subsites, is the known role in surveillance by T-cells and NK cells to detect and destroy potentially cancerous cells [366,367]. While lymphocyte count was higher 1-12 months prior to CRC diagnosis [291,292], this most likely indicates production and recruitment of lymphocytes to the site of pre-cancerous or non-detectable tumours. Overall, my findings, along with those from the

current literature, indicate that elevated lymphocyte count protect against the development of CRC.

As in the case for eosinophil count, future studies could further untangle the biological mechanism through which lymphocytes might prevent CRC development. For example, the counts of a specific lymphocyte type e.g. CD8+ T-cells and NK cells could be studied in relation to CRC risk. Another possibility is to use multiple trait colocalization, a method which can leverage QTL data to provide evidence for the lymphocyte subpopulation that is driving the effect seen in my MR analysis [368]. This can also be used for eosinophil count, to find genetic variants that colocalize between a particular eosinophil population, eosinophil count, and allergic disease, establishing a clearer mechanistic view.

Next, I studied monocyte count, where both the UVMR and MVMR analyses showed little evidence for an effect. However, monocyte count was associated with increased overall and female CRC odds in the observational analysis. Neutrophil count, however, was strongly association with CRC odds in both the univariable and multivariable observational models but again there was little supporting evidence from the MR analyses. Two other observational analyses in UKBB reported a positive association between monocyte and neutrophil count, and cardiovascular disease and lung cancer odds [279,282]. Interestingly, another UKBB study looking at WBC count and blood pressure found neutrophil count to be the strongest association with blood pressure in the observational analysis [281]. However, similarly to what I found in my study, their UVMR and MVMR analyses did not show evidence for an effect by monocyte nor neutrophil count, which the authors attribute to residual confounding, reverse causation or an acute effect influencing WBC count in the observational model [281].

3.4.1. Limitations

Finally, there were a number of limitations in my study. Firstly, sex-specific WBC count instruments were not available for use in the MR analysis, and therefore sex-combined WBC count data was used even in the sex-specific CRC MR analysis. A general observation around the analyses I conducted was the difference in results in the sex-specific CRC. These differences make sense, as the immune systems of men and women are different in their activity, efficiency, and WBC count distributions [369,370]. Women, for example, have been shown to have a more active and efficient immune system [371–373], although this comes with an increase in autoimmune conditions [370]. Interestingly, associations in my observational analyses were different to those from the MR analysis. For example, an increase in lymphocyte count indicated a reduced risk in

men and no evidence in women, while the MR analysis showed evidence for a reduced CRC risk in women and only indicated a protective effect in men. In support of the lymphocyte count MR results, CD4+ and CD8+ T-cell activity and ability to enter tumour tissues has been identified to be better in women than in men [374]. This indicates that the MVMR analysis could better assess the direct effect by lymphocyte count, given the ability to adjust for genetic correlation, rather than an observational study affected by confounding. Moreover, the observational analyses might have had reduced accuracy in comparison to the MR analyses, given the difference in tested cases (N=1,727) compared to the MR approach (N=24,594).

Another explanation is that of sex heterogeneity in instruments used to measure for WBC count is affecting the MR results. Only sex-combined summary statistics for WBC count were available to use in my analysis. These were adjusted for sex, but if there are sex-specific mechanisms through which WBCs act to reduce or increase the risk of CRC, this does not eliminate the potential for sex-combined exposures to bias MR estimates. Three separate MR commentaries do point out this issue [200,375,376], and Gao et al. show that there are differences in MR estimates when studying the relationship between combined and sex-specific BMI instruments on sex-specific outcomes i.e. breast and prostate cancers [377].

Indeed, there are sex-specific differences in CRC incidence (overall and by site) [378], metabolic activity [379], modifiable risk factors [378], and genetic architecture [380]. For example, the role of sex hormones e.g. oestrogen in eosinophil activity has been documented, although the mechanisms through which they act to affect eosinophils, and particularly in cancer, is currently unknown [381]. Given this, the MR results for WBC count and allergic disease on sex-specific CRC risk should be interpreted with this in mind. A future analysis could be done to explore the effect of sex-specific WBC count on sex-specific CRC. The smaller sample-size would likely diminish the number of SNPs used in an MR analysis, leading to a loss in precision, but at the same time could point to a more accurate result.

Secondly, only baseline blood measurements were available for conducting the observational analysis. This assumes that WBC counts were constant and did not allow for establishing of a relationship between a trend in WBC count and its relationship to CRC odds. Nevertheless, baseline WBC count measurements have been previously been shown to be associated with disease risk [279–282,382], making this study into CRC development a worthwhile endeavour.

Thirdly, while there was evidence from the MVMR analyses that lymphocyte count had a protective effect on CRC, it is not known if this is the result of increasing levels of T-cells, B-cells, NK cells, or all of them combined. Furthermore, the analyses here did not study a specific population of WBC cell subtypes, such as $CD8^+$ T lymphocytes or $CD11b^+$ eosinophils, both which have been found to present specific anti-tumour effects in laboratory-based studies [284,308].

Fourthly, the MVMR method used here may not be reliable when investigating traits with very weak instruments [217]. This was the case for basophil count, as the F-statistic was estimated to be between 4.7 and 4.8 (**Table 3-7**). Therefore, despite pointing to an increased detrimental effect compared to the main MVMR analysis, ORs derived from the weak-MVMR analysis should be interpreted with this in mind.

Finally, generally MR, including the analyses done here, assume that the level of exposure is constant over time [184] and that the causal effect is linear and homogeneous [383]. While differences between WBC count in people are thought to be stable through life [384], it is not known how an acute rise in the levels of a certain WBC subtype might lead to CRC tumourigenesis. Therefore, my analyses should be interpreted as looking at how lifetime differences between the counts of WBCs affect the risk of CRC.

3.4.2. Conclusion

In summary, through a combined MR and observational analysis, the results here suggest that the biological consequences of having elevated circulating levels of eosinophils and lymphocytes protect against colorectal cancer development, implicating these cell types and their biological roles in this process. Additionally, a follow-up MR analysis suggested a protective signal detected when assessing allergic disease, which were – as anticipated – similar to the eosinophil count findings. Nevertheless, additional research is needed to disentangle the biological mechanisms and pinpoint a specific pathway through which these WBC subtypes act to reduce the odds of developing CRC, including sex-specific cases.

In this chapter I used particular methodological approaches to study the relationship between WBCs and CRC, a disease of global significance [265]. In **Chapter 5** I have studied the role of neutrophil count in P. falciparum severe malaria [385], another disease of global significance and one that presented a different set of methodological challenges. Principally, the need to identify people in UK Biobank of the African

continental ancestry group. This work is presented in **Chapter 4,** preceding the study presented in **Chapter 5**.

CHAPTER 4. A FRAMEWORK FOR RESEARCH INTO CONTINENTAL ANCESTRY GROUPS OF THE UK BIOBANK

Chapter summary

In this chapter I present my study undertaken on the non-White British populations in UK Biobank (UKBB) [164] which has been published in "BMC Human Genomics" [238]. The main aim was to identify UKBB participants that correspond to the African continental ancestry group (CAG), although those corresponding to the European, South-Asian and East Asian CAGs were also identified due to the identical methodological framework (**Figure 4-1**). The work presented here was a prerequisite for **Chapter 5**, where I generated a genome-wide association study (GWAS) of neutrophil count in people of African ancestry, allowing me to conduct a Mendelian randomization (MR) analysis [169] between neutrophil count and *Plasmodium falciparum* (*P. falciparum*) severe malaria.

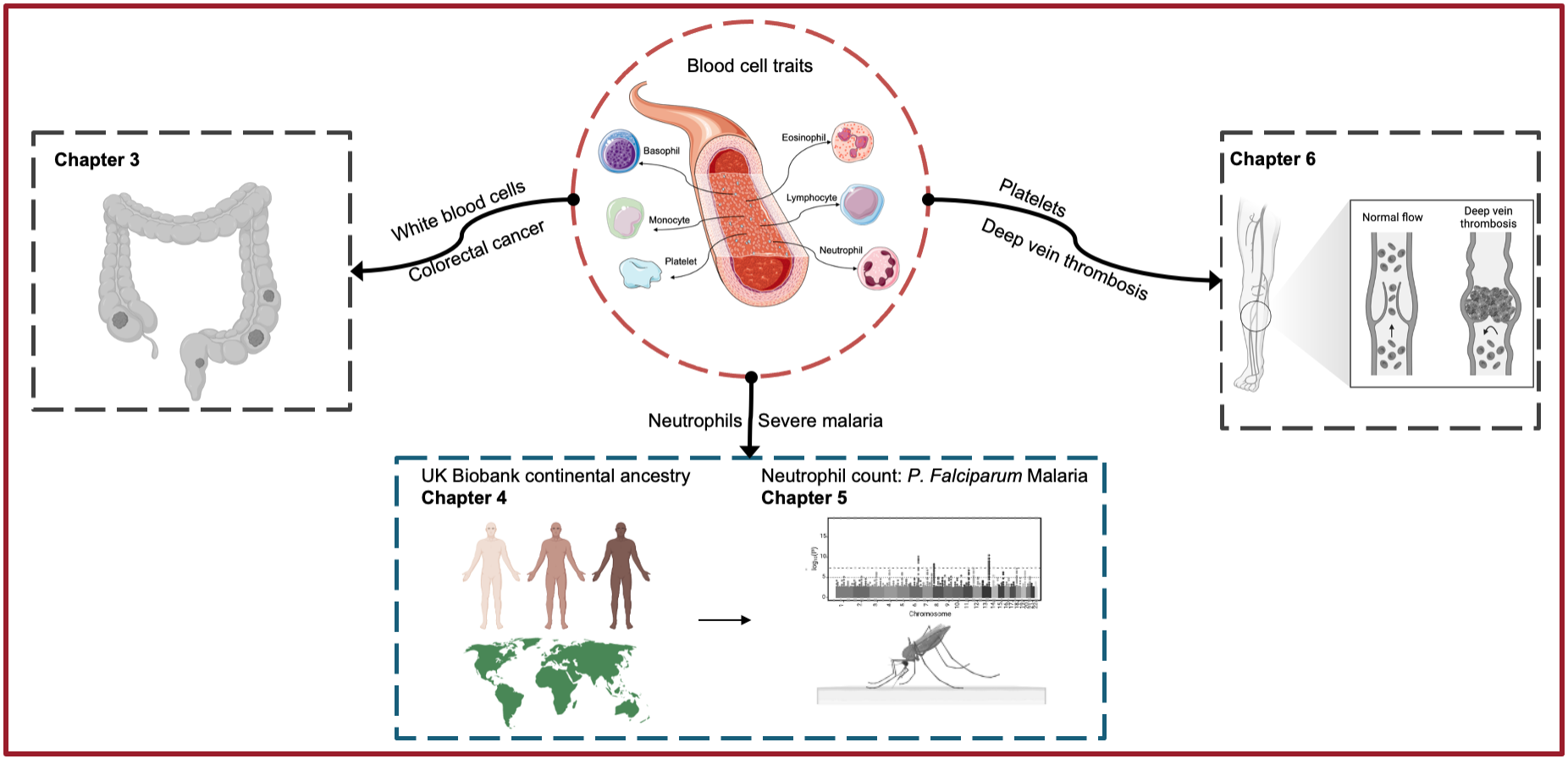


Figure 4-1. *PhD project and current chapter (4 - coloured).*
Created with Microsoft PowerPoint and BioRender.com.

4.1. Introduction

As the research community strives to understand the genetic architecture of disease [347], it has increasingly realised the necessity of inclusion and diversity – of ethnically, ancestrally, environmentally, and geographically diverse populations [386–389]. Not simply to enhance knowledge about health and disease, but to ensure health equity. As alluded to in **Chapter 1**, epidemiological studies including GWAS have been overwhelmingly conducted in European populations [386].

4.1.1. Current studies in diverse populations

Funding efforts and studies including the Human Heredity and Health in Africa (H3Africa) Initiative [390], the Population Architecture using Genomics and Epidemiology (PAGE) Consortium [391], Trans-Omics for Precision Medicine Consortium [392], Hispanic Community Health Study / Study of Latinos (SOL) [393], and the All of Us Research Program [394] are making concerted efforts to include and increase the number of under-represented populations in genomic epidemiology studies. More recent initiatives have focused on improving the understanding of genetic variation across ancestries using large-scale approaches, such as the Million Veterans Program [162] in the US, BioBank Japan (BBJ) [395] and UKBB [161,164].

4.1.2. UK Biobank

The UKBB project has phenotypic and genomic data from a prospective cohort of approximately 500,000 individuals from across the United Kingdom (see **Chapter 2**) [161,164]. It has become an outstanding resource for studies of health and disease, and genetic diversity within the United Kingdom. While it is made up of around 430,000 “white British ancestry” individuals, as defined by UKBB, it also contains a wealth of diversity from other self-described ethnicities (~78,000) [164]. This is a resource that should be utilized to help expand inclusion and diversity in epidemiological studies. The Pan-UK Biobank, or the Pan-ancestry genetic analysis of the UKBB, has leveraged the diversity present in UKBB and is freely providing GWAS summary statistics for over seven thousand phenotypes in six continental ancestry groups (https://pan.ukbb.broadinstitute.org). Studies and public resources like Pan-UK Biobank are vital to the goal of increasing under-represented populations and the larger goal of describing and understanding the genetic architecture of phenotypic traits and disease.

4.1.3. Current limitations

One of the aims of my thesis is to explore the relationship between neutrophil count and the severity of P. falciparum malaria using MR analysis. The most at-hand approach would be to use the summary statistics for neutrophil count generated from the UKBB European sample of ~400,000 people, as described in studies such as Astle et al. [149]. However, one of the requirements of two-sample MR (2SMR) is that the exposure and outcome datasets come from the same underlying population (**Chapter 2**) [200]. Indeed, the genetic architecture between ancestral populations can differ due to linkage disequilibrium (LD) block size or allele frequency differences [396–398], and this applies to traits such as blood cell traits (BCTs, including neutrophil count) [166,399,400]. This makes it unlikely that using SNP instruments for neutrophil count from an European GWAS would yield reliable MR results with the outcome data coming from an African ancestry GWAS of severe malaria caused by *P. falciparum* [245].

However, there has been limited information on intra-population structure in the Pan-UK Biobank, which would be valuable given that, for example, Africa harbours more genetic diversity than any other continent [401]. Moreover, the GWAS models were run in a non-specific manner (e.g. same covariates used across traits) without taking into account the biological particularities of a trait such as BCTs. These might bias association effect-sizes and in turn could then have a downstream effect on post-hoc analyses, such as MR [339,402–404].

Therefore, a description of the continental diversity and population structure present in the non-white British participants of UKBB, specifically the African CAG, would aid me in conducting a GWAS of neutrophil count and run a MR analysis. More broadly, this would aid future study designs, methodological choice(s) and ultimately improve the understanding of how genotype influences phenotype.

4.1.4. Main study objective

Given these challenges, my main objective was to identify a relatively homogenous group of individuals corresponding to the African CAG that would approach a population consistent with a Hardy-Weinberg equilibrium (HWE) model (i.e. homogeneous) [168] and is resultantly more appropriate for many of the assumptions built into many of the methods used in genomic epidemiology studies, such as MR [405,406]. Given the identical methodological approach to complete this objective, I also expanded my work to find individuals from the non-white British segment of UKBB that correspond to the European (EUR), South-Asian (SAS) and East-Asian (EAS) CAGs. Nevertheless, the prime focus will be the AFR CAG, as this is the dataset that I will use in **Chapter 5**.

4.1.5. Study aims

I have divided this chapter's main objective into three separate aims I will try to address:

1) Identify individuals from UKBB that could be assigned to the AFR CAG i.e. would be similar to a population sampled on the African continent
2) Describe the AFR CAG in the context of its population structure and identification of homogeneous clusters within this CAG
3) Assess the reliability of the methods to ensure that the AFR CAG dataset can be used in **Chapter 6** for conducting a GWAS

As a final note, genetic "ancestry" groups identified within my study refer to groups of individuals with a shared genetic ancestry and demographic history. I define "ancestry" here as genetic ancestry or the complex inheritance of one's genetic material, but in practice I will be using methodologies that use genetic similarity to identify groups of individuals with high genetic affinity or likeness [407].

4.2. Methods

4.2.1. Study design

Public data from the 1000 Genomes Project (1KG) [152] was used to provide reference populations from four CAGs – namely, AFR, EUR, SAS and EAS. First, an ADMIXTURE analysis was done to assign non-white British UKBB participants into one of the four CAGs (**Aim 1**). Next, I performed a principal component analysis (PCA, see **Chapter 2** for details) to study the degree of population structure and identified homogeneous sub-clusters within each CAG (**Aim 2**). Afterwards, I assessed the reliability of the data by comparing it with the 1KG reference panel (**Aim 3**) (**Figure 4-2**). The groups and clusters identified here are used as discrete units, but ancestry does not have decisive boundaries and is a continuum [408–411]. Therefore, the use of discrete units is an analytical simplification.

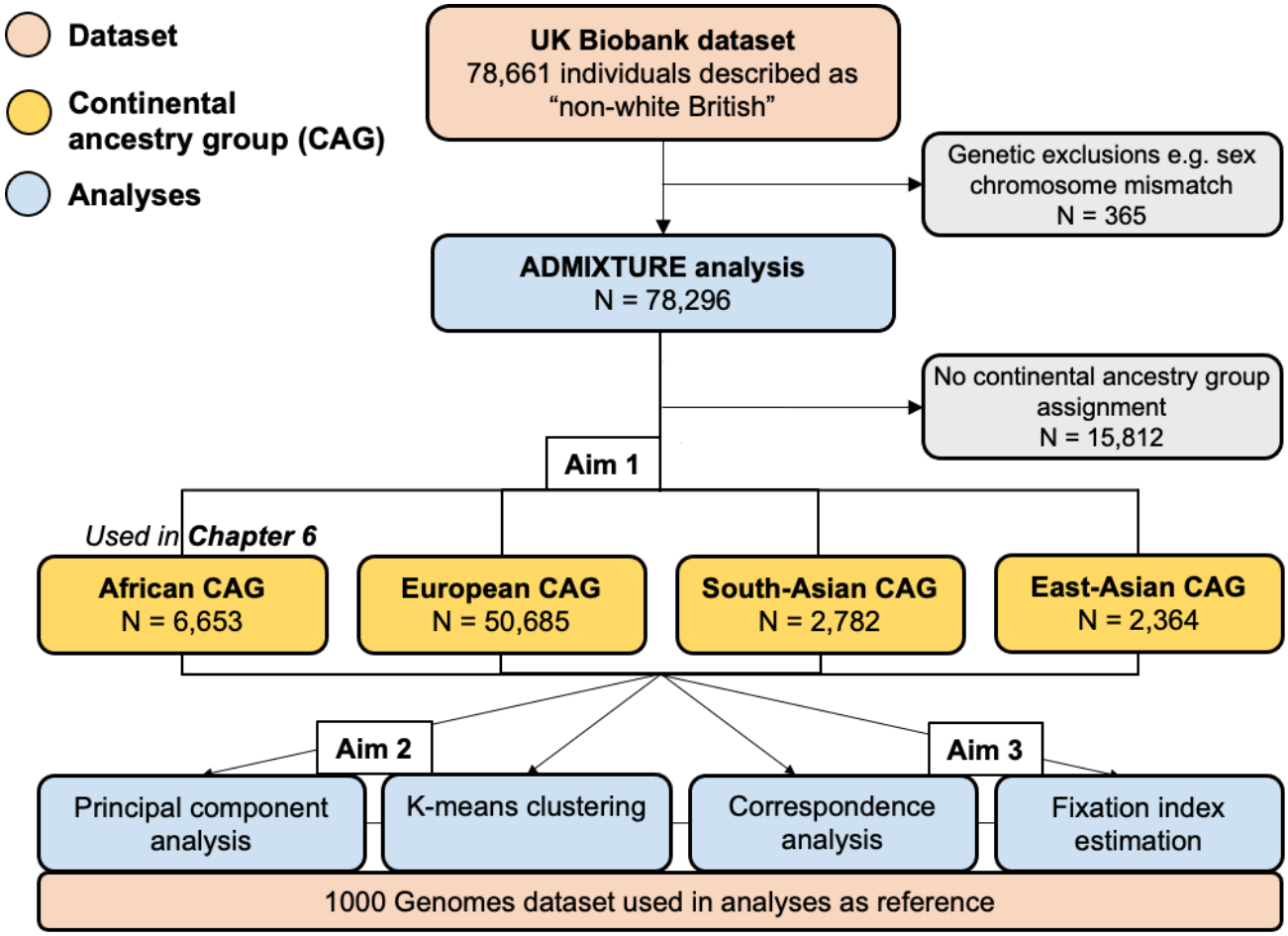


Figure 4-2. *Study design of the chapter.*

4.2.2. UK Biobank genetic data

I used the directly genotyped SNP data from UKBB (N=784,256 SNPs) (see **Chapter 2** for more details). It includes data for a total of 78,661 individuals identified by UKBB as "non-white British" participants – my analyses were restricted to this subset. In addition to genotypic data, I also acquired several variables of interest (self-reported ancestry, country of birth) data for this subset of individuals. 365 exclusions were made when filtering those with sex chromosome mismatch and/or aneuploidy, and outliers with high genetic heterozygosity (HWE test P-value < 0.0001) and missing rates (>0.015, refers to the proportion of missing SNP information for an individual as a proportion of all genotyped SNPs) [412].

4.2.3. 1000 Genomes data

We used genetic data (v5a.20130502) from phase three of the 1KG, which includes data from 5 1KG described super-populations [Europe (EUR), East Asia (EAS), South Asia (SAS), Africa (AFR), and the Americas (AMR)] to provide reference populations for admixture analyses and

population structure inferences ([413] http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/). Our analyses did not include populations from the AMR superpopulation. This is to maintain a simplified analysis that avoided the complicating factors of the potentially recent admixture events that occurred in the Americas, such those from the European colonial period [414]. Included in our analyses are five populations from 1KG super-population label: (AFR), also known as the continental Africa ancestry group (1) Yoruba in Ibadan, Nigeria (YRI); (2) Luhya in Webuye, Kenya (LWK); (3) Gambian in Western Division, The Gambia - Mandinka (GWD); (4) Mende in Sierra Leone (MSL) and (5) Esan in Nigeria (ESN). Five populations from the super-population label EUR or the European CAG: (1) Utah residents with Northern and Western European ancestry (CEU); (2) Toscani in Italia (TSI); (3) British in England and Scotland (GBR); (4) Finnish in Finland (FIN) and (5) Iberian populations in Spain (IBS). Five populations from the super-population label SAS or the South Asian CAG: (1) Gujarati Indian in Houston, Texas (GIH); (2) Punjabi in Lahore, Pakistan (PJL); (3) Bengali in Bangladesh (BEB); (4) Sri Lankan Tamil in the UK (STU) and (5) Indian Telugu in the UK (ITU). Finally, five populations from the super-population label EAS or the East Asian CAG: (1) Han Chinese in Beijing, China (CHB); (2) Japanese in Tokyo, Japan (JPT); (3) Han Chinese South (CHS); (4) Chinese Dai in Xishuangbanna, China (CDX) and (5) Kinh in Ho Chi Minh City, Vietnam (KHV).

4.2.4. Merging UK Biobank and 1000 Genomes

The directly genotyped data from UKBB was used to identify SNPs with the same SNP identifier (RefSNP ID) present in the 1KG data set. I identified a total of 718,711 SNPs with the same ID and extracted them from both data sets using PLINK v2.0, after which the two datasets were merged. After removing problematic SNPs (e.g. multi-allelic, duplicate) in the merge step, a total of 718,487 SNPs remained.

4.2.5. Linkage disequilibrium pruning

SNPs are in linkage disequilibrium (LD) when the probability of an allele in SNP A is correlated to another allele in SNP B i.e. they are not random [415]. This is shown by the r^2 correlation factor, and a cut-off r^2 value is used to state that two or more SNPs are in LD with each other [415]. In programs such as PLINK, the top SNP with the highest minor allele frequency (MAF) in an LD block (i.e. many SNPs in LD over a genomic region) is kept, with the other SNPs being pruned/excluded [416]. Pruning is done in genetic analyses to speed-up the computational time for an analysis (as multiple SNPs in LD can be represented by the top SNP) [221] or to get more accurate results from a PCA by avoiding potential bias of eigenvectors towards high LD regions [223]. Prior to ancestry estimation, I reduced the merged data to a set of independent SNPs based on LD estimates using the PLINK v2.0 function and parameters “--indep-pairwise

50 10 0.025", indicating an r^2 threshold of 0.025, a window size of 50 kilobases (kb) and a window step size of 10 kb. In addition, 24 previously identified genomic regions with extensive linkage disequilibrium were also excluded [398,417]. A total of 30,320 SNPs remained following LD pruning.

4.2.6. Estimating African, European, South Asian, and East Asian ancestry

We included four 1KG populations as reference populations in a supervised ADMIXTURE (v1.3.0) analysis. They were (1) British in England and Scotland (GBR), of the EUR ancestry superpopulation, (2) Yoruba in Ibadan, Nigeria (YRI), of the AFR superpopulation, (3) Indian Telugu in the UK (ITU), of the SAS superpopulation, and (4) Han Chinese South (CHS), of the EAS superpopulation. These singular population samples were chosen to broadly represent each of their four respective continental (superpopulation) ancestry groups. The supervised ADMIXTURE analysis provides, for each UKBB sample, a proportion of ancestry for each of the four reference populations. Those individuals with at least 80% of their ancestry attributed to one CAG were carried forward into further analyses.

4.2.7. Derivation of continental principal components

Afterwards, I performed a PCA analysis to determine the degree of population structure in the African and three other CAGs (see **Chapter 2**). First, I identified unrelated individuals in each CAG and the 1KG datasets using all 718,487 SNPs in the overlapping data set. This was done with the PLINK (v1.9) function --rel-cutoff, and a minor allele frequency (MAF) filter of 0.05 (--maf 0.05) was applied. Second, for each CAG and using all (1KG + UKBB) unrelated individuals assigned to the CAG, a list of approximately 40 thousand LD independent SNPs were identified using the PLINK (v2.0) function --indep-pairwise 50 10 0.025 (--indep-pairwise 50 10 0.02 for AFR and --indep-pairwise 50 10 0.05 for SAS) along with a MAF filter of 0.01, and the exclusion of the 24 previously identified genomic regions with extensive linkage disequilibrium [398,417].

Third, new PLINK files including only the LD independent SNPs identified in step two were subsequently generated. I used the smartrel package from the EIGENSOFT (https://github.com/DReichLab/EIG) software to generate a new list of related individual pairs, after which I produced a list of related individuals to exclude from the PCA [226,228]. An exception in this step was made for the European CAG, as its sample-size was too large to run smartrel. Instead, the list of unrelated individuals generated from step one was used. Finally, smartpca

of the EIGENSOFT package was used to estimate principal components (PC) using only unrelated UKBB samples.

Related and 1KG samples were subsequently projected upon these PCs by smartpca. I excluded sample outliers from the PC analysis by smartpca with the following parameters: using 10 PCs to identify outliers (numoutlierevec), at six standard deviations (SD) from the mean (outliersigmathresh), and with 5 outlier removal iterations (numoutlieriter). **Appendix 11** provides numbers for each of these steps, for each CAG. The EUR CAG was treated uniquely due to its larger sample-size. I ran smartpca twice as described above, once with "fastmode=NO" and then with "fastmode=YES". The former provided estimates of the eigenvalues but not the eigenvectors, while the latter provided eigenvectors but not eigenvalues.

4.2.8. K-means clustering of principal components

Due to the large degree of population structure seen in the PCA of the CAGs, I conducted a K-means clustering analysis [418] with the aim of dividing each CAG into clusters that would resemble more homogeneous sub-groups (K-pops).

However, how does one know how many clusters to divide a dataset into? While this process is arbitrary in the end, there are multiple methods that can estimate the appropriate number of *k* clusters to use in a dataset. One such method is called the Silhouette analysis, which determines how well each data point lies within its cluster, and computes an average value of that value [419]. This can be done over *k* clusters to identify which cluster has the largest average silhouette width [419]. The K-means clustering analysis will also aid in **Chapter 5** to assess the reliability of SNPs associated with neutrophil count in my GWAS.

For each CAG, I estimated the variance explained by each principal component (PC) by dividing the eigenvalue of each PC by the sum of all eigenvalues. To identify the number of top PCs I generated a scree plot, using the variance explained estimates, and visually identified the elbow or valley in each plot. I then used the top PCs in an unsupervised K-means clustering analysis (*k* set from 2 to 20; using the function "kmeans()" from the R stats package) to identify clusters of UKBB individuals that maximize between cluster sums of squares and minimize within cluster sums of squares. Afterwards, an optimum number of clusters (k) was identified by silhouette analysis [419].

4.2.9. Correspondence analysis

Next, I aimed to assess the validity of the K-means clustering analysis. One would expect that if the PCs are indicative of population structure [226], then the K-means clustering algorithm should theoretically provide clusters that represent geographical regions. To test this, I used the self-report country of birth and K-means cluster data to perform a correspondence analysis (CA). A CA is similar to a PCA, although it simplifies a large dataset to inform on patterns between rows and columns [420] – in this case between K-pops and COB.

Each UKBB study participants' COB information matched with United Nations (UN) defined geographic regions (**Appendix 12**) to assign a region of birth (ROB) for each participant. To determine if the K-means population clusters have any relationship with an individual's COB or ROB, I performed CAs using the function "ca()" from the R package "ca", for each CAG [421]. In addition, a chi-square test was performed on the contingency table used in the correspondence analysis. Any COB or ROB with fewer than 10 observations was excluded and individuals for which COB information was not available were also excluded.

4.2.10. Population differentiation among K-means population clusters

Afterwards, I computed the population differentiation value a.k.a. fixation index (Fst) between each CAG [422]. For each CAG, I took the best K-means population clusters, as defined by the silhouette analysis, and re-ran smartpca. However, on this run I had smartpca provide for only an estimation of the average Fst for each pair of populations in the data set, including 1KG populations and UKBB K-means clusters. This was done with the inclusion of the paramaters "fstonly" and "phylipoutname" [423], the latter of which provides a distance matrix of mean Fst values between populations. For the African CAG specifically, I calculated the mean and maximum Fst value in a pair-wise manner between each K-means cluster to better understand the intrapopulation variability and detect outliers or large differences between K-pops.

4.2.11. Description of working environment

All analyses were performed in a Linux environment supported by the University of Bristol's Advanced Computing Research Centre (ACRC) using the following publicly available software packages: PLINK v1.9 and v2.0 [247,416], ADMIXTURE v1.3.0 [222,424], and EIGENSOFT v8.0.0 [226,228]. Scripts, analyses, and figures were run and generated in the R environment using version 3.6.2 on the ACRC computer clusters and version 4.0.2 (Taking Off Again) on local computers [345].

4.3. Results

4.3.1. Estimations of continental ancestry

We included each of the 78,296 UKBB "non-white British" individuals in a supervised ADMIXTURE analysis to estimate a proportion of ancestry to each of African (AFR), European (EUR), South Asian (SAS), and East Asian (EAS) continental ancestry groups (**Figure 4-3**). The proportion of continental ancestry is further illustrated for each individual within the context of UKBB population structure on principal components (PC) one and two as provided by the UKBB (**Figure 4-4**). AFR ancestry (**Figure 4-4A**) runs largely parallel with PC1, the major axis of variation. EUR ancestry runs at a roughly 135-degree angle (**Figure 4-4B**) along PC1 and PC2, while SAS (**Figure 4-4C**) and EAS (**Figure 4-4D**) ancestry run, largely, along PC2.

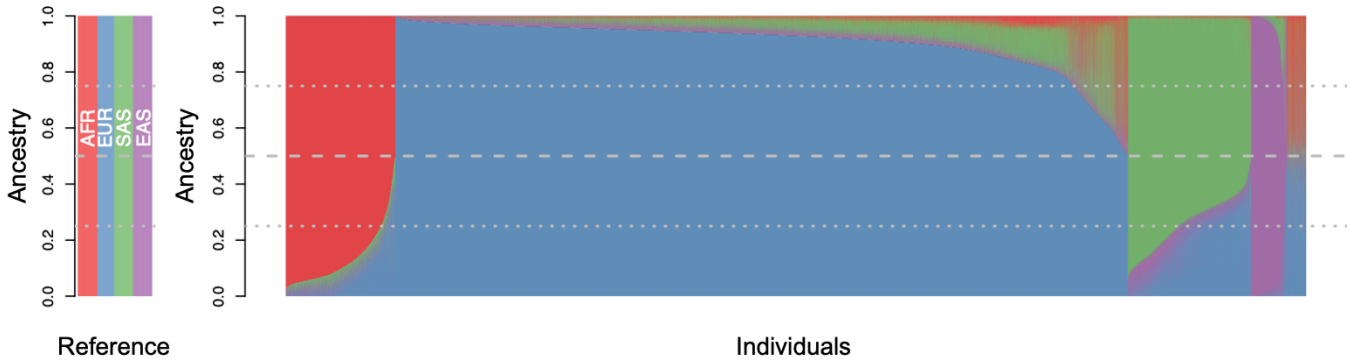


Figure 4-3. *ADMIXTURE analysis.*
The x-axis is each non-white British participant of UKBB, while the y-axis indicates the % continental ancestry by using the 1KG dataset as reference.

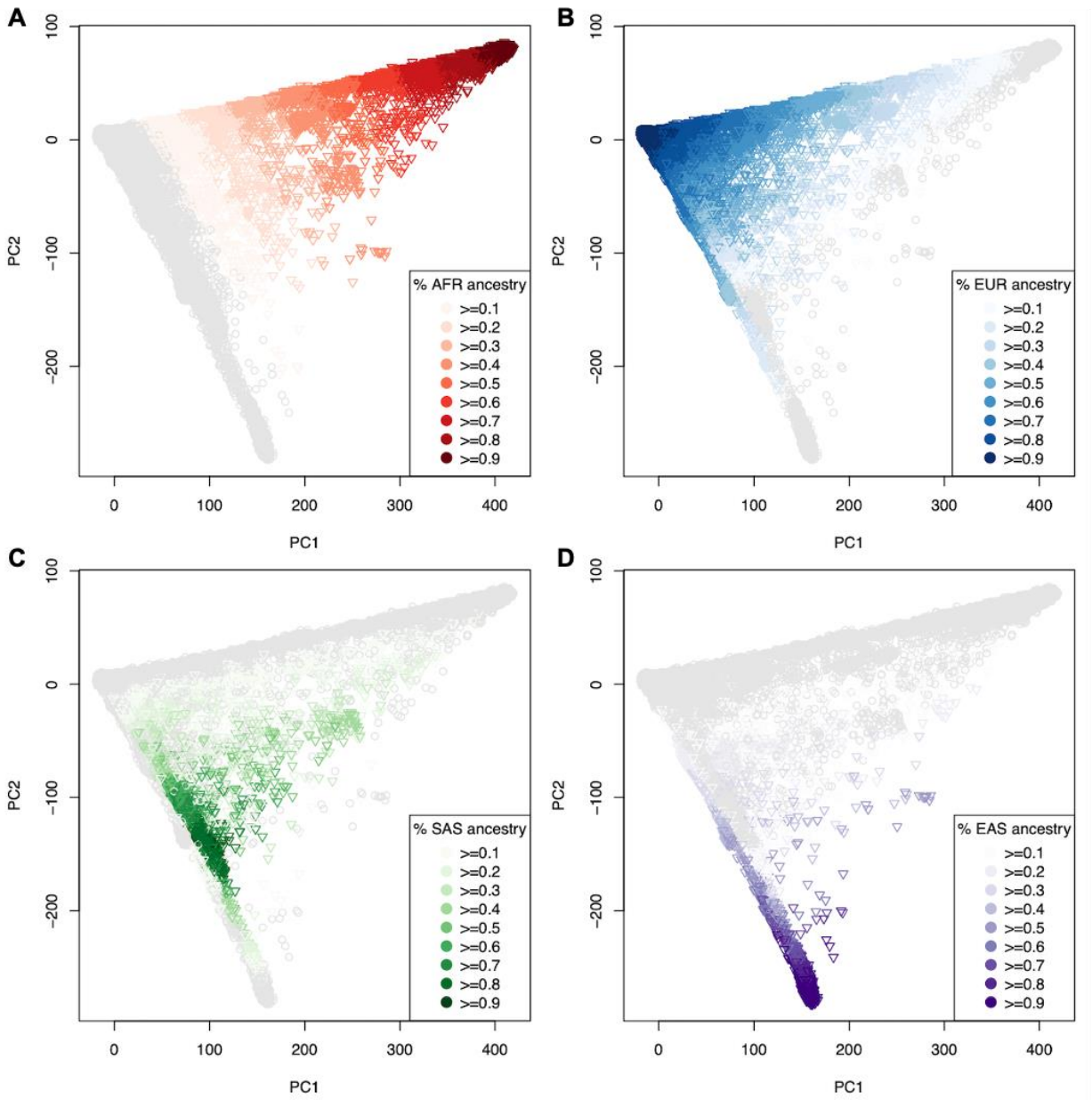


Figure 4-4. *PCA of UKBB non-European samples.*
Ancestry proportions on UKBB PCs: Continental African (A), European (B), South Asian (C), and East Asian (D) ancestry proportions placed on principal components one and two supplied by the UK Biobank.

Of the approximately 78,000 UKBB samples included in the ADMIXTURE analysis 50,685, 6,653, 2,782, and 2,364 individuals had 80% or more of their ancestry attributed to the EUR, AFR, SAS, and EAS continental super-populations respectively. I carried these individuals into further analyses of population structure within these CAGs. The 80% threshold was chosen to allow some error in the broader continental classification while also placing a limit on the complex structure and admixture evaluated in these

subsets. A total of 15,812 “non-white British” UKBB study participants were not included in any of the four CAGs, given the methods and cut-offs used here.

I zoomed in on the African CAG sample and explored the degree of population admixture in the AFR CAG dataset using the data generated by the ADMIXTURE and PCA analyses. East Asian and South Asian ancestry did not seem to follow a pattern on the PC1 and PC2 plane, and the median admixture was only around 2% (**Figure 4-5**). On the other hand, the degree of European vs. African admixture followed a linear pattern on the PC2 axis, indicating structured admixture (**Figure 4-5**). More information on PCs for all CAGs is available in **Appendix 13**.

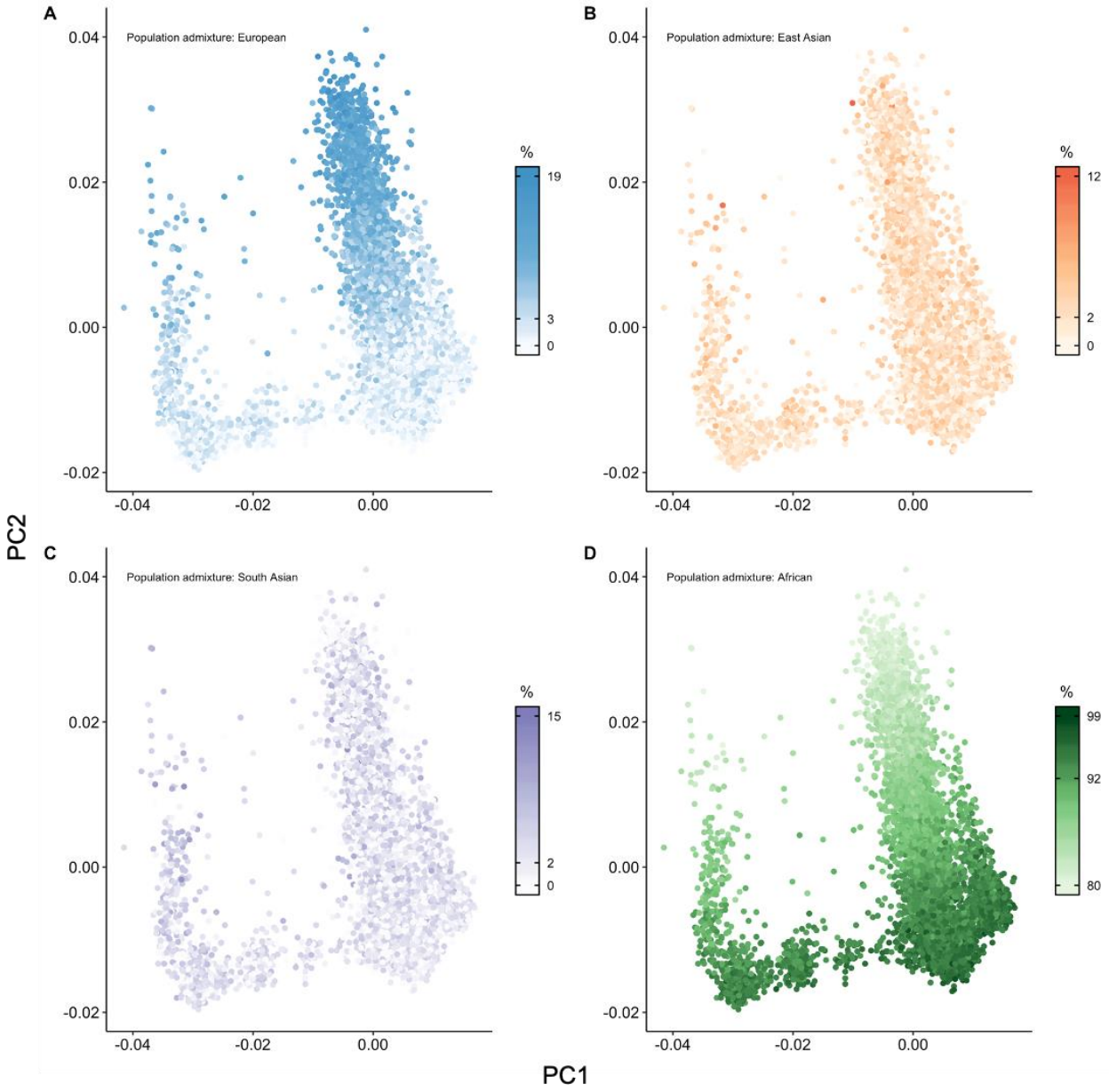


***Figure 4-5**. Admixture in the AFR CAG sample.*

Ancestry proportions on UKBB PCs: European (A), East Asian (B), South Asian (C), and African (D) ancestry proportions placed on principal components one and two from the ADMIXTURE analysis.

4.3.2. Population structure within continental regions

To evaluate the level of population structure among the UKBB CAGs, I first re-estimated PCs for each CAG, while also projecting individuals from 1KG populations from each super-population respectively onto the newly derived PCs (**Figure 4-6, Appendix 11**). For each CAG there was considerable overlap between UKBB individuals and 1KG populations, providing some context for the diversity that is present within the UKBB. For example, in the AFR CAG, PC1 distinguishes West African from East African 1KG populations, while PC2 distinguishes among populations of West Africa (**Figure 4-6**). In the EUR continental ancestry group, the PCs and 1KG populations illustrate a strong North-South axis along PC2, with a similar but less distinctive trend on PC1 (**Appendix 14**).

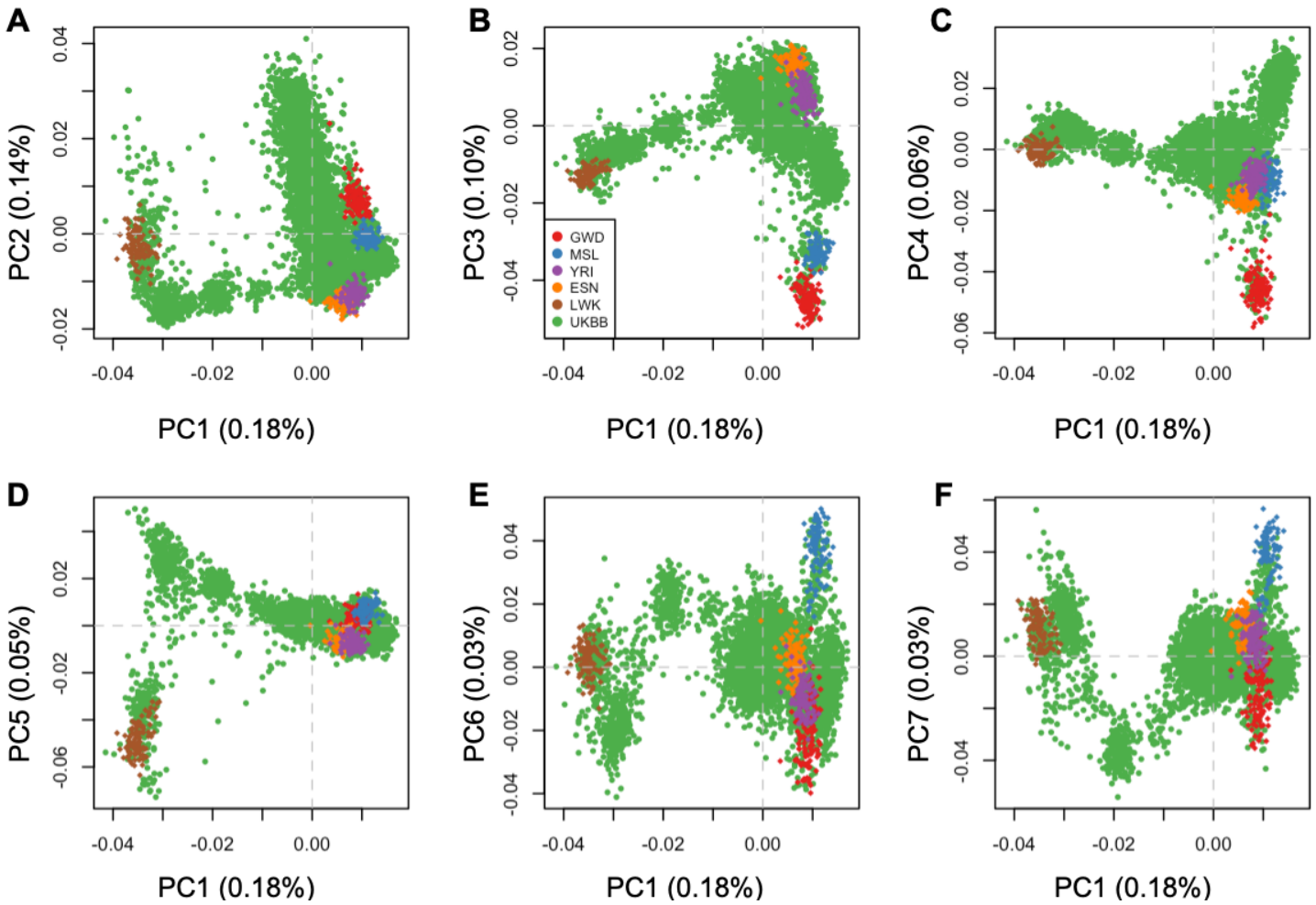


Figure 4-6. *UKBB AFR CAG sample with 1KG African sub-populations projected on PCA plot.*

PC1 on PCs 2-7 (A-F) with variance explained. GWD = Gambian in Western Division – Mandinka; MSL = Mende in Sierra Leone; YRI = Yoruba in Ibadan, Nigeria; ESN = Esan in Nigeria; LWK = Luhya in Webuye, Kenya; UKBB = African CAG identified in my study.

4.3.3. K-means clustering of PCs

Given that many population genetics and epidemiological analyses, such as GWAS, depend on limited population structure, a common desire is to have a relatively homogeneous population sample for these analyses [154]. As such, I used an the K-means unsupervised algorithm to identify groups of individuals that approach HWE population assumptions [168].

To do so I performed a K-means analysis on the top PCs from each CAG to identify 'K' subclusters/groups. An optimum number of K-clusters was determined by a silhouette analysis. For the African CAG I identified seven K-clusters (**Figure 4-7**), and two, four and three for the EUR, SAS and EAS CAGs respectively (**Appendix 15**). A *k* of six (second best fit) was used for the EUR CAG instead of two to better inform on the structure present in this sample.

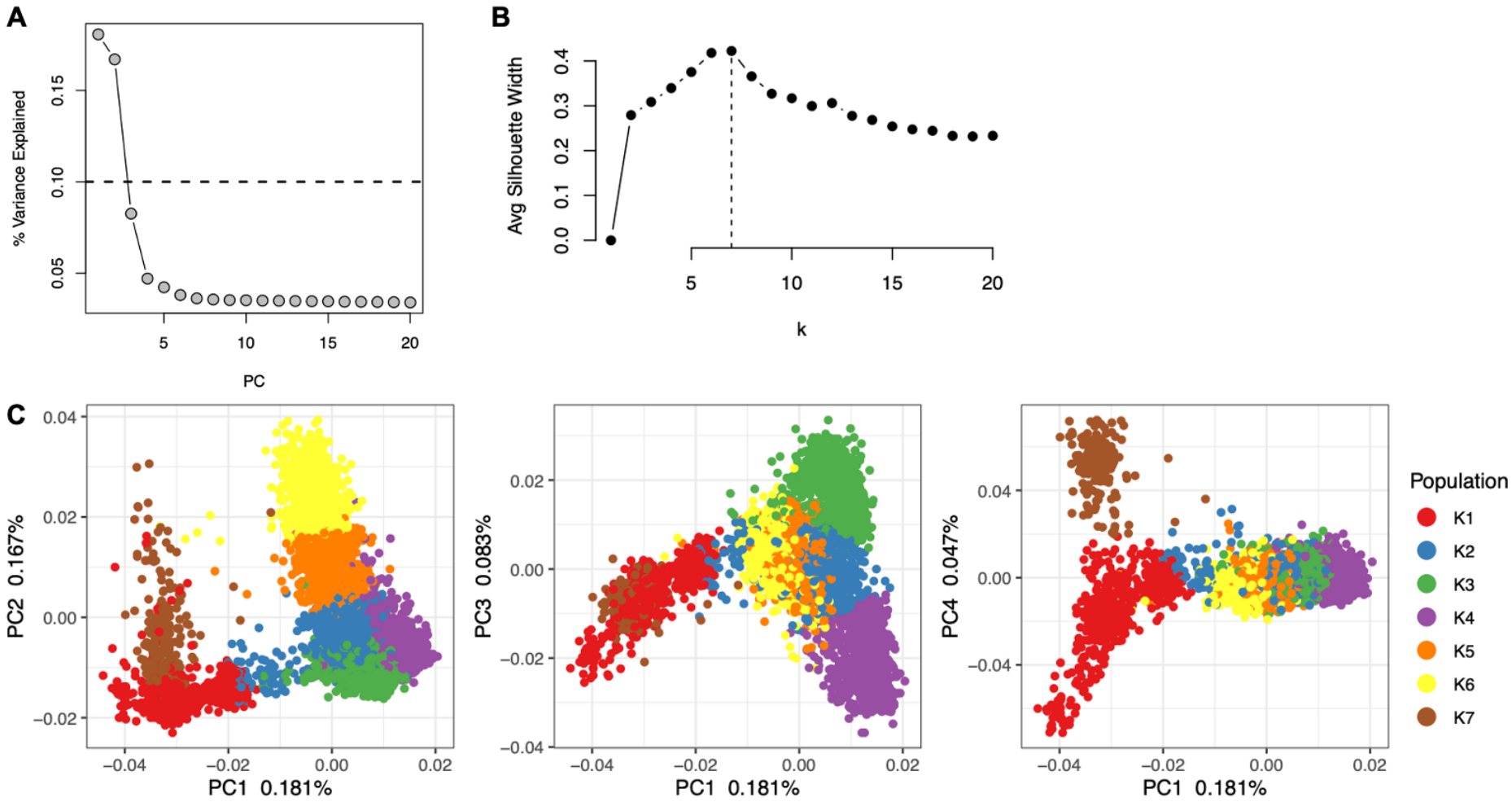


Figure 4-7. *K-means clustering on the AFR CAG sample.*

Scree-plot outlining the variance explained by the top 20 PCs (A). Silhouette analysis to determine the optimal K-cluster number (B). K-means clusters K1-K7 coloured on the PC1~PC2 plane (C).

4.3.4. Country of birth

To evaluate the informativeness of these K-clusters, I mapped onto the PCs the African CAG individuals' COB and ROB data (**Figure 4-8**), along with the other three CAGs. These figures further illustrate the diversity and structure present in the sample. Each CAG presents an observable degree of population structure, and ROB data illustrate non-specific associations between CAGs and ROB. Nevertheless, ROB data illustrates structure across principal components for each CAG. To assess if there is a correlation among the K-clusters identified above and the self-reported place of birth data i.e. test the reliability of K-means approach, I performed a CA for each CAG. The analyses indicate a correlation between K-means clusters and the ROB the African CAG (Dim1 53.29%, Dim2 41.88%, **Figure 4-8**) and the EUR (Dim1 58.25%, Dim2 28.67%), SAS (Dim1 80.00%, Dim2 18.2%), EAS (Dim1 92.11%, Dim2 7.89%) CAGs (**Appendix 16**). When COB was used instead of ROB, an attenuated but correlated structure remained: AFR (Dim1 28.32%, Dim2 25.02%, **Figure 4-8**), EUR (Dim1 40.43%, Dim2 31.89%), SAS (Dim1 61.60%, Dim2 25.31%), EAS (Dim1 50.49%, Dim2 49.51%) (**Appendix 16**).

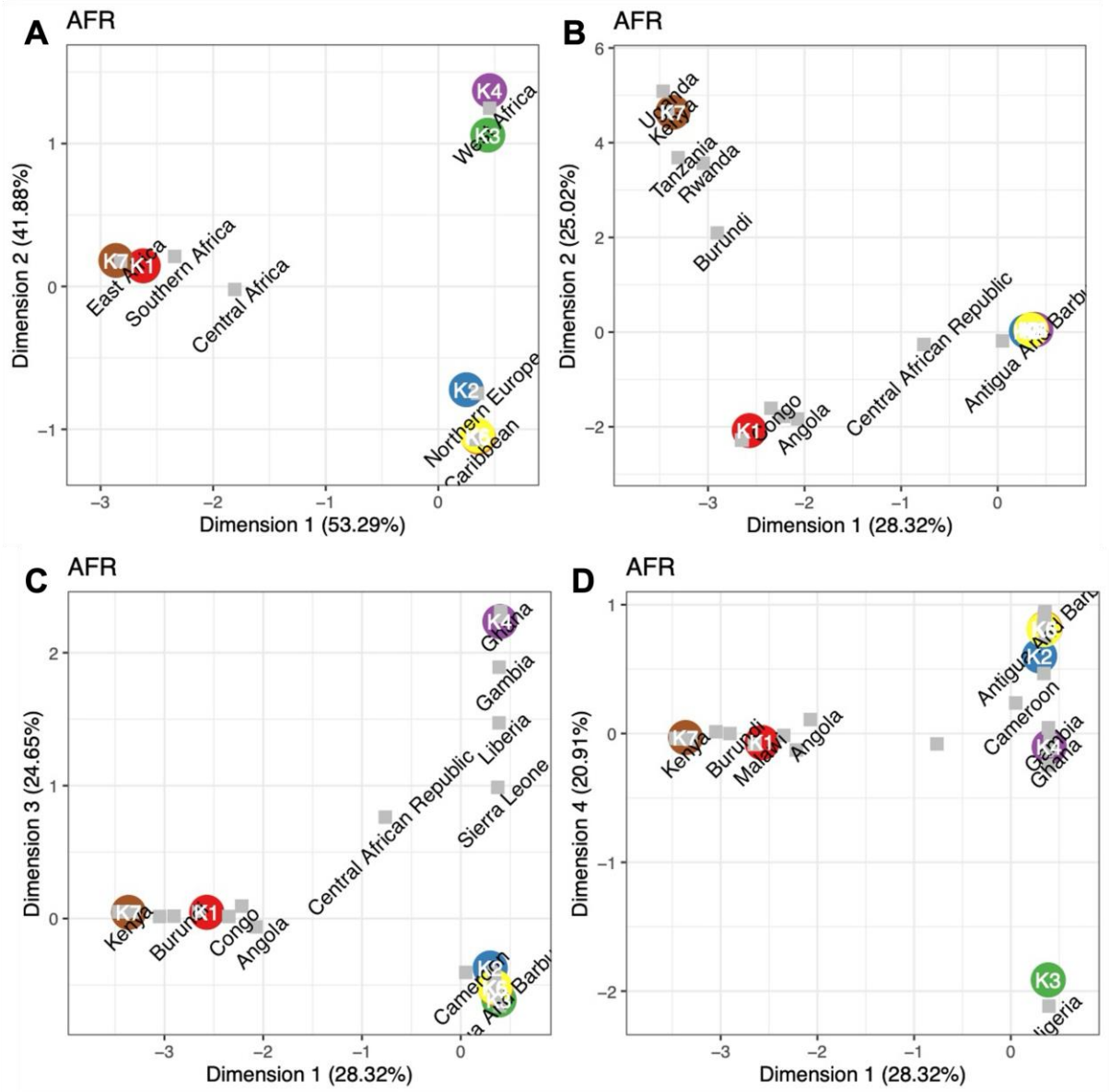


Figure 4-8. *Correspondence analysis.*
K-means clusters were studied in the context of their reliability of identifying homogeneous groups in the structured AFR CAG dataset. This was done using ROB data (A), where the first two dimensions explained most variance, and COB data (B-D), where the first four dimensions explained most variance.

4.3.5. Population differentiation

An evaluation of the degree of population differentiation within each CAG was performed by estimating Fst between each pair of K-cluster groups and 1KG populations. All single-nucleotide polymorphisms (SNPs) that were included in each CAG's principal component analysis were used here. An average, minimum, and maximum estimate was used to summarize the distribution of estimates between pairs (**Figure 4-9**). Relative to the population differentiation observed in the 1KG sample populations I observed a small

degree of population differentiation among AFR and EUR K-means clusters and larger average estimates among SAS and EAS groups. Among the UKBB samples, average Fst estimates indicate that the EAS CAG has the largest amount of population differentiation with an average Fst of 0.0133. This is followed by SAS with an average estimate of 0.0092, EUR with 0.0037, and finally AFR with the smallest average estimate of 0.003.

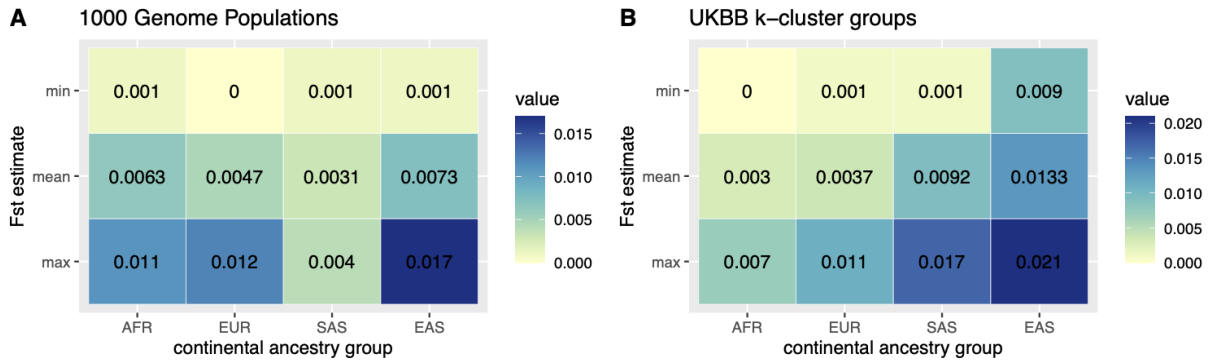


Figure 4-9. *Population differentiation.*
Fst values between sub-populations of 1KG (A) and the K-clusters from UKBB (B). The X-axis indicates the studied CAG, while the Y-axis displays the minimum, average, and maximum Fst value.

Finally, I performed a closer inspection of the AFR CAG sample, where I studied the K-pops in the AFR CAG in a pair-wise manner in terms of their average and maximum Fst values. Here, the average Fst values ranged between 0.0004 and 0.0061, while the maximum value (SNP with highest Fst value) was 0.67 (**Figure 4-10**).

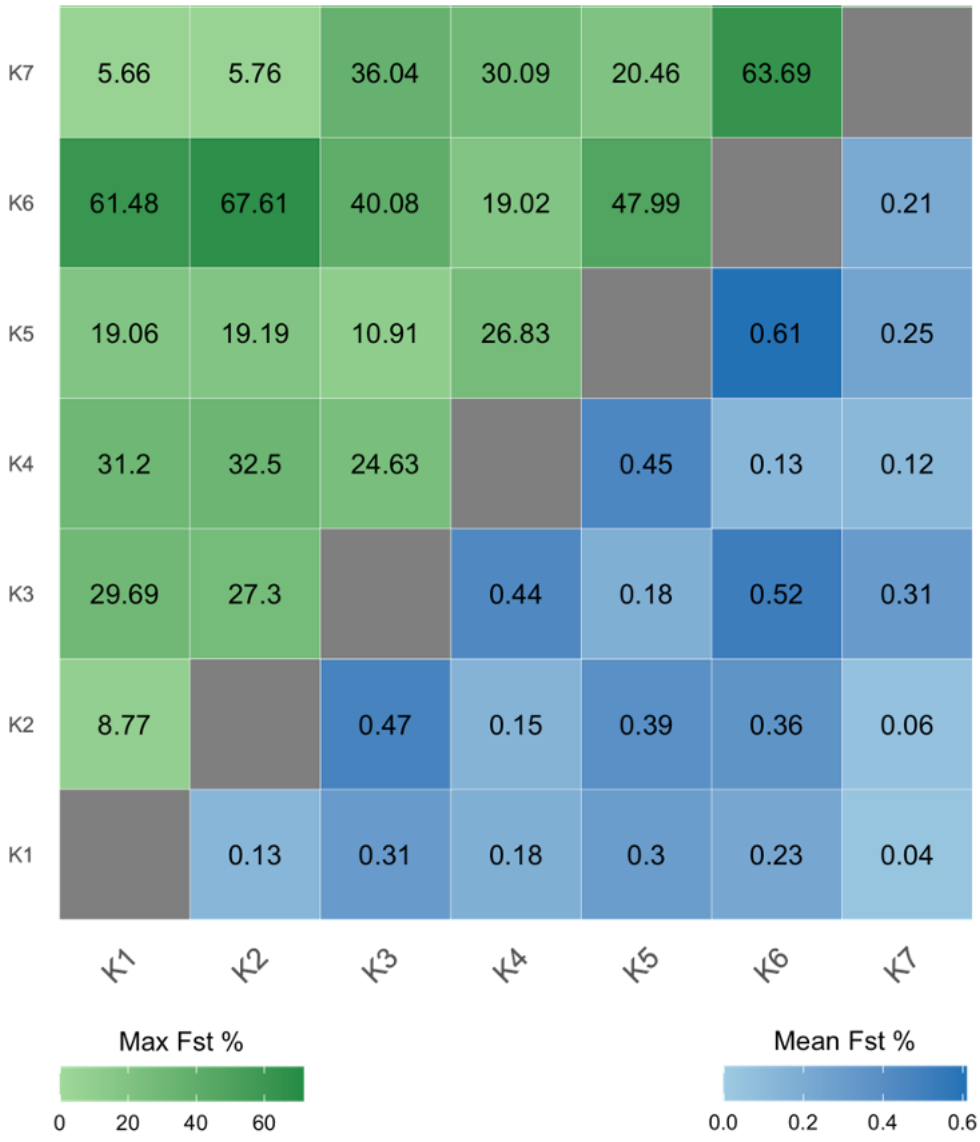


Figure 4-10. *Intrapopulation Fst analysis for the African CAG.*
The Fst value was estimated between each Kpop. In this case, I presented the Fst values as percentages due to the low values in some of the pair-wise results.

4.4. Discussion

Here I present an analytical pipeline to identify participants of the UKBB study with diverse and under-represented ancestries to be used in genomic epidemiology studies. While cohort studies centred in diverse geographic locations are essential for elucidating the effect of environment and genotype on disease, the diversity present in deeply phenotyped studies such as the UKBB should be utilized where possible. While in my thesis I focused on the African CAG specifically, the methodological likeness allowed me to present a description of some of the diversity present in the UKBB.

In summary, we assigned individuals to CAGs (**Aim 1**), after which I illustrated the structure present among individuals within each CAG and identified unsupervised clusters (groups of individuals) within each CAG (**Aim 2**). I then demonstrated that those clusters have an affinity to regions and countries of birth – i.e. the K-means clusters are consistent with geographic structure and isolation by distance models [425,426] (**Aim 3**). Notably, each CAG presents extensive structure, inconsistent with a randomly mating population, but rather with the sampling of unique, geographically distant populations. In particular, East Asian, South Asian, and African CAGs have isolated, or discontinuous groups of individuals in the UKBB sample, exemplified in the K-means clustering analysis [410,411].

In contrast with the Pan-UKBB approach, I have carefully studied the genetic admixture present within the African CAG. More importantly, the methods I presented here provide an approach to identify subsets of individuals to help broaden, inform, and improve the relevance of genetic epidemiological studies, such as the GWAS of neutrophil count I will run on the AFR CAG in the next chapter. This will serve as an example of how the data generated here can be used to improve the understanding of how BCTs can affect the risk of disease associated with non-European populations (**Figure 4-11**).

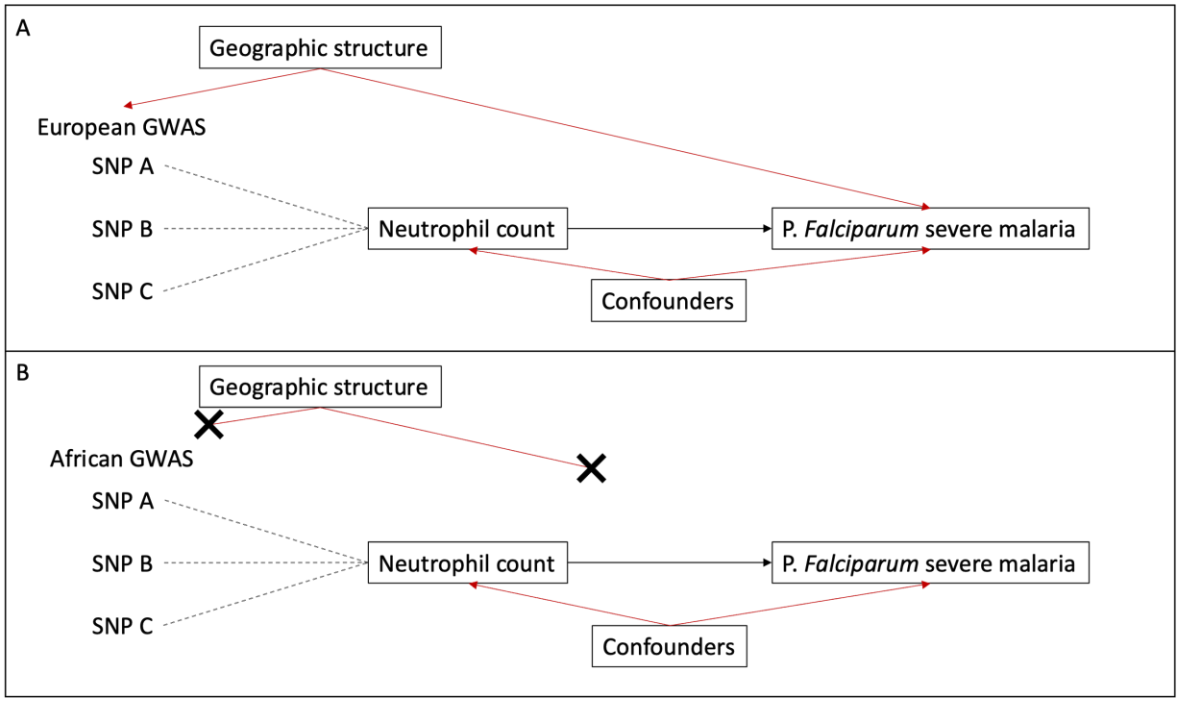


Figure 4-11. *AFR CAG usage example.*
Suppose one might want to use Mendelian randomization to study the relationship between neutrophil count and severe malaria caused by P. Falciparum (as I plan to in Chapter 6). Using summary statistics from a neutrophil count GWAS using individuals with European ancestry (A) may affect estimates due to geographic structure (Ancestry

+ Demography + Environment). This can be overcome by running a GWAS in people of African ancestry (B).

Throughout the paper whenever ancestry is mentioned, I am referring to "genetic ancestry", or individuals who share a demographic history [407,427,428]. They should, at the population level, share a history of mutation, genetic drift, recombination, migration, natural selection, environment, and culture (niche construction [429]). As a product, they should have different genetic variants, allele frequencies, and patterns of LD across their genomes [248,430,431].

The need to perform analyses like association studies, separately in unique ancestral populations, largely comes from the need to avoid correlations between phenotype and genetic ancestry, or differences in allele frequencies among populations – i.e. population structure or population stratification [407,432,433]. For example, if a disease (or environmentally influenced trait) is more frequent in ancestral population 'A' than it is in 'B' and your association analysis pools these ancestral populations together you may erroneously identify any allele that is more frequent in population 'A' as a genetic variant associated with the disease. To avoid these confounding issues, analyses are commonly limited to relatively homogenous populations [428].

In GWAS, the aim is to derive accurate unbiased effect estimates for a genetic variant on a trait. However, the task becomes increasingly challenging, as variation in genetic ancestry comes with different allele frequencies, genetic backgrounds and environments [434]. Methods such as the inclusion of relatedness matrixes and principal components [226,228,234,435] are used to account for cryptic relatedness and undetected fine-scale population stratification. In addition, they are also used to account for correlations between phenotype and genetic ancestry [227,436]. However, is the inclusion of relatedness matrixes or principal components enough to control the structure present in the CAGs presented here? Or would smaller (K-means clusters) more homogenous populations be better suited to epidemiological analyses, like GWAS? As the Pan-UKBB study did not address this question, I aimed to generate these clusters myself to further assess (in **Chapter 5**) the reliability of a GWAS of neutrophil count in the AFR CAG sample in the context of performing a MR analysis between neutrophil count and *P. falciparum* severe malaria.

The problems introduced by population stratification persist even in populations like the "white British" subset of the UKBB, where individual genetic variants and polygenic scores for individual traits can retain correlations with geography, even after correcting

for population structure [402,404]. Moreover, when sampling populations across Europe, where genetic ancestry is known to mirror geography [421,437] effect estimates appear to retain a bias introduced by population structure [438,439]. This is also the case when meta-analysing independently run GWASs [440]. These fine scale issues exemplify some of the reasons for performing separate epidemiological analysis like GWAS for populations with deeper population differentiations, i.e. unique ancestries, demographic histories, and environments. Other challenges and opportunities of population structure in biobank scale data are discussed further in Lawson et al. [339].

The complications of population stratification and opportunities for improving health outcomes for a broader segment of the population, even at the continental level, are precisely why a description of the structure within each CAG was provided here. However, the structure present within the CAGs I defined here might also be too great to be properly accounted for with common methodologies, hence the K-means clustering approach. At the very least, careful consideration is warranted when interpreting results where CAGs are used - because structure matters [403]. Other techniques like uniform manifold approximation and projection [441] or more explicit leveraging of self-described ethnicity could help improve the identification of homogenous groups. Self-described ethnicity is not a synonym for genetic ancestry though, as it is a sociocultural construct [442], although it would help inform cultural, social, and other environmental influences associated with a “population” on phenotypes and disease [428].

4.4.1. Limitations

The methods employed here have several limitations. First, a single 1KG population was used to represent each of four continental ancestry groups evaluated – Africa, Europe, South Asia, and East Asia. One population is a poor proxy for all the variation present in a vast geographical area, such as a continent. However, as the 1KG project does not have optimal population coverage [248], including more or all the 1KG populations of a CAG would still poorly represent all the variation present and would complicate the assignment of individuals to a single ancestry group. Future studies performing whole-genome sequencing on a specific landmass such as Africa will contribute greatly to mapping the global human genome variation [387].

Second, our analyses were limited to four (sub-)continental ancestry groups, to the exclusion of the Americas (AMR, a 1KG superpopulation). Populations from the Americas often have a large and varying amount of recent admixture from various European and African populations [248,414,443–446]. As such, including an AMR population in

the ADMIXTURE analysis, as a reference population, could confound the genetic ancestries being estimated. However, while we limit this study to a few, broad, well characterized CAGs, the approach presented here can be generalised to other specific ancestries.

Third, I note that these estimates were derived from SNPs with a European ascertainment and as such they may not coincide with analyses using an unbiased set of genetic variants. The UKBB Axiom array used to genotype all UKBB participants was designed to optimize imputation of a European population while also including genetic variants previously associated with disease and other phenotypic traits derived from studies primarily conducted in European populations [161,164]. Given this, the genomic data used here will have an ascertainment bias [447] that would affect allele frequency distributions, estimates of LD and diversity and divergence within and among populations. Each of these may influence estimations of Fst, PC estimates and the inferences made from them [448,449]. Specific study designs [229,450] have been made to remove ascertainment bias in genotype arrays so that unbiased inferences could be made for a wider range of genetic ancestries, but this was not available here.

Fourth, the principal components illustrated and used in the unsupervised K-means clustering analyses were derived from the UKBB participants only and resultantly represents the diversity and genetic ancestry found in that data set. The inclusion or use of other public data sets with more numerous sample populations, that better represent regional, or continental diversity, will provide alternative patterns of structure.

Fifth, I was limited by the reference population used in the analyses. While the 1KG data set shall remain an essential reference panel for broad analyses like those conducted here, researchers with specific continental or geographically specific research questions could strengthen and refine the observations made here by including other geographically specific data sets.

Finally, the unsupervised K-means clustering analysis is dependent upon the number of PCs included in it. Here the number of PCs chosen did have an element of subjectivity. While analytical methods are available to select a number of informative PCs [423], I did not implement such methods here. Nevertheless, given that the K-means algorithm weights each PC equally, I tried to limit the PCs included to only those with the largest proportions of variance explained and not necessarily all that are analytically estimated to be informative. Nevertheless, I was unable to determine with certainty whether the population clusters generated here met the desired Hardy-Weinburg Equilibrium criteria.

4.4.2. Conclusion

The approach presented here demonstrates a method to leverage the deeply phenotyped and widely used UKBB data set to help improve the inclusion and equity of epidemiological studies for under-represented populations. While the methods presented here do not describe a perfect solution to identify populations, I hope that they provide an avenue to leverage the diverse data available in UKBB and a methodological platform to improve and build upon. Given the thousands of individuals present in the genetic ancestry groups identified here, the UKBB data set will prove insightful for studies of health and disease not only for the **Chapter 5** of my thesis, but also for future studies in populations beyond the British Isles.

In this chapter I outlined an approach to associate people in large-scale mixed ancestry datasets to specific CAGs. Having identified 6,653 people of African ancestry in UK Biobank, I then focused on one of my thesis aims – using MR [169] to study the relationship between BCTs and disease, namely neutrophil count and *P. falciparum* severe malaria.

CHAPTER 5. NEUTROPHIL COUNT AND *P. FALCIPARUM* SEVERE MALARIA

Chapter summary

The aim of this chapter was to assess the causal link between circulating neutrophils and severe malaria caused by P. falciparum. As mentioned in the background (**Chapter 1**), genetic epidemiology studies on neutrophil count and severe malaria have been sparse, and none have been done using genetic data from people of African ancestry. Therefore I used univariable (UV) [252] two-sample Mendelian randomization (2SMR) [253] to advance the current understanding of how neutrophil count affects the severity of P. falciparum malaria. This required conducting a GWAS of neutrophil count in people of African ancestry for appropriate consideration of population structure in data with the aim of generating reliable allocation/use of genetic associations as MR instruments. To do this, I used the African continental ancestry group (CAG) from UK Biobank (UKBB) which I generated in **Chapter 4** (**Figure 5-1**).

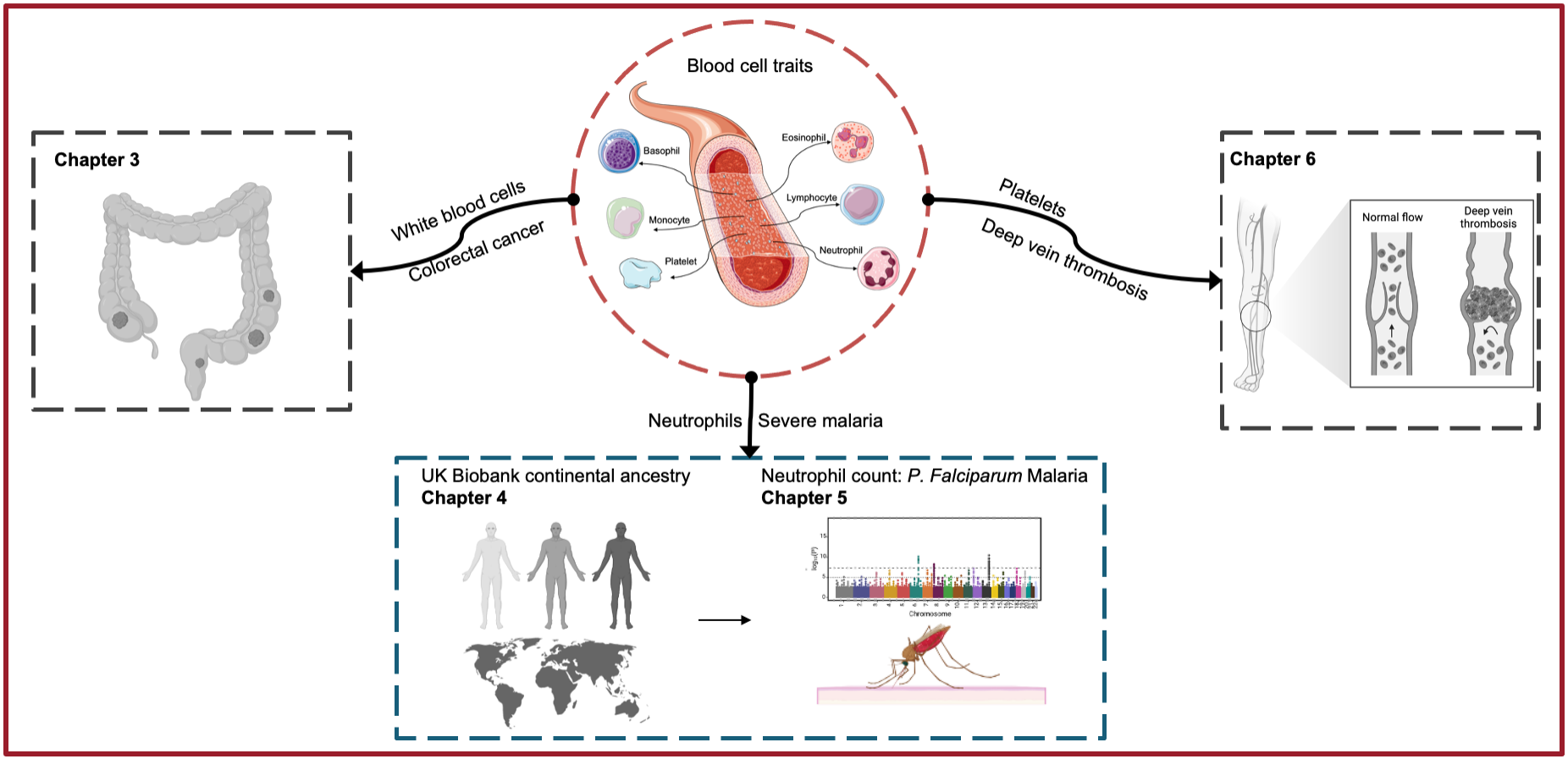


Figure 5-1. *PhD project and current chapter (5 - coloured).*
Created with Microsoft PowerPoint and BioRender.com.

5.1. Introduction

Malaria is a mosquito-transmitted disease that annually affects approximately 215 million people [385,451] and has been the biggest cause of childhood deaths over the past 5000 years [452]. Malaria is caused by protozoan parasites belonging to the *Plasmodium* genus [385] and represents the deadliest disease in human history [453].

Unlike bacteria and viruses, protozoans are unicellular eukaryotes that belong to the Protista Kingdom. Over 65,000 species of protozoans have been identified [454], some of them in the most extreme environments on Earth, such as the permafrost of Russia [455] or industrial acid main drainage systems [456]. Protozoans have been studied most extensively due to their role in human pathogenesis, the most notable species being *Leishmania* [457], *Trypanosoma cruzi* [458], *Toxoplasma gondii* [459], and lastly, *Plasmodium*, responsible for malaria [385]. Plasmodium is an obligate parasite, meaning that it cannot complete its normal life cycle without being parasitic to its host [460]. In most cases, it enters its host through the Anopheles mosquito [461], although zoonotic infections have also been reported [462].

5.1.1. Plasmodium species

There are five notable *Plasmodium* species which are known to cause malaria in humans: *Plasmodium malariae* (*P. malariae*), *Plasmodium ovale* (*P. ovale*), *Plasmodium knowlesi* (*P. knowlesi*), *Plasmodium vivax* (*P. vivax*) and *Plasmodium falciparum* (*P. falciparum*) [463].

P. malariae is endemic in sub-Saharan Africa, south-east Asia (SEA) and the Western Pacific [464]. It rarely leads to severe manifestations of malaria, and its prevalence in a recent meta-analysis was estimated at 3% [464]. *P. ovale* is one of the first *Plasmodium* species identified to infect humans and is present in sub-Saharan Africa, SEA and Western Pacific [465]. However, its pooled prevalence was estimated at 0.03% in a recent meta-analysis [465] and it predominantly causes benign tertian malaria as opposed to severe disease [466]. *P. knowlesi* was first found in patients misdiagnosed with *P. malariae* [467]. It is usually present only in SEA areas habituated by monkeys, which represent the natural vector of *P. knowlesi* [467]. In endemic regions, it had a pooled prevalence of 19%, and its infection has been sometimes linked with more severe manifestations of malaria than *P. falciparum* [467]. *P. vivax* is the most widespread *Plasmodium* species [468]. It is found in both South and Central America, as well as SEA and the Western Pacific, with

little prevalence in sub-Saharan Africa [468]. While it has been historically linked to mild disease, severe malaria caused by *P. vivax* is of increasing concern as registered cases are growing annually [469]. Finally, *P. falciparum* is endemic in the same geographical areas as P. malariae, and cases of co-infection have been reported [464]. However, *P. falciparum* is by far the most life-threatening disease out of all those caused by *Plasmodium* species, and accounts for over 90% of the global malaria deaths [451,470], and is therefore the disease that I will focus on in this chapter.

5.1.2. The life cycle of *P. falciparum*

Plasmodium's life cycle involves both a host and a vector, which in this case is the *Anopheles* mosquito [471]. The mosquito bites the host when it feeds and releases sporozoites in the dermis, which then travel through the lymph and blood to the liver [472]. Here, the sporozoites cross the liver sinusoid and infect hepatocytes [473]. They then multiply thousands of times in a process called schizogony [474]. The resulting parasites, merozoites, enter the bloodstream and detect erythrocytes (red blood cells, RBCs) through a complex called apicomplexan and then infect them [475]. In *P. falciparum*, merozoites enter erythrocytes through ligands called glycophorins [476,477], while in P. vivax the Duffy surface antigen is used [478]. A small number of these merozoites become gametocytes and leave (i.e. egress) the RBCs [479], which are necessary for sexual reproduction of *Plasmodium* [480]. Next, an intracellular trophozoite is formed, in which the merozoites multiply many times through asexual schizogony to form a schizont [481]. Merozoites can then egress the erythrocytes and repeat the cycle of infection [482], which happens every 48 hours for *P. falciparum* and *P. vivax*, but does vary by the *Plasmodium* species [463] (**Figure 5-2**).

Interestingly, parasitised RBCs have been found to have an internal clock through which they are able to rupture at the same time to cause malarial fever [483], showing that even after decades of research, much is still being uncovered on *P. falciparum* pathogenesis. Furthermore, the aforementioned gametocytes are taken by a mosquito during its feeding process and reach its gut [463]. Sensing the new environment, gametocytes transform into gametes and fuse to ultimately form motile ookinetes that cross the epithelial cells of the gut to form oocysts [484,485]. After many sporozoites are produced through mitosis, the oocysts rupture, releasing the sporozoites into the haemolymph, which subsequently reach the mosquito's salivary glands [486,487]. This final event marks the end of a cycle for the *Plasmodium* parasite, which is then able to infect another host (**Figure 5-2**).

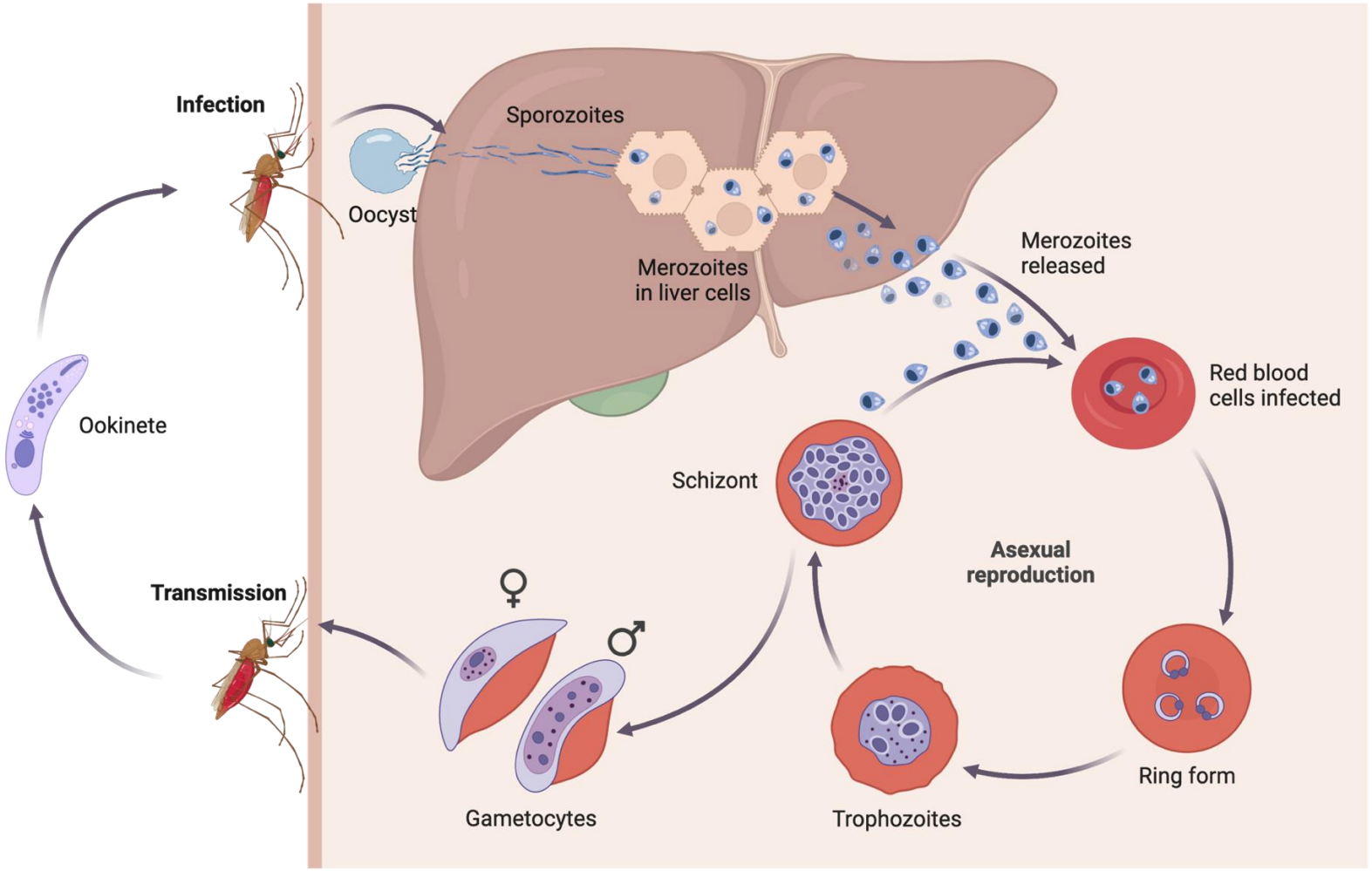


Figure 5-2. *Plasmodium falciparum life cycle.*
Adapted from "Malaria Transmission Cycle", by BioRender.com (2022). Retrieved from https://app.biorender.com/biorender-templates.

5.1.3. The health burden of *P. falciparum*

P. falciparum malaria causes approximately 400,000 deaths each year, primarily in African children under the age of five [385]. The majority of *P. falciparum* malaria cases consist of uncomplicated febrile illness, however a portion of nonimmune infected individuals succumb to severe malaria (SM), which can manifest as cerebral malaria (CM), severe malaria anaemia (SMA), acute respiratory distress or kidney injury [246,488]. *Plasmodium* resides and proliferates in RBCs and pathology is triggered by cytoadherence of infected RBCs (iRBCs) to microcapillary endothelia in different organs, which can lead to vascular obstruction [488]. Inflammation plays a key role in both facilitating iRBC sequestration [489] and in tissue damage [488,490,491]. In cerebral malaria, the deadliest form of the disease, iRBCs sequester in the neurovasculature, provoking blood brain barrier permeabilization, vascular leak and brain swelling [488].

Given the health burden of *P. falciparum* malaria, the World Health Organisation (WHO) has setup a plan for 2030 to reduce the number of malaria cases by 90% [492]. However, while the attempts in the last decades have shown promising results, the global malaria incidence since 2015 has seen a reduction of only 2% [493]. Given the current efforts to accomplish WHO's 2030 goal, it is desirable to use novel methods to identify new risk factors for *P. falciparum* severe malaria and therefore potential therapeutic strategies.

Malaria has been affecting humans for thousands of years [452], and it is estimated that it may have been a health burden 40,000 years ago [470]. As such, it has exerted the strongest known selective pressure on the human genome and has resulted in the selection of various polymorphisms that confer *Plasmodium* tolerance or resistance. Among the most prominent examples are haemoglobin S (Hbs; sickle cell trait) [494] and alpha-thalassemia variants [495], both of which are common in malaria endemic regions despite causing disease in the homozygous state [452]. The HbS polymorphism in the heterozygous state confers the greatest protection (effect size >80%; [452,476]). The heritability of SM is estimated to be around 30% [496,497] but the cumulative effect of the above mentioned variants is thought to only be 2% [452,496], suggesting that polygenic interactions may account for a large part of the missing heritability of this complex disease.

The genomic revolution in the past 25 years has allowed researchers to investigate the genetic mechanisms of complex traits using sample-sizes of thousands of people [498]. In the last decade, this further reduction in costs and advances in technology has led to the generation of consortia with hundreds of thousands of participants, such as UK Biobank

(UKBB) [161,164]. Naturally, this has come with an increase in the number of genetic studies aiming to decipher the biology of malaria, both in relation to the *Plasmodium* [499–501] and human genomes [245,502]. One topic of interest has been genetic polymorphisms and how they affect the risk of severe malaria in susceptible populations, which have been finding increasingly more loci associated with severity of the disease [245].

5.1.4. Neutrophil count and severe malaria

Interestingly, individuals living in malaria-endemic regions, as well as those descended from them, often have reduced numbers of neutrophils in their circulation [503]. This heritable phenomenon is called 'benign ethnic neutropenia' (BEN) and is distinct from life-threatening severe neutropenia [503]. BEN is prominent in South Mediterranean, Middle Eastern, sub-Saharan African and West Indies populations [503]. BEN is estimated to occur in 25-50% of Africans [140,503,504] and 10.7% of Arabs [505] but in less than 1% of people of European ancestry living in the Americas [506]. Neutrophils are essential for immune defence against bacteria and fungi [507], however BEN does not lead to significantly greater susceptibility to infection in the United States [503].

Nevertheless, it remains curious that selection for lower neutrophil counts occurred in sub-Saharan Africa, a region associated with a high infectious disease burden. This observation is partly explained by the finding that in populations of African and Yemenite Jewish ancestry, BEN is strongly associated with a polymorphism in the Duffy antigen receptor for chemokines (DARC) gene, which encodes the Fy/Duffy antigen, a surface receptor utilized by *P. vivax* to invade RBCs [160]. This variant, called rs2814778, abolishes expression of DARC on RBCs and contributes to low prevalence of *P. vivax* in sub-Saharan Africa, where the polymorphism is found at levels close to fixation [452]. DARC, in addition to serving as one of the entry points for *P. vivax*, controls circulating levels of chemokines [508], which also regulate blood neutrophil numbers [508]. While other familial conditions affecting relative neutrophil count have been documented, it is unclear to what extent other polymorphisms contribute to neutrophil count variation in individuals living in malaria endemic regions [509].

Several observational studies have assessed the link between neutrophils and disease severity. In their study of *P. falciparum* malaria patients (mild = 47, severe = 8), Kho et al. found increased neutrophil extracellular trap (NET) count in those with severe malaria [510]. Additionally, Wolfswinkel et al. recorded the white blood cell (WBC) count of patients (N=440) with *P. falciparum* malaria 24 hours post-diagnosis and identified a higher neutrophil count in those with severe manifestations of the disease [511]. Similarly, Berens-

Riha et al. found that a higher neutrophil-to-lymphocyte ratio (NLR) and neutrophil count were both associated with increased parasitaemia and severity of *P. falciparum* malaria [512].

Biologically, neutrophils have recently been shown to have a detrimental role in malaria, promoting pathogenesis by enhancing sequestration of iRBCs in NETs [489] and contributing to inflammatory tissue damage [491,513,514]. A gene expression study found that genes encoding for neutrophil granule proteins were highly expressed in those with severe malaria [515], and the neutrophil chemokines CXCL1 and CXCL8 were also highly present in severe manifestations of the disease [513]. Genetic studies in sub-Saharan Africa found associations with severe malaria in genetic polymorphisms at the genes that encode for receptors that neutrophils use for phagocytosis (Fcγ II/III) [516,517]. On the other hand, neutrophils have also been suggested to participate in removal of iRBCs and in shaping the Plasmodium antigenic repertoire [518]. Overall, there is no consensus on whether severe malaria increases neutrophil count, if a higher neutrophil count increases the risk of developing a severe manifestation of the disease, or if they are simply associated and are not causally linked in either direction.

5.1.5. What can MR add in the context of neutrophils and severe malaria?

These studies raise the possibility that neutrophil count in malaria endemic regions may modulate severity of *P. falciparum* malaria. However, observational studies, such as the ones referenced above, are prone to confounding and reverse causation (**Chapter 2**) [169,191,519]. It is therefore essential to employ additional methods, such as those in genetic epidemiology and population genetics, to study the link between neutrophil count and *P. falciparum* SM, which can aid with our understanding of how BCTs affect disease. Band et al. from the Malaria Genomic Epidemiology Network (MalariaGEN) has been the only study to date to conduct a MR study between neutrophil count and *P. falciparum* severe malaria [245]. However, they used instruments for neutrophil count that were generated from Astle et al.'s GWAS in Europeans from UKBB [149]. Two-sample MR assumes that the exposure and outcome datasets come from the same underlying population [200]. In addition, the genetic architecture of BCTs is known to differ between those of African vs. European ancestry [166,396,399]. As discussed in **Chapter 4**, generating African ancestry-specific instruments for neutrophil count is important for assessing the causal relationship between neutrophil count and SM in a reliable MR framework.

5.1.6. Current GWAS of neutrophil count

Recent efforts have resulted in the generation of hundreds of GWAS using UKBB non-European participants for many traits in a hypothesis-free manner (https://pan.ukbb.broadinstitute.org/). However, the same covariates were used for each trait, and the impact of population structure was not studied, which represent a potential limitation in having reliable instruments for a MR analysis [337]. In addition to this, a recent study by Chen et al. has also used people of non-European ancestry in UKBB as part of their project to perform trans-ancestry GWAS of blood cell traits (BCTs) [166]. However, I have shown in **Chapter 4** that people in the African continental ancestry groups (CAGs) of UKBB display strong population structure [238]. Would therefore running a GWAS of a complex trait such as neutrophil count result in associations that could be linked to a biological mechanism, or would the GWAS associations be a product of residual population structure? In order to answer these questions, a more thorough investigation of the sampled dataset is warranted. This becomes even more important when aiming to conduct causal inference analyses in genetic epidemiology, such as two-sample Mendelian randomization [169,252].

5.1.7. Main study objective

The overarching objective of the work described in this chapter was to assess the relationship between neutrophil count and severe *P. falciparum* malaria, which posed specific methodological challenges in studying the relationship between blood cell traits and disease.

5.1.8. Study aims

I have divided this chapter's main objective into three separate aims I will try to address:

1) Perform a GWAS of neutrophil count in people of African ancestry
2) Identify SNPs from the GWAS that can be reliable instruments for neutrophil count in a MR analysis
3) Conduct a MR analysis between neutrophil count and *P. falciparum* severe malaria

5.2. Methods

5.2.1. Study design

The UK Biobank African CAG from **Chapter 4** was used to conduct a GWAS of neutrophil count (**Aim 1**). This was then followed by sensitivity and post-hoc analyses to assess the

reliability of the findings (**Aim 2**). Finally, I conducted a MR analysis between neutrophil count and severe malaria [245] (**Aim 3**) and interpreted the results (**Aim 4**) (**Figure 5-3**).

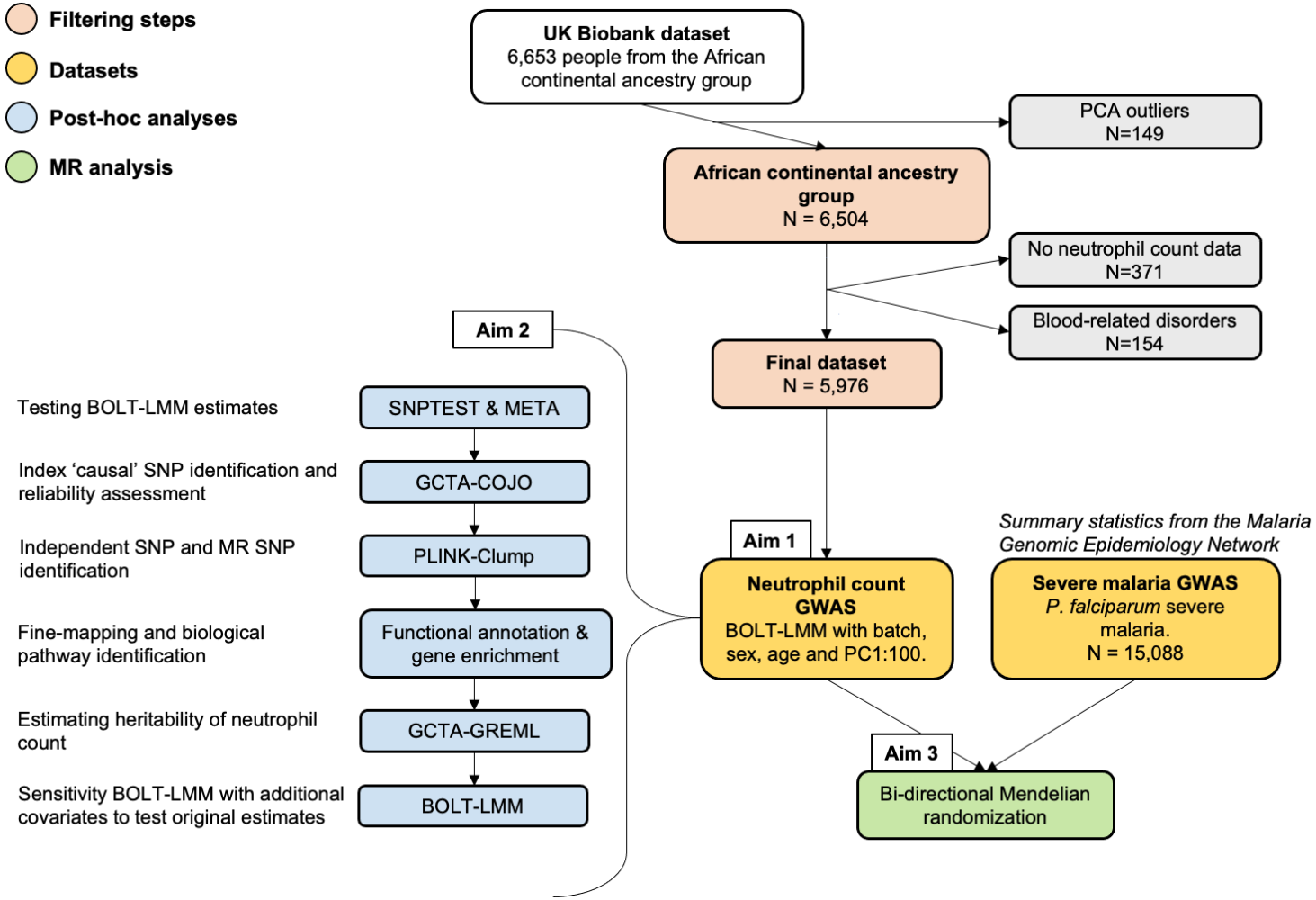


Figure 5-3. *Study design of the project.*

6,653 people representing the UKBB CAG underwent PCA outlier and neutrophil count data filtering, resulting in a final sample of 5,976. The main analysis was a GWAS of neutrophil count run with BOLT-LMM. This was the dataset based on which other post-hoc analyses were conducted. Finally, a MR analysis was performed between neutrophil count and severe malaria caused by P. falciparum using data from MalariaGEN.

5.2.2. UK Biobank genetic data

UK Biobank's "non-white" British data was studied previously in **Chapter 4**, where 6,653 people corresponded to the African CAG, of which 6,504 remained (5,989 unrelated; 515 related) after filtering for principal component analysis (PCA) outliers [238]. These were further assigned into seven clusters based on a K-means clustering algorithm (K1=527; K2=1,177; K3=1,176; K4=1,001; K5=1,206; K6=862; K7=184; see **Chapter 4** for more information) [238]. This dataset (N=6,504) included both directly genotyped (N=784,256) and imputed (N=29,363,284) SNPs filtered with a minor allele count of at least 20. More details on the UK Biobank dataset are found in **Chapter 2**.

5.2.3. UK Biobank phenotypic data

Haematological samples were analysed using four Beckman Coulter LH750 instruments [236]. Total white blood cell (WBC) count and neutrophil percentage (%) were measured through the Coulter method (see **Chapter 2**), with neutrophil count derived as "neutrophil % / 100 x total WBC" and expressed as 10^9 cells/Litre [236]. This is the neutrophil count data that I am using in this project. Afterwards, I split the date variable into year, month, day, and minutes (passed since the start of the day of the appointment visit), while the neutrophil count measurement variable was log-transformed into a variable named "nc_log", which was used as the default neutrophil count variable throughout the study. Other variables that were used in the main analyses were: age, genetic sex, blood sample device ID, UKBB assessment centre and principal components (PCs) 1 to 100, which were generated in **Chapter 4** with EIGENSOFT [226,228]. Filtering was done based on the selection criteria described by Astle et al. [149] and Chen et al. [166]. Briefly, individuals with disorders/diseases that could affect blood counts (e.g. HIV, leukaemia, congenital anaemias, cirrhosis) were removed, bringing the final sample size to 5,976. This dataset is referred to as "AFR_CAG" throughout the chapter.

5.2.4. Pre-GWAS investigative analyses

Descriptive analyses of nc_log were performed. To study the amount of potential population admixture in the AFR_CAG that could affect the test statistics from a GWAS, an analysis was conducted in R on the Duffy SNP rs2814778 [160] for each Kpop. The preponderance of the Duffy SNP rs2814778 allele distribution was outlined in a PCA plot (PC1~PC2), and its association with nc_log in the AFR_CAG dataset was studied with and without PCs. Tracy-Widom statistics [226] were computed to estimate the number of PCs that are significant i.e. that could be added into the GWAS as covariates. Finally, a

power calculation was done assuming a linear model GWAS on the AFR_CAG sample to discuss if there would be enough power to detect a signal.

5.2.5. BOLT-LMM GWAS

BOLT-LMM was used as the software to run the primary (main) GWAS. Before running BOLT-LMM, linkage disequilibrium (LD) scores were generated from the directly genotyped dataset (binary-ped PLINK format) that are required by BOLT-LMM to calibrate the test statistics. This was done with the LDSC package using the following command:

python ldsc.py --bfile $bed_file --l2 --ld-wind-cm 1 –out mergedAll.l2.ldscore.gz

After preparing the phenotypic data to match the desired input, BOLT-LMM was run on AFR_CAG adjusting for age, genetic sex, UKBB assessment centre, blood sampling device, sampling year, sampling month, sampling day, minutes passed in sampling day and the first 100 principal components (PCs).

5.2.6. SNPTEST and META GWAS

To test whether the effect estimates from the BOLT-LMM GWAS were biased due to residual population structure that would characterise a population from the African CAG, a number of “sensitivity” GWAS were conducted. This was done with SNPTEST using a linear model algorithm [520,521], with 16 GWAS conducted as follows: 8 GWAS were run on each K-means cluster (Kpop) + the whole sample with the same parameters as in the BOLT-LMM run, and another 8 in the same manner, but with rs2814778 as an additional covariate. To minimise the chance of errors and to reduce the time needed to run each GWAS, a linear model was first conducted in R using the command “lm(nc_log ~ my_covariates)”. The residuals were then pulled with the “residuals()” function and 16 GWAS were run on AFR_CAG with SNPTEST. The Kpop GWAS were then meta-analysed with META [522] under an inverse-variance method based on a fixed-effects model. The result were two meta-analyses: one without accounting for the Duffy SNP rs2814778 called “META-WOD”, and one where the Duffy SNP was included as a covariate, called “META-WD”.

5.2.7. Conditional & joint association analysis

GCTA-COJO [523,524] was employed to identify independent signals from the BOLT-LMM GWAS, as well as detect any possible secondary signals arising from a stepwise

selection model. SNPs which are close together are usually in LD i.e. their alleles are not random, but correlated [525]. Before running GCTA-COJO, genetic variants with an INFO score < 0.3 were filtered out of the AFR_CAG dataset with QCTOOL. PLINK was then used on this resulting output to filter out related individuals. Following this step, GCTA-COJO was run on the AFR_CAG filtered dataset to identify causal SNPs. These were referred to as "index" in the text. Plots similar to those generated by LocusZoom [526,527] were created in R with the "LocusZooms" package [528].

5.2.8. Post-GWAS sensitivity analyses

Summary statistics for a meta-analysis of neutrophil count in people of African ancestry were downloaded [166] (http://www.mhi-humangenetics.org/en/category/general/), and SNPs passing the GWAS significance threshold (P<5e-8) were filtered in. These variants were then clumped with PLINK default parameters (--clump-p1=5e-8, --clump-r2=0.5, --clump-kb=250), and their dosage data was pulled out from the AFR_CAG genetic dataset. A linear model in R was conducted on each variant (N=13,139) using the same parameters that Chen et al. [166] used in their GWAS. Another linear model was run with the same parameters as in the BOLT-LMM run.

From this latter run, the top 10 associations were kept and were used to conduct two sets of linear models in each Kpop of the AFR_CAG – Set 1, similar to how BOLT-LMM was conducted, and Set 2, similar to how SNPTEST/META were conducted:

Set 1: lm(nc_log ~ SNP_dosage + covariates)

Set 2: lm(residuals ~ SNP_dosage),

residuals = residuals from a linear model conducted on nc_log adjusting for GWAS covariates.

The Kpop results were then meta-analysed in R. The top 10 associations were also pulled from the BOLT-LMM and Chen et al. [166] summary statistics, and a forest plot was generated with the SNP effect estimates for comparison.

5.2.9. Genomic inflation

The genomic inflation factor lambda (λ) [529] was calculated for the BOLT-LMM and SNPTEST meta-analysis runs. This was complemented by generating quantile-quantile (QQ) [530] to investigate any early deviation of the expected P-values from the observed. Additionally, a Manhattan plot [530] was generated to highlight the BOLT-LMM index SNPs,

and two more plots were created to mirror the BOLT-LMM signals with those from a GWAS of neutrophil count in people of African [166] and European [149] ancestry.

5.2.10. PLINK clumping

After GCTA-COJO, I used PLINK to perform clumping with three different thresholds. The first two represent the thresholds for defining independent SNPs in the well-known online platform Functional Mapping and Annotation (FUMA) [531], while the latter being the clumping conditions usually set for conducting a Mendelian randomization analysis [532,533].

1. --clump-p1=5e-8, --clump-r2=0.6, --clump-kb=250
2. --clump-p1=5e-8, --clump-r2=0.1, --clump-kb=250
3. --clump-p1=5e-8, --clump-r2=0.001, --clump-kb=10000

The “bioconductor" R package was used to connect to dbSNP [534] and map genes for the clumps in step 1, which included those in steps 2 and 3.

5.2.11. Characterization of functional loci

A query was placed through the variant effect predictor (VEP) [535] and FUMA [531] on the SNPs in the AFR_CAG filtered dataset. A further, broader literature search was conducted on the index and MR clumping SNPs using Ensembl [536], GeneCards [537], GWAS Catalog [167], The Human Protein Atlas [538], and the Genotype-Tissue Expression (GTEx) project [361].

5.2.12. Heritability analysis

An analysis was conducted with GCTA to estimate the proportion of variance in neutrophil count explained by all genetic variants present in the filtered AFR_CAG dataset [539]. First, a power calculation was done to assess whether the sample-size of unrelated people with neutrophil count data (N=5509) would be enough to detect genetic covariance [540]. Default power calculation parameters were used: $\alpha = 0.05$, $h^2 = 0.3$, var π = 2e-5; α = P-value significance threshold, h^2 = combined genetic heritability for the trait, var π = variance of the off-diagonal elements of a genetic relationship matrix (GRM) [540]. Afterwards, a GRM was generated from the whole filtered AFR_CAG with the following command. UKBB phenotypic data was then used to run GCTA-GREML, with and without adjusting for the Duffy SNP rs2814778. Yang et al. propose a way for estimating heritability while accounting for potential LD bias [541]. In brief, segment-based LD scoring was done on each chromosome. SNPs were stratified in R by LD scores in

four groups for each chromosome [542], yielding 88 SNP groups in total. A GRM was generated for each SNP group, and GCTA-GREML was run similarly to the previous run.

5.2.13. *P. falciparum* severe malaria genetic data

GWAS summary statistics for severe malaria were downloaded from a case-control study that spanned nine African and two Asian countries [245]. In brief, controls samples were gathered from cord blood, and in some cases, from the general population. Cases were diagnosed according to WHO definitions of severe malaria [246] and were categorised according to CM, SMA and other severe malaria (OTHER) symptoms (**Table 5-1**). The majority of the RSIDs in the MalariaGEN dataset used older identifiers, and some of them had the "kgp" prefix that comes with the Illumina-HumanOmni2.5M array. Ideally, in a two-sample MR setting, the two samples would have a perfect match in the available genetic variants. It is desirable to at least maximise the number of matching variants to test. Therefore, RSID information for the MalariaGEN variants was updated in R by using the filtered AFR_CAG dataset (PLINK .bim file) as a reference panel.

Table 5-1. *Description of MalariaGEN cases and controls by country.*

Country	Cases	Controls	Total
Gambia	2,461	2,518	4,979
Mali	259	163	422
Burkina Faso	711	583	1,294
Ghana	391	315	706
Nigeria	112	21	133
Cameroon	583	634	1,217
Malawi	1,161	1,310	2,471
Tanzania	410	388	798
Kenya	1,529	1,539	3,068
TOTAL	7,617	7,471	15,088

5.2.14. Meta-analysis of severe malaria African populations

Summary statistics for severe malaria and its sub-phenotypes were generated from a meta-analysis which included individuals from two non-African countries – Vietnam and Papua New Guinea. The inclusion of SNP effect sizes from GWAS conducted in heterogenous population might bias MR estimates [339]. Therefore, per-population

summary statistics were downloaded (https://www.malariagen.net/sppl25/) for each African country in the study and a meta-analysis was conducted on them using METAL [233,543,544] with the command “metal config.txt” and the following configuration: “TRACKPOSITIONS ON”, “SCHEME STDERR”, “AVERAGEFREQ ON” and “MINMAXFREQ ON”.

5.2.15. Mendelian randomization analysis

The “TwoSampleMR” R package [213,219] was used to perform the MR analyses. The two datasets were harmonised i.e. orientated on the same strand and if SNPs were not found in the outcome dataset, SNP proxies would be searched for. A bi-directional MR analysis was conducted, where the effect of neutrophil count on overall severe malaria, along with the three sub-phenotypes was estimated and vice-versa. The main analysis was conducted using an IVW model [202]. Additionally, a sensitivity MR analysis was conducted to outline the effect estimates of each SNP on the desired outcome, with IVW and MR-Egger [173,205] estimates where the number of instruments was larger than two and three, respectively. MR methods are described in **Chapter 2**.

5.2.16. GWAS with additional covariates

Several analyses were conducted to investigate and describe the phenotypic data in the AFR_CAG dataset. Descriptive statistics for neutrophil count were generated to provide information on the sample that the GWAS were run on. Missing data for additional variables were investigated, and an analysis was conducted to test whether missing data in each of these variables showed evidence of affecting neutrophil count. Moreover, a univariable, multivariable ANOVA type II and multivariable ANOVA type III were conducted to assess the variance explained by environmental, multifactorial and immutable (e.g. place of birth) variables. Following the results from the descriptive analyses, another GWAS was run in BOLT-LMM. “Genetic sex”, “time since last menstruation” and “menopause” variables were combined in a single discrete variable called “menstrual_status” and was created as follows: males, quartiles 1-4 of days since last menstruation, menopause, had hysterectomy. The covariates used in this run were sampling device, sample year, sample month, sample day, minutes passed in sample day, UN region of birth, K-means cluster, smoking status, alcohol drinker status, “menstrual_status”, age, body mass index and PCs 1 to 100. 669 individuals were filtered out for missing values and/or preferred not to answer in these variables, bringing the sample-size to 5,310.

5.2.17. Description of working environment

All analyses were performed in a Linux environment supported by the University of Bristol's Advanced Computing Research Centre (ACRC) using the following publicly available software packages: PLINK v1.9 and v2.0 [247,416], QCTOOL v2.0.7 (https://www.well.ox.ac.uk/~gav/qctool/), LDSC v1.0.1 [545], SNPTEST v2.5.4 [520,521], BOLT-LMM v2.3.6 [435], META v1.7 [522], METAL v2011-03-25 [233,543,544], and GCTA v1.94.0 [523]. All other scripts, analyses, and figures were run and generated in the R environment using version 4.1.2 (Bird Hippie) [345] and Python environment using version 3.7.7 [546] on the ACRC computer clusters.

5.3. Results

5.3.1. Analysis of study sample

Several steps were undertaken prior to running the GWAS. First, I investigated the descriptive statistics of the sample that I wanted to perform the GWAS on. This was done to assess the study population characteristics and to aid in conducting and interpreting the GWA and MR results.

5,976 out of 6,504 individuals in AFR_CAG remained after filtering for missing data and traits affecting blood cells. The mean value for neutrophil count was 2.9 x 10^9 cells/Litre, as expected this was lower compared to a European sample (4.21 x 10^9 cells/Litre) [149,166]. The individuals in the GWAS sample had a larger proportion of females (57%) and was of a higher mean age (39 vs. 58.1 years) [547] and slightly higher body mass index (BMI) (27,6 vs. 29.8 kg/m^2) [548] than the general UK population (**Table 5-2**).

Table 5-2. *Description of GWAS sample.*

Characteristic	**N = 5,976**[1]
Neutrophil count (10^9 cells/Litre)	2.9 (1.2)
Genetic sex	
Female	3,399 / 5,976 (57%)
Male	2,577 / 5,976 (43%)
Age (years)	51.8 (8.1)
BMI (kg/m^2)	29.8 (5.3)
K-means cluster	
K1	519 / 5,976 (8.7%)

Characteristic	N = 5,976[1]
K2	1,142 / 5,976 (19%)
K3	1,150 / 5,976 (19%)
K4	967 / 5,976 (16%)
K5	1,174 / 5,976 (20%)
K6	841 / 5,976 (14%)
K7	183 / 5,976 (3.1%)

1Mean (SD); n / N (%)

Next, I aimed to assess if neutrophil count was normally distributed because most genomic association tests assume that a quantitative traits, such as neutrophil count, have properties consistent with this [549]. Therefore, I performed a Shapiro-Wilk test on the neutrophil count variable, which can provide evidence for the presence of a non-normal distribution [550,551]. This is given by the W-statistic of the Shapiro-Wilk test, where a W = 1 is used as an indicator of a normal distribution.

The untransformed neutrophil count variable had a W-statistic of 0.95, indicating a slight deviation from normality. After applying a natural log-transformation, nc_log (log of neutrophil count) there was strong evidence for a normal distribution (Shapiro–Wilk W = 0.999). Indeed, the median and mean values of the transformed variable were equal, and the histogram appeared to be normally distributed (**Figure 5-4A**). Given these results, I decided to use this transformed nc_log variable in the GWAS. There was some variation in nc_log between each K-means cluster (Kpop) (**Figure 5-4B**), although this was low, with the median hovering around 1.

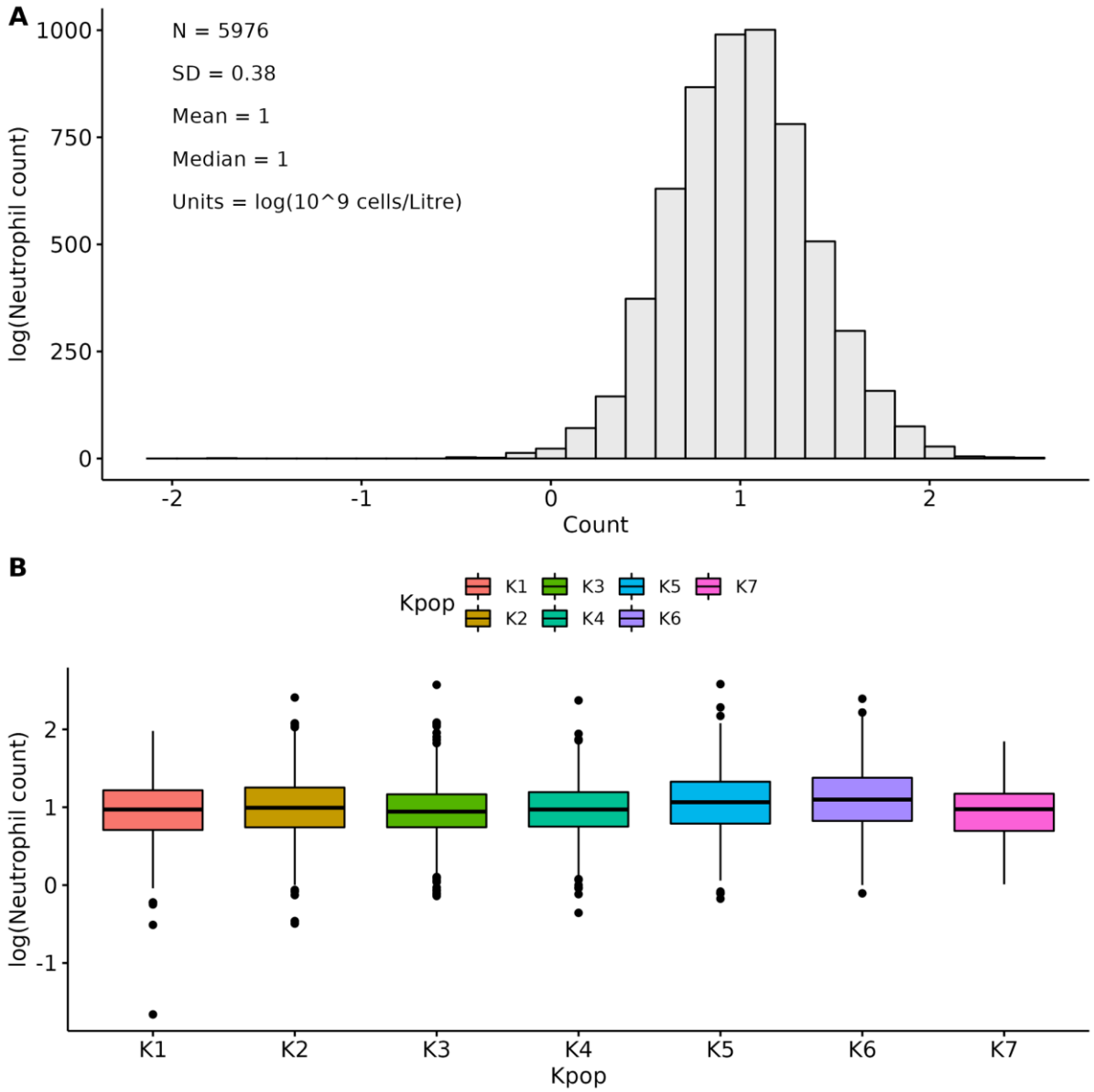


Figure 5-4. *Neutrophil count variation in the GWAS sample.*
Histogram outlining the distribution of neutrophil count levels is shown in the whole AFR_CAG population (A), along with representative boxplots describing neutrophil count variation by K-means cluster sample (B).

As mentioned previously, GWAS are usually done using individuals of a similar genetic background to avoid SNP-Trait associations that are biased or are false-positives due to a confounding effect by ancestry [552]. However, even in white British individuals from UKBB, latent population structure can still affect SNP effect sizes, which requires adjusting for PCs [402]. The analyses in **Chapter 4** indicated prominent population structure in the AFR CAG, given by the estimated 7 Kpops and ADMIXTURE analysis [238]. Therefore, I investigated the number of PCs that should be added into the GWAS to control for population structure.

First, I used the Tracy-Widom statistic from the EIGENSOFT package [226], which provides a quantitative estimate of the PCs that might be added as covariates. This analysis indicated over 100 significant PCs. However, there is no exact way to establish how many PCs should be added into a GWAS, although an unnecessary number of PCs can lead to a reduction in power, while too little might bias GWAS effect sizes due to residual population structure [339]. Previous studies have added 40 to 100 PCs in their UKBB GWA analyses [402,404,436], and inclusion of the first 100 PCs was found useful in adjusting for the effect of population structure in rare SNPs [553].

I next aimed to study if including the first 100 PCs as covariates would be sensible, or whether they would lead to over correction and lead to false negative results [554]. As a test, I picked the rs284778 to study because of its well-known association with lower neutrophil count from the T allele, which is most common in Africans [160,166]. Moreover, the T allele is predominantly absent in Europeans [508], making it a SNP particularly liable to bias in a African-European admixed population. In the previous chapter (**Chapter 4**), I showed that K-means clusters K5 and K6 mostly overlap the Caribbean and Northern Europe regions using 'self-report place of birth' data from UK Biobank [238], indicating population admixture. Indeed, the highest degree of heterozygosity for the rs2814778 SNP was within these two Kpops (**Figure 5-5**) [160].

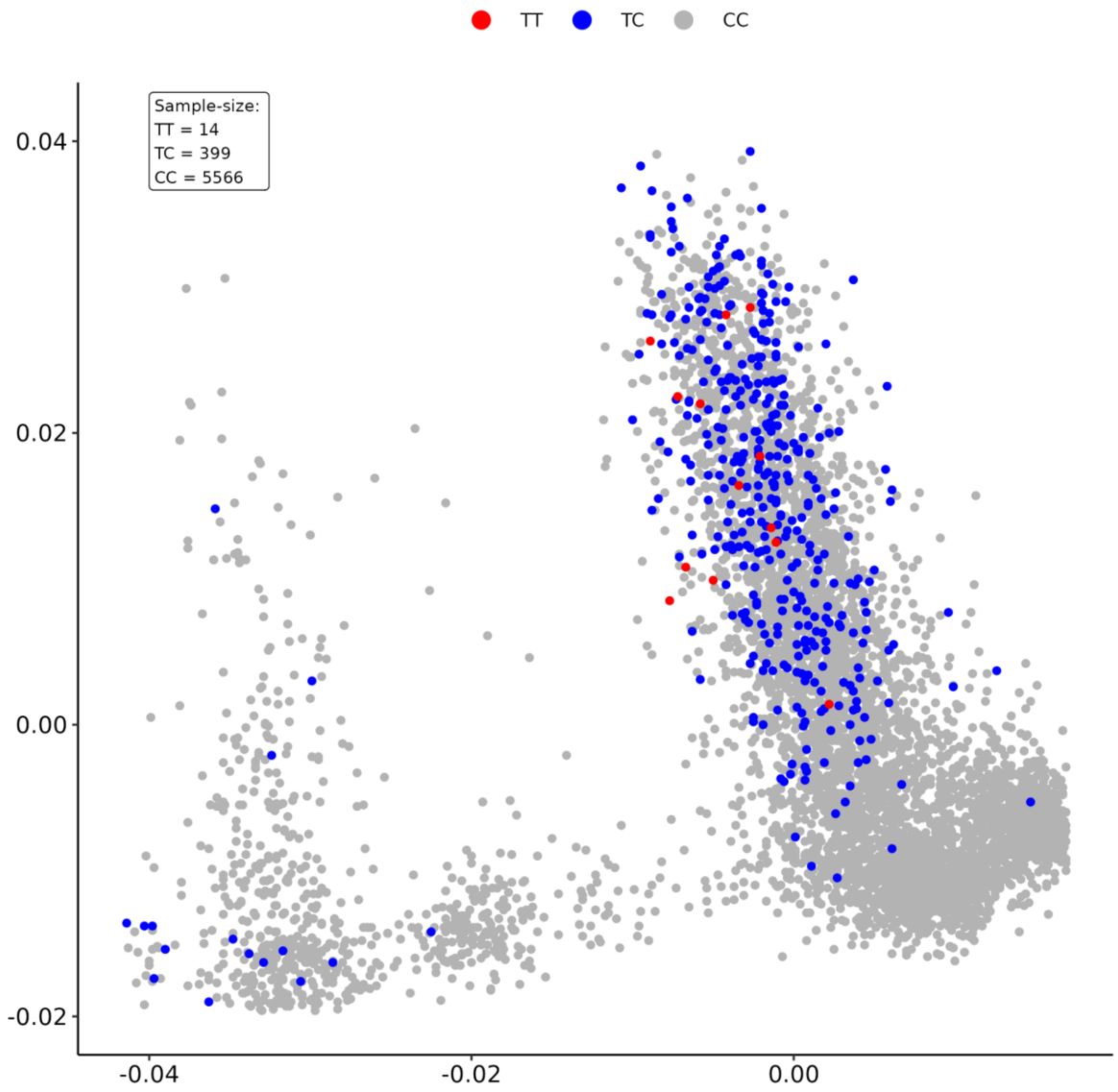


Figure 5-5. *Scatterplot of rs2814778 genotype on PC1~PC2 plane.*
Grey dots are the CC genotype, which is most common in those of African ancestry, blue are the TC genotype, while red is the TT genotype, most common in Europeans.

Next, I ran a linear model in R to establish the effect on rs2814778 on neutrophil count when adding the first 100 PCs as covariates. Each Kpop was investigated separately and showed similar point estimates (**Figure 5-6**). The confidence intervals (CIs) of Kpop K7 were large, but this was due to only one T allele being present in the cluster. While not necessarily generalisable to all SNPs, this result was encouraging as evidence of an association was still present after adjusting for 100 PCs. The findings in **Chapter 4** and the example shown here, along with previous studies outlining the rich genomic diversity in sub-Saharan Africa [397,555], prompted me to use the first 100 PCs as covariates.

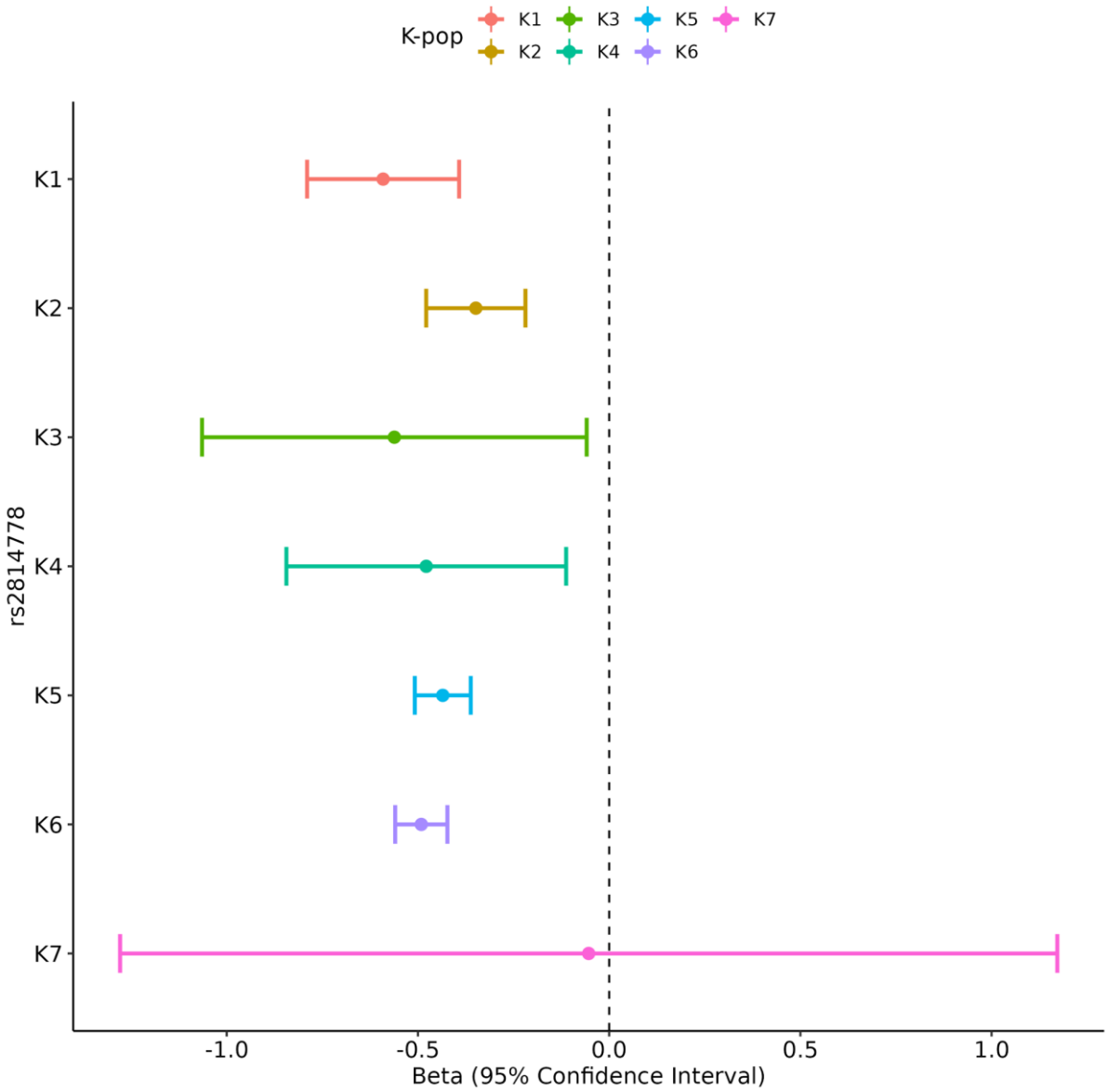


Figure 5-6. *Forest plot of rs2814778 association with log neutrophil count in each Kpop. Results for the model in R adjusting for sex, age, assessment centre, batch and PCs 1:100. Effect sizes are displayed with 95% CIs.*

Finally, I conducted a power calculation. Statistical power is the probability that the null hypothesis (that there is no association) is false [556,557]. Sample-size is an important factor in GWAS [556,557], and the higher the sample-size, the higher the power to detect SNPs which explain a smaller proportion of the variance (heritability) in a particular trait [540].

Therefore, my aim was to establish how liable the AFR_CAG sample would be to false negatives. Statistical power for association testing of all SNPs on neutrophil count was assessed at different values of heritability (h_2) in the AFR_CAG dataset (N=5,976). The calculation indicated over 80% power when the h_2 would be between 0.7-1.0% for

neutrophil count (**Figure 5-7**). This means that the sample-size is large enough to detect SNPs which have a medium to large effect but might be prone for false negatives for SNPs which only contribute little to neutrophil count variation. However, this calculation was done assuming a GWAS run with a traditional linear model.

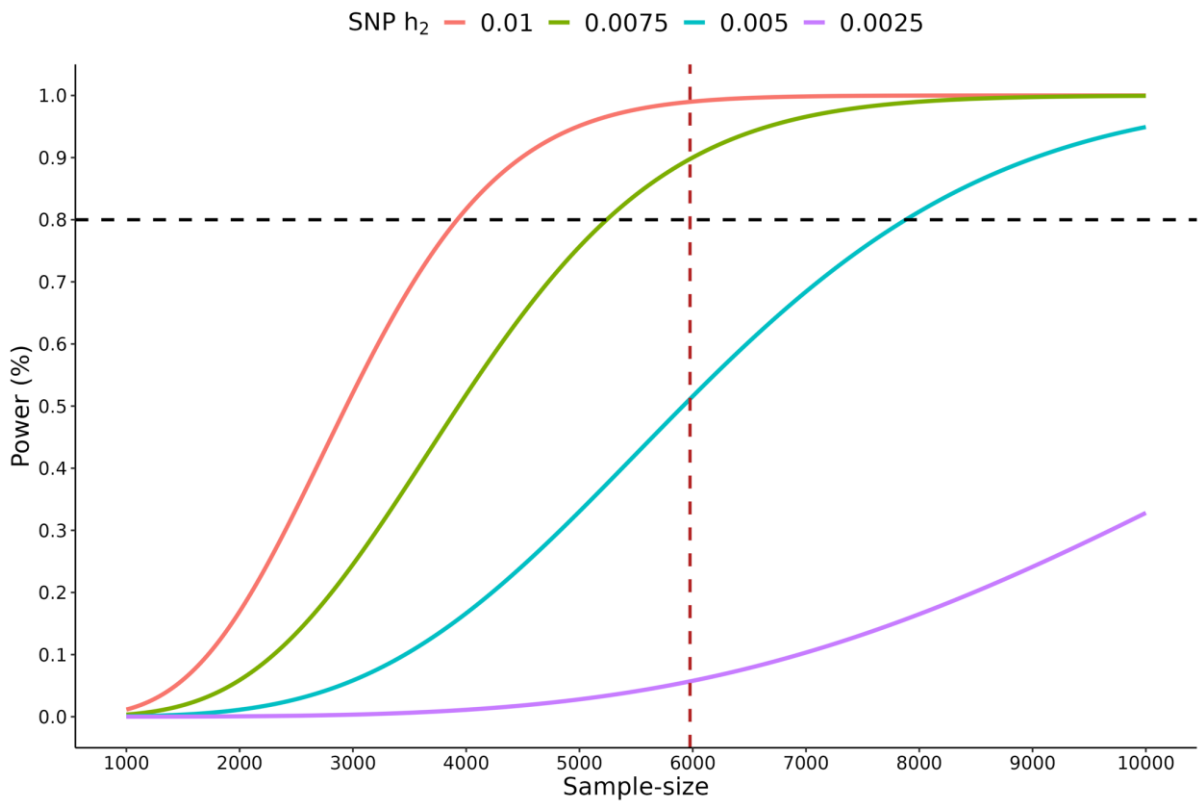


Figure 5-7. *Power calculation of a GWAS AFR_CAG sample.*
The x-axis indicates the sample-size, while the y-axis is the statistical power of an association test. Each curved line shows how power varies by sample-size at different degrees of the variance explained by all the SNPs on neutrophil count (0.01, 0.0075, 0.005, 0.0025). A black horizontal line is fixed at Power=80%, and a red vertical line is drawn at the GWAS sample size of 5,976.

5.3.2. Genome-wide association study

Given the results from the investigative analyses, I decided to use BOLT-LMM for the main GWAS, which employs a linear-mixed model algorithm for conducting association testing [435]. It has been used extensively due to its ability to attempt to account for population structure when running GWAS, particularly in people of European ancestry in UK Biobank [234]. Moreover, BOLT-LMM has been shown to lead to an increase in power compared to linear models [435].

However, it is unknown how well BOLT-LMM performs in non-European populations due to the nature of SNPs and the lineages that derive them to lead to differences in LD or allele frequencies [558]. To ensure that the BOLT-LMM results are reliable, I also aimed to conduct two additional GWAS using a standard linear model. Therefore, three GWAS were performed on the AFR_CAG sample: the main one with BOLT-LMM, and two with SNPTEST/META, with (META-WD) and without (META-WOD) including rs2814778 as a covariate. META-WD was conducted to serve as a negative control.

I performed three additional filtering steps prior to running the post-hoc analyses. This was done to add evidence to the reliability of the GWAS results. As mentioned in **Chapter 2**, most UKBB SNPs were imputed based on ~800K directly genotyped SNPs, and the imputation software provided an "INFO" score [164], which indicates the quality of the imputation [559]. I picked an INFO score threshold of 0.3, as it gives the best balance between data quality and quantity. Another filtering process was a Hardy-Weinberg equilibrium (HWE) test, which is usually used to find SNPs with poor genotyping [525]. Finally, related individuals from the dataset were removed, resulting in 5,509 unrelated people in the filtered AFR_CAG dataset. SNPs with a minor allele count of less than 17 (corresponding to the new sample-size from 20) were removed. 23,530,028 SNPs remained after filtering by INFO score, HWE test and minor allele count.

This AFR_CAG filtered sample was taken forward for further analyses. The BOLT-LMM GWAS run was treated as the main analysis. Here, 704 genetic variants passed the GWAS significance threshold of P<5e-8. Most of these signals were in chromosome 1, in the proximity of rs2814778, which had the lowest P-value across the genome (2.7E-87) (**Figure 5-8**). The META-WOD GWAS had 373 variants passing the threshold, while the META-WD GWAS had 31 significant SNPs, evidencing that most of the identified top signals in META-WOD were likely in LD with rs2814778.

Next, I aimed to identify which SNPs might causally associate with neutrophil count. To do this, I used a conservative GCTA-COJO approach [524], which yielded 10 index SNPs: rs12747038, rs138163369, rs140048432, rs144109344, rs183362544, rs186218882, rs2814778, rs28734019, rs527921556, rs530475031, rs557482905, rs558204720 (**Figure 5-8, Table 5-3**). Genomic location context of each index SNP is available in **Appendix 17, Appendix 18 and Appendix 19**.

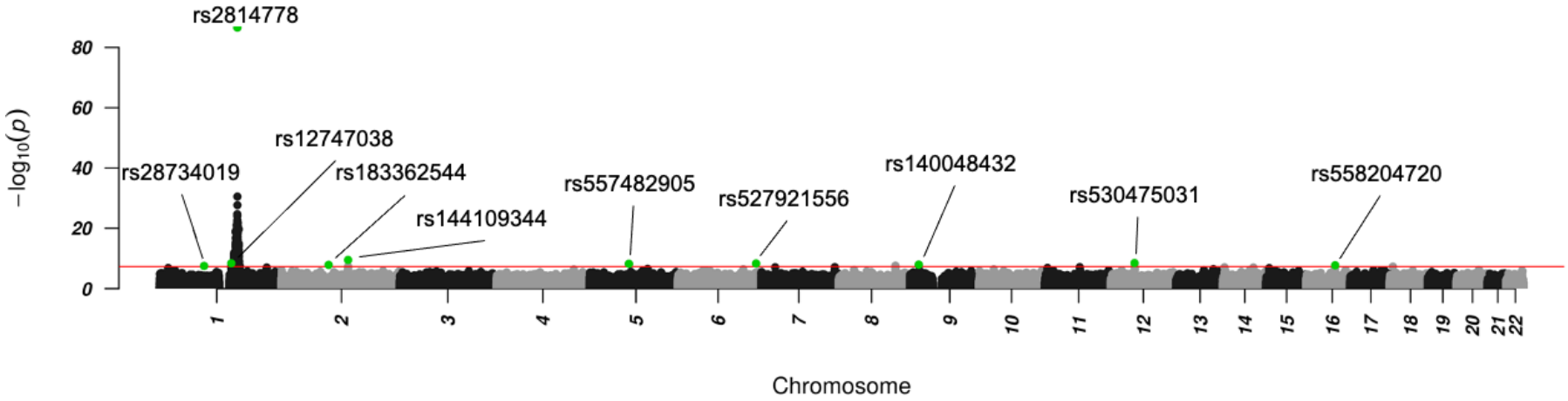


Figure 5-8. *Manhattan plot of neutrophil count GWAS.*

The x-axis is the base-pair position inside each chromosome, while the y-axis is the -log of the association P-value. A GWAS significance line is drawn to correspond to P=5e-8 on the -log(P) axis (A). Index SNPs from the GCTA-COJO run are highlighted in green. QQ-Plot of observed vs. expected P-values for each SNP, along with the genomic inflation factor on the top-left (B).

Table 5-3. *GCTA-COJO index SNPs.*

CHR	SNP	BETA.BOLT	SE.BOLT	P.BOLT	BETA.META-WOD	BETA.META-WD	N.META	Nr.K.META
1	rs2814778	0.43	0.02	2.66E-87	0.28	0	5,793	6
2	rs144109344	-0.12	0.02	3.12E-10	-0.06	-0.06	5,976	7
12	rs530475031	0.73	0.12	3.16E-09	0.47	0.46	4,952	5
1	rs12747038	-0.22	0.04	3.89E-09	-0.13	-0.08	5,976	7
6	rs527921556	0.4	0.07	4.48E-09	0.33	0.31	5,793	6
5	rs557482905	0.55	0.1	5.79E-09	0.42	0.38	3,778	4
9	rs140048432	-0.33	0.06	1.11E-08	-0.25	-0.25	5,976	7
2	rs183362544	0.61	0.11	1.27E-08	0.28	0.28	2,717	4
16	rs558204720	0.52	0.09	1.67E-08	0.37	0.33	1,486	2
1	rs28734019	-0.65	0.12	2.89E-08	-0.53	-0.5	4,124	4

The effect sizes of the BOLT-LMM index SNPs were compared with the ones from SNPTEST/META GWAS. Direction was consistent and effect sizes were similar between the three GWAS, with those generated from the BOLT-LMM run being slightly larger, most likely due to the improved power of the linear-mixed model (**Figure 5-9**). As expected, the META-WD effect size for the rs2814778 SNP was zero when in its inclusion as a covariate, suggesting no errors in the R linear model prior to integration into SNPTEST/META.

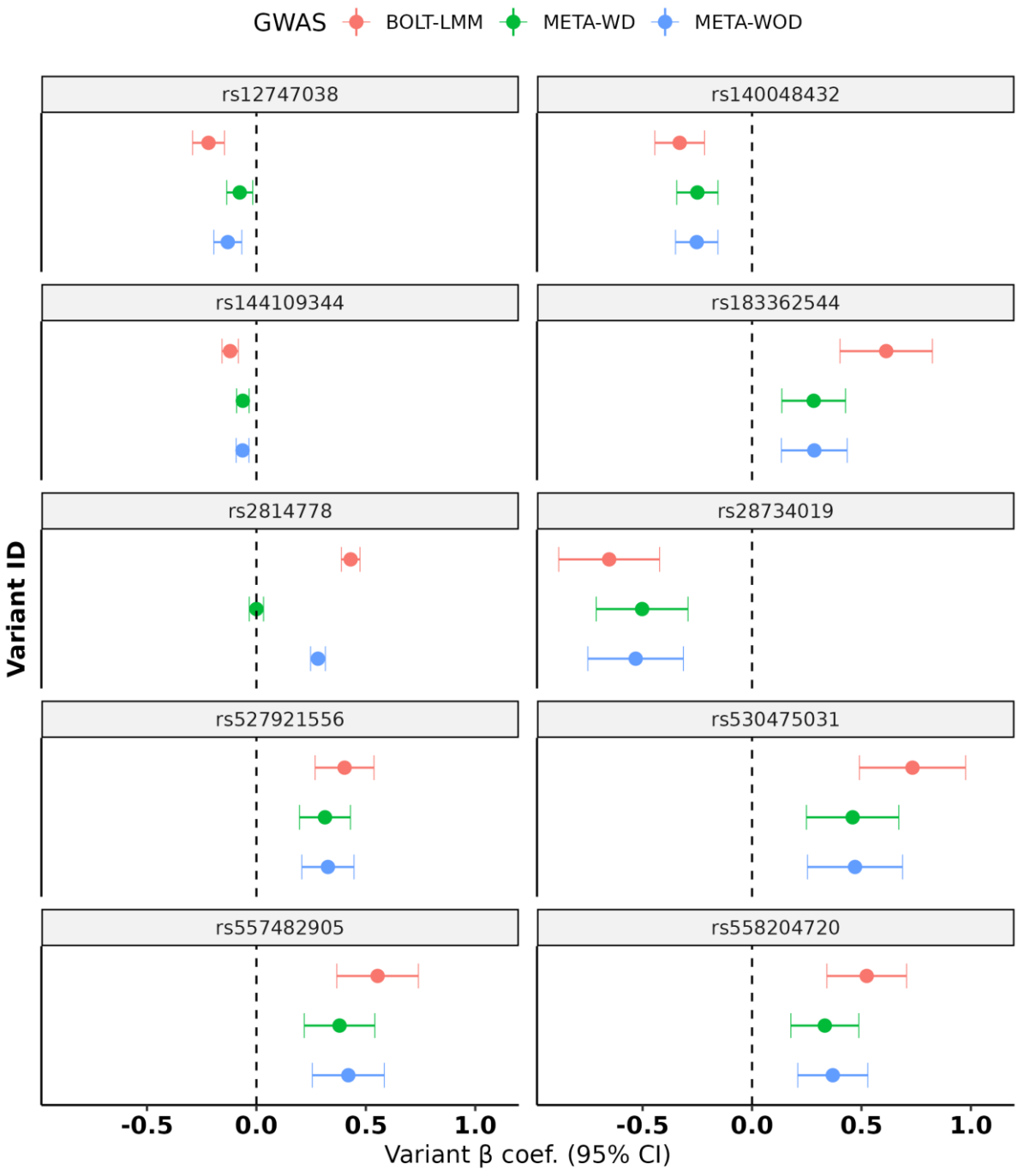


Figure 5-9. *Effect estimates of the index SNPs.*

The beta coefficient for each index SNP is displayed along with 95% CIs. These are displayed for the BOLT-LMM, META-WOD and META-WD GWAS.

Next, a sensitivity analysis was done to check that the GWAS results were not affected by human error and/or software bugs. I used 10 independent SNPs from the Chen et al. study [166] and compared their effect sizes in AFR_CAG dataset across the different methods I employed. There was consistency of directionality and effect sizes between all approaches, including with those from the Chen study (**Figure 5-10**). This indicates that the R scripts and software runs were not affected by technical artifacts.

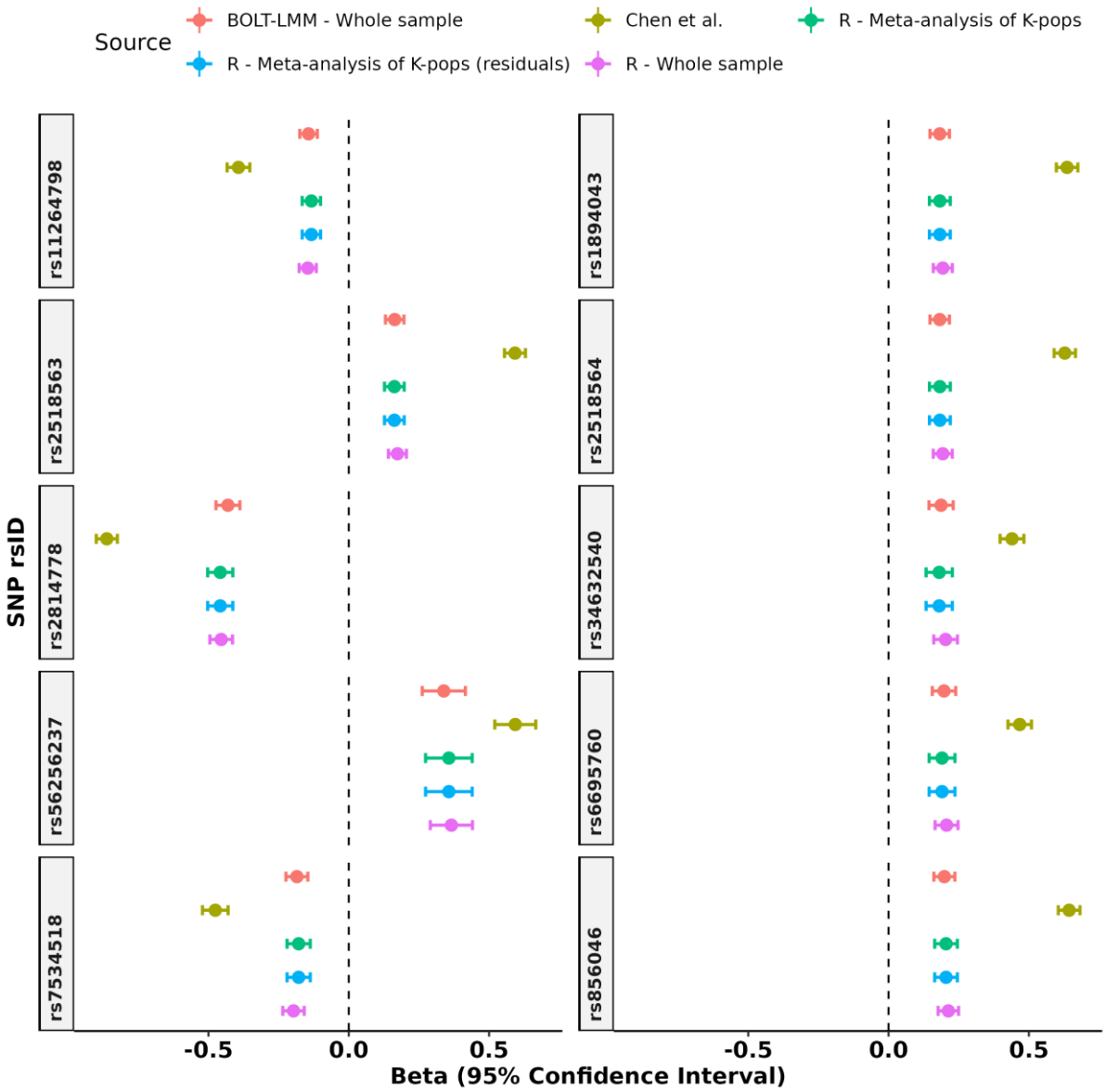


Figure 5-10. *Sensitivity analysis of methods.*
The effect sizes of the top 10 SNPs that were replicated from Chen et al. (beige) in the AFR_CAG dataset were plotted for the following analyses: BOLT-LMM (red), meta-

analysis of Kpops using R (green), meta-analysis of Kpops using residuals in R (blue), association test using the whole AFR_CAG sample in R (pink).

Furthermore, I investigated the association statistics of the index SNPs in each Kpop. This was done to detect discrepancies in directionality and effect sizes, which could indicate residual population structure or a SNP association with a specific Kpop. Overall, there was agreement in direction, and some variation in effect sizes was detected across Kpops (**Figure 5-11**).

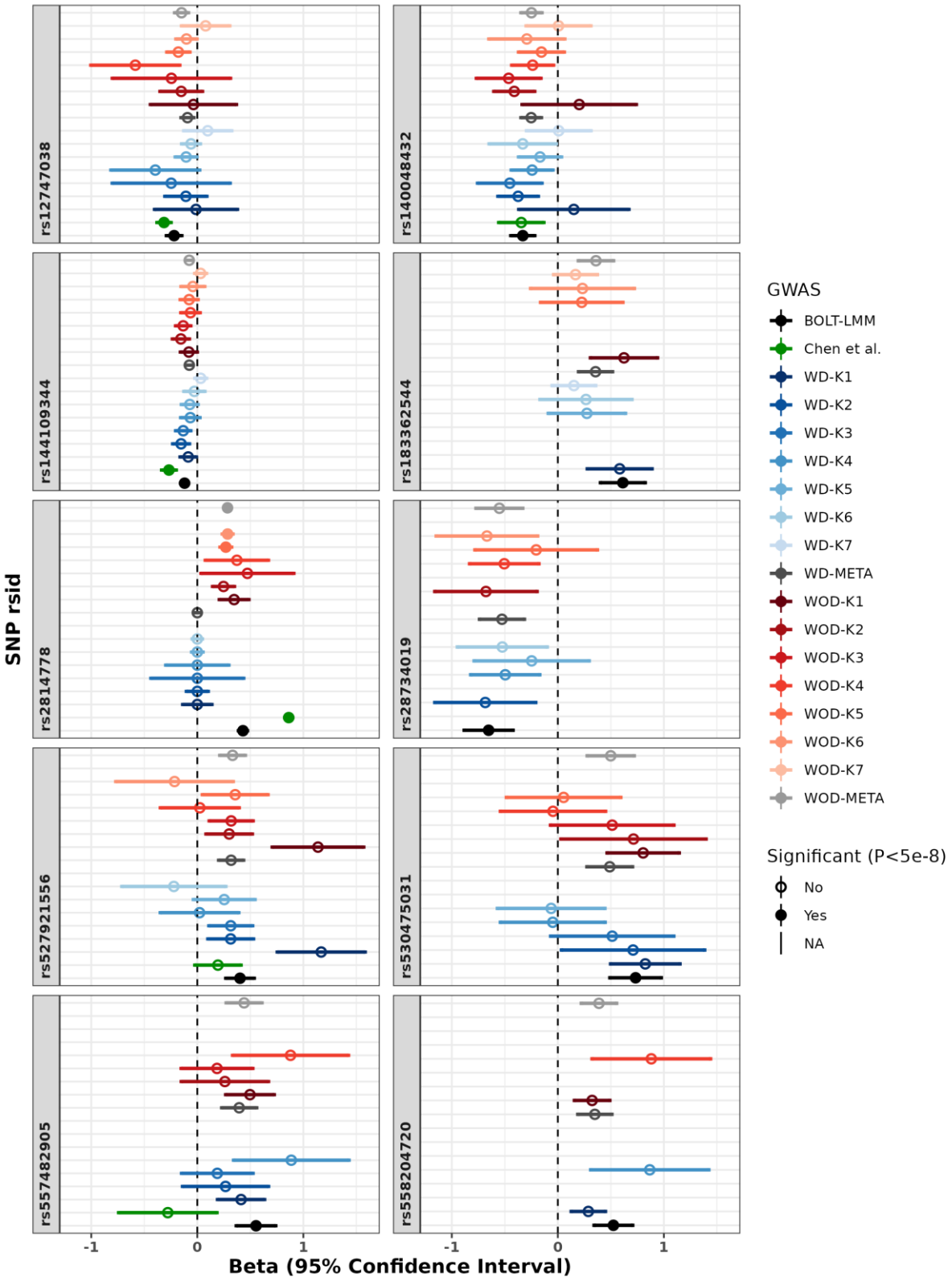


Figure 5-11. *Forest plot of index SNPs by K-means cluster.*

The effect-size of each BOLT index SNP was compared to that from SNPTEST/META, by-Kpop runs and Chen et al GWAS. Effect-sizes for each SNP across GWAS are present in the respective boxes of the figure. The x-axis indicates the effect-size (beta coefficient) of each SNP with 95% CIs, while the y-axis is the type of GWAS (indicated by the figure legend colouring). Some effect sizes were not displayed, either due to a low

minor allele count in the case of the Kpop GWAS, or due to not being present in the summary statistics, in the case of the Chen GWAS. Non-signif. WD = adjusting for rs2814778; WOD = without adjusting for rs2814778.

The GCTA-COJO analysis was also run on the two SNPTEST/META GWAS. The META-WOD analysis identified rs2814778, rs138163369 and rs570518709 as index SNPs. Similarly, the META-WD analysis identified rs138163369 and rs570518709. These two latter SNPs were not identified as index SNPs in the BOLT-LMM analysis, but their P-values were similar (rs138163369 – 4.90E-08, 2.28E-08, 1.22E-08; rs570518709 – 8.10E-08, 1.07E-09, 3.03E-09) (**Table 5-4**).

Table 5-4. *GCTA-COJO independent SNPs between all three GWAS.*
GWAS are BOLT-LMM, META-WOD, META-WD on neutrophil count.

SNP	Shared by	BETA.BOLT	P.BOLT	P.META-WOD	P.META-WD
rs12747038	BOLT	-0.22	3.90E-09	9.31E-06	0.00290611
rs138163369	META-WD|META-WOD	0.53	4.90E-08	2.28E-08	1.22E-08
rs140048432	BOLT	-0.33	1.10E-08	9.78E-07	3.56E-07
rs144109344	BOLT	-0.12	3.10E-10	1.47E-06	9.45E-07
rs183362544	BOLT	0.61	1.30E-08	2.26E-05	1.39E-05
rs186218882	META-WD	0.55	1.50E-06	6.80E-08	1.91E-08
rs2814778	BOLT|META-WOD	0.43	2.70E-87	3.53E-60	0.999999
rs28734019	BOLT	-0.65	2.90E-08	9.57E-07	1.11E-06
rs527921556	BOLT	0.40	4.50E-09	6.92E-08	1.17E-07
rs530475031	BOLT	0.73	3.20E-09	1.07E-05	8.44E-06
rs557482905	BOLT	0.55	5.80E-09	3.03E-07	2.83E-06
rs558204720	BOLT	0.52	1.70E-08	5.22E-06	2.14E-05
rs570518709	META-WD|META-WOD	0.73	8.10E-08	1.07E-09	3.03E-09

As another sensitivity analysis to test the reliability of the BOLT-LMM results, the effect sizes of all GCTA-COJO SNPs were compared in pair-wise manner across the three GWAS. A regression line was fit through the scatter plots, showing a large degree of correlation between the BOLT-LMM effect sizes and the SNPTEST/META runs (META-WOD R^2 = 0.91, META-WD R^2 = 0.93) (**Figure 5-12**).

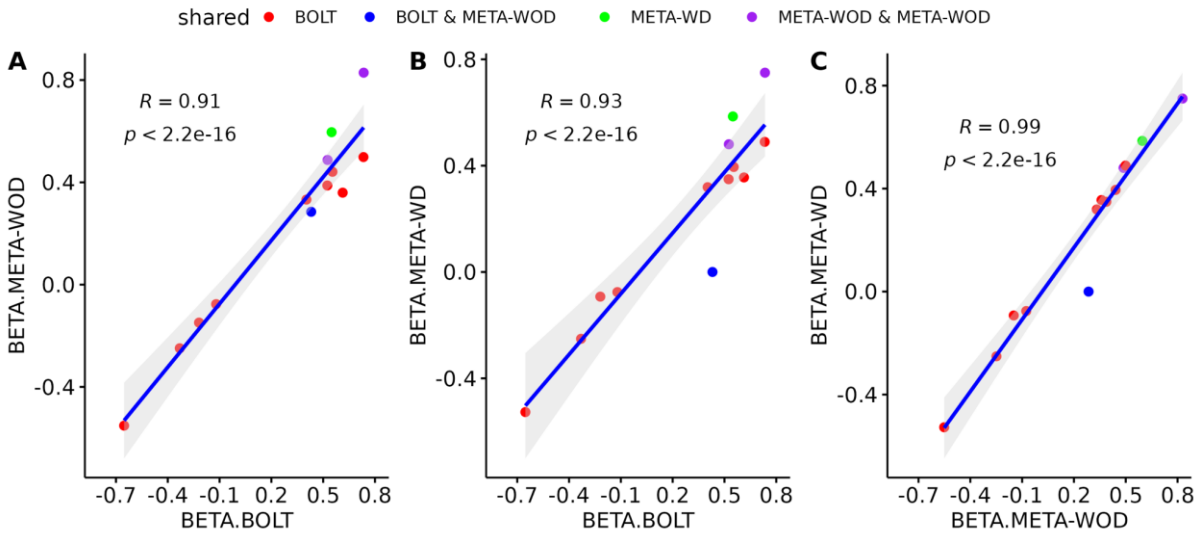


Figure 5-12. *Scatter plot of GCTA-COJO effect sizes.*
Comparison of effect sizes of all GCTA-COJO independent signals of BOLT-LMM with META-WOD (A), BOLT-LMM with META-WD (B) and META-WOD with META-WD (C).

Afterwards, I calculated the genomic inflation for the three GWAS. This was done to assess if the inflation seen in the main GWAS was predominantly due to population structure, technical errors or from polygenicity i.e. multiple independent SNPs associated with neutrophil count [529]. The GC inflation factor λ was lower in the SNPTEST/META GWAS runs (1.016, 1.013) compared the BOLT-LMM run (**Table 5-5**). Moreover, the QQ-plot of the BOLT-LMM and META-WOD GWAS were similar and did not display an early deviation from the expected P-value, indicating no systemic bias in association statistics [560] (**Figure 5-13**). The lower λ in the SNPTEST/META runs combined with the QQ-plot results indicate that the higher inflation in the BOLT-LMM GWAS is most likely due to polygenicity i.e. detection of increased top loci (SNPs) across the genome compared with the linear models, increasing the evidence that the main GWAS results are reliable.

Table 5-5. *Genomic control of all three GWAS.*

GWAS type	GC lambda
BOLT	1.047

META-WOD	1.016
META-WD	1.013

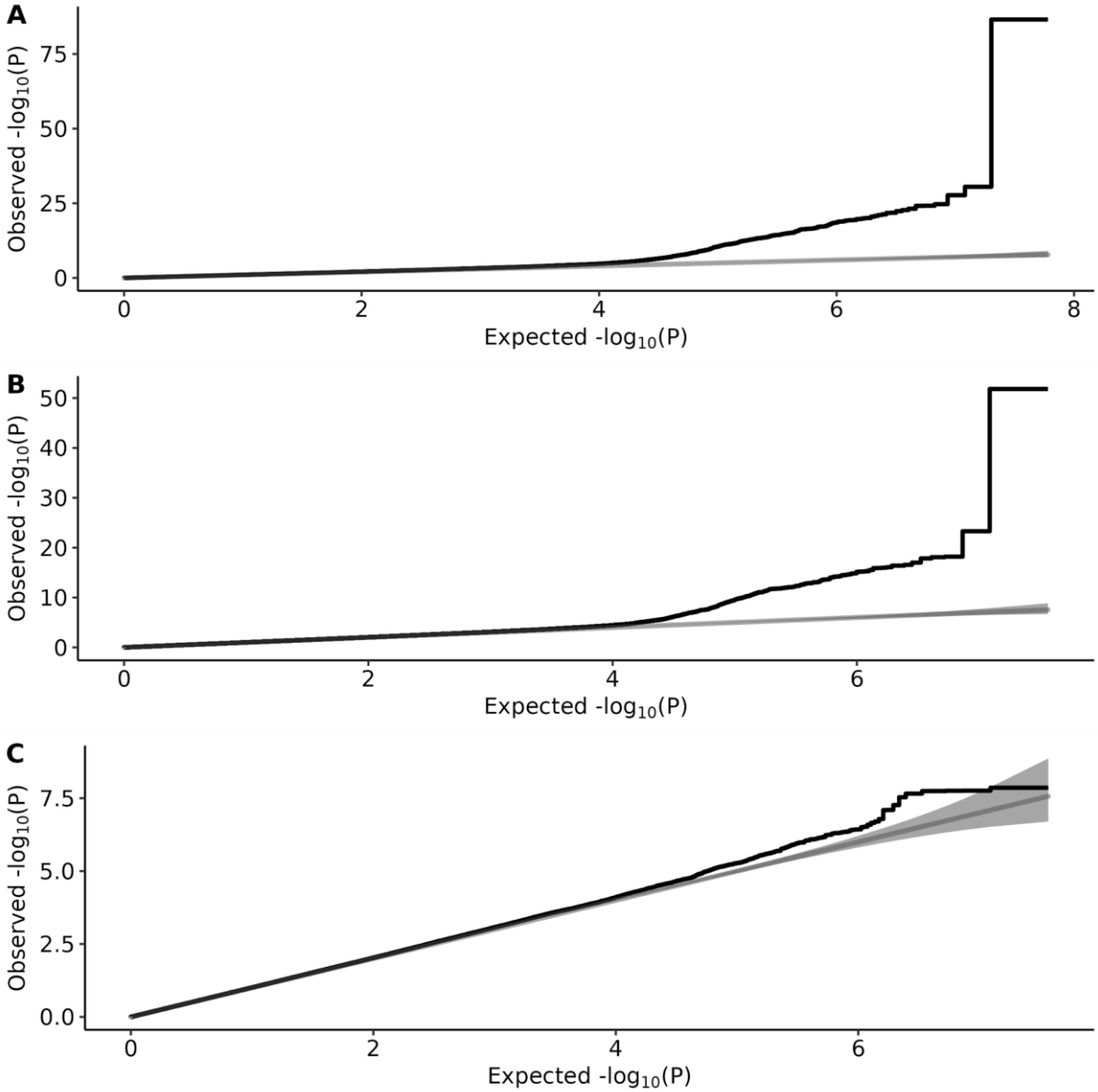


Figure 5-13. *QQ-Plots of all three GWAS.*
The X-axis is the expected P-value (-log) and the Y-axis represents the observed P-value (-log).

A common step in GWAS is the identification of independent SNPs through clumping i.e. SNPs with the lowest P-value in a particular genomic window [525]. This is generally less conservative than the COJO approach [561], but is useful in understanding the genetic architecture of the studied trait [154]. Two PLINK clumping analyses were performed on the filtered AFR_CAG summary statistics using the same clumping parameters on the

well-known FUMA platform [531]. Here, 193 SNPs were identified as loci at the relaxed threshold of r^2=0.6, 73 independent loci at the stringent threshold of r^2=0.1 (**Appendix 20**). Finally, 12 top loci were identified at r^2=0.001 and a 10Mb window, which are the very conservative MR clumping parameters [532,533].

Furthermore, a FUMA analysis was run on the filtered AFR_CAG dataset for the top loci (r^2=0.1). This was done to visualise which genomic locations are affecting neutrophil count and if they are more likely to have a particular genetic function compared to the whole genome i.e. functional variants [562]. Seventeen genomic risk loci were identified (**Figure 5-14**). The ANNOVAR analysis [563] showed evidence for changes in genetic function enrichment relative to all SNPs in the reference panel – intronic [-log(E) = 0.678, P = 5.91e-11], non-coding intronic RNA [-log(E) = 1.42, P = 2.05e-4], upstream of gene [-log(E) = 2.59, P = 3.06e-4], three prime untranslated region (UTR3) [-log(E) = 2.57, P = 4.86e-4], downstream of gene [-log(E) = 2.02, P = 1.38e-2], exonic [-log(E) = 1.73, P = 4.86e-2], non-coding exonic RNA [-log(E) = 1.82, P = 3.68e-2] (**Figure 5-14**).

Next, I investigated if the independent SNPs in the main GWAS were present in the GWAS Catalog [167], as I aimed to see if they have been previously associated with WBC count or immunity. Here, SNPs predominantly showed associations with white blood cell count variation, further improving the reliability of the GWAS (**Appendix 21**).

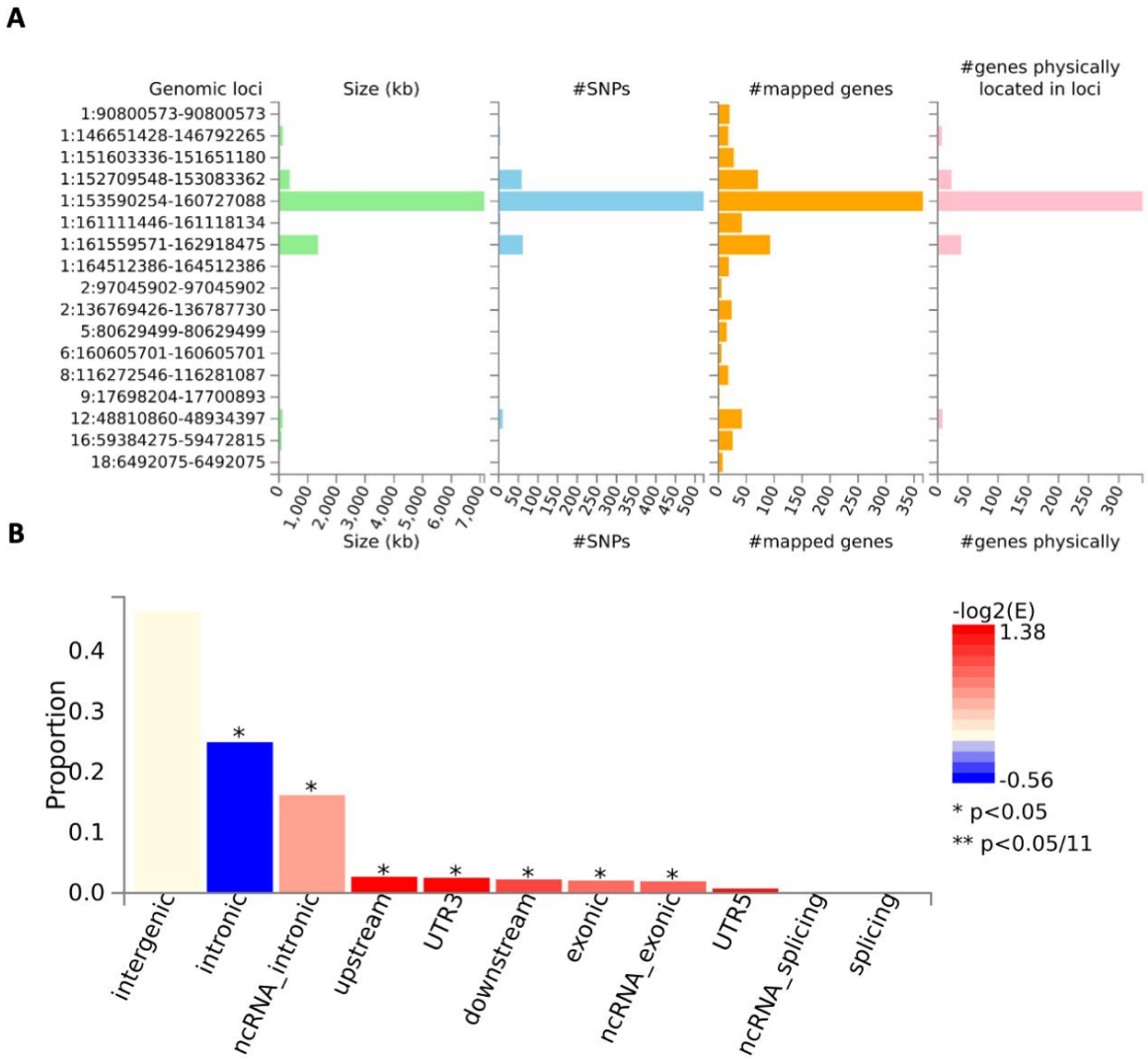


Figure 5-14. *Description of genomic risk loci.*
FUMA analysis results for SNPs passing the GWAS significance threshold in the BOLT-LMM filtered GWAS.

The AFR_CAG GWAS was contrasted with a neutrophil count GWAS in Africans from Chen et al. [166], where a quantitative analysis, 81.71% of the GWAS significant SNPs from Chen et al. were replicated (using the same covariates) in the AFR_CAG dataset (P<0.05) (**Table 5-6**). The Manhattan plots also visually showed a good degree of overlap (**Figure 5-15**). As Similarly, the AFR_CAG GWAS was compared with one done in Europeans from Astle et al. [149], highlighting the difference in the genetic architecture of neutrophil count between Africans and Europeans and displaying the difference in signal strength when conducting GWAS in hundreds of thousands of people compared to ~6000, as was the case with the AFR_CAG sample (**Figure 5-16**). Finally, SNPs that were top loci at r^2=0.1 were investigated in the Astle and Chen summary statistics, as

well as in the GWAS Catalog. Nineteen genetic variants were not present in these two datasets, 7 of which were index SNPs (**Table 5-7**).

Table 5-6. *Replication analysis of Chen independent loci.*

Summary	**P-value no PC covariates**	**P-value with PC1:10 as covariates**
N SNPs	13139	13139
Min	2.47E-158	6.53E-130
1st Quartile	2.20E-07	3.61E-05
Median	0.000124486	0.00225415
Mean	0.013520426	0.043460521
3rd Quartile	0.003708045	0.027477984
Max	0.936191471	0.997954198
Percentage P < 0.05	92.97%	81.71%

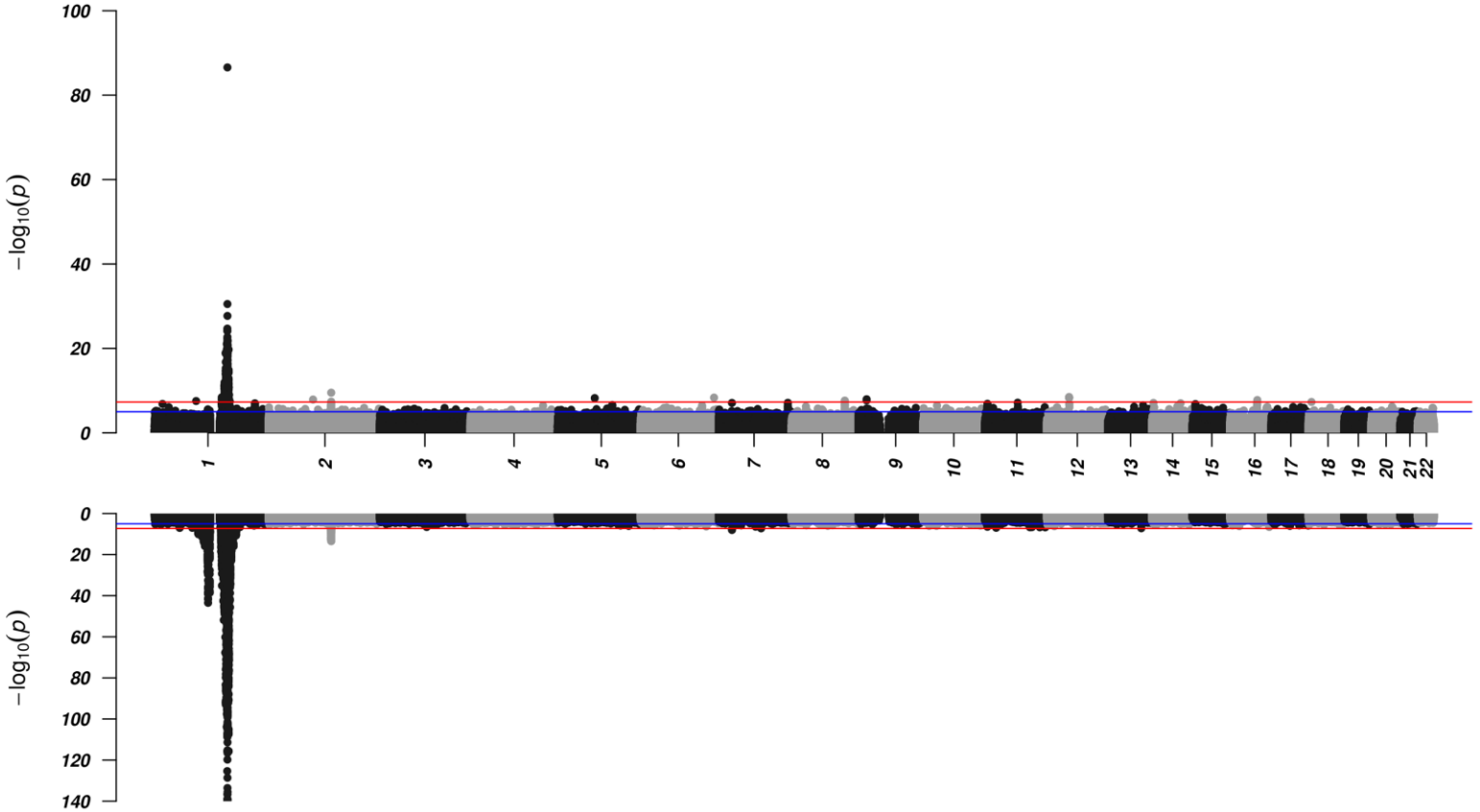


Figure 5-15. *Comparison of GWAS results for neutrophil count in Africans.*

Manhattan plot of BOLT-LMM neutrophil count GWAS from my study (top) mirrored with another Manhattan plot generated using summary statistics from a GWAS of neutrophil count done in people of African ancestry [166]*.*

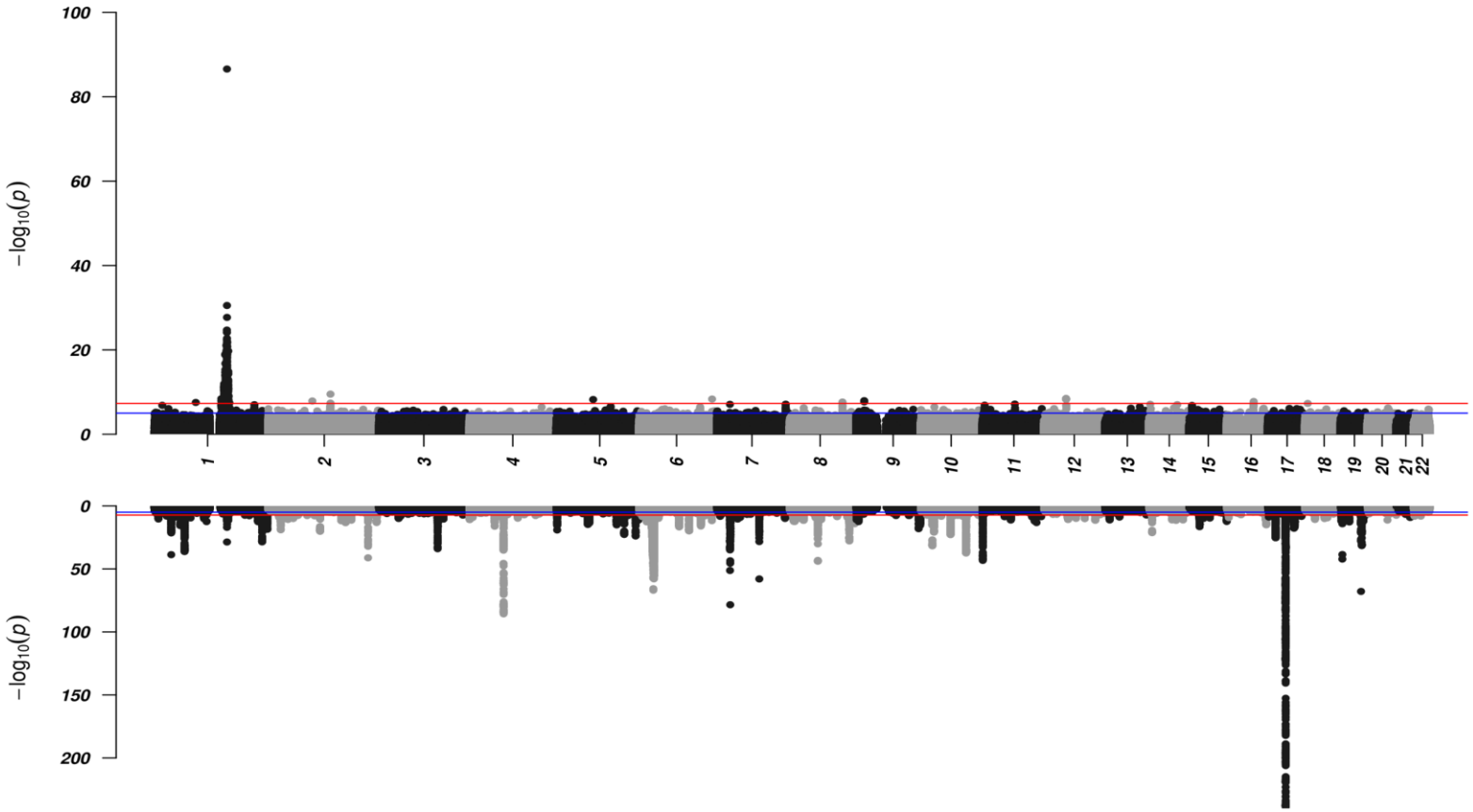


Figure 5-16. *Comparison of GWAS results for neutrophil count in Europeans.*

Manhattan plot of BOLT-LMM neutrophil count GWAS from my study (top) mirrored with another Manhattan plot generated by a GWAS of neutrophil count done in people of European ancestry in UK Biobank [149]*.*

Table 5-7. *Top loci not found in Astle or Chen.*

Only independent SNPs clumped at r = 0.1 are shown.

SNP	CHR	BP (GRCh37)	r0.001 lead?	cojo_index	In Astle/Chen?	Nearest gene	Type
rs28734019	1	90800573	Yes	Yes	No	RNU6-695P	Intergenic
rs61823703	1	159542164	No	No	No	OR10AE1P	Intergenic
rs539456851	1	158731459	No	No	No	OR6N1	Intergenic
rs371178711	1	158186653	No	No	No	RP11-404O13.5	Intergenic
rs146677619	1	158995984	No	No	No	IFI16	Intronic
rs11576058	1	161111446	No	No	No	UFC1	Intergenic
1:158777618_CT_C	1	158777618	No	No	No	OR10AA1P	Downstream
rs183362544	2	97045902	Yes	Yes	No	NCAPH	Intergenic
rs11422063	1	159799599	No	No	No	SLAMF8	Intronic
rs112483667	1	151651180	No	No	No	SNX27	Intronic
rs12406899	1	157540651	No	No	No	FCRL4	Intergenic
rs1103805	1	158924741	No	No	No	PYHIN1	Intronic
rs557482905	5	80629499	Yes	Yes	No	ACOT12	Intronic
rs527921556	6	160605701	Yes	Yes	No	SLC22A2	Intronic
rs10096834	8	116281087	Yes	No	No	TRPS1	Intergenic
rs140048432	9	17700893	Yes	Yes	No	SH3GL2	Intronic
rs530475031	12	48810860	Yes	Yes	No	C12orf54	Intronic
rs558204720	16	59472815	Yes	Yes	No	LOC105371298	Intronic
rs138163369	18	6492075	Yes	No	No	CTD-2124B20.2	Intergenic

5.3.3. Heritability analysis

A heritability analysis was conducted with GCTA-GREML to estimate the variance explained by the genetic component on neutrophil count. Without adjusting for rs2814778, the genetic variance was estimated at 0.101 (10.1%) (SE = 0.018), and the phenotypic variance at 0.133 (13.3%) (SE = 0.003) with an analysis P-value of 2.29e-09. When adjusting for the Duffy SNP, the genetic variance was estimated at 0.050 (5%) (SE = 0.017), twice as low as in the previous analysis, and the phenotypic variance was estimated at 0.123 (12.3%) (SE = 0.002), with the analysis P-value of 1.36E-03 (**Table 5-8**).

Table 5-8. *Estimated heritability of neutrophil count.*

Source	Variance.WOD	SE.WOD	Variance.WD	SE.WD
V(G)	0.101	0.018	0.050	0.017
V(e)	0.032	0.017	0.073	0.017
Vp	0.133	0.003	0.123	0.002
V(G)/Vp	0.761647	0.132462	0.406716	0.135124
P-value	2.29E-09		1.36E-03	
N	5509		5509	

V(G) = genetic variance

V(e) = environmental variance

Vp = total variance

V(G)/Vp = proportion of genetic variance from total variance

A second run of GREML was done by stratifying on LD regions for each chromosome to account for potential LD bias. This was not successful, as the standard errors when estimating heritability were too high.

5.3.4. Descriptive analyses of neutrophil count

Next, I aimed to assess if the index SNPs were still associated with neutrophil count when conditioning on variables such as BMI and smoking status. This was done to investigate the reliability of the index SNPs in the context of their relationship with neutrophil count. Moreover, I wanted to assess if including PCs was indeed accounting for possible population structure.

First, the descriptive statistics of the AFR_CAG dataset were studied with these additional variables (**Table 5-9**).

Table 5-9. *Detailed descriptive statistics.*

Characteristic	**N = 5,976**[1]
Menopause status	
Male	2,600 / 5,976 (44%)
Prefer not to answer	45 / 5,976 (0.8%)
No	1,279 / 5,976 (21%)
Hysterectomy	365 / 5,976 (6.1%)
Not sure - other	208 / 5,976 (3.5%)
Yes	1,479 / 5,976 (25%)
BMI (kg/m^2)	29.8 (5.3)
Missing	100
Smoking status	
Prefer not to answer	57 / 5,937 (1.0%)
Never	4,361 / 5,937 (73%)
Previous	897 / 5,937 (15%)
Current	622 / 5,937 (10%)
Missing	39
Alcohol drinker status	
Prefer not to answer	50 / 5,937 (0.8%)
Never	1,112 / 5,937 (19%)
Previous	349 / 5,937 (5.9%)
Current	4,426 / 5,937 (75%)
Missing	39
1Mean (SD); n / N (%)	

Several variables had missing data or had values assigned as “prefer not to answer” / “not sure” in the case of self-reported traits. There was no evidence of a difference in neutrophil count between these data types and those that were kept in the dataset (**Table 5-10**). 5,310 individuals remained in the dataset after filtering out these data types.

Table 5-10. *Association between excluded variable data and neutrophil count.*

Exposure	**Type**	**BETA**	**SE**	**P-value**	**N**
Body mass index	Missing	0.04	0.04	0.29	100

UN region of birth	Missing	0.00	0.03	0.89	138
Alcohol drinker status	Missing	-0.06	0.06	0.30	39
Alcohol drinker status	Prefer not to answer	-0.02	0.05	0.75	50
Menopause	Missing	-0.09	0.08	0.23	23
Menopause	Not sure	0.05	0.03	0.08	208
Menopause	Prefer not to answer	0.05	0.06	0.42	45
Smoking status	Missing	-0.06	0.06	0.31	39
Smoking status	Prefer not to answer	0.03	0.05	0.58	57

Next, the variance explained by each variable was studied. This was done to assess which variables might be added into a sensitivity GWAS. Several traits still explained a notable amount of variance in neutrophil count even in the type III ANOVA analysis. These were sample year (0.15%), sample month (0.29%), menopause (0.49%), self-reported UN region of birth (0.37%), assessment centre (0.40%), BMI (0.15%), smoking status (%1.58) and the rs2814778 SNP (9.56%) (**Figure 5-17**).

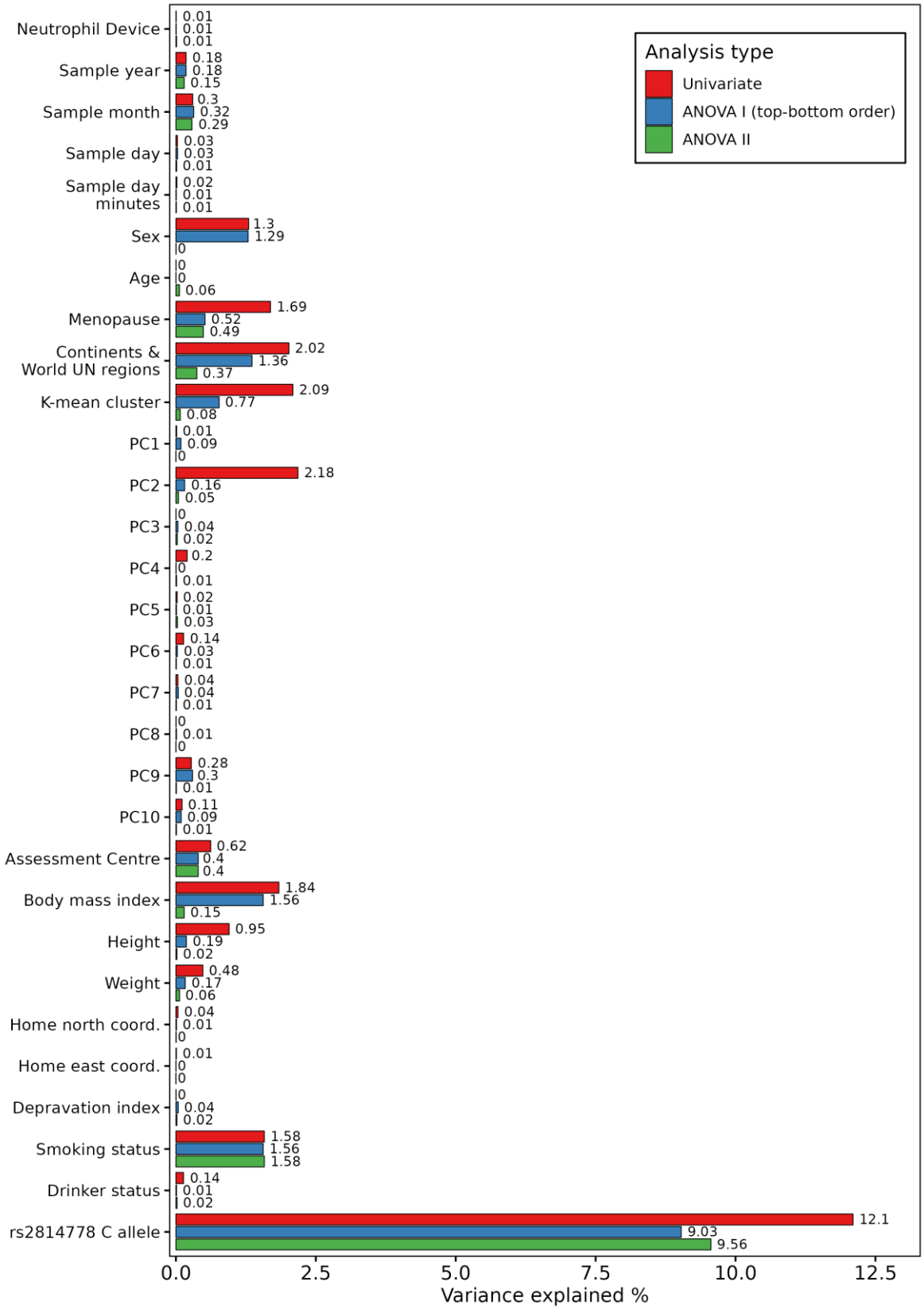


Figure 5-17. *Variance explained on neutrophil count by traits.*

5.3.5. BOLT-LMM run with additional covariates

The ANOVA results showed that the main GWAS results might have been affected to a degree by the variables listed in **Figure 5-17**, and that the PCs might have not captured the whole population structure present in the AFR_CAG sample. Therefore, to assess the reliability of the main GWAS, a sensitivity BOLT-LMM GWAS was done with six additional covariates on 5,310 individuals: UN region of birth, K-means cluster, smoking status, alcohol drinker status, menstrual status and BMI. The association statistics of this sensitivity run and the main BOLT-LMM GWAS run were compared, showing very similar results (**Table 5-11**). This provided evidence that the effect of these additional variables on the main GWAS were modest, and that the PCs added into the model largely counteracted the effects of population structure.

Table 5-11. *Comparison between main BOLT-LMM and BOLT-LMM with additional covariates.*

SNP	CHR	BP (GRCh37)	BETA.BOLT	SE.BOLT	P.BOLT	BETA.sensitivity	SE.sensitivity	P.sensitivity
rs28734019	1	90800573	-0.65	0.12	2.90E-08	-0.70	0.12	1.40E-08
rs12747038	1	146651428	-0.22	0.04	3.90E-09	-0.24	0.04	1.30E-09
rs2814778	1	159174683	0.43	0.02	2.70E-87	0.49	0.02	1.40E-90
rs183362544	2	97045902	0.61	0.11	1.30E-08	0.63	0.12	3.10E-07
rs144109344	2	136787730	-0.12	0.02	3.10E-10	-0.13	0.02	2.30E-10
rs557482905	5	80629499	0.55	0.10	5.80E-09	0.59	0.10	6.00E-09
rs527921556	6	160605701	0.40	0.07	4.50E-09	0.48	0.07	1.30E-10
rs10096834	8	116281087	0.04	0.01	2.30E-08	0.04	0.01	1.60E-08
rs140048432	9	17700893	-0.33	0.06	1.10E-08	-0.35	0.06	4.70E-08
rs530475031	12	48810860	0.73	0.12	3.20E-09	0.84	0.13	3.30E-10
rs558204720	16	59472815	0.52	0.09	1.70E-08	0.55	0.10	8.60E-08
rs138163369	18	6492075	0.53	0.10	4.90E-08	0.59	0.11	2.60E-08

5.3.6. Mendelian randomization

Finally, after establishing the reliability of the GWAS through several post-hoc and sensitivity analyses, I conducted the MR analysis. Here, a bi-directional MR was done between neutrophil count and SM. For the latter, I used summary statistics from the MalariaGEN [245]. Only 3 SNPs were available to proxy for neutrophil count after data harmonization with the malaria dataset. For SM as an exposure, 7 SNPs were available for overall SM, 2 for CM and 3 for other SM.

There was little evidence of an effect of neutrophil count on overall severe malaria (IVW OR: 1.03, 95% CI: 0.98 to 1.07; P = 0.24), CM (IVW OR: 1.00, 95% CI: 0.94 to 1.06; P = 0.98), SMA (IVW OR: 1.08, 95% CI: 0.99 to 1.18; P = 0.08) and OTHER SM (IVW OR: 1.03, 95% CI: 0.98 to 1.09; P = 0.26), although the effect estimates were trending towards an increased risk of severity (**Figure 5-18, Table 5-12**). Similarly, there was little evidence of an effect of overall severe malaria (IVW OR: 2.03, 95% CI: 0.70 to 5.84; P = 0.19), CM (IVW OR: 2.14, 95% CI: 0.70 to 6.57; P = 0.18) and OTHER SM (IVW OR: 2.08, 95% CI: 0.59 to 7.34; P = 0.25) on neutrophil count. However, there was a direction agreement in effect estimates towards neutrophil count increase (**Figure 5-18, Table 5-12**). There were no SNPs instrumenting for SMA which passed the GWAS significance threshold, and therefore could not be analysed.

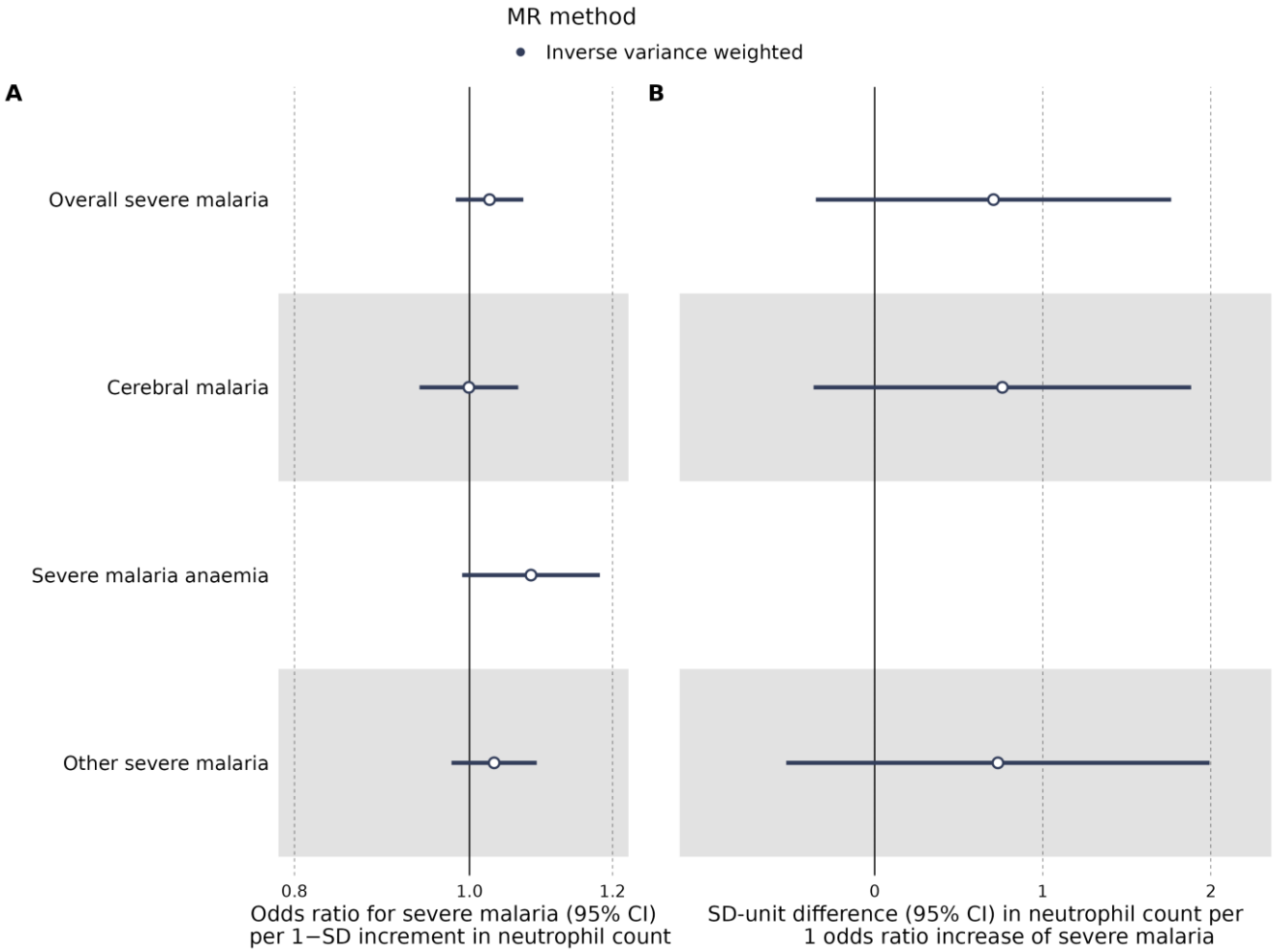


Figure 5-18. *Bi-directional Mendelian randomization.*

Forest plot of the IVW MR analysis with neutrophil count as an exposure (A) and severe malaria as an exposure (B). Overall severe malaria and its sub-phenotypes are listed on the y-axis, with the effect estimates on the x-axis. In the first instance, the MR results are interpreted as an OR increase severe malaria per 1-SD increase in neutrophil count, while in the latter as a 1-SD unit difference in neutrophil count per 1-OR increase in severe malaria risk.

Table 5-12. *MR analysis between neutrophil count and P. falciparum severe malaria.*

Exposure	Outcome	Method	No. SNPs	BETA	SE	P-value	OR	OR.lci95	OR.uci95
Cerebral malaria	Neutrophil count	Inverse variance weighted	2	0.76	0.57	0.18			
Other severe malaria	Neutrophil count	Inverse variance weighted	3	0.73	0.64	0.25			
Overall severe malaria	Neutrophil count	Inverse variance weighted	7	0.71	0.54	0.19			
Neutrophil count	Cerebral malaria	Inverse variance weighted	3	0.00	0.03	0.98	1.00	0.94	1.06
Neutrophil count	Other severe malaria	Inverse variance weighted	3	0.03	0.03	0.26	1.03	0.98	1.09
Neutrophil count	Overall severe malaria	Inverse variance weighted	3	0.03	0.02	0.24	1.03	0.98	1.07
Neutrophil count	Severe malaria anaemia	Inverse variance weighted	3	0.08	0.04	0.08	1.08	0.99	1.18

A single-SNP MR analysis was done to study the effect each genetic variant on the outcome. For neutrophil count as exposure, SNPs rs2325919 (proxy for rs2814778), rs7460611 (proxy for rs10096834), and rs144109344 were used. There was little evidence of an effect by any single SNP, although the general direction was towards increasing the risk of severe malaria (**Figure 5-19, Appendix 22**). The estimated conditional F-statistic for SNPs rs2325919, rs7460611 and rs144109344 were 182, 16 and 36.

For severe malaria as an exposure, SNPs rs113892119, rs116423146, rs1419114, rs553707144, rs557568961, rs57032711, rs8176751 were used to proxy for overall severe malaria, rs113892119 and rs543034558 for CM, and rs113892119, rs116423146, rs557568961 for OTHER (**Figure 5-20, Appendix 23**). The estimated conditional F-statistic for SNPs rs113892119, rs116423146, rs1419114, rs553707144, rs557568961, rs57032711 and rs8176751 were 96, 32, 30, 38, 119, 32 and 44.

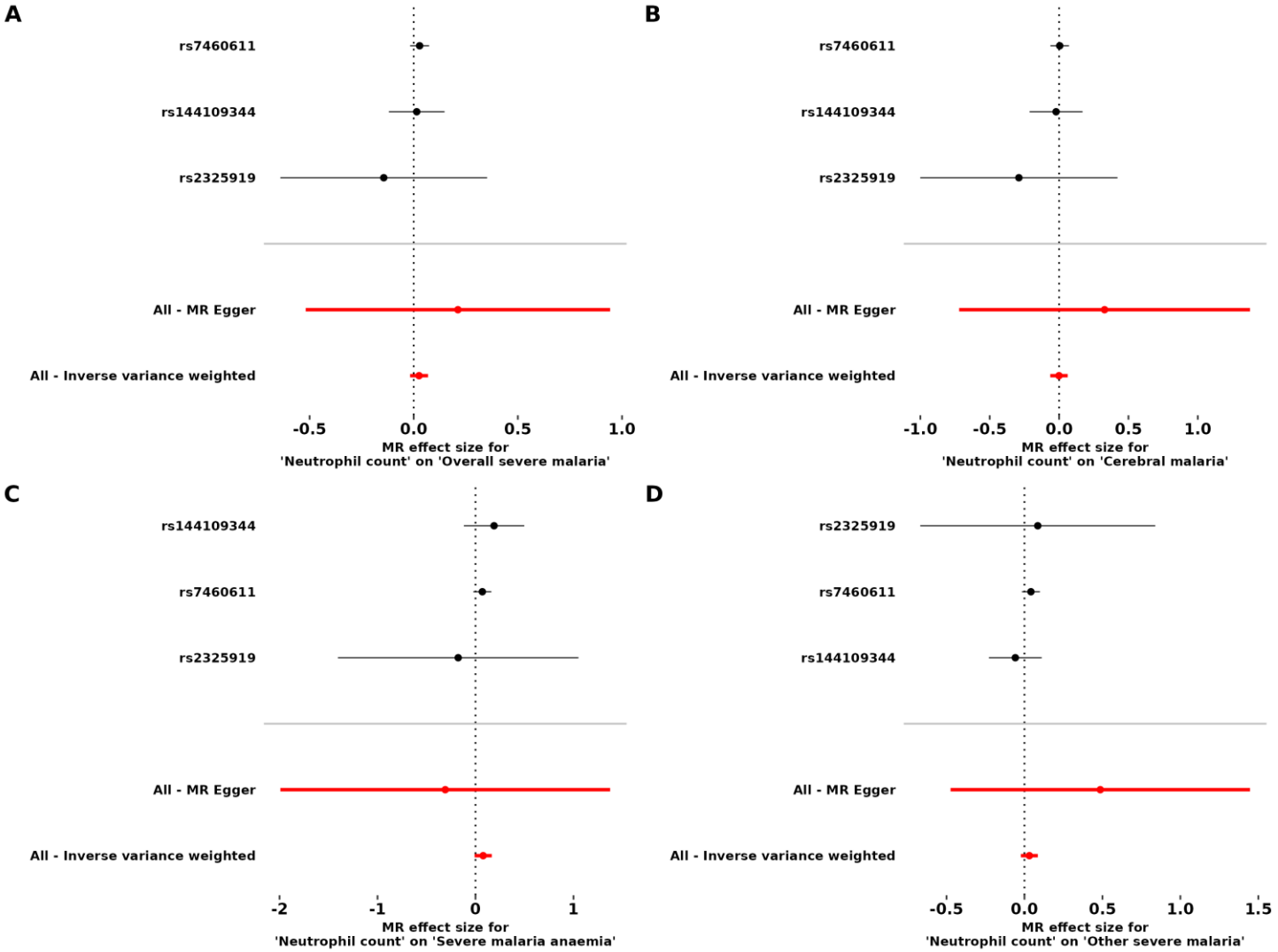


Figure 5-19. *Single-SNP MR analysis of neutrophil count on severe malaria.*

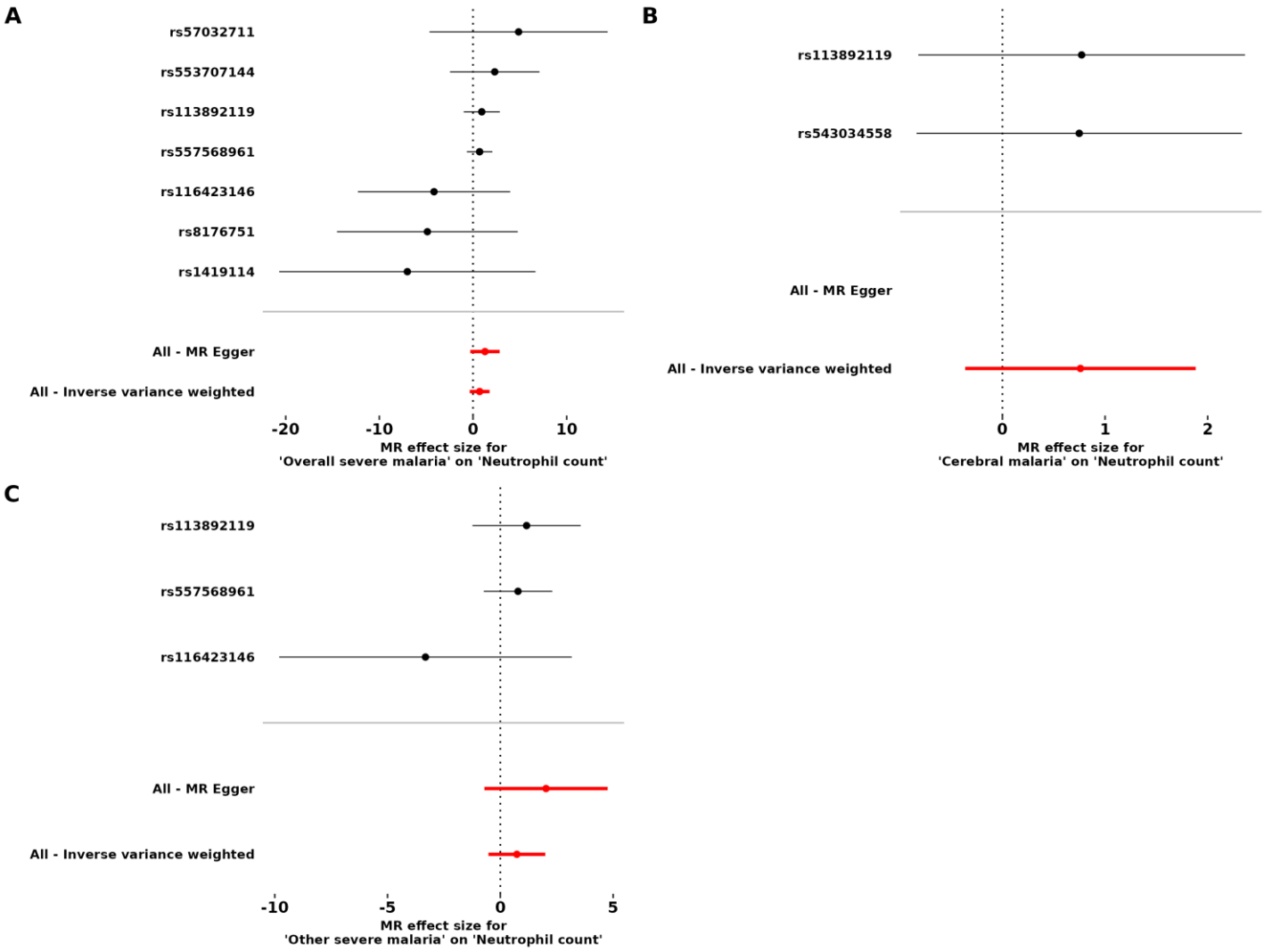


Figure 5-20. *Single-SNP MR analysis of severe malaria on neutrophil count.*

5.4. Discussion

Here, I used the data generated in **Chapter 4** to conduct a GWAS of neutrophil count in individuals from the AFR CAG in UKBB. Seventy-three independent loci were identified, of which nineteen were novel. Ten index SNPs were found using the conservative GCTA-COJO approach, and another two through MR clumping. Moreover, BOLT-LMM was found to be reliable in conducting GWAS on people of African ancestry. Ultimately, this allowed me to run a MR analysis between neutrophil count and *P. falciparum* severe malaria.

The overarching aim of this chapter was to establish if higher neutrophil count is a causal factor in the severity of *P. falciparum* malaria using genetic epidemiology methods, thereby improving our understanding of blood cell traits (BCTs) and disease. I chose this study design to best address the challenges that this posed. First, I ran a GWAS of neutrophil count with BOLT-LMM to overcome the possible issues of population structure and smaller sample-size in the AFR CAG dataset (Aim 1). Afterwards, sensitivity analyses were undertaken at each step of the study to ensure the reliability of the BOLT-LMM GWAS results (Aim 2). This was followed by my initial aim, to run a MR analysis between neutrophil count and SM caused by *P. falciparum* (Aim 3). Finally, I triangulated my findings with those from previous GWAS and performed literature searches to assess the validity and biological significance of my results (Aim 4).

One of the questions I aimed to answer in this study was whether BOLT-LMM would be able to provide reliable results when performing GWAS in people of non-European ancestry, such as those in the UKBB AFR CAG. For example, in their meta-analysis of BCTs in non-European datasets, Chen at al. used a linear model in PLINK to run their GWAS, restricting BOLT-LMM only to the European dataset [166]. Here, the visual comparison in Manhattan plots between my GWAS and Chen at al.'s showed a large degree of overlap. Compared to the META-WD and META-WOD GWAS, the BOLT-LMM approach was more similar with that of Chen et al conducted with a larger sample-size (N=15,171). These findings indicate that BOLT-LMM can be reliably used to conduct GWAS in non-European populations, which could be advantageous in identifying more causal ancestry-specific SNPs for BCTs in future studies, as the power of BOLT-LMM scales with increasing GWAS sample-size [435].

There was a stark contrast between the genetic architecture of neutrophil count in people of African vs. European ancestry [149]. Interestingly, tissue expression for BCTs has been

found to vary between ancestries as well [399], further showing the importance of conducting GWAS in diverse populations to improve the understanding of BCT biology.

Next, I investigated the GCTA-COJO index SNPs. Genetic variants passing the MR clumping parameters were also investigated, as these were used in the MR analysis. Two additional SNPs were identified here, which might have not been found in the more conservative GCTA-COJO run due to conditioning on SNPs across the genome, and these small effects might have pulled their P-values (2.30e-08 and 4.90e-08) below the GWAS significance threshold [561].

The first index SNP was rs28734019 (1p22.2), mapping to the pseudogene *RNU6-695P*. Another SNP close to the same pseudogene is rs10922833. A study using 287 liver samples to identify protein quantitative loci (pQTL) found rs10922833 to be a trans-acting pQTL for Tenascin C, an extracellular matrix protein [564]. Due to their close location, an analysis with LD Link [565] was done to test whether the two variants were in LD. However, this was not the case (r^2 = 8.41e-05, D' = 0.43), making it unlikely for rs28734019 to be regulating Tenascin C levels by proxy through rs10922833.

The next identified index SNP was rs12747038, located on chromosome 1 (1q21.1). As a confirmation of my analysis, Chen et al. and Hu et al. had also identified rs12747038 to be associated with neutrophil count and found a similar effect size (AFR_CAG BETA = -0.22, P-value = 3.90e-09; Chen BETA = -0.31, P-value = 3e-20; Hu BETA = -0.21, P-value = 8e-36) [166,400]. A GTEx search of rs12747038 showed its role as an expression quantitative trait locus (eQTL) i.e. associated with levels of gene expression [566]. Here, the strongest association was with decreasing the expression of *CHD1L* [normalised effect size (NES) = -0.25, P-value = 1.9e-20] in the whole blood. The chromodomain helicase DNA binding protein 1 like (CHD1L) protein is involved in a multitude of biological processes, such as DNA repair, gene transcription and translation [567]. However, its role in blood cell traits has not been explored. As an eQTL, rs12747038 was also associated with increased expression of a nearby upstream pseudogene *NBPF13P* (NES = 0.35, P-value = 9.0e-11) in the whole blood. Additionally, rs12747038 has a role as a splicing QTL (sQTL) i.e. affecting alternative splicing to make different protein isoforms [568], which can be more relevant mechanistically to a phenotype [569]. The strongest association as a sQTL was with *NBPF12* (NES = 0.49, P-value = 2.9e-9) in the thyroid, known as neuroblastoma breakpoint family member 12. McCartney et al. had found that rs11239931, a SNP mapping to the *NBPF13P* pseudogene and sQTL for *NBPF12*, was also associated with a decrease in granulocyte count (BETA = -0.23, P-value = 4e-12) in people of African ancestry (N=6,152) [570]. *NBPF12* is part of the

neuroblastoma breakpoint family, which has been associated with an array of traits, such as autism, psoriasis and various cancers [571]. While previous GWAS have replicated the association between rs12747038 and neutrophil count, the mechanism is unknown.

The rs2814778 (chromosome 1q23.2) index SNP has been the most replicated genetic variant in people of African ancestry known to affect neutrophil count [166,572–577], with the CC genotype (most common in Africans) associated with decreased neutrophil count [508]. The exact location of rs2814778 is inside a promoter upstream of the *ACKR1*/*DARC* (Atypical Chemokine Receptor 1/Duffy Antigen Receptor for Chemokines) gene [503]. The CC genotype inhibits the binding of the GATA transcription factor and therefore *ACKR1* expression, preventing the production of a glycosylated transmembrane receptor [508]. This receptor is predominantly found on erythrocytes and is heavily involved in chemokine signalling, such as CXCL8 and CCL5 [503]. Interestingly, those who suffer from BEN do not have a worse immune response compared to those with the TC or TT genotypes [503]. While many studies have replicated the association of rs2814778, the exact mechanism of BEN is still under investigation, although a proposed mechanism is the reduced differentiation of granulocytes under homeostatic conditions, and a heightened response under stress [503].

The next index SNP was rs183362544, found on chromosome 2 (2q11.2). Several other SNPs associated with WBC count have been mapped to the same *NCAPH* (Non-SMC Condensin I Complex Subunit H) gene: rs10209780 [578], rs111162559 [579], rs34063378 [156] with eosinophil count and rs561539268 [156], rs584811 [580] with monocyte count. Biologically, the NCAPH protein is known to play a role in cell mitosis, DNA repair and regulation of transcription [581]. A study found that increased *NCAPH* expression in lung adenocarcinoma was associated with increased Th2 T-cell infiltration, decreased innate immune cell numbers, and decreased survival [582]. Similarly, a lookup in The Human Protein Atlas [583] showed that *NCAPH* had higher expression in granulocytes and regulatory T-cells, further showing a role in immunity for *NCAPH*. These findings provide further evidence that rs183362544 could have a role in regulating neutrophil count levels. However, there was no data available on how rs183362544 affects gene expression, and its exact genomic location is not inside the *NCAPH* gene. Therefore, while nearby SNPs have also been found to associate with WBC count, it is uncertain if this occurs through the hypothesised mechanism of regulating *NCAPH* expression.

rs144109344 is another index genetic variant on chromosome 2 (2q21.3), and its association was similar to that in the studies of Chen et al. and Soremekun et al. (N=17,802 Africans): AFR_CAG BETA = -0.12, P-value = 3.10e-10; Chen BETA = -0.27,

P-value = 3.39e-14; Soremekun BETA = -0.21, P-value = 2e-13) [166,575]. Similarly, other SNPs mapping to the *DARS*/*CXCR4* (Aspartyl-TRNA Synthetase 1/C-X-C Motif Chemokine Receptor 4) genes have been associated with neutrophil and monocyte count [149,156,166,578,580,584]. Biologically, DARS1 is an enzyme that is part of the multi-tRNA synthetase complex (MSC) [585]. This complex serves many functions, ranging from DNA repair, transcription and translation and immune signalling [585], although DARS itself has not been implicated in regulating blood cell traits. On the other hand, CXCR4 is a chemokine receptor which binds to CXCL12 [586], and is known to regulate the release of neutrophils from the bone marrow during both homeostasis and infections [587]. Interestingly, CXCR4 has been involved in *P. falciparum* pathogenesis. Macrophage migration inhibitory factor (MIF) can interact with CXCR2 and CXCR4 to recruit neutrophils [588], and the Plasmodium falciparum parasite is known to produce MIF (PfMIF) as well [589]. A previous laboratory study using both murine (*P berghei*) and human (*P falciparum*) models found impairment of the parasite liver-cycle in both knocked-out and drug-targeted CXCR4 [590]. Moreover, MIF was found to be released from erythrocytes infected with P falciparum, which together with CXCR4 triggered the recruitment of neutrophils and formation of NETs [591]. PfMIF has been recently identified to possess DNase properties [592], which can interfere with NET formation and activity [591].

The next index SNP is rs557482905 (chromosome 5q14.1) which is inside the *ACOT12* (Acyl-CoA Thioesterase 12) gene. There was no evidence for an association in the Chen et al. GWAS (AFR_CAG BETA = 0.55, P-value = 5.80e-09; Chen BETA = -0.28, P-value = 0.24). The ACOT12 enzyme is predominantly found in the liver and plays a role in cellular metabolism and activity through catalysing the hydrolysis of acetyl-coenzyme A (acetyl-CoA) into CoA and fatty acids [593,594]. A GWAS Catalog query showed that some other SNPs mapping to this gene were associated with body size measurements, like height and BMI [580,595], which makes sense given the role of ACOT12 in metabolism. Interestingly Xing et al. found that rs7735423 (A/G) in *ACOT12* was associated with higher psoriasis rates in a Han Chinese population (N=1,027) [596]. Moreover, a study looking at COVID-19 severity in a transcriptomic analysis (N=66) found that fatty acids were associated with an increase in neutrophil-to-lymphocyte ratio, and *ACOT12* expression increased with disease severity [597]. Nevertheless, the current evidence of *ACOT12* in immunity is limited, and a further replication analysis is needed given the contrast with previous GWAS.

The rs527921556 index SNP is found on chromosome 6q25.3 and is inside the *SLC22A2* (Solute Carrier Family 22 Member 2) gene, encoding the OCT2 (organic cation transporter 2) protein [598]. OCT2 is found in the renal tubule where it was discovered to

play a role in eliminating drugs such as metformin [599]. In terms of the GWAS results, Chen et al. did not find evidence for an association. Moreover, most SNPs mapping to this gene associate with lipoprotein A levels, and no SNPs were identified to influence WBC count [600]. The next index SNP was rs140048432 (chromosome 9p22.2), found inside the *SH3GL2* (SH3 Domain Containing GRB2 Like 2, Endophilin A1) gene, encoding the endophilin 1 protein [601]. Endophilin A1 is found in the brain and has a role in intracellular processes such as tyrosine kinase activation and apoptosis [601]. However, endophilin 1 has not been studied in relation to immunity, and the Chen at al. GWAS did not find an association of this SNP with neutrophil count. rs530475031 is found on chromosome 12q13.11 and is inside the *C12orf54* (Chromosome 12 Open Reading Frame 54) gene. Another study found rs11458 mapping to this gene and was associated with haemoglobin levels [602], although the role of the encoded protein is unknown. rs558204720 (chromosome 16) is an intronic SNP, but no SNPs were found to associate with blood cell traits at the *LOC105371298* gene.

rs10096834 (MR clump SNP, chromosome 8q23.3) is an intergenic SNP, with the closest gene being *TRPS1* (Transcriptional Repressor GATA Binding 1). The zinc finger transcription repressor encoded by this gene plays many roles, such as in embryonal development and chondrocyte cell cycle regulation [603]. However, there is evidence of this transcription factor's role in immunity. A GWAS Catalog search identified other SNPs close to rs10096834 that associate with neutrophil and monocyte count [149,156,166,580]. A further query on The Human Protein Atlas [583] showed that *TRPS1* had the highest expression in monocytes out of all other immune cells. Its expression has been found to regulate the recruitment of tumour-infiltrating lymphocytes at the site of breast cancer cells [604] and another GWAS found an association between rs2049865, mapped to *TRPS1*, and abdominal infections [605]. Moreover, *TRPS1* was found to play a role in Th17 cell differentiation [606]. Interestingly, a study looking at hunter-gatherer African populations identified positive selection of certain SNPs at the *TRPS1* gene, which the authors speculate might be due to immune advantages in the rainforest environment [607]. Six genetic variants (chr8:116702422-116802422) which were fixed in Europeans had non-zero allele frequencies (0.01-0.80) in the three hunter-gatherer populations [607]. However, rs10096834 was not inside this genetic region, and its allele frequency was similar in the AFR CAG sample compared with European populations. Overall, these findings point to a role for *TRPS1* in immunity, but its effect might not be specific to neutrophils.

rs138163369 (MR clump SNP, chromosome 18p11.31) maps to a region of the genome that encodes a lncRNA transcript *CTD-2124B20.2*, and there was no evidence of

association in Chen et al.'s study (AFR_CAG BETA = 0.53, P-value = 4.90e-08; Chen BETA = 0.18, P-value = 0.13). Yang et al. showed that expression of *CTD-2124B20.2* was positively correlated with worse lung adenocarcinoma prognosis, but they did not find an association between its expression levels and WBC count [608]. Further evidence on SNPs mapping to *CTD-2124B20.*2 is limited, most likely due to the fixed alleles in non-African populations [536].

The aim of the description above was to provide a possible biological mechanism through which these SNPs might causally affect the levels of neutrophil count, and therefore aid in the biological interpretation should there be evidence for a causal effect given by the MR analysis. However, mapping SNPs to a functional process i.e. fine-mapping is notoriously difficult [609], and it will be interesting to see what future studies discover on the genetic architecture [347] of neutrophil count in people of African ancestry.

In terms of the sensitivity analyses undertaken after the main GWAS, the combined evidence suggested that the instruments generated from my GWAS could be used in a MR analysis. However, only three SNPs were also present inside the severe malaria dataset and therefore also estimated the single-SNP effects through the reduced-form estimator i.e. Wald ratio method [557].

In the MR analysis, increased neutrophil count showed limited evidence of increasing the risk of SM, with some evidence seen on the SMA sub-phenotype, where rs10096834 (proxied by rs7460611) showed the most evidence for an effect. Band et al. also performed a MR analysis between neutrophil count and *P. falciparum* SM [245]. However, they used SNPs for neutrophil count generated from a GWAS in Europeans from UKBB [149], where they found no evidence of an effect on SM (AFR_CAG BETA = 0.03, P-value = 0.24; Band BETA = 0.00, P-value = 0.87) [245]. When doing the MR in the other direction, there was little evidence for an effect by SM on neutrophil count. The results here are contrary to the expected outcome of increased neutrophil count leading to an increased risk of severe malaria, given the present literature described in the introduction [491,510–514].

5.4.1. Limitations

Nevertheless, my study has certain limitations. Firstly, a possible limitation for the novel genetic variants identified here is Winner's curse [610]. All GWAS use a P-value significance threshold to affirm if there is evidence that a SNP is associated with a trait, which is commonly set at 5e-8 [611,612], as per the recommendations of Risch and Merikangas back in 1996 [613]. However, one consequence of a "significance" threshold in

GWAS is that some SNPs can pass this threshold by chance in the first discovery study, which is then not replicated in subsequent studies [614,615]. Due to the sample-size constraints of the AFR_CAG sample, generating a replication sample was not possible. However, the AFR_CAG summary statistics were compared to those from Chen et al. [166] and showed a good degree of nominal replicability. Nevertheless, future GWAS in sub-Saharan African with large sample-sizes will be able to identify variants with more common alleles and smaller effect sizes, diminishing the possible effect of Winner's curse [614].

Secondly, only a limited number of instruments were available to proxy for neutrophil count in the MR analysis. Seven index SNPs had a very high effect allele count, which might have been fixed in the MalariaGEN study population and so could not be used in the MR analysis. Similarly, the rs2814778 SNP most likely had a very small allele frequency and might have been eliminated, although I was able to use another SNP in LD with it as a proxy. While LD proxies are useful, they can also come with the caveat of not precisely instrumenting the trait [252]. Moreover, the MR-Egger method is not reliable with a small number of instruments [204].

Thirdly, severe manifestations of *P. falciparum* malaria are more common in children and young adults [616,617], and the immune system has been observed to be less effective in terms of neutrophil activity with increasing age [618]. Given that the average age for the AFR_CAG sample was lower than that in sub-Saharan Africa (58.1 vs 16.8 years) [619], the results of the GWAS should be interpreted with this in mind, while those from the MR analysis are likely a violation of the 2SMR assumptions [200].

Finally, the most impactful limitation in this study is the small sample-size and hence statistical power. As mentioned previously, I have chosen to use BOLT-LMM here to best address the issues of a small sample-size and the presence of population structure. Current studies done on people living in sub-Saharan Africa have been small [166,573–575] compared to what is currently being done in regions such as Europe, East Asia and the US [162,164,584]. The approach I have taken in **Chapter 4** and in this chapter has provided a valuable resource and could be pooled together with other GWAS done in people of African ancestry to create a larger dataset. In any case, having a large-scale study like UKBB in sub-Saharan African would be very useful in terms of finding common SNPs with smaller effect sizes that could be used reliably for polygenic risk score generation or MR analyses. Not only is this important from an equity standpoint but is also helpful in understanding the biology of complex traits, such as BCTs, and that can ultimately have a positive impact in the health outcomes of all.

5.4.2. Conclusion

In this study I conducted a GWAS of neutrophil count in people from the UKBB African CAG. I identified several top SNPs that associated with neutrophil count, which allowed me to address the initial aim of running a MR analysis between neutrophil count and SM caused by *P. falciparum*. While the MR results did not display a concrete result, this only shows the importance of conducting large-scale biobank studies in Africa.

Similarly to **Chapter 3**, in this chapter I used specific methodological approaches to study the biological relationship between neutrophil count and severe malaria, a disease of global significance [385]. On the same narrative thread, the work in **Chapter 6** is focused on another important disease linked with BCTs [620], deep vein thrombosis.

CHAPTER 6. PHENOME-WIDE ANALYSIS OF DEEP VEIN THROMBOSIS AETIOLOGY

Chapter summary

In the final results chapter of my thesis I focused on identifying novel causes for deep vein thrombosis (DVT) [249], a disease of global prominence (**Figure 6-1**) [621]. Like the other diseases I have studied, BCTs play a role in the development of the disease, particularly platelets, a subset of BCTs. In this chapter, I used a hypothesis-free Mendelian randomization (MR) [169] approach to discover novel risk factors for DVT. This work aimed to inform on the mechanism through which platelets might lead to DVT development.

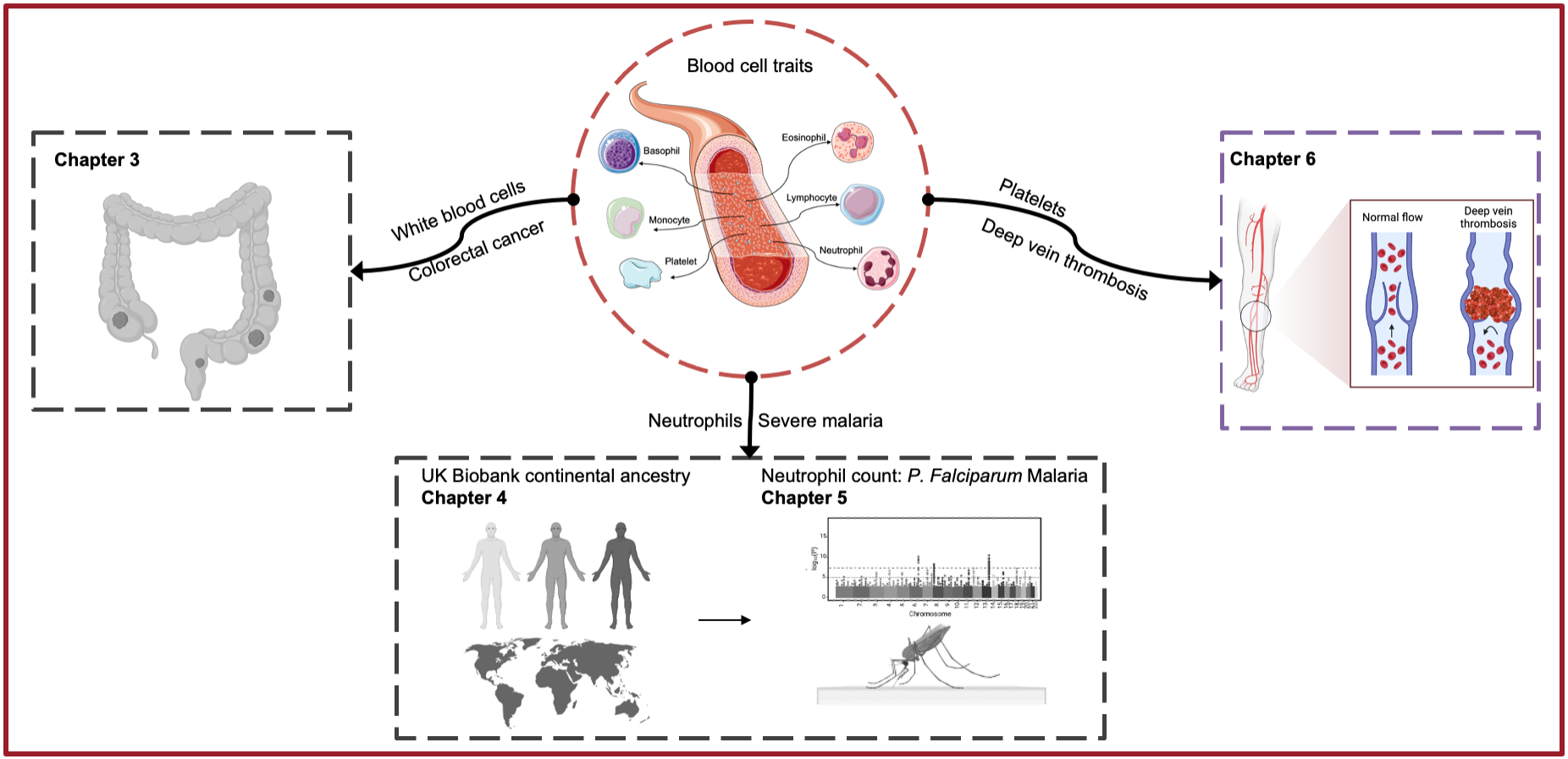


Figure 6-1. *PhD project and current chapter (6 - coloured).*
Created with BioRender.com.

6.1. Introduction

Under normal physiological conditions, platelets and fibrin form clots to prevent blood loss at the site of vessel injury [622]. However, when clots (or thromboses) form abnormally they can disrupt blood flow [623,624]; when this occurs in the deep veins of the limbs or pelvis, this is known as deep vein thrombosis (DVT) (**Figure 6-2**). A complication of DVT is pulmonary embolism (PE), where a clot breaks away from a deep vein wall and becomes lodged in a pulmonary blood vessel, obstructing blood flow to the lungs and causing respiratory dysfunction. In 2021, there were approximately one million incident cases of venous thromboembolism (VTE) in the United states alone [625]. DVT accounts for approximately two-thirds of VTE events and PE is the primary contributor to mortality. While VTE was a primary cause for 10,511 deaths in the UK in 2020 [626], the actual contribution of VTE to annual deaths is estimated to be 2-3 fold higher [627].

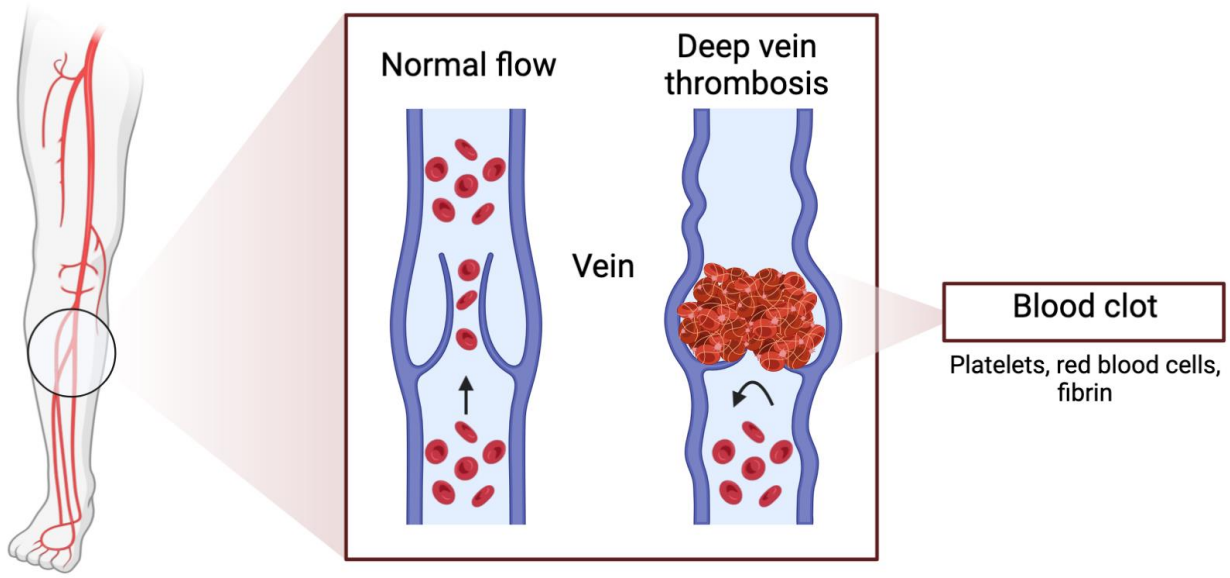


Figure 6-2. *Deep vein thrombosis of the lower leg.*
Made with BioRender.com.

6.1.1. Current treatments for DVT

To prevent acute and chronic complications it is essential to establish an accurate diagnosis of DVT. The symptoms of DVT alone are often not specific or sufficient to make a diagnosis, and about half of those suffering DVT will have no symptoms [628]. Symptoms are considered in conjunction with known risk factors to help estimate the likelihood of

DVT and determine whether thromboprophylaxis is required [624]. Pharmacological thromboprophylaxis includes the use of anticoagulants, such as intravenous heparin and oral warfarin (a vitamin K antagonist), which have been used in combination to treat DVT for over 50 years, but require constant maintenance and monitoring [624]. More recently direct oral anticoagulants (DOAC), such as dabigatran (which inhibits thrombin) or rivaroxaban (which inhibits factor Xa), have been employed with reduced economic costs relative to traditional treatments [629].

6.1.2. DVT aetiology

Environmental/acquired risk factors for DVT include age, obesity, immobility (e.g. hospitalization) and pregnancy [146,623,630,631]. Genetic factors which increase the risk of DVT are those such as deficiencies in the anticoagulation proteins antithrombin, protein C, protein S and Factor V Leiden [623,630,631]. Studies implementing traditional epidemiology methods have identified potential plasma proteins as biomarkers for DVT, including von Willebrand Factor (vWF) and the cell adhesion molecule, P-selectin, both of which are positively associated with DVT and with platelet levels [632]. Other proteins which may be involved in the development of DVT include those that regulate platelet function [622], coagulation factors, as well as proteins secreted upon activation of platelets [633]. Recent technological developments such as the SomaScan by SomaLogic [634] and Olink's proximity extension assay allow the detection of a broad range of plasma proteins, enabling the assessment of a wider range of plasma proteins in DVT risk. Identification of proteins involved in the causality of DVT is important as the majority of pharmacological targets are proteins [635].

Given the potential link between platelets and DVT, studies have also explored the role of platelets on DVT risk in a traditional epidemiological framework. Pana-Noeva et al. conducted a cross-sectional analysis using VTE cases (N=159) and controls (N=140) selected from two prospective studies in Germany [636]. Here, they identified that both platelet count (PLT) and mean platelet volume (MPV) had predictive value for identifying VTE [636]. Similarly, Xiong et al. looking at pre-operative Chinese elderly patients (N=1,391) identified a higher PLT at the time of the hospital visit in those with DVT [637]. In their nested case-control study, Edvardsen et al. studied the relationship between PLT, MPV and vWF levels with incident VTE (predominantly DVT) over an average follow-up to event of 7.5 years in a Norwegian population (cases = 403, controls = 816) [638]. Here, they found increased odds for VTE with increases in both PLT and MPV, as well as from their interaction with vWF, indicating a role for both platelet count and function in DVT/VTE [638].

6.1.3. Current limitations on platelet mechanisms and DVT risk

While there are known risk factors for DVT and platelets are one of the main cells involved in thrombus formation (**Chapter 1**), much is still unknown about the mechanisms of DVT aetiology. Additionally, most observational studies on DVT aetiology have been done with a hypothesis in mind, which might not explore novel traits involved in DVT risk. MR can address the limitations of observational epidemiology such as confounding and reverse causation (**Chapter 2**) [169,219,335,519]. Moreover, as most GWAS summary statistics are publicly available and MR is undertaken on a computer, establishing novel pathways involved in disease development through hypothesis-free investigations becomes an attainable goal [335]. The identification of potential novel risk factors associated with platelets and therefore drug targets is required for improved DVT prophylaxis [624], which is essential given the global burden of the disease. Here, I have employed two-sample MR, which uses data from separate genome-wide association studies (GWAS) for exposures and outcomes of interest (**Chapter 2**) [200] to consider the effect of multiple exposures (phenotypes) on DVT risk.

6.1.4. Main objective

The aim of my study was to find novel risk factors for DVT that are associated with platelet count and function and establish a biological mechanism through which platelets might affect the risk of DVT. I therefore decided to undertake a hypothesis-free MR approach i.e. an MR phenome-wide association study (MR-PheWAS) [185] on DVT risk.

6.1.5. Study aims

I have divided my study into three aims to better address the main objective:

1. Perform a MR-PheWAS on DVT risk
2. Assess the findings and their biological relevance to DVT aetiology
3. Investigate a biological mechanism through which these risk factors could affect the risk of DVT

6.2. Methods

6.2.1. Study design

With the aim to identify novel risk factors for DVT associated with platelets, I performed a MR-PheWAS to estimate the effects of 973 exposures on DVT risk. As 24 of the 57 exposures estimated to influence DVT were adiposity-related, I next decided to investigate potential mediators of this mechanistic relationship further, focusing the mechanistic investigations on circulating proteins altered by adiposity [639,640] and performed a two-sample mediation MR to estimate the effect of BMI on DVT with BMI-associated proteins as mediators. An overview of the study design is shown in **Figure 6-3**. All analyses were conducted using R version 3.6.1. The MR-PheWAS was conducted using the TwoSampleMR R package [219]. STROBE-MR [337] reporting guidelines were followed (**Appendix 24**).

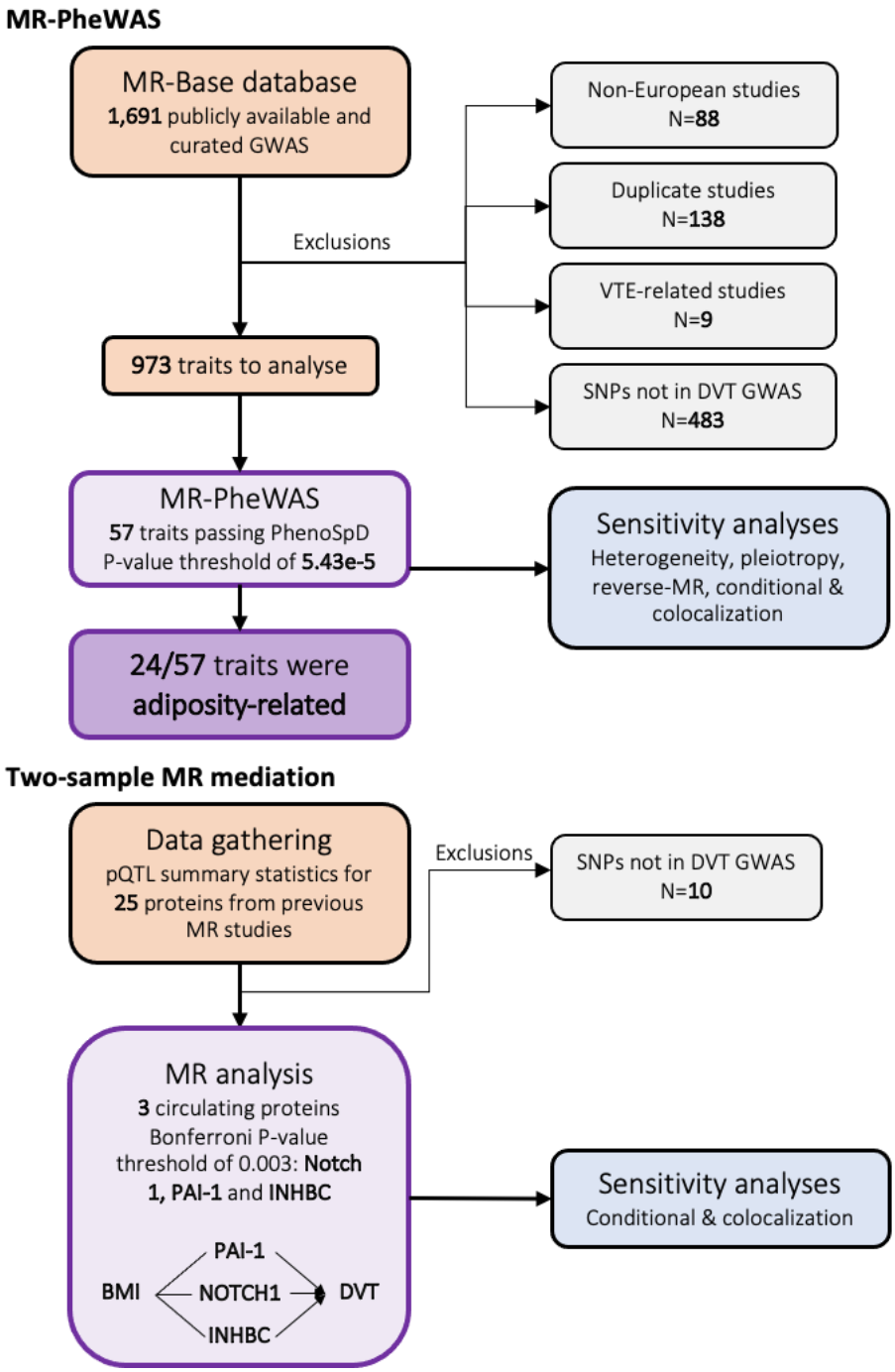


Figure 6-3. *Study design.*

First, an MR-PheWAS analysis to identify risk factors for DVT was done using the MR-Base database. 24 out of 57 traits identified were associated with adiposity, therefore, I followed up the PheWAS with a two-sample mediation MR between BMI-associated pQTL data on DVT risk. MR = mendelian randomization; GWAS = genome-wide association study; VTE = venous thromboembolism; DVT = deep vein thrombosis; SNP = single-nucleotide polymorphism; pQTL = protein quantitative trait loci; PAI-1 = Plasminogen activator inhibitor-1; NOTCH1 = Neurogenic locus notch homolog protein 1; INHBC = Inhibin Subunit Beta C.

6.2.2. Deep vein thrombosis data

The outcome of interest (DVT) was presented in OpenGWAS as “Non-cancer illness code self-reported: deep venous thrombosis (dvt)”; these summary results describe a GWAS of Europeans (6,767 cases and 330,392 controls) performed using the PHEnome Scan ANalysis Tool (PHESANT), followed by genotypic data selected through SNP quality control (QC) [164,641] (http://www.nealelab.is/uk-biobank).

6.2.3. GWAS data for exposures

Genetic data for exposures were obtained from the OpenGWAS database of harmonised GWAS summary data [219]. The exposures encompassed lifestyle, disease and biological traits. The MR-Base package TwoSampleMR R package permits the hypothesis-free analysis of all catalogued exposures to DVT. Non-European (N=88) and duplicate (N=138) studies were excluded. In the case of duplicate studies, those with the highest sample size were retained. VTE (DVT and PE) and VTE-related (e.g. phlebitis and thrombophlebitis) traits were removed (N=9). The genetic instruments used for the analysis were single-nucleotide polymorphisms (SNPs) associated with each of the exposures at a genome-wide level of significance (P<5e-8). As genetic confounding may bias MR estimates if SNPs are correlated [187], linkage disequilibrium (LD) clumping in PLINK [247] was conducted to ensure the SNPs used to instrument exposures were independent [radius = 10,000 kilobases (kb); $r^2 = 0.001$] using the 1000 Genomes (1KG) European reference panel [248]. I also used the 1000 Genomes European dataset [248] to identify potential SNP proxies (with which the initial SNP is in LD with, $r^2>0.8$) for those SNPs absent in the DVT summary statistics. The reported effect size for a given SNP was expressed along with the standard error (SE) in standard deviation (SD) units of the level of the risk factor for a continuous exposure, or as a unit change in the exposure on the log-odds scale for a binary trait.

6.2.4. Protein quantitative trait locus data

I aimed to determine whether BMI-associated proteins were mediating the relationship between adiposity and DVT. A list of BMI-associated proteins was obtained from two previous MR studies investigating the effect of BMI on the circulating proteome [639,640]. I used protein quantitative trait loci (pQTL) data [362,642] to identify SNPs associated with circulating protein levels at a genome wide level of significance ($P \leq 5e\text{-}08$). Protein detection platforms for the pQTL data included the SOMAScan® by SomaLogic and Olink (ProSeek CVD array I) [643–646]. Twenty-five proteins were identified using these criteria (**Appendix 25**). PLINK clumping (radius = 10,000kb; $r^2 = 0.001$) was performed to ensure

the genetic variants used to instrument protein levels were independent. Proxy SNPs for those SNPs that were not present in the DVT data were identified through the 1KG European dataset [248].

Another MR-PheWAS was conducted to establish if thyrotoxicosis affects the levels of circulating proteins from pQTL data (N=3,370) curated withthin OpenGWAS. The pQTL data available to study were gathered from the same consortia described for the BMI-associated proteins [643–646]. PLINK clumping (radius = 10,000kb; r^2 = 0.001) was performed to ensure the genetic variants used to instrument thyrotoxicosis were independent. Proxy SNPs for those SNPs that were not present in the pQTL data were identified through the 1KG European dataset [248].

6.2.5. Data harmonisation

The majority of GWAS present the effects of a SNP on a trait in relation to the allele on the forward strand. However, the allele present on the forward strand can change as reference panels get updated. This requires correction (harmonisation) so that both exposure and outcome data reference the same strand [250]. For exposure and outcome data harmonisation, incorrect but unambiguous alleles were corrected, while ambiguous alleles were removed. In the case of palindromic SNPs (A/T or C/G), allele frequencies were used to solve ambiguities. Harmonisation was not possible for 483 exposures (variants were not present in the DVT GWAS), resulting in a final list of 973 exposures to include in the MR-PheWAS. For my pQTL analysis, 15 out of 25 proteins had genetic variants (including proxies) available in the DVT GWAS (**Appendix 26**). Finally, PhenoSpD was used for multiple testing correction in the MR-PheWAS analysis (P=5.43e-5), while Bonferroni correction was used in the pQTL MR (P= 0.05 / 15 = 0.003).

6.2.6. MR-PheWAS

I conducted hypothesis-free MR-PheWAS using the TwoSampleMR R package [647]. The effect of a given exposure on DVT was estimated using the inverse-variance weighted (IVW) method for exposures with more than one SNP [202], while Wald ratios (WRs) were derived for exposures with a single SNP [189] (**Chapter 2**).

6.2.7. MR sensitivity analyses

Horizontal pleiotropy occurs when a SNP influences the outcome via a pathway other than the exposure of interest, thus violating a key assumption of MR (**Chapter 2**) [191]. MR methods which make differing assumptions regarding pleiotropy were performed as sensitivity analyses where genetic instruments were comprised of more than 3 SNPs: MR-Egger regression, simple mode, weighted mode, and weighted median (**Chapter 2**) [204,205,212,214]. While conventional MR methods assume effect homogeneity, large numbers of genetic instruments associated with an exposure can describe heterogenous effects (e.g. variants associated with BMI may be associated with DVT via a number of alterations to the circulating proteome) [648]. To test for genetic heterogeneity, I used the maximum likelihood [557] estimator and MR-Egger [205] for the exposures which were proxied by 2 or more variants.

6.2.8. Two-sample MR pQTL mediation analysis

In the follow-up MR mediation analysis, I estimated the effect of BMI-associated proteins on DVT using the TwoSampleMR R package [647]. An IVW MR analysis was performed for FABP4, for which 3 SNPs were available to use as instruments. WRs were derived for the remaining proteins. Where proteins were estimated to have a causal effect on DVT, a MR mediation analysis was performed to estimate the proportion mediated by a protein in the BMI-DVT link [218].

The method I used to calculate the proportion mediated by each protein was the product of coefficients method (see **Chapter 2**) [218]. Here, the effect of BMI on protein levels is estimated, after which another MR is done to estimate the effect of protein levels on DVT. These are then multiplied to get the indirect effect, which is then divided by the total effect (BMI to DVT) to estimate the proportion mediated (**Figure 6-4**) [218]. This assumes that the indirect and total effect are in the same direction (both negative or both positive) [218].

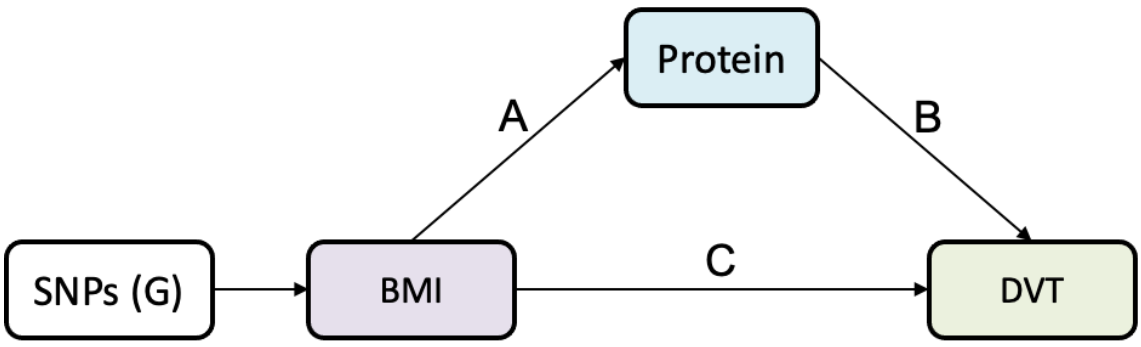


Figure 6-4. *Outline of mediation analysis for BMI-associated proteins.*

*A MR analysis is run between BMI and DVT, providing the total effect estimate (C). Two MR analyses are then conducted: BMI on protein levels (A) and protein levels on DVT (B). The proportion mediated can be estimated through the product of coefficients method A*B / C * 100, assuming A*B and C are both in the same direction.*

6.2.9. Multiple testing correction

As my MR-PheWAS estimated the causal relationship between a large number of exposures and DVT, I used PhenoSpD to estimate the number of independent traits in order to correct for multiple testing [649], which can adjust the P-value significance threshold better than through traditional methods, especially when many traits are being tested. I used GWAS summary data describing the top 1000 associated SNPs for each exposure to create a phenotypic correlation matrix by Pearson correlation. This correlation matrix was used as an input for PhenoSpD to assess the number of independent exposures through matrix spectral decomposition [650,651], generating a P-value threshold of 5.43e-5. For the pQTL analysis, I used a Bonferroni correction, accounting for 15 independent tests (P=0.003), which would yield a similar threshold to an FDR correction given the small number of independent tests.

6.2.10. Beta coefficient transformation

Linear mixed model (LMM) methodology has gained popularity in GWAS due to its ability to control for population structure and deal with large datasets [435]. Regression (beta) coefficients from MR analyses are usually converted to odds ratios (ORs) or risk ratios (RRs) to make results interpretable. However, as GWAS software such as BOLT-LMM still use a linear model (rather than a logistic model) when analysing case-control traits, beta coefficients cannot be calculated directly when using thus must be approximated [435]. Using previously described methodology [652], I approximated logRRs for my MR estimates with the following formula: $\beta / (\mu * (1 - \mu))$, where μ = case fraction.

6.2.11. Bidirectional MR

Where there was evidence of an association with exposures tested in the MR-PheWAS, I performed a bidirectional MR analysis to assess the direction causality between a given exposure and DVT. This was conducted to identify potential pathways of reverse causation, which would invalidate MR assumptions [197].

6.2.12. Colocalization analysis

Only one genetic instrument was available for some of the exposures investigated (N=10). As the Wald ratio estimator is susceptible to genetic confounding, I performed a colocalization analysis for each single-SNP trait. Colocalization analysis uses Bayesian statistics to estimate whether an exposure and outcome share a causal signal in a region of the genome [653], which can then strengthen the evidence that there is a causal relationship. I used the R package "coloc" (https://cran.r-project.org/web/packages/coloc/) approximate Bayes factor (coloc.abf) function with default settings for prior probabilities to conduct a colocalization analysis with the following hypotheses: H0 (no causal variant), H1 (causal variant for trait 1 only), H2 (causal variant for trait 2 only), H3 (two distinct causal variants) and H4 (one common causal variant) [653]. I then used LocusZoom (https://locuszoom.org/) to provide visual evidence for the presence of a shared signal between my exposures and DVT.

6.2.13. Conditional analysis

I performed a conditional analysis for each single-SNP trait using the GCTA-COJO software [523] to identify any potential shared secondary signals in a 1MB region [524], with the aim of performing an additional colocalization analysis on those secondary signals if the primary colocalization analysis did not find a shared causal signal. Secondary signals are SNPs which pass the GWAS significance threshold when a top SNP in its vicinity is conditioned on [524]. I downloaded summary statistics for these traits from OpenGWAS (https://gwas.mrcieu.ac.uk/) [654] and used genotypic data from the Avon Longitudinal Study of Parents and Children (ALSPAC) as a reference panel. Further details of the cohort are described elsewhere [655,656], in brief: 14,541 pregnancies to women with an expected delivery date of April 1, 1991, to December 31, 1992, were enrolled. I used the genotypic data of 8,890 mothers to perform the conditional analysis. Ethical approval for the study was obtained from the ALSPAC Ethics and Law Committee and the Local Research Ethics Committee. The study website contains details of all available data through a fully searchable data dictionary and variable search tool (http://www.bristol.ac.uk/alspac/researchers/our-data/).

6.3. Results

6.3.1. MR-PheWAS

Of the 973 exposures investigated, 945 were identified as independent using PhenoSpD, setting the P-value threshold for my MR analysis at 5.43e-5. Fifty-seven exposures were estimated to influence DVT risk (**Figure 6-5**, **Table 6-1)**.

I observed strong causal evidence for a number of exposures including: "Hyperthyroidism/thyrotoxicosis" (IVW Log RR: 2.39, 95% CI: 1.88 to 2.90; P = 8.69e-18); "Treatment/medication code: carbimazole" (IVW Log RR: 3.60, 95% CI: 2.70 to 4.50, P = 2.41e-12); "Chronic obstructive airways disease/chronic obstructive pulmonary disease (COPD)" (WR Log RR: 3.72, 95% CI: 1.39 to 4.37; P = 9.21e-07); "Varicose veins" (IVW Log RR: 1.90, 95% CI: 1.30 to 2.50; P = 2.36e-07) and "Varicose veins of the lower extremities" (IVW Log RR: 3.40, 95% CI: 2.31 to 4.49; P = 5.13e-07) (**Figure 6-5**).

Adiposity, an established risk factor for DVT [657], and its related traits (N=24, **Table 6-1**) were all positively associated with DVT. These include traits identified in previous MR studies, such as "Body Mass Index" (IVW Log RR: 0.40, 95% CI: 0.32 to 0.47; P = 1.60e-22), fat mass e.g. "Whole body fat mass" (IVW Log RR: 0.44, 95% CI: 0.36 to 0.51; P = 4.65e-27) and fat-free mass e.g. "Whole body fat-free mass" (IVW Log RR: 0.41, 95% CI: 0.31 to 0.50; P = 3.90e-14) [658] (**Figure 6-5**). Another previously-associated trait is "Height" (IVW Log RR: 0.15, 95% CI: 0.08 to 0.21; P = 5.92e-06) [659]. Other associated height-related traits not previously investigated in an MR framework include "Standing height" (IVW Log RR: 0.17, 95% CI: 0.09 to 0.24; P = 4.61e-06) and "Comparative height size at age 10" (IVW Log RR: 0.30, 95% CI: 0.20 to 0.40; P = 1.93e-06) (**Figure 6-5**).

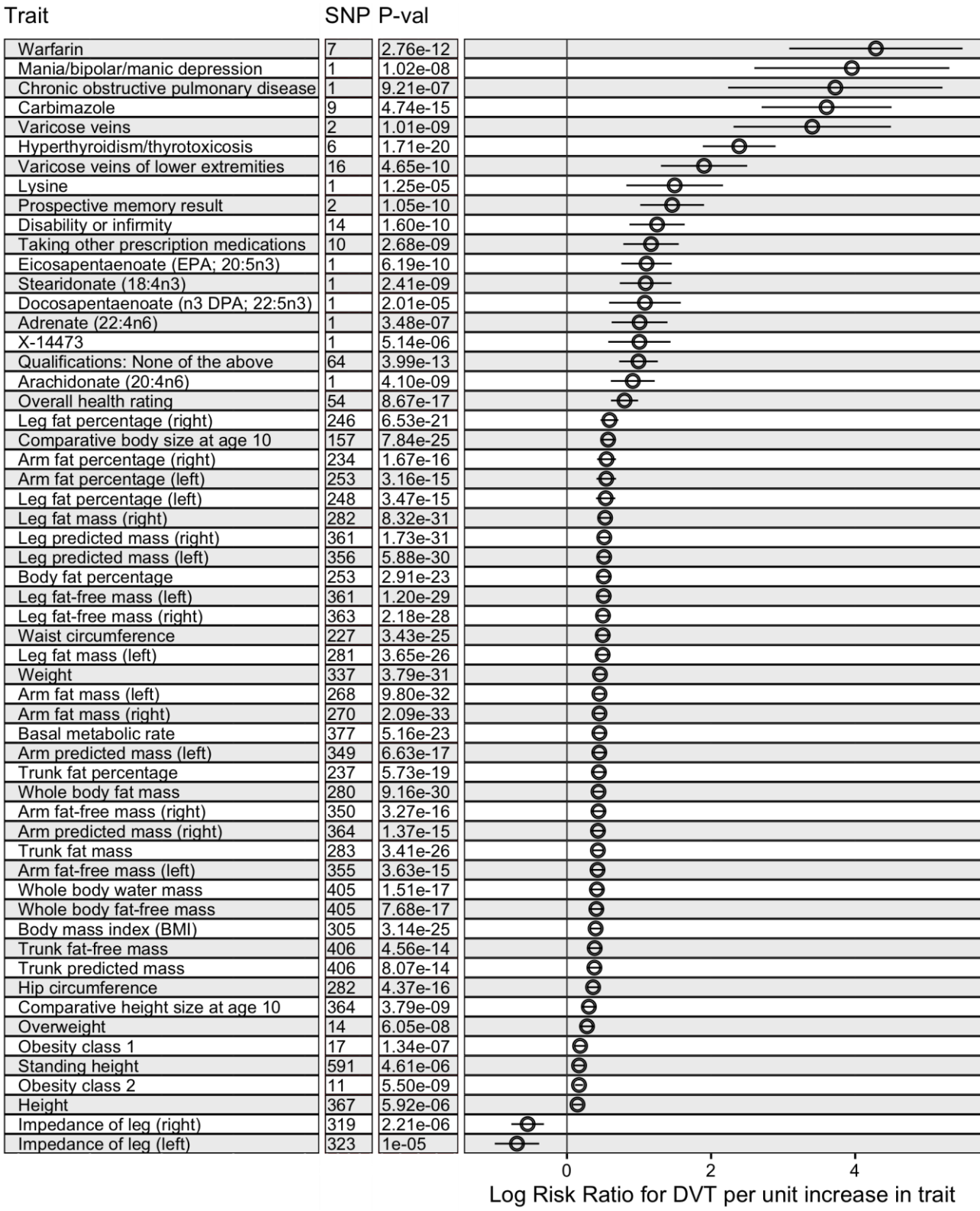


Figure 6-5. *MR-PheWAS results.*

Only traits passing the PhenoSpD significance threshold are shown here. A many-to-one forest plot of the exposures which passed the P-value threshold following multiple testing correction (5.43e-5). Each trait is accompanied by two additional descriptive columns (No. SNPs and P-value), while log risk ratio (RR) is displayed to the right, alongside with the confidence intervals. MR methods: IVW (SNP > 1) and WR (SNP = 1).

Over 50% of the exposures (N=31) which passed my P-value threshold for multiple testing were found to have heterogenous effects between instruments using the maximum likelihood method. Of these, most (N=24) were traits related to body size (mass and adiposity). The remaining heterogenous traits were: “basal metabolic rate” (PHet: 3.71e-03); “warfarin treatment” (PHet: 5.66e-40); “Height” (PHet: 1.58e-03); “Standing height” (PHet = 4.61e-06); “Comparative height size at age 10” (PHet = 1.93e-06); “Impedance of leg (right)” (PHet: 4.23e-06) and “Impedance of leg (left)” (PHet: 9.96e-21). These findings are consistent with my IVW and MR-Egger heterogeneity analyses (**Table 6-1**).

MR-Egger estimates indicated strong evidence of horizontal pleiotropy for “Qualifications: None of the above” (intercept = -5.69e-04, P = 3.35e-02), “Impedance of leg (right)” (intercept = 2.58e-04, P = 3.22e-04) and “Impedance of leg (left)” (intercept = 2.22e-04, P = 7.24e-03) (**Table 6-1**). I was unable to assess whether the “Prospective memory result” trait was pleiotropic, as this exposure was instrumented using only 2 SNPs. In bidirectional MR analyses, DVT was estimated to increase warfarin treatment [“Treatment/medication code: warfarin” (beta = 0.29; SE = 0.02; P = 1.79e-30)], implying reverse causation, and therefore violating MR assumptions (**Appendix 27**).

Table 6-1. *MR-PheWAS results. Adiposity-related traits are coloured in orange.*

Exposure	No. SNP	MR method	Log Risk Ratio*	CI (95%)		SE	P-value	$P_{Het\ (ML)}$	P_{Plt}
Treatment/medication code: warfarin	7	IVW	4.29	3.09	5.49	0.61	1.40E-09	5.66E-40	0.4260
Mania/bipolar disorder/manic depression	1	WR	3.95	2.60	5.30	0.69	5.18E-06	NA	NA
Chronic obstructive airways disease/copd	1	WR	3.72	1.39	4.37	0.76	9.21E-07	NA	NA
Treatment/medication code: carbimazole	9	IVW	3.60	2.70	4.50	0.46	2.41E-12	5.21E-01	0.1048
Varicose veins	2	IVW	3.40	2.31	4.49	0.56	5.13E-07	4.42E-01	NA
Hyperthyroidism/thyrotoxicosis	6	IVW	2.39	1.88	2.90	0.26	8.69E-18	6.69E-01	0.3874
Varicose veins of lower extremities	16	IVW	1.90	1.30	2.50	0.31	2.36E-07	1.91E-01	0.5039
Lysine	1	WR	1.50	0.61	1.96	0.34	1.25E-05	NA	NA
Prospective memory result	2	IVW	1.46	1.02	1.90	0.23	5.33E-08	4.61E-01	NA
Long-standing illness disability or infirmity	14	IVW	1.25	0.87	1.63	0.20	8.13E-08	2.17E-01	0.4463
Taking other prescription medications	10	IVW	1.17	0.79	1.55	0.20	1.36E-06	4.83E-01	0.4399
Eicosapentaenoate (EPA; 20:5n3)	1	WR	1.10	0.75	1.45	0.18	3.14E-07	NA	NA
Stearidonate (18:4n3)	1	WR	1.09	0.73	1.45	0.18	1.22E-06	NA	NA

Exposure	No. SNP	MR method	Log Risk Ratio*	CI (95%)		SE	P-value	$P_{Het\ (ML)}$	P_{Plt}
Docosapentaenoate (n3 DPA; 22:5n3)	1	WR	1.08	0.47	1.46	0.25	2.01E-05	NA	NA
Adrenate (22:4n6)	1	WR	1.01	0.55	1.32	0.20	3.48E-07	NA	NA
X-14473	1	WR	1.01	0.48	1.35	0.22	5.14E-06	NA	NA
Qualifications: None of the above	64	IVW	0.99	0.72	1.26	0.14	2.03E-10	6.18E-01	0.0335
Arachidonate (20:4n6)	1	WR	0.91	0.61	1.22	0.16	2.08E-06	NA	NA
Overall health rating	54	IVW	0.80	0.61	0.99	0.10	4.40E-14	5.14E-01	0.6398
Leg fat percentage (right)	246	IVW	0.59	0.47	0.71	0.06	3.32E-18	2.87E-03	0.2399
Comparative body size at age 10	157	IVW	0.57	0.46	0.68	0.06	3.98E-22	5.18E-01	0.1954
Arm fat percentage (right)	234	IVW	0.55	0.42	0.68	0.07	8.48E-14	8.47E-17	0.6940
Arm fat percentage (left)	253	IVW	0.55	0.41	0.68	0.07	1.61E-12	1.32E-24	0.6983
Leg fat percentage (left)	248	IVW	0.54	0.40	0.67	0.07	1.76E-12	7.00E-04	0.7261
Leg fat mass (right)	282	IVW	0.53	0.44	0.62	0.05	4.23E-28	9.07E-03	0.4978
Leg predicted mass (right)	361	IVW	0.52	0.43	0.60	0.04	8.79E-29	1.34E-02	0.6652
Leg predicted mass (left)	356	IVW	0.52	0.43	0.60	0.05	2.99E-27	5.18E-03	0.8052
Body fat percentage	253	IVW	0.51	0.41	0.61	0.05	1.48E-20	4.79E-02	0.6346
Leg fat-free mass (left)	361	IVW	0.51	0.42	0.60	0.05	6.10E-27	4.73E-03	0.8069
Leg fat-free mass (right)	363	IVW	0.50	0.41	0.59	0.05	1.11E-25	5.05E-03	0.5560
Waist circumference	227	IVW	0.50	0.40	0.59	0.05	1.74E-22	1.65E-02	0.5222

Exposure	No. SNP	MR method	Log Risk Ratio*	CI (95%)		SE	P-value	$P_{Het\ (ML)}$	P_{Plt}
Leg fat mass (left)	281	IVW	0.50	0.40	0.59	0.05	1.85E-23	3.71E-02	0.5530
Weight	337	IVW	0.46	0.38	0.54	0.04	1.93E-28	1.33E-03	0.8573
Arm fat mass (right)	270	IVW	0.45	0.38	0.52	0.04	1.06E-30	3.60E-01	0.2818
Arm fat mass (left)	268	IVW	0.45	0.38	0.53	0.04	4.98E-29	1.93E-01	0.1348
Basal metabolic rate	377	IVW	0.45	0.36	0.54	0.05	2.62E-20	3.71E-03	0.7064
Arm predicted mass (left)	349	IVW	0.45	0.34	0.55	0.05	3.37E-14	1.53E-05	0.2577
Trunk fat percentage	237	IVW	0.44	0.35	0.54	0.05	2.91E-16	2.43E-03	0.6180
Whole body fat mass	280	IVW	0.44	0.36	0.51	0.04	4.65E-27	1.75E-01	0.1772
Arm fat-free mass (right)	350	IVW	0.44	0.33	0.54	0.05	1.66E-13	2.95E-04	0.2180
Arm predicted mass (right)	364	IVW	0.43	0.32	0.54	0.05	6.96E-13	9.35E-05	0.2660
Trunk fat mass	283	IVW	0.43	0.35	0.51	0.04	1.73E-23	2.90E-03	0.6360
Arm fat-free mass (left)	355	IVW	0.42	0.32	0.53	0.05	1.84E-12	3.14E-05	0.1920
Whole body water mass	405	IVW	0.42	0.32	0.51	0.05	7.67E-15	1.32E-04	0.3436
Whole body fat-free mass	405	IVW	0.41	0.31	0.50	0.05	3.90E-14	2.06E-04	0.3422
Body mass index (BMI)	305	IVW	0.40	0.32	0.47	0.04	1.60E-22	6.81E-02	0.5286
Trunk fat-free mass	406	IVW	0.39	0.29	0.48	0.05	2.32E-11	2.46E-06	0.0575
Trunk predicted mass	406	IVW	0.38	0.28	0.48	0.05	4.10E-11	9.09E-06	0.0513
Hip circumference	282	IVW	0.36	0.28	0.45	0.04	2.22E-13	2.92E-04	0.0876

Exposure	No. SNP	MR method	Log Risk Ratio*	CI (95%)		SE	P-value	$P_{Het\ (ML)}$	P_{Plt}
Comparative height size at age 10	364	IVW	0.30	0.20	0.40	0.05	1.93E-06	1.56E-05	0.1080
Overweight	14	IVW	0.28	0.18	0.38	0.05	3.07E-05	3.44E-01	0.1711
Obesity class 1	17	IVW	0.18	0.11	0.25	0.03	1.34E-07	7.33E-01	0.2392
Standing height	591	IVW	0.17	0.09	0.24	0.04	4.61E-06	3.14E-05	0.1018
Obesity class 2	11	IVW	0.17	0.11	0.22	0.03	2.79E-06	5.45E-01	0.6859
Height	367	IVW	0.15	0.08	0.21	0.03	5.92E-06	1.58E-03	0.3372
Impedance of leg (right)	319	IVW	-0.55	-0.80	-0.35	0.12	2.21E-06	4.23E-06	0.0003
Impedance of leg (left)	323	IVW	-0.69	-1.05	-0.43	0.16	1.00E-05	9.96E-21	0.0072

*Methods: Inverse Variance Weighted (SNP > 1) and Wald Ratio
(SNP = 1).

*LogRiskRatio is the logged value of the beta coefficient of the MR analysis into risk ratios. It can be read as an increase in the LogRisk of DVT per unit increase in trait.

*PHET ML is the P-value of the Maximum Likelihood analysis looking at heterogeneity between genetic variants used to instrument a trait. H0 is that there is no heterogeneity present.

*PPlt is the P-value of the MR-Egger analysis looking at the presence of horizontal pleiotropy. H0 is that there is no pleiotropy present.

6.3.2. Blood cell traits and DVT risk

A major motivation for exploring risk factors for DVT were the potential identification of a causal relationship between BCTs and DVT. There were two platelet traits that were part of the MR-PheWAS: platelet count (Log RR: -1.90; P = 1.00) and Mean Platelet Volume (Log RR: 0.001; P = 1.00), neither which showed evidence for a causal effect on DVT risk. This was unexpected, given the observational findings described in the introduction. Several explanations exist for this, one being an incomplete repertoire of SNPs available to proxy for platelet traits. Another explanation is that the SNPs proxying for platelet traits might have positive and negative effects on DVT risk, which cancel out when running the MR analysis. Additionally, the number of GWAS summary statistics in OpenGWAS might have been a limiting factor at the time of running the analysis. Indeed, there were only instruments available for platelet count and mean platelet volume when conducting the MR-PheWAS.

6.3.3. Estimated effects of BMI-driven proteins on DVT risk

Of the 57 traits estimated to increase risk of DVT, 24 were adiposity related (**Table 6-1**). While adiposity is an established risk factor for DVT, the biological mechanisms underlying the effect of adiposity on DVT are not well understood. I therefore used a two-sample MR mediation analysis to test whether altered levels of 15 circulating blood proteins, driven by adiposity, are responsible for this association (**Appendix 26, Appendix 28**). Blood-circulating proteins were investigated as they have the potential to alter platelet activity or act as a component of the platelet clotting cascade. Two recent MR studies have demonstrated that BMI causally affects the levels of 15 circulating proteins [639,640]. Three of these proteins were estimated to influence DVT risk: Neurogenic locus notch homolog protein 1 (NOTCH1; WR Log RR: 0.57, 95% CI: 0.45 to 0.68; P = 1.12e-23), Plasminogen activator inhibitor-1 (PAI-1; WR Log RR: 0.42, 95% CI: 0.30 to 0.54; P = 4.27e-12) and Inhibin beta C chain (INHBC; WR Log RR: -1.18, 95% CI: -2.18 to -0.69; P = 0.002), all three associated with platelet function. Mediation analysis was performed for PAI-1 (the only protein where BMI-protein and protein-DVT effect estimates were consistent in directionality): the proportion of the BMI-DVT effect mediated by PAI-1 was estimated to be 18.56% (**Figure 6-6, Table 6-2**).

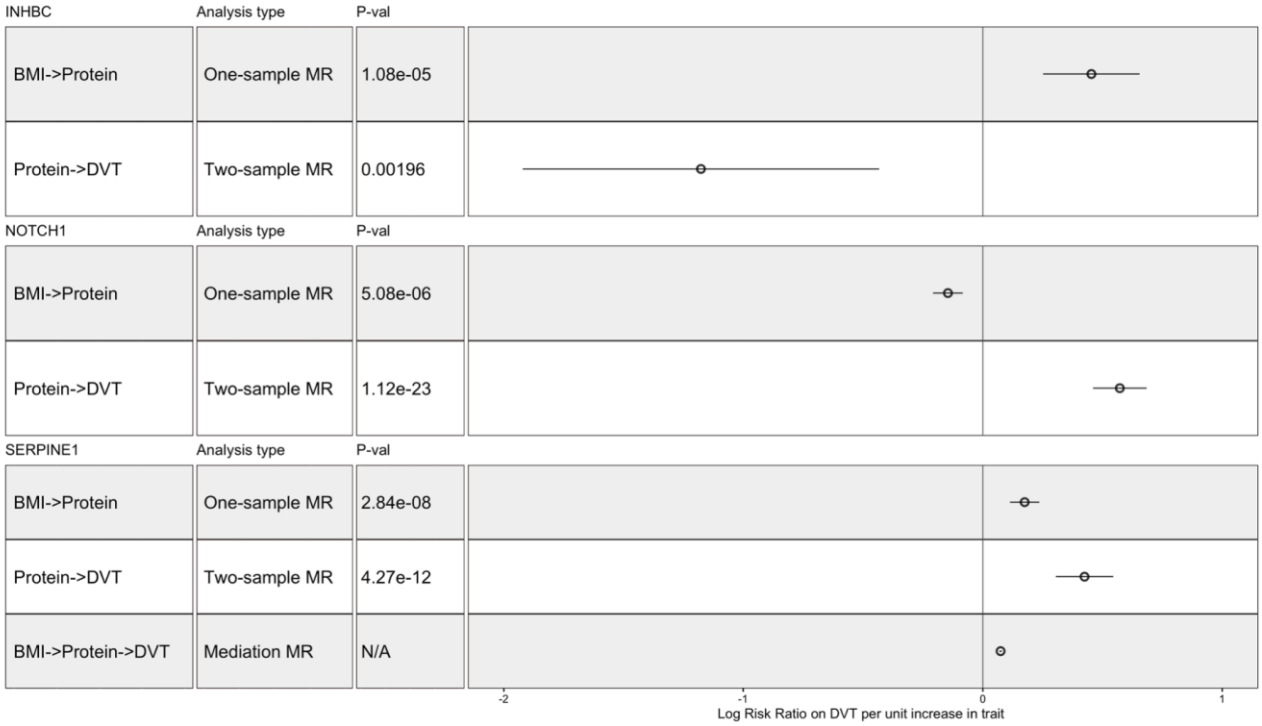


Figure 6-6. *A many-to-one forest plot of the three BMI-associated proteins which passed the multiple-testing corrected P-value threshold (0.003) in the MR analysis.*

Each protein is accompanied by two additional descriptive columns (type of analysis conducted and P-value), while the effect is displayed to the right, alongside with the confidence intervals (Beta coefficient/Log RR ± 95% CI).

Table 6-2. *pQTL MR mediation analysis.*

Exposure	MR method	Log Risk Ratio*	CI (95%)		P-value	Beta coefficient - BMI to protein*	Proportion (%) mediated by protein
Neurogenic locus notch homolog protein 1	Wald ratio	0.57	0.45	0.68	1.12E-23	-0.15	Effect not consistent
Plasminogen activator inhibitor 1	Wald ratio	0.42	0.30	0.54	4.27E-12	0.17	18.56
Inhibin beta C chain	Wald ratio	-1.18	-2.18	-0.69	1.96E-03	0.45	Effect not consistent

LogRiskRatio is the logged value of the beta coefficient of the MR analysis into risk ratios. It can be read as an increase in the LogRisk of DVT per increase in cirulating protein levels.

*BMI-Protein MR effect estimates from Goudswaard et al (https://doi.org/10.1038/s41366-021-00896-1) and Zaghlool et al (https://doi.org/10.1038/s41467-021-21542-4)

6.3.4. Conditional and colocalization analyses

Several of the exposures in my MR analyses could be instrumented using only one genetic variant, and therefore required a conditional and colocalization analysis to provide additional evidence of causality. There were no secondary signals after conditioning on the top SNP for each exposure-DVT pair. There was evidence of a shared causal variant for PAI-1 (PP.S = 97.5%), strengthening the evidence that there is a true causal relationship between the levels of this protein and DVT (**Figure 6-7**, **Appendix 29**).

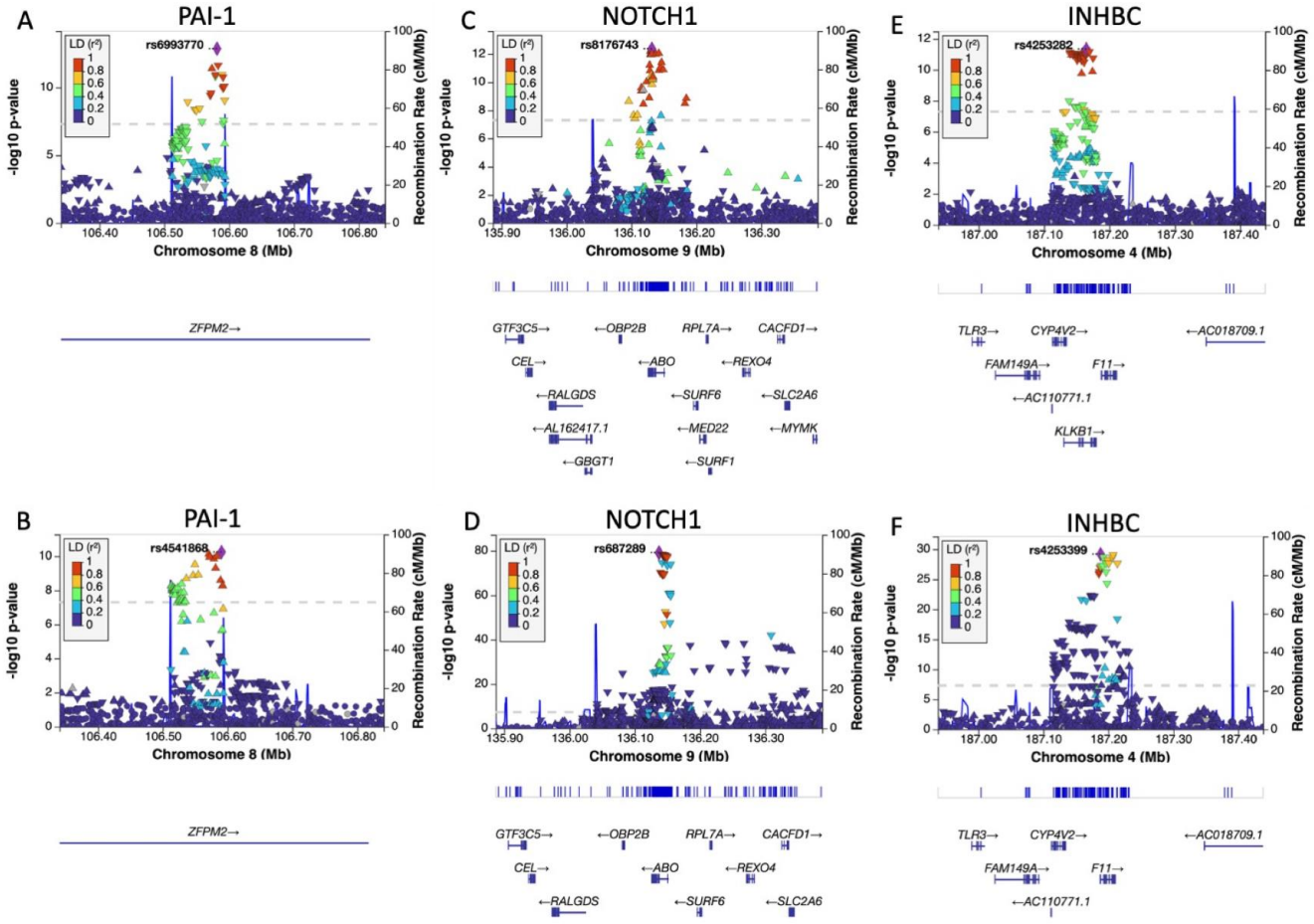


Figure 6-7. *LocusZoom plots of pQTLs with evidence of an effect on DVT risk.*

LocusZoom plots in a 1Mb region of the SNP used to instrument for each protein in both exposure (A,C,E) and outcome (DVT: B,D,F) data: PAI-1 (A,B), NOTCH1 (C,D), INHBC (E,F). The top signal in the region is labelled in each figure. The x-axis represents the position within the chromosome, while the y-axis is the -log$_{10}$ of the P-value. Each dot is a SNP, and the colours indicate how much LD there is between the reference SNP and the other genetic variants.

6.3.5. Enrichment analysis of MR-PheWAS traits

While the results from the analysis of BMI-associated proteins on DVT highlighted a valuable result, I wanted to assess whether the fact that half of the traits identified in the MR-PheWAS were adiposity-related due to the “Anthropometric” category being more common in the initial dataset with 973 traits. In addition to the "Anthropometric” category (which included adiposity-related traits only), the initial dataset also included the following categories: “Fatty acid”, “Health”, “Cardiovascular”, “Anthropometric-impedance”, “Anthropometric-fat-free”, “Psychiatric / neurological”, “Lung”, “Anthropometric-height”, “Unknown metabolite”, “Hormone”, “Amino acid”, “Medication”, “Intelligence”, “Education”, “Energy” (**Appendix 30**).

To do this, I performed an enrichment analysis between the initial dataset with 973 traits and the results dataset with 58 traits. First, I calculated the expected and observed counts for each category within the initial dataset and results dataset, respectively. Following this, for categories with five or more counts, I conducted a Chi-Squared test, while for those with less than five I ran a Fisher’s Exact test, to determine whether a category was under- or over-represented compared to its frequency in the initial dataset with 973 traits (**Appendix 30**).

After applying a Bonferroni multiple testing correction, there was evidence that the “Anthropometric” and “Anthropometric-fat-free” categories were over-represented in the results dataset of 58 traits (**Figure 6-8**).

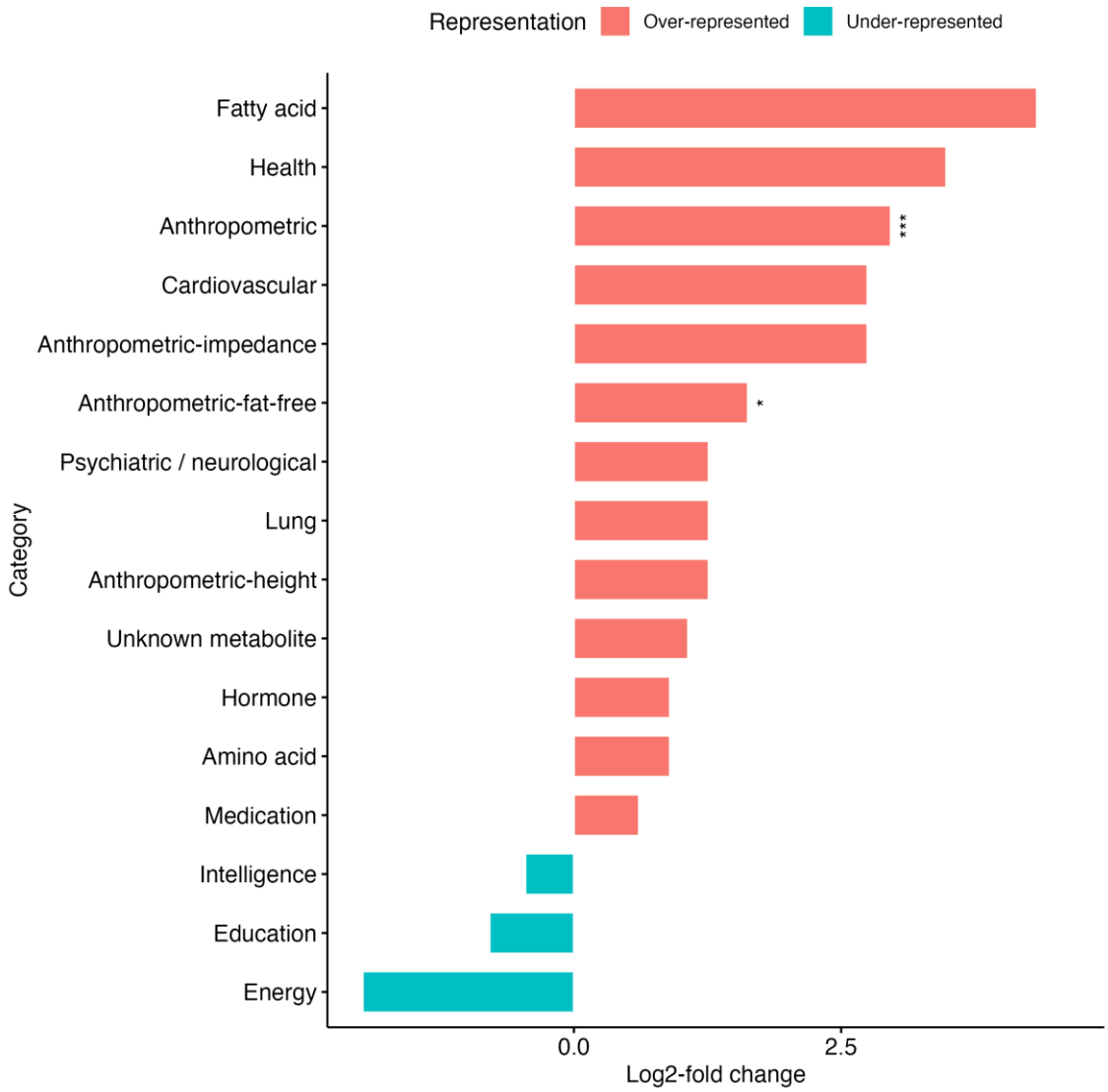


Figure 6-8. *Enrichment analysis of MR-PheWAS categories.*
The counts for each category were compared between the initial 973-trait dataset and the 58-trait results dataset. The y-axis indicates each studied category, while the x-axis represents the log2-fold change, which is the log2 of the observed counts divided by the expected counts. (P < 0.05, ** P < 0.01, *** P < 0.001)*

6.3.6. Thyrotoxicosis and DVT risk

Nevertheless, several thyroid-related traits were also present in the MR-PheWAS results: hyperthyroidism, Carbimazole treatment, and thyrotoxicosis. I decided to explore the relationship between thyrotoxicosis and DVT in the context of finding mediating proteins in a similar maner to what was done in the previous analysis for BMI and DVT.

There were no studies which had MR data available for the effect of thyrotoxicosis on blood proteins levels. Therefore, I undertook a follow-up MR-PheWAS on the blood

proteome to identify if thyrotoxicosis had a causal effect on the levels of circulating proteins using data from OpenGWAS (N=3,370). After adjusting the P-value threshold for multiple testing even with a less conservative False Discovery Rate correction, there was no evidence of an effect by thyrotoxicosis on any of the proteins available in OpenGWAS (**Appendix 31**, database version Aug 2023). While this can be a consequence of thryrotoxicosis having no actual effect on the blood proteome, it can also be that the limited number of instruments for thyrotoxicosis (N=13) made it less likely for an effect to be detected.

6.4. Discussion

With the aim to identify novel causal risk factors for DVT, I performed a hypothesis-free MR-PheWAS of 973 exposures to DVT, of which 57 passed a conservative P-value threshold for evidence of causality. I confirmed causality for several previously established risk factors for DVT (such as BMI and height) and have identified several novel putative causal risk factors (such as hyperthyroidism and varicose veins). Of the 57 exposures estimated to influence DVT risk, 24 were adiposity-related traits. Therefore, I investigated whether the impact of adiposity on DVT is mediated by circulating proteins known to be altered by BMI [639,640]. Here, I provide novel evidence that the platelet-associated circulating protein, PAI-1 has a causal role in DVT aetiology and is involved in mediating the BMI-DVT relationship.

The MR-PheWAS approach suggested a number of traits that could affect DVT development. One of these is standing height, which has been previously associated with increased DVT risk [660] and my results align with this finding. With increased height, a greater volume of blood is required which can increase the stress on blood vessels, disrupting haemostasis [660]. Fat-free mass was also estimated to increase risk of DVT in my study. While counterintuitive, this effect could be mediated through height, as taller people usually have more fat-free mass [658,659]. As expected, many body size related traits showed evidence of heterogeneity, likely due to the large number of SNPs used to instrument these traits and the many underlying biological pathways explaining variation in adiposity.

Venous blood stasis caused by immobility is also a known risk factor for DVT [624]. Here, I report evidence that long standing illness, disability, or infirmity increases DVT risk. A proposed mechanism is stasis of blood flow in the veins which can be either due to a particular neurological condition or due to the paralysis of the lower limbs [661].

My study also provides evidence for novel DVT risk factors. Hyperthyroidism has previously been proposed to contribute to DVT, as indicated by a recent systematic review and meta-analysis of cohort studies showing association with DVT (RR: 1.33, 95% CI: 1.28 to 1.39; I^2 = 14%) [662]. In the present study, I provide novel evidence for a causal effect of hyperthyroidism/thyrotoxicosis on DVT risk (IVW RR: 10.91, 95% CI: 3.97 to 18.17; P = 3.14e-25). Hyperthyroidism has been found to enhance platelet function [663], although the underlying mechanism is not fully understood. It may involve thyroid hormones (THs) promoting a hypercoagulable state and venous thrombi formation, by increasing plasma concentration of factor VIII, fibrinogen, PAI-1 and vWF [664]. TH T4 may also directly enhance platelet function through integrin $\alpha_v\beta_3$ [665]. In addition, THs enhance basal metabolic rate (BMR) and thermogenesis, both of which affect body weight. Indeed, I found that an increase in basal metabolic rate is associated with DVT. While a higher BMR should lead to lower BMI and thus lower DVT risk, it is likely that my results may be explained by the hyperthyroidism-associated mechanisms outlined above.

My MR estimates also support evidence of a causal association between varicose veins and increased risk of DVT. Varicose veins can result in the inability of the blood to fully return to the heart, leading to the enlargement of the veins, and in time, potentially an increased risk of DVT due to stasis [666]. Varicose veins have been outlined as a possible risk factor in general practice patients in Germany [667], as well as in a Chinese retrospective study of over 100K people [666].

COPD was also associated with an increased risk of DVT. COPD is a severe chronic respiratory disease, having been studied extensively for its role in PE [668]. Indeed, both PE and DVT are more prevalent and underdiagnosed in people with COPD [669]. My colocalization analysis did not provide evidence that would support my MR estimates. Moreover, as the SNP used to proxy for COPD (rs9579496) is intergenic i.e. in-between genes, I was unable to compare my results with any locus-specific experimental studies.

Afterwards, as adiposity is an established risk factor for DVT, the estimates I observe between adiposity-related traits and DVT most likely reflect true causal relationships. The estimate I report here for BMI (RR: 1.49, 95% CI: 1.38 to 1.60; P = 3.14e-25) is consistent with a previous MR study conducted in individuals of Danish descent (OR: 1.57, 95% CI: 1.08 to 1.97; P = 3e-03) [631]. In addition, my results are in agreement with the estimated effect of BMI on VTE in the FinnGen consortium (MR OR: 1.58, 95% CI: 1.28 to 1.95; P = 2.00e-05) [658]. Higher adiposity is associated with dysregulated metabolism, which is

one factor that can promote a hypercoagulable state and impair venous return, increasing the chance of thrombi formation [670]. Given that 42% of the traits I found to be associated with DVT were adiposity-related, and that previously I and others found that adiposity is associated with changes to the circulating proteome [639,640], I hypothesised that adiposity-driven changes to the circulating proteome may promote DVT. BMI-driven candidates include proteins that can modulate coagulation (anti-thrombin III, PAI-1) [671,672], platelet function (PAI-1, adiponectin, IGFBP/IGF) [671–673] and/or thrombosis (galectin-3) [674].

Using my MR approach, I was able to estimate the effect of 15 BMI-driven circulating proteins on DVT risk. My analyses suggest a causal role for 3 of these proteins (NOTCH1, PAI-1 and INHBC). Given the established role of some of the circulating proteins in coagulation and thrombosis, the lack of evidence for an estimated effect is surprising e.g. anti-thrombin III [671]. This could represent a true result or my limited ability to instrument circulating proteins using single SNPs.

PAI-1 was the only protein for which evidence was directionally consistent with mediation of the BMI-DVT relationship (circulating levels of PAI-1 were positively associated with BMI and with DVT). This adds to the evidence that platelet traits are involved in DVT, given that 90% of PAI-1 is present in platelets [675]. A study using data from the Million Veterans Program to identify novel VTE risk factors has also confirmed colocalization with DVT for the same PAI-1 SNP (rs6993770, *ZFPM2* locus) used in my analysis [676]. Klarin et al. previously identified in their MR analysis that rs4602861 (*ZFPM2* locus) increased the risk of VTE (OR: 1.08, CI: 1.03-1.15) [677], which is in LD with the PAI-1 SNP in my study ($R^2 = 0.93$). On top of replicating this previous finding, I also showed that this locus increases DVT risk through regulating PAI-1 levels. Moreover, PAI-1 has been associated with an increase in VEGF levels [678–680], which was found to increase the risk of VTE in a previous MR study [681], further adding to the evidence that PAI-1 is involved in DVT development. A follow-up analysis in a murine model found that PAI-1-overexpressing mice had 1.5-fold larger thrombus size compared to PAI-$1^{-/-}$ mice [676].

Moreover, a recent observational study done in inhabitants of Tromsø, Norway (cases = 383, controls = 782) found that PAI-1 increased the risk of future VTE, and that PAI-1 mediated ~15% of the obesity-VTE relationship [682], a number comparable to my MR estimate (18.6%). These results are consistent with the known role for PAI-1 in inhibiting fibrinolysis (breakdown of a clot) [683]. In addition, PAI-1 expression has been previously found to be associated with DVT formation in mice [683] and in humans after total hip arthroplasty [672]. PAI-1 overexpression is enhanced in visceral fat tissue [684], and while

waist-to-hip ratio (WHR) is highly correlated with visceral fat [685], I did not find evidence of an effect of WHR on DVT. Finally, there has been extensive research into PAI-1 drug targets, ranging from synthetic peptides, RNA aptamers to monoclonal antibodies [686]. Rosuvastatin, an HMG-CoA reductase inhibitor, has been found to inhibit PAI-1 *in vitro* [687]. Randomised clinical trials using rosuvastatin have confirmed that it reduced occurrence of symptomatic venous thromboembolism [688] and increased plasma fibrinolytic potential [689], supporting a role for statins in VTE treatment and prevention, possibly via altered PAI-1.

Although I found evidence for a role of INHBC and NOTCH1 in DVT risk, estimates were inconsistent with mediation of the BMI-DVT relationship. I found that circulating INHBC levels were negatively associated with DVT, suggesting circulating levels of INHBC may have a protective effect. Inhibins are part of the growth and differentiation superfamily of transforming growth factor beta (TGF-β) [690] and play a role in inhibiting the levels of follicle-stimulating hormone (FSH) produced by the pituitary gland [691]. Although I did not find evidence of causality between FSH and DVT, a recent study showed that FSH can enhance thrombin generation [692]. This discrepancy could be due to INHBC acting through a different pathway compared to FSH. With regards to NOTCH1, I found that higher expression was associated with an increased risk of DVT. NOTCH1 plays a role in responses to microenvironmental conditions, vascular development and is a shear stress and flow sensor in the vasculature [693]. Interestingly, a recent study found that NOTCH1 and its ligand Delta-like ligand 4 (DLL-4) are present on platelets and are involved in their activation and thrombus formation [694]. While NOTCH targeting has not been done in relation to VTE, current small molecular drugs such as Crenigacestat [695] and targeting antibodies such as Brontictuzumab [696] are being used in clinical trials to inhibit NOTCH signalling for the treatment of T-cell acute lymphoblastic leukaemia and solid tumours, respectively [697].

Finally, I undertook a follow-up blood proteome MR-PheWAS analysis to assess if thyrotoxicosis had an effect on the levels of circulating proteins, aiming to then perform a similar analysis to that of the adiposity-associated pQTL approach in the previous section. Unfortunately, there was no evidence that thyrotoxicosis had an effect on any of the 3,370 proteins in OpenGWAS, even with a more relaxed Benjamini-Hochberg correction, making it not possible to go forward with a two-step mediation MR analysis.

6.4.1. Limitations

There are some limitations to my approach. Firstly, although the number of traits in MR-Base is large and continues to grow, and the approach was undertaken in a hypothesis-free manner, I was limited by the traits available in the platform at the time of the analysis. In addition, the availability of genetic instruments for some traits within the platform are limited, meaning a false null finding could be reported. Moreover, some of the exposures did not have a SNP or proxy present in the outcome (DVT) dataset, making it infeasible to perform MR analysis. Finally, I have chosen to investigate risk factors for DVT as opposed to PE (which is observed in about 40% of DVT cases [698]) to increase the power to detect causal risk factors for DVT. Future analyses could focus on PE specifically to identify predictive risk factors for this outcome.

6.4.2. Conclusion

In summary, I have confirmed estimates of previously identified traits on DVT (e.g. adiposity-related, height), and identified novel risk factors that could act through platelet activity, such as hyperthyroidism. I also provide evidence that the relationship between adiposity and DVT is mediated by dysregulated levels of circulating proteins associated with platelet count and function, such as PAI-1. These findings improve the understanding of DVT aetiology and have notable clinical significance regarding platelet traits.

This chapter marks the end of my thesis results. In the next chapter, the discussion, I zoom-out and describe my findings in the broader context of the current literature. I also provide my opinion on topics such as diverse ancestry GWAS and the future of MR in BCTs and disease.

CHAPTER 7. DISCUSSION

Chapter summary

In this chapter I provide a brief recap of the thesis results. Afterwards, I relate my findings back into the broader context of literature to explore the contribution, potential impact, and limitations of my work. Finally, I end with a personal note on where I see the field of genetics and Mendelian randomization (MR) going forward.

7.1. Synthesis of the findings

Moving back to the overarching objective from **Chapter 1**, I set out to explore how the use of genetic proxies for blood cell traits (BCTs) can be used to expand the current knowledge on BCTs and disease through MR [169]. While randomised controlled trials (RCTs) would still be the "gold standard" of establishing causality [183], its limitations in the context of BCTs discussed in **Chapter 1** and **Chapter 2** make MR the next most desirable method. To exemplify the usefulness of MR, I selected three diseases to serve as my aims: colorectal cancer (CRC, **Chapter 3**), severe malaria (SM) caused by *Plasmodium falciparum* (*P. falciparum*, **Chapter 5**) and deep vein thrombosis (DVT, **Chapter 6**).

7.1.1. Relationship between white blood cell count and CRC risk (Chapter 3)

Here, I explored how variation in WBC subtype count could affect the risk of CRC, a disease of global importance which has been previously linked with inflammation [266,382]. Previous studies had been few and limited in scope [291,292,300], which made it difficult to estimate if and how each WBC subtype count might affect CRC development. Due to the genetic correlation between WBC subtype counts [149,166], I used multivariable MR (MVMR) [215] to estimate their direct effect on CRC risk of each WBC subtype. Additionally, I used the UK Biobank (UKBB) dataset [161,164] to conduct a cohort analysis of WBC count and incident CRC to complement the MR analyses. When assessed in this way, there was evidence that the risk of disease was reduced with increasing cell count for both increasing eosinophil count and lymphocyte count. Finally, in an analysis using a different strategy where I followed-up the eosinophil finding, I performed a MR analysis between allergic disease (associated with eosinophils [51]) and CRC risk. In this case, it was suggested that allergic disease had a potentially protective effect on CRC

development. This extended analysis outlined how a MR of BCTs can lead to new hypothesis generation and investigation.

7.1.2. People from UK Biobank associated with the African continent (Chapter 4)

Afterwards, I set out to perform a MR analysis between neutrophil count and P. *falciparum* SM. Being a two-sample MR analysis, this required that both the exposure and outcome GWAS data come from the same underlying population [200]. It was apparent that it was important to take into account the population structure (**Chapter 4**) that might be present in a sample, as the diversity within Africa is greater than in any other continent [387,555,699]. As the degree of intra-population structure had not been explored by previous studies, I aimed to use the UKBB dataset to identify a group of individuals similar to one sampled in Africa and generate sub-clusters inside this continental ancestry group (CAG) that would resemble the more homogeneous populations typically used in GWAS. Using a combination of software tools and valuable resources such as the 1000 Genomes Project [248], I was able to identify four CAGs in UKBB, including 6,653 people as part of the African CAG (and 7 clusters) [238]. This ultimately helped me conduct **Chapter 5**, as I was able to run a neutrophil count GWAS in people of African ancestry while accounting for intrapopulation structure that could affect association statistics.

7.1.3. Relationship between neutrophil count and P. *falciparum* SM (Chapter 5)

The work in **Chapter 4** allowed me to complete analysis of the relationship between neutrophil count and P. falciparum SM. Using the African CAG data generated in **Chapter 4**, I conducted a GWAS of neutrophil count, identifying 73 loci associated with the trait and 12 SNPs that could be used in a MR analysis. Afterwards, I employed several sensitivity analyses to ensure the validity of GWAS and potential MR SNPs. In the end, however, the MR analysis suggested little evidence of an effect in either direction, most likely due to the low number of instruments for both neutrophil count and SM.

7.1.4. Establishing platelet-associated risk factors for DVT (Chapter 6)

The final chapter of my thesis was aimed at understanding the exposures that might cause DVT, a disease known to be related to platelet measurements and function [146,700].

As the link between platelets and DVT is well-known [622,632,633], I aimed to perform a MR phenome-wide association study (PheWAS) [197]. This was done to identify a possible mechanism through which platelets could affect the risk of DVT in a comprehensive manner which could then lead to new mechanistic investigations on platelets and DVT. Here, 57 exposures were found to influence DVT risk, of which half were associated with adiposity [249]. Therefore, I investigated if any BMI-associated circulating proteins might mediate the relationship between adiposity and BMI through a MR mediation analysis. Blood-circulating proteins were investigated as they have the potential to alter platelet activity and/or act as a component of the platelet clotting cascade. Here, plasminogen activation inhibitor 1 (PAI-1), a protein predominantly present in platelets (>90%) [672], was identified as a mediator [249], and akin to the follow-up analysis in **Chapter 3**, suggested a mechanistic pathway though which platelet levels might affect DVT development.

7.2. Placing my research in context

Overall, the findings of my thesis have addressed the overarching objective I set out in **Chapter 1**. The work done in **Chapter 3** identified eosinophil count and lymphocyte count as novel risk factors for CRC. Similarly, **Chapter 6** was marked by identification of novel risk factors for DVT, such as the platelet-associated PAI-1 protein. **Chapter 4** and **Chapter 5** were successful in showing how one might construct MR instruments for BCTs in understudied populations, and although there was no evidence of an effect in the subsequent MR analysis of neutrophil count to malaria, this emphasised the need for more genetic studies in Africa.

7.2.1. Recent developments

Since the completion of the analyses, several new studies have been published; the data from which could be used in future analyses on the BCT-disease pairs I have studied in this thesis.

One of these is the newest iteration of the CRC risk meta-analysis, with a total sample-size of 265,791 individuals (100,204 cases and 154,587 controls) which identified 50 new loci associated with CRC risk [583]. This not only included a GWAS, but a transcriptome-wide association study (TWAS) and a methylation-wide association study (MWAS), further describing the genetic architecture [347] of CRC development [583]. Therefore, an analysis of WBC subtype count on this larger dataset could be used as a

replication study, and integrating the TWAS and MWAS results could also pinpoint how differences in WBC gene expression might influence CRC risk.

Moreover, the recent release of the Uganda Genome Resource included summary statistics for BCTs, such as neutrophil count, for almost ~5,000 Ugandans [701]. Therefore, one way to overcome the low instrument number in the **Chapter 5** MR analysis could be through a meta-analysis of these summary statistics with those generated in my GWAS of neutrophil count, raising the sample-size to over 11,000.

Finally, a recent trans-ancestry meta-analysis of venous thromboembolism (VTE, DVT + pulmonary embolism) was done on ~80,000 individuals (47,822 Europeans) across 30 studies, identifying 48 novel associations with VTE, increasing the number of independent SNPs associated with VTE to 135 [700]. Here, platelets were found to be the largest contributing factor apart from the known coagulation pathways, further showing that platelets traits play an important role in DVT risk [700]. Using this new dataset in a follow-up MR-PheWAS analysis could outline new novel risk factors for DVT related to platelet count and activity.

7.2.2. Public health and clinical implications

As discussed in **Chapter 1**, BCTs do not themselves affect the risk of disease but are rather used as indicators/flags for an increase or decrease in a biological mechanism that can then affect the risk of disease. These can then be assessed if they make biological sense by referring to the literature. However, this approach does not determine the specific pathways through which the cells associated with these BCT measurements act downstream to influence disease risk. Nevertheless, BCTs have been successfully used previously in both traditional epidemiological approaches and MR studies [149,279,281,300], outlining their use as traits to investigate disease aetiology and aiding in the generation of new research.

One potential way to investigate the pathways involved in disease aetiology could be through a colocalization analysis [702] making use of publicly available datasets, such as eqtlGEN [703] or GTEx [361]. These include expression quantitative trait loci (eQTL) data for BCT subsets such as CD8+ T-cells based on single-cell eQTL data, or specific genes such as *RNASE2* encoding for the eosinophil-derived neurotoxin (EDN) protein [361,703]. A shared signal between the BCT and a gene coding for an effect protein would be helpful in mapping pathways explaining a MR result. Another way would be to run a two-step MR analysis [218], where the effect of e.g. eosinophil count is estimated on an eQTL or

protein QTL (pQTL) trait, outlining genes or proteins affected by eosinophil count. The next step would be an MR of these highlighted eQTL/pQTL onto a disease, allowing for establishing if these act as mediating factors on disease development, as well as allowing to quantify the proportion mediated by each intermediate eQTL/pQTL trait in the eosinophil count to disease relationship [218].

The study of BCTs and disease has the potential to contribute not only to the current literature, but also to the improvement of people's health outcomes. While BCTs themselves are not likely to be directly targetable, untangling the biological mechanisms through which a BCT affects the risk of disease could lead to the development of novel therapeutic approaches. For example, in **Chapter 3** I discussed how the results of my analyses could be taken forward, such as through RNA-Seq [363] or SNP clustering [704]. These would then have the potential to pinpoint a more specific mechanism that affects the risk of disease, opening up the possibility of drug target identification [705]. One such example is present in **Chapter 6**, where I studied specific proteins and how they affect the risk of DVT (e.g. plasminogen activation inhibitor-1, PAI-1).

7.2.3. Population structure

Another point of discussion is population structure and how it affects the study of BCTs and disease. In **Chapter 4**, I observed prominent structure in the African CAG compared to the European CAG. The most common way of dealing with population structure in GWAS is through adjustment of principal components (PCs) [228]. At the same time, however, adjusting for structure can also lead to over correction and loss of signals which could then be informative in a MR study [339]. Other scientists, such as Eran Elhaik, have pushed back entirely on the usage of PCs in genetics, accepting only the first PC as a potential covariate in a GWA model [706].

In the context of neutrophil count and severe malaria, ancestry is tied with the environment, exposure and outcome simultaneously [245,452,707]. This makes it difficult to untangle as adjusting for PCs might bias GWAS effect-sizes due to either under or over correction [339]. One solution would be to sample a population of individuals of European ancestry living in sub-Saharan Africa and assess their severity of malaria compared to the native population, which would (in theory) allow for controlling of the environmental factors.

As most studies have been conducted in European populations, it is important to focus on undertaking more GWAS in diverse populations [153,386]. One of the advantages of

genomic analyses in diverse populations is that they can find novel phenotypes not seen anywhere else. For example, an early study by Kenny et al. had discovered back in 2012 that a SNP in *the tyrosinase-related protein 1* (*TYRP1*) gene causing an amino-acid change was responsible for a blond hair phenotype in people living in Papua New Guinea [708]. From a practical healthcare point of view, genetic studies in non-European populations have helped inform on how to improve the health outcomes of a broader segment of the population. For example, studies done in people of African ancestry have allowed for the generation of novel and effective treatments for ancestry-associated diseases such as malaria, sickle cell disease and certain cardiomyopathies [386].

However, the need for GWAS in diverse ancestries also comes from a need to improve the health outcomes of specific populations [386]. Sub-Saharan Ancestry individuals have smaller linkage disequilibrium (LD) blocks (regions of the genome where allele variation in SNPs is highly correlated i.e. $r^2 > 0.8$) compared to those of European ancestry [709]. This can limit the translational ability of polygenic risk scores (PRSs), as different SNPs might be representative for LD blocks with different sizes [709]. Therefore, it becomes important for genomic studies to be conducted in non-European populations to identify ancestry-specific signals that can inform on health traits, as has been the case with BCTs [166,399].

This call for diverse initiatives should not be seen as an expense from a public policy point of view, but rather as an investment for the improvement of everyone's health outcomes. For example, those of African ancestry have lower cholesterol levels and a lower risk of heart disease due to differences in allele frequencies for particular SNPs compared to Europeans, allowing for the investigation of potential novel cholesterol-lowering drugs [710] that could aid in treatments benefitting everyone.

7.3. Biobanks – variation or power?

The recent initiative by OurFutureHealth (OFH) to create a biobank of multi-omic data in the UK in partnership with the National Health Service (NHS) is one answer to addressing the issues of diversity [711]. However, it remains to be seen how the ambitious target of 5-million samples by 2025 will be met [712]. Nevertheless, having a large sample-size can be a success that comes with its own problems. UKBB is known to suffer from selection bias because it is a cohort study where participants are gathered through voluntary participation, and therefore the study population is likely not representative of the general UK population [713]. Therefore, given the volunteer-based approach to OFH,

these issues might persist even with an increase in sample-size, as was the case with the infamous 1936 Literary Digest poll blunder [714]. The Literary Digest newspaper tried to poll who would likely win the 1936 US presidential election and sent over 10 million mock ballots to US citizens, of which around ~2.5 million posted back a filled ballot [714].

One might think at first that having such an unprecedented large sample-size would make the polling results extremely reliable. However, as those who answered the poll were more likely to vote for the Republican candidate Landon, the polling results showed that he would win against the Democratic incumbent Roosevelt [714]. In reality, Roosevelt had an overwhelming win over Landon, which was the opposite of what the polling results suggested [714]. Similarly, if a sample of the population is invited to take part in a biobank study, and a subset of this sample is more likely to participate in the study, it might make results from both observational and GWA studies less generalizable to the whole UK population. Indeed, this is the case with UKBB, where participants were more likely to be older, female, healthier, and with a higher socioeconomic status compared to the general UK population [713].

In contrast, studies done in people living in sub-Saharan Africa have been small in sample-size compared to those in Europeans. Fortunately, the recent announcement by the pan-African Bioinformatic Network on the establishment of eight genomics centres in Africa is of great importance [715], as it reflects back to the discussion in **Chapter 4** and **Chapter 5** where I highlighted the importance of expanding genetic research in sub-Saharan Africa. Therefore, a trade-off between variation and power might not be needed if the current trend of increasing diversity and sample-size holds. Collaborations in the form of data sharing to undertake meta-analyses from datasets from OFH, UKBB, the Million Veterans Program [162] and African initiatives could lead to both an increase in power and an improved detection of SNP variation that might affect BCTs.

7.4. Mendelian randomization – past, present and future

By employing a state-of-the-art method in genetic epidemiology known as MR, I showed how genetically proxied BCTs can be relevant to identifying risk factors for disease. Specific methodological approaches for each chapter were made, such as employing MVMR or MR-PheWAS [215,219]. However, the development of MR methods is relatively recent.

In their 2003 paper on MR that I mentioned in **Chapter 1**, Davey-Smith and Ebrahim make the case for the potential of MR in improving the understanding of disease by using genetic proxies for exposures [169]. At the time, this seemed overly optimistic, especially since the first actual GWAS was not conducted until 2 years later [157]. Just over a decade after the MR debut paper, Burgess et al. discuss in their editorial the advances that had been made since 2003 [716]. Here, they confirm the success of MR and its popularity fuelled by the rise of large-scale biobank studies and methodological advances [716], such as MR-Egger [205].

This year (2023) marks the 20th anniversary of the contemporary conceptualisation of MR. In the context of BCTs, the 2020s have seen a promising increase in the number of MR studies between BCTs and disease [352,575,717–720]. New sensitivity MR methods, such as median and mode based estimates allow for better assessment of MR results [204,212]. Outlier removal methods, such as MR-PRESSO, are able to detect and correct for horizontal pleiotropic SNPs that might bias MR estimates [207].

However, these new methodological approaches are not limited to just testing the validity of MR assumptions. More recent advances MR methods have outlined a polynomial approach to studying the effect of an exposure on an outcome [383]. For example, BMI has been shown to make a “U” curve with cholesterol levels, which might have previously been interpreted as a linear increase using standard MR [383]. It would be interesting to assess if BCTs affect disease risk in a non-linear pattern. Finally, as datasets become larger and more comprehensive in the amount of phenotypes studied, a forward MR-PheWAS [185] of BCTs on all outcomes would be interesting. This approach could identify BCTs as novel risk factors for traits or diseases without any prior hypothesis, thereby allowing for the generation of new hypotheses and mechanistic investigations.

7.5. Conclusion

Overall, the ascending trend of discovery, impact, and methodological advances during the last decade in the domain of genomics has continued in the 2020s [154]. Even so, much is still left to be discovered on the topic of BCTs and disease. I personally view the current unknowns and potential limitations outlined above as an untapped resource that can be mined through the current toolsets that have generously been provided by past and current initiatives. The exploration of BCTs and disease is just the tip of the iceberg, and much more will likely be discovered on the role of blood cells in regulating disease risk in the upcoming years.

BIBLIOGRAPHY

1. Kalachanis K, Michailidis IE. The Hippocratic View on Humors and Human Temperament. *European Journal of Social Behaviour*. 2015;2(2):1-5.
2. Friedland G. Discovery of the function of the heart and circulation of blood. *Cardiovasc J Afr*. 2009;20(3):160.
3. Stelmack RM, Stalikas A. Galen and the humour theory of temperament. *Pers Individ Dif*. 1991;12(3):255-263. doi:10.1016/0191-8869(91)90111-N
4. Ribatti D. William Harvey and the discovery of the circulation of the blood. *J Angiogenes Res*. 2009;1(1):3. doi:10.1186/2040-2384-1-3
5. Davis IM. “Round, red globules floating in a crystalline fluid” – Antoni van Leeuwenhoek’s observations of red blood cells and hemocytes. *Micron*. 2022;157:103249. doi:10.1016/J.MICRON.2022.103249
6. Doyle D. William Hewson (1739–74): the father of haematology. *Br J Haematol*. 2006;133(4):375-381. doi:10.1111/J.1365-2141.2006.06037.X
7. Ribatti D, Crivellato E. Giulio Bizzozero and the discovery of platelets. *Leuk Res*. 2007;31(10):1339-1341. doi:10.1016/j.leukres.2007.02.008
8. Dean L. Blood and the cells it contains. Published online 2005.
9. Ferretti E, Hadjantonakis AK. Mesoderm specification and diversification: from single cells to emergent tissues. *Curr Opin Cell Biol*. 2019;61:110. doi:10.1016/J.CEB.2019.07.012
10. Orkin SH, Zon LI. Hematopoiesis: An Evolving Paradigm for Stem Cell Biology. *Cell*. 2008;132(4):631-644. doi:10.1016/J.CELL.2008.01.025
11. Cvejic A. Mechanisms of fate decision and lineage commitment during haematopoiesis. *Immunol Cell Biol*. 2016;94(3):230-235. doi:10.1038/ICB.2015.96
12. Birbrair A, Frenette PS. Niche heterogeneity in the bone marrow. *Ann N Y Acad Sci*. 2016;1370(1):82. doi:10.1111/NYAS.13016
13. Gurevitch O, Slavin S, Feldman AG. Conversion of red bone marrow into yellow – Cause and mechanisms. *Med Hypotheses*. 2007;69(3):531-536. doi:10.1016/J.MEHY.2007.01.052
14. Mattiucci D, Naveiras O, Poloni A. Bone Marrow “Yellow” and “Red” Adipocytes”: Good or Bad Cells? *Current Molecular Biology Reports 2018 4:3*. 2018;4(3):117-122. doi:10.1007/S40610-018-0098-6
15. Gurkan UA, Akkus O. The mechanical environment of bone marrow: A review. *Ann Biomed Eng*. 2008;36(12):1978-1991. doi:10.1007/S10439-008-9577-X/TABLES/3

16. Morrison SJ, Scadden DT. The bone marrow niche for haematopoietic stem cells. *Nature 2014 505:7483*. 2014;505(7483):327-334. doi:10.1038/nature12984
17. Belyavsky A, Petinati N, Drize N. Hematopoiesis during Ontogenesis, Adult Life, and Aging. *Int J Mol Sci*. 2021;22(17). doi:10.3390/IJMS22179231
18. Guillerman RP. Marrow: Red, yellow and bad. *Pediatr Radiol*. 2013;43(SUPPL. 1):181-192. doi:10.1007/S00247-012-2582-0/FIGURES/18
19. Liggett LA, Sankaran VG. Unraveling Hematopoiesis through the Lens of Genomics. *Cell*. 2020;182(6):1384-1400. doi:10.1016/J.CELL.2020.08.030
20. Qin P, Pang Y, Hou W, et al. Integrated decoding hematopoiesis and leukemogenesis using single-cell sequencing and its medical implication. *Cell Discovery 2020 7:1*. 2021;7(1):1-17. doi:10.1038/s41421-020-00223-4
21. Siracusa MC, Kim BS, Spergel JM, Artis D. Basophils and allergic inflammation. *J Allergy Clin Immunol*. 2013;132(4):789. doi:10.1016/J.JACI.2013.07.046
22. Karasuyama H, Miyake K, Yoshikawa S, Yamanishi Y. Multifaceted roles of basophils in health and disease. *Journal of Allergy and Clinical Immunology*. 2018;142(2):370-380. doi:10.1016/j.jaci.2017.10.042
23. Marone G, Schroeder JT, Mattei F, et al. Is There a Role for Basophils in Cancer? *Front Immunol*. 2020;11:2103. doi:10.3389/fimmu.2020.02103
24. Hamey FK, Lau WWY, Kucinski I, et al. Single-cell molecular profiling provides a high-resolution map of basophil and mast cell development. *Allergy*. 2021;76(6):1731-1742. doi:10.1111/ALL.14633
25. St. John AL, Rathore APS, Ginhoux F. New perspectives on the origins and heterogeneity of mast cells. *Nature Reviews Immunology 2022 23:1*. 2022;23(1):55-68. doi:10.1038/s41577-022-00731-2
26. Shah H, Eisenbarth S, Tormey CA, Siddon AJ. Behind the scenes with basophils: an emerging therapeutic target. *Immunotherapy Advances*. 2021;1(1). doi:10.1093/IMMADV/LTAB008
27. Miyake K, Shibata S, Yoshikawa S, Karasuyama H. Basophils and their effector molecules in allergic disorders. *Allergy*. 2021;76(6):1693-1706. doi:10.1111/ALL.14662
28. Huang H, Li Y, Liu B. Transcriptional Regulation of Mast Cell and Basophil Lineage Commitment. *Semin Immunopathol*. 2016;38(5):539. doi:10.1007/S00281-016-0562-4
29. Zhang N, Zhang ZM, Wang XF. The roles of basophils in mediating the immune responses. *https://doi.org/101177/20587392211047644*. 2021;19. doi:10.1177/20587392211047644
30. McEuen AR, Calafat J, Compton SJ, et al. Mass, charge, and subcellular localization of a unique secretory product identified by the basophil-specific

antibody BB1. *Journal of Allergy and Clinical Immunology*. 2001;107(5 SUPPL.):842-848. doi:10.1067/mai.2001.114650

31. Peng J, Siracusa MC. Basophils in antihelminth immunity. *Semin Immunol*. 2021;53(November):101529. doi:10.1016/j.smim.2021.101529
32. Rigoni A, Colombo MP, Pucillo C. Mast cells, basophils and eosinophils: From allergy to cancer. *Semin Immunol*. 2018;35:29-34. doi:10.1016/J.SMIM.2018.02.001
33. Blank U, Huang H, Kawakami T. The high affinity IgE receptor: a signaling update. *Curr Opin Immunol*. 2021;72:51-58. doi:10.1016/j.coi.2021.03.015
34. Maddur MS, Kaveri S V., Bayry J. Basophils as antigen presenting cells. *Trends Immunol*. 2010;31(2):45-48. doi:10.1016/J.IT.2009.12.004
35. Yoshikawa S, Miyake K, Kamiya A, Karasuyama H. The role of basophils in acquired protective immunity to tick infestation. *Parasite Immunol*. 2021;43(5):e12804. doi:10.1111/PIM.12804
36. Karasuyama H, Shibata S, Yoshikawa S, Miyake K. Basophils, a neglected minority in the immune system, have come into the limelight at last. *Int Immunol*. 2021;33(12):809-813. doi:10.1093/INTIMM/DXAB021
37. Sastre B, Rodrigo-Muñoz JM, Garcia-Sanchez DA, Cañas JA, Del Pozo V. Eosinophils: Old Players in a New Game. *J Investig Allergol Clin Immunol*. 2018;28(5):289-304. doi:10.18176/JIACI.0295
38. Valent P, Klion AD, Horny HP, et al. Contemporary consensus proposal on criteria and classification of eosinophilic disorders and related syndromes. *Journal of Allergy and Clinical Immunology*. 2012;130(3):607-612.e9. doi:10.1016/J.JACI.2012.02.019
39. Gleich GJ, Klion AD, Lee JJ, Weller PF. The Consequences of Not Having Eosinophils. *Allergy*. 2013;68(7):829. doi:10.1111/ALL.12169
40. Masterson JC, Menard-katcher C, Larsen LD, Furuta GT, Spencer LA. Heterogeneity of Intestinal Tissue Eosinophils: Potential Considerations for Next-Generation Eosinophil-Targeting Strategies. *Cells*. 2021;10(2):1-21. doi:10.3390/CELLS10020426
41. Mack EA, Pear WS. Transcription factor and cytokine regulation of eosinophil lineage commitment. *Curr Opin Hematol*. 2020;27(1):27. doi:10.1097/MOH.0000000000000552
42. Schwartz JT, Magier AZ, Marshall SA, Fulkerson PC. Eosinophil progenitor cell blood levels inversely correlate with disease control in pediatric patients with asthma. *Journal of Allergy and Clinical Immunology*. 2019;143(2):AB5. doi:10.1016/j.jaci.2018.12.017

43. Klion AD, Ackerman SJ, Bochner BS. Contributions of Eosinophils to Human Health and Disease. *https://doi.org/101146/annurev-pathmechdis-012419-032756*. 2020;15:179-209. doi:10.1146/ANNUREV-PATHMECHDIS-012419-032756
44. Berdnikovs S. The twilight zone: plasticity and mixed ontogeny of neutrophil and eosinophil granulocyte subsets. *Semin Immunopathol*. 2021;43(3):337-346. doi:10.1007/S00281-021-00862-Z/FIGURES/2
45. Freire PC, Muñoz CH, Stingl G. IgE autoreactivity in bullous pemphigoid: eosinophils and mast cells as major targets of pathogenic immune reactants. *British Journal of Dermatology*. 2017;177(6):1644-1653. doi:10.1111/BJD.15924
46. MacGlashan D. Autoantibodies to IgE and FcεRI and the natural variability of spleen tyrosine kinase expression in basophils. *Journal of Allergy and Clinical Immunology*. 2019;143(3):1100-1107.e11. doi:10.1016/J.JACI.2018.05.019
47. Gaur P, Zaffran I, George T, Rahimli Alekberli F, Ben-Zimra M, Levi-Schaffer F. The regulatory role of eosinophils in viral, bacterial, and fungal infections. *Clin Exp Immunol*. 2022;209(1):72-82. doi:10.1093/CEI/UXAC038
48. Jacobsen EA, Jackson DJ, Heffler E, et al. Eosinophil Knockout Humans: Uncovering the Role of Eosinophils Through Eosinophil-Directed Biological Therapies. *https://doi.org/101146/annurev-immunol-093019-125918*. 2021;39:719-757. doi:10.1146/ANNUREV-IMMUNOL-093019-125918
49. Abengózar MÁ, Fernández-Reyes M, Salazar VA, et al. Essential Role of Enzymatic Activity in the Leishmanicidal Mechanism of the Eosinophil Cationic Protein (RNase 3). *ACS Infect Dis*. 2022;8(7):1207-1217. doi:10.1021/ACSINFECDIS.1C00537/ASSET/IMAGES/LARGE/ID1C00537_0007.JPEG
50. Amber KT, Chernyavsky A, Agnoletti AF, Cozzani E, Grando SA. Mechanisms of pathogenic effects of eosinophil cationic protein and eosinophil-derived neurotoxin on human keratinocytes. *Exp Dermatol*. 2018;27(12):1322-1327. doi:10.1111/EXD.13782
51. Aoki A, Hirahara K, Kiuchi M, Nakayama T. Eosinophils: Cells known for over 140 years with broad and new functions. *Allergology International*. 2021;70(1):3-8. doi:10.1016/J.ALIT.2020.09.002
52. Zhang D, Yang X, Wang T, Ji X, Wu X. Advances in organic fluorescent probes for bromide ions, hypobromous acid and related eosinophil peroxidase-A review. *Anal Chim Acta*. Published online November 14, 2022:340626. doi:10.1016/J.ACA.2022.340626

53. Wechsler ME, Munitz A, Ackerman SJ, et al. Eosinophils in Health and Disease: A State-of-the-Art Review. *Mayo Clin Proc*. 2021;96(10):2694-2707. doi:10.1016/J.MAYOCP.2021.04.025
54. Miller JFAP. The function of the thymus and its impact on modern medicine. *Science (1979)*. 2020;369(6503):1-8. doi:10.1126/SCIENCE.ABA2429/ASSET/C64A2808-E722-47CC-AEA5-1676C1A3065E/ASSETS/GRAPHIC/369_ABA2429_F6.JPEG
55. Cooper MD, Peterson RDA, Good RA. Delineation of the Thymic and Bursal Lymphoid Systems in the Chicken. *Nature 1965 205:4967*. 1965;205(4967):143-146. doi:10.1038/205143a0
56. Miller JFAP. The discovery of thymus function and of thymus-derived lymphocytes. *Immunol Rev*. 2002;185(1):7-14. doi:10.1034/J.1600-065X.2002.18502.X
57. Kiessling R, Klein E, Wigzell H. „Natural" killer cells in the mouse. I. Cytotoxic cells with specificity for mouse Moloney leukemia cells. Specificity and distribution according to genotype. *Eur J Immunol*. 1975;5(2):112-117. doi:10.1002/EJI.1830050208
58. Farber DL, Yudanin NA, Restifo NP. Human memory T cells: generation, compartmentalization and homeostasis. *Nat Rev Immunol*. 2014;14(1):24. doi:10.1038/NRI3567
59. Reynaldi A, Smith NL, Schlub TE, et al. Fate mapping reveals the age structure of the peripheral T cell compartment. *Proc Natl Acad Sci U S A*. 2019;116(10):3974-3981. doi:10.1073/PNAS.1811634116/-/DCSUPPLEMENTAL
60. Jones DD, Wilmore JR, Allman D. Cellular dynamics of memory B cell populations: IgM+ and IgG+ memory B cells persist indefinitely as quiescent cells. *J Immunol*. 2015;195(10):4753. doi:10.4049/JIMMUNOL.1501365
61. Wu SY, Fu T, Jiang YZ, Shao ZM. Natural killer cells in cancer biology and therapy. *Mol Cancer*. 2020;19(1):1-26. doi:10.1186/S12943-020-01238-X/TABLES/3
62. Kumar B V., Connors TJ, Farber DL. Human T Cell Development, Localization, and Function throughout Life. *Immunity*. 2018;48(2):202-213. doi:10.1016/J.IMMUNI.2018.01.007
63. Park JE, Botting RA, Conde CD, et al. A cell atlas of human thymic development defines T cell repertoire formation. *Science (1979)*. 2020;367(6480). doi:10.1126/SCIENCE.AAY3224/SUPPL_FILE/AAY3224_TABLE-S5.CSV
64. Wang Y, Liu J, Burrows PD, Wang JY. B Cell Development and Maturation. *Adv Exp Med Biol*. 2020;1254:1-22. doi:10.1007/978-981-15-3532-1_1/FIGURES/4

65. Pieper K, Grimbacher B, Eibel H. B-cell biology and development. *Journal of Allergy and Clinical Immunology*. 2013;131(4):959-971. doi:10.1016/J.JACI.2013.01.046
66. Liao S, von der Weid PY. Lymphatic system: An active pathway for immune protection. *Semin Cell Dev Biol*. 2015;38:83-89. doi:10.1016/J.SEMCDB.2014.11.012
67. Randolph GJ, Ivanov S, Zinselmeyer BH, Scallan JP. The Lymphatic System: Integral Roles in Immunity. *https://doi.org/101146/annurev-immunol-041015-055354*. 2017;35:31-52. doi:10.1146/ANNUREV-IMMUNOL-041015-055354
68. Willard-Mack CL. Normal Structure, Function, and Histology of Lymph Nodes. *Toxicol Pathol*. 2006;34(5):409-424. doi:10.1080/01926230600867727
69. Wieczorek M, Abualrous ET, Sticht J, et al. Major histocompatibility complex (MHC) class I and MHC class II proteins: Conformational plasticity in antigen presentation. *Front Immunol*. 2017;8(MAR):292. doi:10.3389/FIMMU.2017.00292/BIBTEX
70. Pishesha N, Harmand TJ, Ploegh HL. A guide to antigen processing and presentation. *Nature Reviews Immunology 2022 22:12*. 2022;22(12):751-764. doi:10.1038/s41577-022-00707-2
71. Andersen MH, Schrama D, Thor Straten P, Becker JC. Cytotoxic T cells. *Journal of Investigative Dermatology*. 2006;126(1):32-41. doi:10.1038/sj.jid.5700001
72. Raphael I, Nalawade S, Eagar TN, Forsthuber TG. T cell subsets and their signature cytokines in autoimmune and inflammatory diseases. *Cytokine*. 2015;74(1):5-17. doi:10.1016/J.CYTO.2014.09.011
73. Zhu J. T helper cell differentiation, heterogeneity, and plasticity. *Cold Spring Harb Perspect Biol*. 2018;10(10). doi:10.1101/cshperspect.a030338
74. Crotty S. T Follicular Helper Cell Biology: A Decade of Discovery and Diseases. *Immunity*. 2019;50(5):1132-1148. doi:10.1016/J.IMMUNI.2019.04.011
75. Fairfax BP, Vannberg FO, Radhakrishnan J, et al. An integrated expression phenotype mapping approach defines common variants in LEP, ALOX15 and CAPNS1 associated with induction of IL-6. *Hum Mol Genet*. 2010;19(4):720-730. doi:10.1093/HMG/DDP530
76. Zhu X, Zhu J. CD4 T Helper Cell Subsets and Related Human Immunological Disorders. *International Journal of Molecular Sciences 2020, Vol 21, Page 8011*. 2020;21(21):8011. doi:10.3390/IJMS21218011
77. Cyster JG, Allen CDC. B Cell Responses: Cell Interaction Dynamics and Decisions. *Cell*. 2019;177(3):524-540. doi:10.1016/J.CELL.2019.03.016
78. Prager I, Watzl C. Mechanisms of natural killer cell-mediated cellular cytotoxicity. *J Leukoc Biol*. 2019;105(6):1319-1329. doi:10.1002/JLB.MR0718-269R

79. Gilchrist JJ, Makino S, Naranbhai V, et al. Natural Killer cells demonstrate distinct eQTL and transcriptome-wide disease associations, highlighting their role in autoimmunity. *Nature Communications 2022 13:1*. 2022;13(1):1-13. doi:10.1038/s41467-022-31626-4
80. Teh YC, Ding JL, Ng LG, Chong SZ. Capturing the fantastic voyage of monocytes through time and space. *Front Immunol*. 2019;10(MAR):834. doi:10.3389/FIMMU.2019.00834/BIBTEX
81. Ziegler-Heitbrock L. Monocyte subsets in man and other species. *Cell Immunol*. 2014;289(1-2):135-139. doi:10.1016/J.CELLIMM.2014.03.019
82. Austermann J, Roth J, Barczyk-Kahlert K. The Good and the Bad: Monocytes’ and Macrophages’ Diverse Functions in Inflammation. *Cells 2022, Vol 11, Page 1979*. 2022;11(12):1979. doi:10.3390/CELLS11121979
83. Ożańska A, Szymczak D, Rybka J. Pattern of human monocyte subpopulations in health and disease. *Scand J Immunol*. 2020;92(1):e12883. doi:10.1111/SJI.12883
84. Canè S, Ugel S, Trovato R, et al. The endless saga of monocyte diversity. *Front Immunol*. 2019;10:1786. doi:10.3389/FIMMU.2019.01786/BIBTEX
85. Wolf AA, Yáñez A, Barman PK, Goodridge HS. The ontogeny of monocyte subsets. *Front Immunol*. 2019;10(JULY):1642. doi:10.3389/FIMMU.2019.01642/BIBTEX
86. Cormican S, Griffin MD. Human Monocyte Subset Distinctions and Function: Insights From Gene Expression Analysis. *Front Immunol*. 2020;11:1070. doi:10.3389/FIMMU.2020.01070/BIBTEX
87. Sebastian A, Sanju S, Jain P, Priya VV, Varma PK, Mony U. Non-classical monocytes and its potential in diagnosing sepsis post cardiac surgery. *Int Immunopharmacol*. 2021;99:108037. doi:10.1016/J.INTIMP.2021.108037
88. Kapellos TS, Bonaguro L, Gemünd I, et al. *Human Monocyte Subsets and Phenotypes in Major Chronic Inflammatory Diseases*. Vol 10. Frontiers Media S.A.; 2019:2035.
89. Zhang L, Hofer TP, Zawada AM, et al. Epigenetics in non-classical monocytes support their pro-inflammatory gene expression. *Immunobiology*. 2020;225(3):151958. doi:10.1016/J.IMBIO.2020.151958
90. Narasimhan PB, Marcovecchio P, Hamers AAJ, Hedrick CC. *Nonclassical Monocytes in Health and Disease*. Vol 37.; 2019:439-456.
91. Hidalgo A, Chilvers ER, Summers C, Koenderman L. The Neutrophil Life Cycle. *Trends Immunol*. 2019;40(7):584-597. doi:10.1016/J.IT.2019.04.013

92. Koenderman L, Tesselaar K, Vrisekoop N. Human neutrophil kinetics: a call to revisit old evidence. *Trends Immunol*. 2022;43(11):868-876. doi:10.1016/J.IT.2022.09.008
93. Manley HR, Keightley MC, Lieschke GJ. The Neutrophil Nucleus: An Important Influence on Neutrophil Migration and Function. *Front Immunol*. 2018;9:2867. doi:10.3389/FIMMU.2018.02867/BIBTEX
94. Overbeeke C, Tak T, Koenderman L. The journey of neutropoiesis: how complex landscapes in bone marrow guide continuous neutrophil lineage determination. *Blood*. 2022;139(15):2285-2293. doi:10.1182/BLOOD.2021012835
95. Carrington EM, Louis C, Kratina T, et al. BCL-XL antagonism selectively reduces neutrophil life span within inflamed tissues without causing neutropenia. *Blood Adv*. 2021;5(11):2550-2562. doi:10.1182/BLOODADVANCES.2020004139
96. Burn GL, Foti A, Marsman G, Patel DF, Zychlinsky A. The Neutrophil. *Immunity*. 2021;54(7):1377-1391. doi:10.1016/J.IMMUNI.2021.06.006
97. Naish E, Wood AJT, Stewart AP, et al. The formation and function of the neutrophil phagosome. *Immunol Rev*. Published online 2022. doi:10.1111/IMR.13173
98. Othman A, Sekheri M, Filep JG. Roles of neutrophil granule proteins in orchestrating inflammation and immunity. *FEBS J*. 2022;289(14):3932-3953. doi:10.1111/FEBS.15803
99. Sienkiewicz M, Jaśkiewicz A, Tarasiuk A, Fichna J. Lactoferrin: an overview of its main functions, immunomodulatory and antimicrobial role, and clinical significance. *https://doi.org/101080/1040839820211895063*. 2021;62(22):6016-6033. doi:10.1080/10408398.2021.1895063
100. Zeng MY, Miralda I, Armstrong CL, Uriarte SM, Bagaitkar J. The roles of NADPH oxidase in modulating neutrophil effector responses. *Mol Oral Microbiol*. 2019;34(2):27. doi:10.1111/OMI.12252
101. Brinkmann V, Reichard U, Goosmann C, et al. Neutrophil Extracellular Traps Kill Bacteria. *Science (1979)*. 2004;303(5663):1532-1535. doi:10.1126/SCIENCE.1092385/SUPPL_FILE/BRINKMANN.SOM.PDF
102. Hidalgo A, Libby P, Soehnlein O, Aramburu IV, Papayannopoulos V, Silvestre-Roig C. Neutrophil extracellular traps: from physiology to pathology. *Cardiovasc Res*. 2022;118(13):2737-2753. doi:10.1093/CVR/CVAB329
103. Özcan A, Boyman O. Mechanisms regulating neutrophil responses in immunity, allergy, and autoimmunity. *Allergy*. 2022;77(12):3567-3583. doi:10.1111/ALL.15505
104. Xie X, Shi Q, Wu P, et al. Single-cell transcriptome profiling reveals neutrophil heterogeneity in homeostasis and infection. *Nature Immunology 2020 21:9*. 2020;21(9):1119-1133. doi:10.1038/s41590-020-0736-z

105. Montaldo E, Lusito E, Bianchessi V, et al. Cellular and transcriptional dynamics of human neutrophils at steady state and upon stress. *Nature Immunology 2022 23:10*. 2022;23(10):1470-1483. doi:10.1038/s41590-022-01311-1
106. Peiseler M, Kubes P. More friend than foe: the emerging role of neutrophils in tissue repair. *J Clin Invest*. 2019;129(7):2629-2639. doi:10.1172/JCI124616
107. Solari FA, Krahn D, Swieringa F, et al. Multi-omics approaches to study platelet mechanisms. *Curr Opin Chem Biol*. 2023;73:102253. doi:10.1016/J.CBPA.2022.102253
108. van der Meijden PEJ, Heemskerk JWM. Platelet biology and functions: new concepts and clinical perspectives. *Nat Rev Cardiol*. 2019;16(3):166-179. doi:10.1038/s41569-018-0110-0
109. Marjory B, Harr K, Seelig D, Wardrop J, Weiss D. *Schlam's Veterinary Hematology*.
110. Morcos MNF, Li C, Munz CM, et al. Fate mapping of hematopoietic stem cells reveals two pathways of native thrombopoiesis. *Nature Communications 2022 13:1*. 2022;13(1):1-13. doi:10.1038/s41467-022-31914-z
111. Kaushansky K. Determinants of platelet number and regulation of thrombopoiesis. *Hematology*. 2009;2009(1):147-152. doi:10.1182/ASHEDUCATION-2009.1.147
112. Stegner D, van Eeuwijk J, Gorelashvili MG, et al. Spatial Regulation of Thrombopoiesis in the Bone Marrow. *Blood*. 2018;132(Supplement 1):SCI-23. doi:10.1182/BLOOD-2018-99-109545
113. Holinstat M. Normal platelet function. *Cancer and Metastasis Reviews*. 2017;36(2):195-198. doi:10.1007/S10555-017-9677-X/FIGURES/1
114. Ambrosio AL, Di Pietro SM. Storage pool diseases illuminate platelet dense granule biogenesis. *http://dx.doi.org/101080/0953710420161243789*. 2016;28(2):138-146. doi:10.1080/09537104.2016.1243789
115. Dupuis A, Bordet JC, Eckly A, Gachet C. Platelet δ-Storage Pool Disease: An Update. *Journal of Clinical Medicine 2020, Vol 9, Page 2508*. 2020;9(8):2508. doi:10.3390/JCM9082508
116. Smith CW. Release of α-granule contents during platelet activation. *https://doi.org/101080/0953710420211913576*. 2021;33(4):491-502. doi:10.1080/09537104.2021.1913576
117. Flaumenhaft R, Sharda A. Platelet Secretion. *Platelets*. Published online January 1, 2019:349-370. doi:10.1016/B978-0-12-813456-6.00019-9
118. Margraf A, Zarbock A. Platelets in Inflammation and Resolution. *The Journal of Immunology*. 2019;203(9):2357-2367. doi:10.4049/JIMMUNOL.1900899
119. Mandel J, Casari M, Stepanyan M, Martyanov A, Deppermann C. Beyond Hemostasis: Platelet Innate Immune Interactions and Thromboinflammation.

International Journal of Molecular Sciences 2022, Vol 23, Page 3868. 2022;23(7):3868. doi:10.3390/IJMS23073868

120. Karampini E, Bierings R, Voorberg J. Orchestration of Primary Hemostasis by Platelet and Endothelial Lysosome-Related Organelles. *Arterioscler Thromb Vasc Biol*. 2020;40:1441-1453. doi:10.1161/ATVBAHA.120.314245
121. Matsuhira T, Sakai H. Artificial oxygen carriers, from nanometer- to micrometer-sized particles, made of hemoglobin composites substituting for red blood cells. *Particuology*. 2022;64:43-55. doi:10.1016/J.PARTIC.2021.08.010
122. Thiagarajan P, Parker CJ, Prchal JT. How Do Red Blood Cells Die? *Front Physiol*. 2021;12:318. doi:10.3389/FPHYS.2021.655393/BIBTEX
123. Menon V, Ghaffari S. Erythroid enucleation: a gateway into a “bloody” world. *Exp Hematol*. 2021;95:13-22. doi:10.1016/J.EXPHEM.2021.01.001
124. Nandakumar SK, Ulirsch JC, Sankaran VG. Advances in understanding erythropoiesis: evolving perspectives. *Br J Haematol*. 2016;173(2):206-218. doi:10.1111/BJH.13938
125. Peter Klinken S. Red blood cells. *Int J Biochem Cell Biol*. 2002;34(12):1513-1518. doi:10.1016/S1357-2725(02)00087-0
126. Y F, NS B, P L. Biochemistry, Hemoglobin Synthesis. *StatPearls*. Published online February 7, 2019.
127. Karsten E, Breen E, Herbert BR. Red blood cells are dynamic reservoirs of cytokines. *Scientific Reports 2018 8:1*. 2018;8(1):1-12. doi:10.1038/s41598-018-21387-w
128. George-Gay B, Parker K. Understanding the complete blood count with differential. *Journal of Perianesthesia Nursing*. 2003;18(2):96-117. doi:10.1053/JPAN.2003.50013
129. Levine RA, Wardlaw SC. A New Technique for Examining Blood. *Am Sci*. 1988;76(6):592-598.
130. Verso ML. The Evolution of Blood-counting Techniques.
131. Vembadi A, Menachery A, Qasaimeh MA. Cell Cytometry: Review and Perspective on Biotechnological Advances. *Front Bioeng Biotechnol*. 2019;7(JUN):147. doi:10.3389/FBIOE.2019.00147/BIBTEX
132. Wintrobe MM, Fred HL. Maxwell Myer Wintrobe: New History and a New Appreciation. *Tex Heart Inst J*. 2007;34(3):328. Accessed February 5, 2023. /pmc/articles/PMC1995040/
133. Luo J, Chen C, Li Q. White blood cell counting at point-of-care testing: A review. *Electrophoresis*. 2020;41(16-17):1450-1468. doi:10.1002/ELPS.202000029

134. Briggs C, Harrison P, Machin SJ. Continuing developments with the automated platelet count. *Int J Lab Hematol*. 2007;29(2):77-91. doi:10.1111/J.1751-553X.2007.00909.X
135. Chabot-Richards DS, George TI. White Blood Cell Counts Reference Methodology. doi:10.1016/j.cll.2014.10.007
136. Graham MD. The Coulter Principle: Foundation of an Industry. Published online 2003. doi:10.1016/S1535-5535(03)00023-6
137. COULTER LH 750 System.
138. Vis JY, Huisman A. Verification and quality control of routine hematology analyzers. *Int J Lab Hematol*. 2016;38:100-109. doi:10.1111/IJLH.12503
139. Bessman JD, Gilmer PR, Gardner FH. Improved Classification of Anemias by MCV and RDW. *Am J Clin Pathol*. 1983;80(3):322-326. doi:10.1093/AJCP/80.3.322
140. Shoenfeld Y, Alkan ML, Asaly A, Carmeli Y, Katz M. Benign familial leukopenia and neutropenia in different ethnic groups. *Eur J Haematol*. 1988;41(3):273-277. doi:https://doi.org/10.1111/j.1600-0609.1988.tb01192.x
141. Cerezo-Wallis D, Ballesteros I. Neutrophils in cancer, a love–hate affair. *FEBS J*. 2022;289(13):3692-3703. doi:10.1111/FEBS.16022
142. Nakagome K, Nagata M. Involvement and Possible Role of Eosinophils in Asthma Exacerbation. *Front Immunol*. 2018;9:2220. doi:10.3389/FIMMU.2018.02220/BIBTEX
143. Watson RA, Tong O, Cooper R, et al. Immune checkpoint blockade sensitivity and progression-free survival associates with baseline CD8+ T cell clone size and cytotoxicity. *Sci Immunol*. 2021;6(64):8825. doi:10.1126/SCIIMMUNOL.ABJ8825/SUPPL_FILE/SCIIMMUNOL.ABJ8825_TABLES_S1_TO_S13.ZIP
144. Müller F, Mutch NJ, Schenk WA, et al. Platelet polyphosphates are proinflammatory and procoagulant mediators in vivo. *Cell*. 2009;139(6):1143. doi:10.1016/J.CELL.2009.11.001
145. Ding P, Zhang S, Yu M, et al. IL-17A promotes the formation of deep vein thrombosis in a mouse model. *Int Immunopharmacol*. 2018;57:132-138. doi:10.1016/J.INTIMP.2018.02.006
146. Navarrete S, Solar C, Tapia R, Pereira J, Fuentes E, Palomo I. Pathophysiology of deep vein thrombosis. *Clinical and Experimental Medicine 2022*. Published online April 26, 2022:1-10. doi:10.1007/S10238-022-00829-W
147. Pillon NJ, Loos RJF, Marshall SM, Zierath JR. Metabolic consequences of obesity and type 2 diabetes: Balancing genes and environment for personalized care. *Cell*. 2021;184(6):1530. doi:10.1016/J.CELL.2021.02.012

148. Gray IC, Campbell DA, Spurr NK. Single nucleotide polymorphisms as tools in human genetics. *Hum Mol Genet*. 2000;9(16):2403-2408. doi:10.1093/HMG/9.16.2403
149. Astle WJ, Elding H, Jiang T, et al. The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. *Cell*. 2016;167(5):1415-1429.e19. doi:10.1016/j.cell.2016.10.042
150. Gibbs RA. The Human Genome Project changed everything. *Nature Reviews Genetics 2020 21:10*. 2020;21(10):575-576. doi:10.1038/s41576-020-0275-3
151. Belmont JW, Hardenbol P, Willis TD, et al. The international HapMap project. *Nature*. 2003;426(6968):789-796. doi:10.1038/nature02168
152. Altshuler DL, Durbin RM, Abecasis GR, et al. A map of human genome variation from population-scale sequencing. *Nature*. 2010;467(7319):1061-1073. doi:10.1038/nature09534
153. Abdellaoui A, Yengo L, Verweij KJH, Visscher PM. 15 years of GWAS discovery: Realizing the promise. *The American Journal of Human Genetics*. 2023;110(2):179-194. doi:10.1016/j.ajhg.2022.12.011
154. Uffelmann E, Huang QQ, Munung NS, et al. Genome-wide association studies. *Nature Reviews Methods Primers 2021 1:1*. 2021;1(1):1-21. doi:10.1038/s43586-021-00056-9
155. Yazar S, Alquicira-Hernandez J, Wing K, et al. Single-cell eQTL mapping identifies cell type-specific genetic control of autoimmune disease. *Science (1979)*. 2022;376(6589). doi:10.1126/SCIENCE.ABF3041/SUPPL_FILE/SCIENCE.ABF3041_MDAR_REPRODUCIBILITY_CHECKLIST.PDF
156. Vuckovic D, Bao EL, Akbari P, et al. The Polygenic and Monogenic Basis of Blood Traits and Diseases. *Cell*. 2020;182(5):1214-1231.e11. doi:10.1016/J.CELL.2020.08.008/ATTACHMENT/347CE04A-7337-4664-BB5B-5ED6234B8F9E/MMC11.DOCX
157. Klein RJ, Zeiss C, Chew EY, et al. Complement Factor H Polymorphism in Age-Related Macular Degeneration. *Science*. 2005;308(5720):385. doi:10.1126/SCIENCE.1109557
158. Andrews NC. Genes determining blood cell traits. *Nature Genetics 2009 41:11*. 2009;41(11):1161-1162. doi:10.1038/ng1109-1161
159. Tong O, Fairfax BP. Dissecting genetic determinants of variation in human immune responses. *Curr Opin Immunol*. 2020;65:74-78. doi:10.1016/j.coi.2020.05.005

160. Reich D, Nalls MA, Kao WHL, et al. Reduced neutrophil count in people of African descent is due to a regulatory variant in the Duffy antigen receptor for chemokines gene. *PLoS Genet*. 2009;5(1). doi:10.1371/journal.pgen.1000360
161. Sudlow C, Gallacher J, Allen N, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med*. 2015;12(3):e1001779-e1001779. doi:10.1371/journal.pmed.1001779
162. Gaziano JM, Concato J, Brophy M, et al. Million Veteran Program: A mega-biobank to study genetic influences on health and disease. *J Clin Epidemiol*. 2016;70:214-223. doi:10.1016/J.JCLINEPI.2015.09.016
163. Huyghe JR, Bien SA, Harrison TA, et al. Discovery of common and rare genetic risk variants for colorectal cancer. *Nat Genet*. 2019;51(1):76. doi:10.1038/S41588-018-0286-6
164. Bycroft C, Freeman C, Petkova D, et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature*. 2018;562(7726):203-209. doi:10.1038/s41586-018-0579-z
165. Boyle EA, Li YI, Pritchard JK. An expanded view of complex traits: from polygenic to omnigenic. *Cell*. 2017;169(7):1177. doi:10.1016/J.CELL.2017.05.038
166. Chen MH, Raffield LM, Mousas A, et al. Trans-ethnic and Ancestry-Specific Blood-Cell Genetics in 746,667 Individuals from 5 Global Populations. *Cell*. 2020;182(5):1198-1213.e14. doi:10.1016/j.cell.2020.06.045
167. Buniello A, Macarthur JAL, Cerezo M, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res*. 2019;47. doi:10.1093/NAR/GKY1120
168. Panoutsopoulou K, Wheeler E. Key Concepts in Genetic Epidemiology. *Methods in Molecular Biology*. 2018;1793:7-24. doi:10.1007/978-1-4939-7868-7_2
169. Davey Smith G, Ebrahim S, Smith GD, Ebrahim S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *Int J Epidemiol*. 2003;32(1):1-22. doi:10.1093/ije/dyg070
170. Ebrahim S, Davey Smith G. Mendelian randomization: Can genetic epidemiology help redress the failures of observational epidemiology? *Hum Genet*. 2008;123(1):15-33. doi:10.1007/S00439-007-0448-6/TABLES/5
171. von Hinke S, Davey Smith G, Lawlor DA, Propper C, Windmeijer F. Genetic markers as instrumental variables. *J Health Econ*. 2016;45:131-148. doi:10.1016/J.JHEALECO.2015.10.007
172. Burgess S, Butterworth A, Malarstig A, Thompson SG. Use of Mendelian randomisation to assess potential benefit of clinical intervention. *BMJ*. 2012;345. doi:10.1136/BMJ.E7325

173. Bowden J, Hemani G, Davey Smith G. Invited Commentary: Detecting Individual and Global Horizontal Pleiotropy in Mendelian Randomization—A Job for the Humble Heterogeneity Statistic? *Am J Epidemiol*. 2018;187(12):2681-2685. doi:10.1093/AJE/KWY185
174. Eiriksdottir G, Harris TB, Launer LJ, et al. Association between C reactive protein and coronary heart disease: mendelian randomisation analysis based on individual participant data. *BMJ*. 2011;342(7794):425. doi:10.1136/BMJ.D548
175. Kaptoge S, Di Angelantonio E, Lowe G, et al. C-reactive protein concentration and risk of coronary heart disease, stroke, and mortality: an individual participant meta-analysis. *Lancet*. 2010;375(9709):132. doi:10.1016/S0140-6736(09)61717-7
176. Voight BF, Peloso GM, Orho-Melander M, et al. Plasma HDL cholesterol and risk of myocardial infarction: A mendelian randomisation study. *The Lancet*. 2012;380(9841):572-580. doi:10.1016/S0140-6736(12)60312-2
177. Yarmolinsky J, Bonilla C, Haycock PC, et al. Circulating Selenium and Prostate Cancer Risk: A Mendelian Randomization Analysis. *JNCI Journal of the National Cancer Institute*. 2018;110(9):1035. doi:10.1093/JNCI/DJY081
178. Lippman SM, Klein EA, Goodman PJ, et al. Effect of Selenium and Vitamin E on Risk of Prostate Cancer and Other Cancers: The Selenium and Vitamin E Cancer Prevention Trial (SELECT). *JAMA*. 2009;301(1):39-51. doi:10.1001/JAMA.2008.864
179. Rothman KJ, Greenland S. Causation and Causal Inference in Epidemiology. *https://doi.org/102105/AJPH2004059204*. 2005;95(SUPPL. 1). doi:10.2105/AJPH.2004.059204
180. Rezigalla AA, Rezigalla AA. Observational Study Designs: Synopsis for Selecting an Appropriate Study Design. *Cureus*. 2020;12(1). doi:10.7759/CUREUS.6692
181. Noordzij M, Dekker FW, Zoccali C, Jager KJ. Study designs in clinical research. *Nephron Clin Pract*. 2009;113(3):218-221. doi:10.1159/000235610
182. DiPietro NA. Methods in epidemiology: Observational study designs. *Pharmacotherapy*. 2010;30(10):973-984. doi:10.1592/PHCO.30.10.973
183. Faraoni D, Schaefer ST. Randomized controlled trials vs. observational studies: why not just live together? *BMC Anesthesiol*. 2016;16(1). doi:10.1186/S12871-016-0265-3
184. Yarmolinsky J, Wade KH, Richmond RC, et al. Causal inference in cancer epidemiology: What is the role of mendelian randomization? *Cancer Epidemiology Biomarkers and Prevention*. 2018;27(9):995-1010. doi:10.1158/1055-9965.EPI-17-1177/70010/AM/CAUSAL-INFERENCE-IN-CANCER-EPIDEMIOLOGY-WHAT-IS

185. Millard LAC, Davies NM, Timpson NJ, Tilling K, Flach PA, Smith GD. MR-PheWAS: hypothesis prioritization among potential causal effects of body mass index on many outcomes, using Mendelian randomization. *Scientific Reports 2015 5:1*. 2015;5(1):1-17. doi:10.1038/srep16645
186. Zhao SS, MacKie SL, Zheng J. Why clinicians should know about Mendelian randomization. *Rheumatology*. 2021;60(4):1577-1579. doi:10.1093/RHEUMATOLOGY/KEAB007
187. Haycock PC, Burgess S, Wade KH, Bowden J, Relton C, Davey Smith G. Best (but oft-forgotten) practices: the design, analysis, and interpretation of Mendelian randomization studies. *Am J Clin Nutr*. 2016;103(4):965-978. doi:10.3945/ajcn.115.118216
188. Holmberg MJ, Andersen LW. Collider Bias. *JAMA*. 2022;327(13):1282-1283. doi:10.1001/JAMA.2022.1820
189. Lawlor Debbie A, Harbord Roger M, Sterne Jonathan AC, et al. Mendelian randomization: Using genes as instruments for making causal inferences in epidemiology. *Stat Med*. 2008;27(8):1133-1163. doi:10.1002/sim.3034
190. Gala H, Tomlinson I. The use of Mendelian randomisation to identify causal cancer risk factors: promise and limitations. *J Pathol*. 2020;250(5):541-554. doi:10.1002/PATH.5421
191. Davey Smith G, Hemani G. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Hum Mol Genet*. 2014;23(R1):R89-R98. doi:10.1093/hmg/ddu328
192. de Leeuw C, Savage J, Bucur IG, Heskes T, Posthuma D. Understanding the assumptions underlying Mendelian randomization. *European Journal of Human Genetics 2022 30:6*. 2022;30(6):653-660. doi:10.1038/s41431-022-01038-5
193. Burgess S, Scott RA, Timpson NJ, Davey Smith G, Thompson SG, Consortium EI. Using published data in Mendelian randomization: a blueprint for efficient identification of causal risk factors. *Eur J Epidemiol*. 2015;30(7):543-552. doi:10.1007/s10654-015-0011-z
194. Davies NM, Holmes M V., Davey Smith G. Reading Mendelian randomisation studies: a guide, glossary, and checklist for clinicians. *BMJ (Online)*. 2018;362:601. doi:10.1136/bmj.k601
195. Burgess S, Davey Smith G, Davies NM, et al. Guidelines for performing Mendelian randomization investigations. *Wellcome Open Research 2020 4:186*. 2020;4:186. doi:10.12688/wellcomeopenres.15555.2
196. Burgess S, Davies NM, Thompson SG. Bias due to participant overlap in two-sample Mendelian randomization. *Genet Epidemiol*. 2016;40(7):597. doi:10.1002/GEPI.21998

197. Zheng J, Baird D, Borges MC, et al. Recent Developments in Mendelian Randomization Studies. *Curr Epidemiol Rep*. 2017;4(4):330-345. doi:10.1007/s40471-017-0128-6
198. Burgess S, Thompson SG. Use of allele scores as instrumental variables for Mendelian randomization. *Int J Epidemiol*. 2013;42(4):1134. doi:10.1093/IJE/DYT093
199. Zhao Q, Wang J, Hemani G, Bowden J, Small DS. STATISTICAL INFERENCE IN TWO-SAMPLE SUMMARY-DATA MENDELIAN RANDOMIZATION USING ROBUST ADJUSTED PROFILE SCORE. 2020;48(3):1742-1769. doi:10.1214/19-AOS1866
200. Lawlor DA. Commentary: Two-sample Mendelian randomization: opportunities and challenges. *Int J Epidemiol*. 2016;45(3):908-915. doi:10.1093/ije/dyw127
201. Sadreev II, Elsworth BL, Mitchell RE, et al. Navigating sample overlap, winner's curse and weak instrument bias in Mendelian randomization studies using the UK Biobank. *medRxiv*. Published online July 1, 2021:2021.06.28.21259622. doi:10.1101/2021.06.28.21259622
202. Burgess S, Dudbridge F, Thompson SG. Combining information on multiple instrumental variables in Mendelian randomization: comparison of allele score and summarized data methods. *Stat Med*. 2016;35(11):1880-1906. doi:https://doi.org/10.1002/sim.6835
203. Bowden J, Fabiola Del Greco M, Minelli C, Smith GD, Sheehan NA, Thompson JR. Assessing the suitability of summary data for two-sample Mendelian randomization analyses using MR-Egger regression: the role of the [... formula ...] statistic. *Int J Epidemiol*. 2016;45(6):1961. doi:10.1093/IJE/DYW220
204. Bowden J, Davey Smith G, Haycock PC, Burgess S. Consistent Estimation in Mendelian Randomization with Some Invalid Instruments Using a Weighted Median Estimator. *Genet Epidemiol*. 2016;40(4):304-314. doi:https://doi.org/10.1002/gepi.21965
205. Bowden J, Davey Smith G, Burgess S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *Int J Epidemiol*. 2015;44(2):512-525. doi:10.1093/ije/dyv080
206. Burgess S, Thompson SG. *Bias in Causal Estimates from Mendelian Randomization Studies with Weak Instruments*.; 2010.
207. Verbanck M, Chen CY, Neale B, Do R. Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. *Nature Genetics 2018 50:5*. 2018;50(5):693-698. doi:10.1038/s41588-018-0099-7

208. Bowden J, Del Greco M F, Minelli C, et al. Improving the accuracy of two-sample summary-data Mendelian randomization: moving beyond the NOME assumption. *Int J Epidemiol*. 2019;48(3):728-742. doi:10.1093/IJE/DYY258
209. Egger M, Smith GD, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. *BMJ : British Medical Journal*. 1997;315(7109):629. doi:10.1136/BMJ.315.7109.629
210. Burgess S, Thompson SG. Interpreting findings from Mendelian randomization using the MR-Egger method. *Eur J Epidemiol*. 2017;32(5):377. doi:10.1007/S10654-017-0255-X
211. Slob EAW, Burgess S. A comparison of robust Mendelian randomization methods using summary data. *Genet Epidemiol*. 2020;44(4):313-329. doi:10.1002/GEPI.22295
212. Hartwig FP, Davey Smith G, Bowden J. Robust inference in summary data Mendelian randomization via the zero modal pleiotropy assumption. *Int J Epidemiol*. 2017;46(6):1985-1998. doi:10.1093/ije/dyx102
213. Hemani G, Tilling K, Davey Smith G. Orienting the causal relationship between imprecisely measured traits using GWAS summary data. *PLoS Genet*. 2017;13(11). doi:10.1371/JOURNAL.PGEN.1007081
214. Hemani G, Bowden J, Haycock PC, et al. Automating Mendelian randomization through machine learning to construct a putative causal map of the human phenome. *bioRxiv*. Published online 2017.
215. Burgess S, Thompson SG. Multivariable Mendelian Randomization: The Use of Pleiotropic Genetic Variants to Estimate Causal Effects. *Am J Epidemiol*. 2015;181(4):251. doi:10.1093/AJE/KWU283
216. Sanderson E, Davey Smith G, Windmeijer F, Bowden J. An examination of multivariable Mendelian randomization in the single-sample and two-sample summary data settings. *Int J Epidemiol*. 2019;48(3):713-727. doi:10.1093/ije/dyy262
217. Sanderson E, Spiller W, Bowden J. Testing and correcting for weak and pleiotropic instruments in two-sample multivariable Mendelian randomization. *Stat Med*. 2021;40(25):5434-5452. doi:10.1002/SIM.9133
218. Carter AR, Sanderson E, Hammerton G, et al. Mendelian randomisation for mediation analysis: current methods and challenges for implementation. *European Journal of Epidemiology 2021 36:5*. 2021;36(5):465-478. doi:10.1007/S10654-021-00757-1
219. Hemani G, Zheng J, Elsworth B, et al. The MR-base platform supports systematic causal inference across the human phenome. *Elife*. 2018;7.

220. Chen W, Wu Y, Zheng Z, et al. Improved analyses of GWAS summary statistics by reducing data heterogeneity and errors. *Nature Communications 2021 12:1*. 2021;12(1):1-10. doi:10.1038/s41467-021-27438-7
221. Alexander DH, Shringarpure SS, Novembre J, Lange K. *Admixture 1.3 Software Manual*.; 2020.
222. Alexander DH, Lange K. Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC Bioinformatics*. 2011;12:246. doi:10.1186/1471-2105-12-246
223. Abegaz F, Chaichoompu K, Génin E, et al. Principals about principal components in statistical genetics. *Brief Bioinform*. 2019;20(6):2200-2216. doi:10.1093/BIB/BBY081
224. Menozzi P, Piazza A, Cavalli-Sforza L. Synthetic maps of human gene frequencies in Europeans. *Science (1979)*. 1978;201(4358):786-792. doi:10.1126/science.356262
225. Li JZ, Absher DM, Tang H, et al. Worldwide human relationships inferred from genome-wide patterns of variation. *Science (1979)*. 2008;319(5866):1100-1104. doi:10.1126/science.1153717
226. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. *PLoS Genet*. 2006;2(12):2074-2093. doi:10.1371/journal.pgen.0020190
227. Price AL, Zaitlen NA, Reich D, Patterson N. New approaches to population stratification in genome-wide association studies. *Nat Rev Genet*. 2010;11(7):459. doi:10.1038/NRG2813
228. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet*. 2006;38(8):904-909. doi:10.1038/ng1847
229. Patterson N, Moorjani P, Luo Y, et al. Ancient Admixture in Human History. *Genetics*. 2012;192(3):1065. doi:10.1534/GENETICS.112.145037
230. Sinaga KP, Yang MS. Unsupervised K-means clustering algorithm. *IEEE Access*. 2020;8:80716-80727. doi:10.1109/ACCESS.2020.2988796
231. Meirmans PG, Hedrick PW. Assessing population structure: FST and related measures. *Mol Ecol Resour*. 2011;11(1):5-18. doi:10.1111/J.1755-0998.2010.02927.X
232. Marchini J, Howie B. Genotype imputation for genome-wide association studies. *Nat Rev Genet*. 2010;11(7):499-511. doi:10.1038/nrg2796
233. Willer CJ, Li Y, Abecasis GR. METAL: Fast and efficient meta-analysis of genomewide association scans. *Bioinformatics*. 2010;26(17):2190-2191. doi:10.1093/BIOINFORMATICS/BTQ340

234. Loh PR, Kichaev G, Gazal S, Schoech AP, Price AL. Mixed-model association for biobank-scale datasets. *Nature Genetics 2018 50:7*. 2018;50(7):906-908. doi:10.1038/s41588-018-0144-6
235. VanRaden PM. Efficient Methods to Compute Genomic Predictions. *J Dairy Sci*. 2008;91(11):4414-4423. doi:10.3168/JDS.2007-0980
236. Sheard S, Nicholls R, Froggatt J. UK Biobank Haematology Data Companion Document.
237. Constantinescu AE, Bull CJ, Jones N, et al. Circulating white blood cell traits and colorectal cancer risk: A Mendelian randomization study. *medRxiv*. 2023;17:2023.03.03.23286764. doi:10.1101/2023.03.03.23286764
238. Constantinescu AE, Mitchell RE, Zheng J, et al. A framework for research into continental ancestry groups of the UK Biobank. *Human Genomics 2022 16:1*. 2022;16(1):1-14. doi:10.1186/S40246-022-00380-5
239. Huyghe JR, Harrison TA, Bien SA, et al. Genetic architectures of proximal and distal colorectal cancer are partly distinct. *Gut*. 2021;70(7):1325-1334. doi:10.1136/GUTJNL-2020-321534
240. Peters U, Jiao S, Schumacher FR, et al. Identification of Genetic Susceptibility Loci for Colorectal Tumors in aGenome-Wide Meta-analysis. *Gastroenterology*. 2013;144(4). doi:10.1053/j.gastro.2012.12.020
241. Schumacher FR, Schmit SL, Jiao S, et al. Genome-wide association study of colorectal cancer identifies six new susceptibility loci. *Nature Communications 2015 6:1*. 2015;6(1):1-7. doi:10.1038/ncomms8138
242. Schmit SL, Edlund CK, Schumacher FR, et al. Novel Common Genetic Susceptibility Loci for Colorectal Cancer. *JNCI: Journal of the National Cancer Institute*. 2019;111(2):146-157. doi:10.1093/JNCI/DJY099
243. Zhang B, Jia WH, Matsuda K, et al. Large-scale genetic study in East Asians identifies six new loci associated with colorectal cancer risk. *Nat Genet*. 2014;46(6):533. doi:10.1038/NG.2985
244. Ferreira MA, Vonk JM, Baurecht H, et al. Shared genetic origin of asthma, hay fever and eczema elucidates allergic disease biology. *Nature Genetics 2017 49:12*. 2017;49(12):1752-1757. doi:10.1038/ng.3985
245. Malaria Genomic Epidemiology Network. Insights into malaria susceptibility using genome-wide data on 17,000 individuals from Africa, Asia and Oceania. *Nat Commun*. 2019;10(1):5732. doi:10.1038/s41467-019-13480-z
246. WHO Africa. Severe Malaria. Published online 2014. doi:10.1111/tmi.12313
247. Purcell S, Neale B, Todd-Brown K, et al. PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*. 2007;81(3):559-575. doi:10.1086/519795

248. Auton A, Abecasis GR, Altshuler DM, et al. A global reference for human genetic variation. *Nature*. 2015;526(7571):68-74. doi:10.1038/nature15393
249. Constantinescu AE, Bull CJ, Goudswaard LJ, et al. A phenome-wide approach to identify causal risk factors for deep vein thrombosis. *bioRxiv*. Published online May 6, 2022:476135. doi:10.1101/476135
250. Wootton RE, Sallis HM. Let's call it the effect allele: a suggestion for GWAS naming conventions. *Int J Epidemiol*. 2020;49(5):1734-1735. doi:10.1093/IJE/DYAA149
251. Elsworth B, Lyon M, Alexander T, et al. The MRC IEU OpenGWAS data infrastructure. *bioRxiv*. Published online August 10, 2020:2020.08.10.244293. doi:10.1101/2020.08.10.244293
252. Hartwig FP, Davies NM, Hemani G, Smith GD. Two-sample Mendelian randomization: avoiding the downsides of a powerful, widely applicable but potentially fallible technique. *Int J Epidemiol*. 2016;45(6):1717-1726.
253. Smith GD, Ebrahim S. SEER. Surveillance, Epidemiology, and End Results (SEER) Program. *Int J Epidemiol*. 2003;32(1):1-22. doi:10.1093/ije/dyg070
254. Demb J, Earles A, Martínez ME, et al. Risk factors for colorectal cancer significantly vary by anatomic site. 2019;6(1).
255. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. 2018;68(6):394-424. doi:10.3322/CAAC.21492
256. Sung H, Siegel RL, Rosenberg PS, Jemal A. Emerging cancer trends among young adults in the USA: analysis of a population-based cancer registry. *Lancet Public Health*. 2019;4(3):e137-e147. doi:10.1016/S2468-2667(18)30267-6/ATTACHMENT/EA91DB97-1C30-4C2F-B604-48BF68319153/MMC1.PDF
257. Mauri G, Sartore-Bianchi A, Russo AG, Marsoni S, Bardelli A, Siena S. Early-onset colorectal cancer in young individuals. *Mol Oncol*. 2019;13(2):109. doi:10.1002/1878-0261.12417
258. Xi Y, Xu P. Global colorectal cancer burden in 2020 and projections to 2040. *Transl Oncol*. 2021;14(10):101174. doi:10.1016/J.TRANON.2021.101174
259. Sawicki T, Ruszkowska M, Danielewicz A, Niedźwiedzka E, Arłukowicz T, Przybyłowicz KE. A Review of Colorectal Cancer in Terms of Epidemiology, Risk Factors, Development, Symptoms and Diagnosis. *Cancers (Basel)*. 2021;13(9). doi:10.3390/CANCERS13092025
260. Hull MA. Nutritional prevention of colorectal cancer. *Proceedings of the Nutrition Society*. 2021;80(1):59-64. doi:10.1017/S0029665120000051

261. Li FY, Lai MD. Colorectal cancer, one entity or three *. *J Zhejiang Univ Sci B*. 2009;10(3):219-229. doi:10.1631/jzus.B0820273
262. Ellis H, Mahadevan V. Anatomy of the caecum, appendix and colon. *Surgery (Oxford)*. 2014;32(4):155-158. doi:10.1016/J.MPSUR.2014.02.001
263. Biller LH, Schrag D. Diagnosis and Treatment of Metastatic Colorectal Cancer: A Review. *JAMA*. 2021;325(7):669-685. doi:10.1001/JAMA.2021.0106
264. Nakagawa-Senda H, Hori M, Matsuda T, Ito H. Prognostic impact of tumor location in colon cancer: The Monitoring of Cancer Incidence in Japan (MCIJ) project. *BMC Cancer*. 2019;19(1):1-9. doi:10.1186/S12885-019-5644-Y/TABLES/3
265. Rawla P, Sunkara T, Barsouk A. Epidemiology of colorectal cancer: incidence, mortality, survival, and risk factors. *Prz Gastroenterol*. 2019;14(2):89. doi:10.5114/PG.2018.81072
266. Keum NN, Giovannucci E. Global burden of colorectal cancer: emerging trends, risk factors and prevention strategies. *Nature Reviews Gastroenterology & Hepatology 2019 16:12*. 2019;16(12):713-732. doi:10.1038/s41575-019-0189-8
267. Safiri S, Sepanlou SG, Ikuta KS, et al. The global, regional, and national burden of colorectal cancer and its attributable risk factors in 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. *Lancet Gastroenterol Hepatol*. 2019;4(12):913-933. doi:10.1016/S2468-1253(19)30345-0
268. Aicr, WCRF. Diet, nutrition, physical activity and colorectal cancer.
269. Xiao W, Huang J, Zhao C, Ding L, Wang X, Wu B. Diabetes and Risks of Right-Sided and Left-Sided Colon Cancer: A Meta-Analysis of Prospective Cohorts. *Front Oncol*. 2022;12:1466. doi:10.3389/FONC.2022.737330/BIBTEX
270. Guttmacher AE, Collins FS, Lynch HT, De La Chapelle A. Hereditary Colorectal Cancer. Guttmacher AE, Collins FS, eds. *https://doi.org/101056/NEJMra012242*. 2003;348(10):919-932. doi:10.1056/NEJMRA012242
271. Peltomäki P, Olkinuora A, Nieminen TT. Updates in the field of hereditary nonpolyposis colorectal cancer. *https://doi.org/101080/1747412420201782187*. 2020;14(8):707-720. doi:10.1080/17474124.2020.1782187
272. Belhadj S, Terradas M, Munoz-Torres PM, et al. Candidate genes for hereditary colorectal cancer: Mutational screening and systematic review. *Hum Mutat*. 2020;41(9):1563-1576. doi:10.1002/HUMU.24057
273. Tomlinson I, Webb E, Carvajal-Carmona L, et al. A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21. *Nature Genetics 2007 39:8*. 2007;39(8):984-988. doi:10.1038/ng2085
274. Newcomb PA, Baron J, Cotterchio M, et al. Colon cancer family registry: An international resource for studies of the genetic epidemiology of colon cancer.

Cancer Epidemiology Biomarkers and Prevention. 2007;16(11):2331-2343. doi:10.1158/1055-9965.EPI-07-0648/347675/P/COLON-CANCER-FAMILY-REGISTRY-AN-INTERNATIONAL

275. Schmitt M, Greten FR. The inflammatory pathogenesis of colorectal cancer. *Nature Reviews Immunology 2021 21:10*. 2021;21(10):653-667. doi:10.1038/s41577-021-00534-x
276. Zhong X, He X, Wang Y, et al. Warburg effect in colorectal cancer: the emerging roles in tumor microenvironment and therapeutic implications. *Journal of Hematology & Oncology 2022 15:1*. 2022;15(1):1-29. doi:10.1186/S13045-022-01358-5
277. Tuomisto AE, Mäkinen MJ, Väyrynen JP. Systemic inflammation in colorectal cancer: Underlying factors, effects, and prognostic significance. *World J Gastroenterol*. 2019;25(31):4383. doi:10.3748/WJG.V25.I31.4383
278. Nicholson LB. The immune system. *Essays Biochem*. 2016;60(3):275-301. doi:10.1042/EBC20160017
279. Wong JYY, Bassig BA, Loftfield E, et al. White Blood Cell Count and Risk of Incident Lung Cancer in the UK Biobank. *JNCI Cancer Spectr*. 2020;4(2). doi:10.1093/JNCICS/PKZ102
280. Watts EL, Perez-Cornago A, Kothari J, Allen NE, Travis RC, Key TJ. Hematologic markers and prostate cancer risk: a prospective analysis in UK Biobank. *Cancer Epidemiology Biomarkers and Prevention*. 2020;29(8):1615-1626. doi:10.1158/1055-9965.EPI-19-1525/70795/AM/HAEMATOLOGICAL-MARKERS-AND-PROSTATE-CANCER-RISK-A
281. Siedlinski M, Jozefczuk E, Xu X, et al. Siedlinski et al White Blood Cells and Blood Pressure Indices. *Circulation*. 2020;141(16):1307. doi:10.1161/CIRCULATIONAHA.119.045102
282. Welsh C, Welsh P, Mark PB, et al. Association of total and differential leukocyte counts with cardiovascular disease and mortality in the UK Biobank. *Arterioscler Thromb Vasc Biol*. 2018;38(6):1415-1423. doi:10.1161/ATVBAHA.118.310945
283. Wei Y, Zhang X, Wang G, et al. The impacts of pretreatment circulating eosinophils and basophils on prognosis of stage Ⅰ–Ⅲ colorectal cancer. *Asia Pac J Clin Oncol*. 2018;14(5):e243-e251. doi:10.1111/AJCO.12871
284. Idos GE, Kwok J, Bonthala N, Kysh L, Gruber SB, Qu C. The Prognostic Implications of Tumor Infiltrating Lymphocytes in Colorectal Cancer: A Systematic Review and Meta-Analysis. *Scientific Reports 2020 10:1*. 2020;10(1):1-14. doi:10.1038/s41598-020-60255-4
285. Lee YJ, Lee HR, Nam CM, Hwang UK, Jee SH. White Blood Cell Count and the Risk of Colon Cancer. *Yonsei Med J*. 2006;47(5).

286. Wu J, Ge X, Zhu W, et al. Values of applying white blood cell counts in the prognostic evaluation of resectable colorectal cancer. *Mol Med Rep*. 2019;19(3):2330-2340. doi:10.3892/MMR.2019.9844
287. Watt DG, Martin JC, Park JH, Horgan PG, McMillan DC. Neutrophil count is the most important prognostic component of the differential white cell count in patients undergoing elective surgery for colorectal cancer. 2015;210(1):24-30.
288. Prizment AE, Vierkant RA, Smyrk TC, et al. Tumor eosinophil infiltration and improved survival of colorectal cancer patients: Iowa Women's Health Study. *Modern Pathology*. 2016;29(5):516-527. doi:10.1038/modpathol.2016.42
289. Rosman Y, Hornik-Lurie T, Meir-Shafrir K, et al. Changes in peripheral blood eosinophils may predict colorectal cancer – A retrospective study. *World Allergy Organization Journal*. 2022;15(10):100696. doi:10.1016/J.WAOJOU.2022.100696
290. Iijima K, Fujibayashi K, Okumura M, Sasabe N, Gunji T. Number of Eosinophils and Incidence of Cancer in a Japanese Population : A Single Institution Study. *Original Article Ningen Dock International*. 2019;6(1):56-61.
291. Goshen R, Mizrahi B, Akiva P, et al. Predicting the presence of colon cancer in members of a health maintenance organisation by evaluating analytes from standard laboratory records. *British Journal of Cancer 2017 116:7*. 2017;116(7):944-950. doi:10.1038/bjc.2017.53
292. Boursi B, Mamtani R, Hwang WT, Haynes K, Yang YX. A Risk Prediction Model for Sporadic CRC Based on Routine Lab Results. *Dig Dis Sci*. 2016;61(7):2076-2086. doi:10.1007/S10620-016-4081-X/TABLES/4
293. Humphry E, Armstrong CE. Physiology of red and white blood cells. *Anaesthesia & Intensive Care Medicine*. 2022;23(2):118-122. doi:10.1016/J.MPAIC.2021.10.019
294. Liu Q, Luo D, Cai S, Li Q, Li X. Circulating basophil count as a prognostic marker of tumor aggressiveness and survival outcomes in colorectal cancer. *Clinical and Translational Medicine 2020 9:1*. 2020;9(1):1-12. doi:10.1186/S40169-019-0255-4
295. Ohmori K, Luo Y, Jia Y, et al. IL-3 Induces Basophil Expansion In Vivo by Directing Granulocyte-Monocyte Progenitors to Differentiate into Basophil Lineage-Restricted Progenitors in the Bone Marrow and by Increasing the Number of Basophil/Mast Cell Progenitors in the Spleen. *The Journal of Immunology*. 2009;182(5):2835-2841. doi:10.4049/JIMMUNOL.0802870
296. Merluzzi S, Betto E, Ceccaroni AA, Magris R, Giunta M, Mion F. Mast cells, basophils and B cell connection network. *Mol Immunol*. 2015;63(1):94-103. doi:10.1016/J.MOLIMM.2014.02.016

297. Sektioglu IM, Carretero R, Bulbuc N, et al. Basophils promote tumor rejection via chemotaxis and infiltration of CD8+ T cells. *Cancer Res*. 2017;77(2):291-302. doi:10.1158/0008-5472.CAN-16-0993/652451/AM/BASOPHILS-PROMOTE-TUMOR-REJECTION-VIA-CHEMOTAXIS
298. Mantovani A, Marchesi F, Malesci A, Laghi L, Allavena P. Tumor-Associated Macrophages as Treatment Targets in Oncology. *Nat Rev Clin Oncol*. 2017;14(7):399. doi:10.1038/NRCLINONC.2016.217
299. De Monte L, Wörmann S, Brunetto E, et al. Basophil recruitment into tumor-draining lymph nodes correlates with Th2 inflammation and reduced survival in pancreatic cancer patients. *Cancer Res*. 2016;76(7):1792-1803. doi:10.1158/0008-5472.CAN-15-1801-T/652039/AM/BASOPHIL-RECRUITMENT-INTO-TUMOR-DRAINING-LYMPH
300. Prizment AE, Anderson KE, Visvanathan K, Folsom AR. Inverse association of eosinophil count with colorectal cancer incidence: Atherosclerosis Risk in Communities study. *Cancer Epidemiology Biomarkers and Prevention*. 2011;20(9):1861-1864. doi:10.1158/1055-9965.EPI-11-0360/66619/AM/INVERSE-ASSOCIATION-OF-EOSINOPHIL-COUNT-WITH
301. Loktionov A. Eosinophils in the gastrointestinal tract and their role in the pathogenesis of major colorectal disorders. *World J Gastroenterol*. 2019;25(27):3503. doi:10.3748/WJG.V25.I27.3503
302. Virdee PS, Patnick J, Watkinson P, Holt T, Birks J. Full Blood Count Trends for Colorectal Cancer Detection in Primary Care: Development and Validation of a Dynamic Prediction Model. *Cancers (Basel)*. 2022;14(19):4779. doi:10.3390/CANCERS14194779/S1
303. Harbaum L, Pollheimer MJ, Kornprat P, Lindtner RA, Bokemeyer C, Langner C. Peritumoral eosinophils predict recurrence in colorectal cancer. *Modern Pathology 2015 28:3*. 2014;28(3):403-413. doi:10.1038/modpathol.2014.104
304. Vayrynen JP, Lau MC, Haruki K, et al. Prognostic Significance of Immune Cell Populations Identified by Machine Learning in Colorectal Cancer Using Routine Hematoxylin and Eosin–Stained Sections. *Clinical Cancer Research*. 2020;26(16):4326-4338. doi:10.1158/1078-0432.CCR-20-0071/77128/AM/PROGNOSTIC-SIGNIFICANCE-OF-IMMUNE-CELL-POPULATIONS
305. Feng X, Jiao X, Xu Y, et al. The predictive value of routine laboratory tests for colorectal polyps: a retrospective study. *J Gastrointest Oncol*. 2022;13(1):256-264. doi:10.21037/JGO-21-933/COIF
306. Legrand F, Driss V, Delbeke M, et al. Human Eosinophils Exert TNF-α and Granzyme A-Mediated Tumoricidal Activity toward Colon Carcinoma Cells. *The*

Journal of Immunology. 2010;185(12):7443-7451. doi:10.4049/JIMMUNOL.1000446

307. Gatault S, Delbeke M, Driss V, et al. IL-18 Is Involved in Eosinophil-Mediated Tumoricidal Activity against a Colon Carcinoma Cell Line by Upregulating LFA-1 and ICAM-1. *The Journal of Immunology*. 2015;195(5):2483-2492. doi:10.4049/JIMMUNOL.1402914
308. Kienzl M, Hasenoehrl C, Valadez-Cosmes P, et al. IL-33 reduces tumor growth in models of colorectal cancer with the help of eosinophils. *Oncoimmunology*. 2020;9(1). doi:10.1080/2162402X.2020.1776059/SUPPL_FILE/KONI_A_1776059_SM1922.DOCX
309. Reichman H, Itan M, Rozenberg P, et al. Activated eosinophils exert antitumorigenic activities in colorectal cancer. *Cancer Immunol Res*. 2019;7(3):388-400. doi:10.1158/2326-6066.CIR-18-0494/471605/P/ACTIVATED-EOSINOPHILS-EXERT-ANTITUMORIGENIC
310. Cooper MD, Alder MN. The Evolution of Adaptive Immune Systems. *Cell*. 2006;124(4):815-822. doi:10.1016/J.CELL.2006.02.001
311. Wu YY, Zhang X, Qin YY, Qin JQ, Lin FQ. Mean platelet volume/platelet count ratio in colorectal cancer: A retrospective clinical study. *BMC Cancer*. 2019;19(1):1-7. doi:10.1186/S12885-019-5504-9/FIGURES/2
312. Yang J, Guo X, Wang M, Ma X, Ye X, Lin P. Pre-treatment inflammatory indexes as predictors of survival and cetuximab efficacy in metastatic colorectal cancer patients with wild-type RAS. *Sci Rep*. 2017;7(1). doi:10.1038/S41598-017-17130-6
313. Tanio A, Saito H, Uejima C, et al. A prognostic index for colorectal cancer based on preoperative absolute lymphocyte, monocyte, and neutrophil counts. *Surg Today*. 2019;49(3):245-253. doi:10.1007/S00595-018-1728-6/TABLES/2
314. Maher J, Davies ET. Targeting cytotoxic T lymphocytes for cancer immunotherapy. *Br J Cancer*. 2004;91(5):817. doi:10.1038/SJ.BJC.6602022
315. Reid FSW, Egoroff N, Pockney PG, Smith SR. A systematic scoping review on natural killer cell function in colorectal cancer. *Cancer Immunology, Immunotherapy*. 2021;70(3):597-606. doi:10.1007/S00262-020-02721-6/TABLES/1
316. Wouters MCA, Nelson BH. Prognostic significance of tumor-infiltrating B cells and plasma cells in human cancer. *Clinical Cancer Research*. 2018;24(24):6125-6135. doi:10.1158/1078-0432.CCR-18-1481/73996/AM/PROGNOSTIC-SIGNIFICANCE-OF-TUMOR-INFILTRATING-B

317. Tosolini M, Kirilovsky A, Mlecnik B, et al. Clinical impact of different classes of infiltrating T cytotoxic and helper cells (Th1, Th2, Treg, Th17) in patients with colorectal cancer. *Cancer Res*. 2011;71(4):1263-1271. doi:10.1158/0008-5472.CAN-10-2907/656702/P/CLINICAL-IMPACT-OF-DIFFERENT-CLASSES-OF
318. Guilliams M, Mildner A, Yona S. Developmental and Functional Heterogeneity of Monocytes. *Immunity*. 2018;49(4):595-613. doi:10.1016/J.IMMUNI.2018.10.005
319. Wen S, Chen N, Peng J, et al. Peripheral monocyte counts predict the clinical outcome for patients with colorectal cancer: A systematic review and meta-analysis. *Eur J Gastroenterol Hepatol*. 2019;31(11):1313-1321. doi:10.1097/MEG.0000000000001553
320. Zhao Y, Ge X, Xu X, Yu S, Wang J, Sun L. Prognostic value and clinicopathological roles of phenotypes of tumour-associated macrophages in colorectal cancer. *J Cancer Res Clin Oncol*. 2019;145(12):3005-3019. doi:10.1007/S00432-019-03041-8/FIGURES/5
321. Chun E, Lavoie S, Michaud M, et al. CCL2 Promotes Colorectal Carcinogenesis by Enhancing Polymorphonuclear Myeloid-Derived Suppressor Cell Population and Function. *Cell Rep*. 2015;12(2):244. doi:10.1016/J.CELREP.2015.06.024
322. Shibutani M, Maeda K, Nagahara H, et al. The peripheral monocyte count is associated with the density of tumor-associated macrophages in the tumor microenvironment of colorectal cancer: A retrospective study. *BMC Cancer*. 2017;17(1):1-7. doi:10.1186/S12885-017-3395-1/FIGURES/4
323. Olingy CE, Dinh HQ, Hedrick CC. Monocyte heterogeneity and functions in cancer. *J Leukoc Biol*. 2019;106(2):309-322. doi:10.1002/JLB.4RI0818-311R
324. Quail DF, Amulic B, Aziz M, et al. Neutrophil phenotypes and functions in cancer: A consensus statement. *Journal of Experimental Medicine*. 2022;219(6):39. doi:10.1084/JEM.20220011/213202
325. Yamamoto T, Kawada K, Obama K. Inflammation-Related Biomarkers for the Prediction of Prognosis in Colorectal Cancer Patients. *International Journal of Molecular Sciences 2021, Vol 22, Page 8002*. 2021;22(15):8002. doi:10.3390/IJMS22158002
326. Saurer L, Zysset D, Rihs S, et al. TREM-1 promotes intestinal tumorigenesis. *Scientific Reports 2017 7:1*. 2017;7(1):1-12. doi:10.1038/s41598-017-14516-4
327. Mizuno R, Kawada K, Itatani Y, Ogawa R, Kiyasu Y, Sakai Y. The role of tumor-associated neutrophils in colorectal cancer. *Int J Mol Sci*. 2019;20(3):1-14. doi:10.3390/ijms20030529
328. Teijeira Á, Garasa S, Gato M, et al. CXCR1 and CXCR2 Chemokine Receptor Agonists Produced by Tumors Induce Neutrophil Extracellular Traps that Interfere

with Immune Cytotoxicity. *Immunity*. 2020;52(5):856-871.e8. doi:10.1016/J.IMMUNI.2020.03.001

329. Governa V, Trella E, Mele V, et al. The interplay between neutrophils and CD8+ T cells improves survival in human colorectal cancer. *Clinical Cancer Research*. 2017;23(14):3847-3858. doi:10.1158/1078-0432.CCR-16-2047/72490/AM/THE-INTERPLAY-BETWEEN-NEUTROPHILS-AND-CD8-T-CELLS
330. Murphy N, Ward HA, Jenab M, et al. Heterogeneity of Colorectal Cancer Risk Factors by Anatomical Subsite in 10 European Countries: A Multinational Cohort Study. *Clinical Gastroenterology and Hepatology*. 2019;17(7):1323-1331.e6. doi:10.1016/J.CGH.2018.07.030
331. Nelson RL, Dollear T, Freels S, Persky V. The relation of age, race, and gender to the subsite location of colorectal carcinoma. *Cancer*. 1997;80(2):193-197. doi:https://doi.org/10.1002/(SICI)1097-0142(19970715)80:2<193::AID-CNCR4>3.0.CO;2-V
332. White A, Ironmonger L, Steele RJC, Ormiston-Smith N, Crawford C, Seims A. A review of sex-related differences in colorectal cancer incidence, screening uptake, routes to diagnosis, cancer stage and survival in the UK. *BMC Cancer*. 2018;18(1):1-11. doi:10.1186/S12885-018-4786-7/TABLES/7
333. Murphy G, Devesa SS, Cross AJ, Inskip PD, McGlynn KA, Cook MB. Sex disparities in colorectal cancer incidence by anatomic subsite, race and age. *Int J Cancer*. 2011;128(7):1668-1675. doi:10.1002/IJC.25481
334. Labadie JD, Savas S, Harrison TA, et al. Genome-wide association study identifies tumor anatomical site-specific risk variants for colorectal cancer survival. 2022;12(1):1-10.
335. Evans DM, Davey Smith G. Mendelian Randomization: New Applications in the Coming Age of Hypothesis-Free Causality. *Annu Rev Genomics Hum Genet*. 2015;16(1):327-350. doi:10.1146/annurev-genom-090314-050016
336. Ramspek CL, Steyerberg EW, Riley RD, et al. Prediction or causality? A scoping review of their conflation within current observational research. *Eur J Epidemiol*. 2021;36(9):889-898. doi:10.1007/S10654-021-00794-W/TABLES/4
337. Skrivankova VW, Richmond RC, Woolf BAR, et al. Strengthening the reporting of observational studies in epidemiology using mendelian randomisation (STROBE-MR): explanation and elaboration. *BMJ*. 2021;375. doi:10.1136/BMJ.N2233
338. Carress H, Lawson DJ, Elhaik E. Population genetic considerations for using biobanks as international resources in the pandemic era and beyond. *BMC Genomics*. 2021;22(1):1-19. doi:10.1186/s12864-021-07618-x

339. Lawson DJ, Davies NM, Haworth S, et al. Is population structure in the genetic biobank era irrelevant, a challenge, or an opportunity? *Hum Genet*. 2020;139(1):23-41. doi:10.1007/s00439-019-02014-8
340. Shim H, Chasman DI, Smith JD, et al. A Multivariate Genome-Wide Association Analysis of 10 LDL Subfractions, and Their Response to Statin Treatment, in 1868 Caucasians. *PLoS One*. 2015;10(4):120758. doi:10.1371/JOURNAL.PONE.0120758
341. Greco M F Del, Minelli C, Sheehan NA, Thompson JR. Detecting pleiotropy in Mendelian randomisation studies with summary data and a continuous outcome. *Stat Med*. 2015;34(21):2926-2940. doi:10.1002/SIM.6522
342. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*. 1995;57(1):289-300. doi:10.1111/J.2517-6161.1995.TB02031.X
343. Townsend P. Deprivation. *J Soc Policy*. 1987;16(2):125-146. doi:10.1017/S0047279400020341
344. Burrows K, Bull CJ, Dudding T, et al. Genome-wide Association Study of Cancer Risk in UK Biobank. Published online 2021. doi:10.5523/bris.aed0u12w0ede20olb0m77p4b9
345. Core R Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Published 2019. Accessed March 2, 2021. http://www.r-project.org
346. Gagliano Taliun SA, Evans DM. Ten simple rules for conducting a mendelian randomization study. *PLoS Comput Biol*. 2021;17(8):e1009238. doi:10.1371/JOURNAL.PCBI.1009238
347. Timpson NJ, Greenwood CMT, Soranzo N, Lawson DJ, Richards JB. Genetic architecture: the shape of the genetic contribution to human traits and disease. *Nature Reviews Genetics 2017 19:2*. 2017;19(2):110-124. doi:10.1038/nrg.2017.101
348. Elgart M, Goodman MO, Isasi C, et al. Correlations between complex human phenotypes vary by genetic background , gender , and environment. *Cell Rep Med*. 2022;3(12):100844. doi:10.1016/j.xcrm.2022.100844
349. Wyse C, O'Malley G, Coogan AN, McConkey S, Smith DJ. Seasonal and daytime variation in multiple immune parameters in humans: Evidence from 329,261 participants of the UK Biobank cohort. *iScience*. 2021;24(4). doi:10.1016/J.ISCI.2021.102255/ATTACHMENT/9F1CD6D8-A481-42C8-9BC6-AB92B331A713/MMC1.PDF

350. Ito Y, Satoh T, Takayama K, Miyagishi C, Walls AF, Yokozeki H. Basophil recruitment and activation in inflammatory skin diseases. *Allergy*. 2011;66(8):1107-1113. doi:10.1111/J.1398-9995.2011.02570.X
351. Wu C, Qiu Y, Zhang R, et al. Association of peripheral basophils with tumor M2 macrophage infiltration and outcomes of the anti-PD-1 inhibitor plus chemotherapy combination in advanced gastric cancer. *J Transl Med*. 2022;20(1):1-15. doi:10.1186/S12967-022-03598-Y/FIGURES/5
352. Sun S, Liu Y, Li L, et al. Mendelian randomization analysis of the association between human blood cell traits and uterine polyps. *Scientific Reports 2021 11:1*. 2021;11(1):1-9. doi:10.1038/s41598-021-84851-0
353. Silva J, Canão P, Espinheira MC, Trindade E, Carneiro F, Dias JA. Eosinophils in the gastrointestinal tract: how much is normal? *Virchows Archiv*. 2018;473(3):313-320. doi:10.1007/S00428-018-2405-2/FIGURES/2
354. Virdee PS, Patnick J, Watkinson P, et al. Trends in the full blood count blood test and colorectal cancer detection: a longitudinal, case-control study of UK primary care patient data. *NIHR Open Research 2022 2:32*. 2022;2:32.
355. Briede I, Strumfa I, Vanags A, Gardovskis J. The Association Between Inflammation, Epithelial Mesenchymal Transition and Stemness in Colorectal Carcinoma. *J Inflamm Res*. 2020;13:15. doi:10.2147/JIR.S224441
356. Benson VS, Hartl S, Barnes N, Galwey N, Van Dyke MK, Kwon N. Blood eosinophil counts in the general population and airways disease: a comprehensive review and meta-analysis. *European Respiratory Journal*. 2022;59(1). doi:10.1183/13993003.04590-2020
357. Morrison J, Knoblauch N, Marcus JH, Stephens M, He X. Mendelian randomization accounting for correlated and uncorrelated pleiotropic effects using genome-wide summary statistics. *Nature Genetics 2020 52:7*. 2020;52(7):740-747. doi:10.1038/s41588-020-0631-4
358. Karim AF, Westenberg LEH, Eurelings LEM, Otten R, Gerth Van Wijk R. The association between allergic diseases and cancer: a systematic review of the literature.
359. Yuan S, Vithayathil M, Kar S, et al. Assessing the protective role of allergic disease in gastrointestinal tract cancers using Mendelian randomization analysis. 2021;76(5):1559-1562. doi:10.1111/ALL.14616
360. Jiang X, Dimou NL, Zhu Z, et al. Allergy, asthma, and the risk of breast and prostate cancer: a Mendelian randomization study. *Cancer Causes and Control*. 2020;31(3):273-282. doi:10.1007/S10552-020-01271-7/TABLES/4

361. Lonsdale J, Thomas J, Salvatore M, et al. The Genotype-Tissue Expression (GTEx) project. *Nature Genetics 2013 45:6*. 2013;45(6):580-585. doi:10.1038/ng.2653
362. Wu L, Candille SI, Choi Y, et al. Variation and Genetic Control of Protein Abundance in Humans. *Nature*. 2013;499(7456):79. doi:10.1038/NATURE12223
363. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*. 2009;10(1):57. doi:10.1038/NRG2484
364. Edin S, Kaprio T, Hagström J, et al. The Prognostic Importance of CD20+ B lymphocytes in Colorectal Cancer and the Relation to Other Immune Cell subsets. *Scientific Reports 2019 9:1*. 2019;9(1):1-9. doi:10.1038/s41598-019-56441-8
365. Berntsson J, Svensson MC, Leandersson K, et al. The clinical impact of tumour-infiltrating lymphocytes in colorectal cancer differs by anatomical subsite: A cohort study. *Int J Cancer*. 2017;141(8):1654-1666. doi:10.1002/IJC.30869
366. Rao S, Gharib K, Han A. Cancer Immunosurveillance by T Cells. *Int Rev Cell Mol Biol*. 2019;342:149-173. doi:10.1016/BS.IRCMB.2018.08.001
367. Meza Guzman LG, Keating N, Nicholson SE. Natural Killer Cells: Tumor Surveillance and Signaling. *Cancers (Basel)*. 2020;12(4). doi:10.3390/CANCERS12040952
368. Giambartolomei C, Liu JZ, Zhang W, et al. A Bayesian framework for multiple trait colocalization from summary association statistics. *Bioinformatics*. 2018;34(15):2538. doi:10.1093/BIOINFORMATICS/BTY147
369. Bain BJ. Ethnic and sex differences in the total and differential white cell count and platelet count. *Clin Pathol*. 1996;49(7):664-666.
370. Chakraborty B, Byemerwa J, Krebs T, Lim F, Chang CY, McDonnell DP. Estrogen Receptor Signaling in the Immune System. *Endocr Rev*. Published online June 16, 2022. doi:10.1210/ENDREV/BNAC017
371. Bouman A, Schipper M, Heineman MJ, Faas MM. Gender Difference in the Non-Specific and Specific Immune Response in Humans. *American Journal of Reproductive Immunology*. 2004;52(1):19-26. doi:10.1111/J.1600-0897.2004.00177.X
372. Scotland RS, Stables MJ, Madalli S, Watson P, Gilroy DW. Sex differences in resident immune cell phenotype underlie more efficient acute inflammatory responses in female mice. *Blood*. 2011;118(22):5918-5927. doi:10.1182/BLOOD-2011-03-340281
373. Kverneland AH, Streitz M, Geissler E, et al. Age and gender leucocytes variances and references values generated using the standardized ONE-Study protocol. *Cytometry Part A*. 2016;89(6):543-564. doi:10.1002/CYTO.A.22855

374. Ray AL, Nofchissey RA, Khan MA, et al. The role of sex in the innate and adaptive immune environment of metastatic colorectal cancer. *British Journal of Cancer 2020 123:4*. 2020;123(4):624-632. doi:10.1038/s41416-020-0913-8
375. Tan VY, Yarmolinsky J, Lawlor DA, Timpson NJ. Letter regarding article, "Associations of obesity and circulating insulin and glucose with breast cancer risk: a Mendelian randomization analysis." *Int J Epidemiol*. 2019;48(3):1014-1015. doi:10.1093/IJE/DYZ013
376. Wang Z, Lu J. Sex-specific exposures and sex-combined outcomes in two-sample Mendelian randomization may mislead the causal inference. *Arthritis Res Ther*. 2022;24(1):1-2. doi:10.1186/S13075-022-02922-7/METRICS
377. Gao Y, Zhang J, Zhao H, Guan F, Zeng P. Instrumental Heterogeneity in Sex-Specific Two-Sample Mendelian Randomization: Empirical Results From the Relationship Between Anthropometric Traits and Breast/Prostate Cancer. *Front Genet*. 2021;12:772. doi:10.3389/FGENE.2021.651332/BIBTEX
378. Lopes-Ramos CM, Quackenbush J, DeMeo DL. Genome-Wide Sex and Gender Differences in Cancer. *Front Oncol*. 2020;10. doi:10.3389/FONC.2020.597788
379. Cai Y, Rattray NJW, Zhang Q, et al. Sex Differences in Colon Cancer Metabolism Reveal A Novel Subphenotype. *Scientific Reports 2020 10:1*. 2020;10(1):1-13. doi:10.1038/s41598-020-61851-0
380. Nazarian A, Kulminski AM. Genome-wide analysis of sex disparities in the genetic architecture of lung and colorectal cancers. *Genes (Basel)*. 2021;12(5). doi:10.3390/GENES12050686/S1
381. Artham S, Chang CY, McDonnell DP. Eosinophilia in cancer and its regulation by sex hormones. *Trends Endocrinol Metab*. 2022;34(1):5-20. doi:10.1016/j.tem.2022.11.002
382. Nøst TH, Alcala K, Urbarova I, et al. Systemic inflammation markers and cancer incidence in the UK Biobank. *Eur J Epidemiol*. 2021;36(8).
383. Sulc J, Sjaarda J, Kutalik Z. Polynomial Mendelian randomization reveals non-linear causal effects for obesity-related traits. *Human Genetics and Genomics Advances*. 2022;3(3):1-9. doi:10.1016/j.xhgg.2022.100124
384. Brodin P, Davis MM. Human immune system variation. *Nat Rev Immunol*. 2017;17(1):21. doi:10.1038/NRI.2016.125
385. WHO Africa. *World Malaria Report 2019*.; 2019.
386. Sirugo G, Williams SM, Tishkoff SA. The Missing Diversity in Human Genetic Studies. *Cell*. 2019;177(1):26-31. doi:10.1016/J.CELL.2019.02.048
387. Bentley AR, Callier SL, Rotimi CN. Evaluating the promise of inclusion of African ancestry populations in genomics. *npj Genomic Medicine 2020 5:1*. 2020;5(1):1-9. doi:10.1038/s41525-019-0111-x

388. Cooke Bailey JN, Bush WS, Crawford DC. Editorial: The Importance of Diversity in Precision Medicine Research. *Front Genet*. 2020;11:875. doi:10.3389/fgene.2020.00875
389. Green ED, Gunter C, Biesecker LG, et al. Strategic vision for improving human health at The Forefront of Genomics. *Nature 2020 586:7831*. 2020;586(7831):683-692. doi:10.1038/s41586-020-2817-4
390. Consortium TH. Enabling the genomic revolution in Africa: H3Africa is developing capacity for health-related genomics research in Africa. *Science*. 2014;344(6190):1346. doi:10.1126/SCIENCE.1251546
391. Matise TC, Study for the P, Ambite JL, et al. The Next PAGE in Understanding Complex Traits: Design for the Analysis of Population Architecture Using Genetics and Epidemiology (PAGE) Study. *Am J Epidemiol*. 2011;174(7):849-859. doi:10.1093/AJE/KWR160
392. Taliun D, Harris DN, Kessler MD, et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature 2021 590:7845*. 2021;590(7845):290-299. doi:10.1038/s41586-021-03205-y
393. Gallo LC, Penedo FJ, Carnethon M, et al. The Hispanic Community Health Study/Study of Latinos Sociocultural Ancillary Study: Sample, Design, and Procedures. *Ethn Dis*. 2014;24(1):77.
394. Investigators TA of URP. The "All of Us" Research Program. *https://doi.org/101056/NEJMsr1809937*. 2019;381(7):668-676. doi:10.1056/NEJMSR1809937
395. Nagai A, Hirata M, Kamatani Y, et al. Overview of the BioBank Japan Project: Study design and profile. *J Epidemiol*. 2017;27(3 Suppl):S2. doi:10.1016/J.JE.2016.12.005
396. Hodonsky CJ, Baldassari AR, Bien SA, et al. Ancestry-specific associations identified in genome-wide combined-phenotype study of red blood cell traits emphasize benefits of diversity in genomics. *BMC Genomics*. 2020;21(1):1-14. doi:10.1186/S12864-020-6626-9/FIGURES/2
397. Martin AR, Teferra S, Möller M, Hoal EG, Daly MJ. The critical needs and challenges for genetic architecture studies in Africa. *Curr Opin Genet Dev*. 2018;53:113-120. doi:10.1016/J.GDE.2018.08.005
398. Price AL, Weale ME, Patterson N, et al. Long-Range LD Can Confound Genome Scans in Admixed Populations. *Am J Hum Genet*. 2008;83(1):132. doi:10.1016/J.AJHG.2008.06.005
399. Wen J, Xie M, Rowland B, et al. Transcriptome-wide association study of blood cell traits in african ancestry and hispanic/latino populations. *Genes (Basel)*. 2021;12(7):1049. doi:10.3390/genes12071049

400. Hu Y, Bien SA, Nishimura KK, et al. Multi-ethnic genome-wide association analyses of white blood cell and platelet traits in the Population Architecture using Genomics and Epidemiology (PAGE) study. *BMC Genomics*. 2021;22(1):1-11. doi:10.1186/S12864-021-07745-5
401. Tucci S, Akey JM. The long walk to African genomics. *Genome Biol*. 2019;20(1):1-3. doi:10.1186/S13059-019-1740-1/METRICS
402. Haworth S, Mitchell R, Corbin L, et al. Apparent latent structure within the UK Biobank sample has implications for epidemiological analysis. *Nat Commun*. 2019;10(1). doi:10.1038/S41467-018-08219-1
403. Barton N, Hermisson J, Nordborg M. Why structure matters. *Elife*. 2019;8. doi:10.7554/ELIFE.45380
404. Abdellaoui A, Hugh-Jones D, Yengo L, et al. Genetic correlates of social stratification in Great Britain. *Nature Human Behaviour 2019 3:12*. 2019;3(12):1332-1342. doi:10.1038/s41562-019-0757-5
405. Rodriguez S, Gaunt TR, Day INM. Hardy-Weinberg Equilibrium Testing of Biological Ascertainment for Mendelian Randomization Studies. *Am J Epidemiol*. 2009;169(4):505-514. doi:10.1093/AJE/KWN359
406. Graffelman J, Weir BS. On the testing of Hardy-Weinberg proportions and equality of allele frequencies in males and females at biallelic genetic markers. *Genet Epidemiol*. 2018;42(1):34-48. doi:10.1002/GEPI.22079
407. Mathieson I, Scally A. What is ancestry? *PLoS Genet*. 2020;16(3):e1008624. doi:10.1371/JOURNAL.PGEN.1008624
408. Rosenberg NA, Pritchard JK, Weber JL, et al. Genetic structure of human populations. *Science (1979)*. 2002;298(5602):2381-2385. doi:10.1126/SCIENCE.1078311/SUPPL_FILE/ROSENBERG.SOM.PDF.PDF
409. Berezovskiĭ ND, Giria VN. [Estimation of combining ability of specialized types of the big white breed]. *Tsitol Genet*. 1991;25(6):56-60.
410. Serre D, Pääbo S. Evidence for Gradients of Human Genetic Diversity Within and Among Continents. *Genome Res*. 2004;14(9):1679. doi:10.1101/GR.2529604
411. Rosenberg NA, Mahajan S, Ramachandran S, Zhao C, Pritchard JK, Feldman MW. Clines, Clusters, and the Effect of Study Design on the Inference of Human Population Structure. Allison D, ed. *PLoS Genet*. 2005;1(6):e70. doi:10.1371/JOURNAL.PGEN.0010070
412. Mitchell RE, Hemani G, Dudding T, Corbin L, Harrison S, Paternoster L. UK Biobank Genetic Data: MRC-IEU Quality Control, version 2, 18/01/2019.
413. Sudmant PH, Rausch T, Gardner EJ, et al. An integrated map of structural variation in 2,504 human genomes. *Nature*. 2015;526(7571):75-81. doi:10.1038/nature15394

414. Ongaro L, Scliar MO, Flores R, et al. The Genomic Impact of European Colonization of the Americas. *Current Biology*. 2019;29(23):3974-3986.e4.
415. Joiret M, Mahachie John JM, Gusareva ES, Van Steen K. Confounding of linkage disequilibrium patterns in large scale DNA based gene-gene interaction studies. *BioData Min*. 2019;12(1):1-23. doi:10.1186/S13040-019-0199-7/TABLES/6
416. Chang CC, Chow CC, Tellier LCAM, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: Rising to the challenge of larger and richer datasets. *Gigascience*. 2015;4(1). doi:10.1186/s13742-015-0047-8
417. Weale ME. Quality Control for Genome-Wide Association Studies. In: Barnes MR, Breen G, eds. *Genetic Variation: Methods and Protocols*. Humana Press, New York, NY; 2010:31.
418. Hartigan JA, Wong MA. A K-Means Clustering Algorithm. *Journal of the Royal Statistical Society*. 1979;28(1):100-108. doi:10.9756/bijdm.1106
419. Batool F, Hennig C. Clustering with the Average Silhouette Width. *Comput Stat Data Anal*. 2021;158:107190. doi:10.1016/J.CSDA.2021.107190
420. Katariya A, Detroja KP. Pattern matching using correspondence analysis. *Proceedings of the American Control Conference*. Published online 2013:2662-2667. doi:10.1109/ACC.2013.6580236
421. Lao O, Lu TT, Nothnagel M, et al. Correlation between Genetic and Geographic Structure in Europe. *Current Biology*. 2008;18(16):1241-1248. doi:10.1016/J.CUB.2008.07.049
422. Nagylaki T. Fixation Indices in Subdivided Populations. *Genetics Society of America*. 1997;148(3):1325–1332.
423. Reich D, Thangaraj K, Patterson N, Price AL, Singh L. Reconstructing Indian population history. *Nature*. 2009;461(7263):489-494. doi:10.1038/nature08365
424. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res*. 2009;19(9):1655-1664. doi:10.1101/gr.094052.109
425. Morton NE. Isolation by Distance. *Genetics*. 1943;28(2):114. doi:10.1016/B978-0-12-374984-0.00820-2
426. Slatkin M. ISOLATION BY DISTANCE IN EQUILIBRIUM AND NON-EQUILIBRIUM POPULATIONS. *Evolution*. 1993;47(1):264-279. doi:10.1111/J.1558-5646.1993.TB01215.X
427. Birney E, Inouye M, Raff J, Rutherford A, Scally A. The language of race, ethnicity, and ancestry in human genetic research.
428. Peterson RE, Kuchenbaecker K, Walters RK, et al. Genome-wide Association Studies in Ancestrally Diverse Populations: Opportunities, Methods, Pitfalls, and Recommendations. *Cell*. 2019;179(3):589-603. doi:10.1016/J.CELL.2019.08.051

429. Laland KN, Odling-Smee J, Myles S. How culture shaped the human genome: bringing genetics and the human sciences together. *Nature Reviews Genetics 2010 11:2*. 2010;11(2):137-148. doi:10.1038/nrg2734
430. Przeworski M, Wall JD. Why is there so little intragenic linkage disequilibrium in humans? *Genet Res*. 2001;77(2):143-151. doi:10.1017/S0016672301004967
431. Ptak SE, Voelpel K, Przeworski M. Insights into recombination from patterns of linkage disequilibrium in humans. *Genetics*. 2004;167(1):387. doi:10.1534/GENETICS.167.1.387
432. Lander ES, Schork NJ. Genetic dissection of complex traits. *Science (1979)*. 1994;265(5181):2037-2048. doi:10.1126/SCIENCE.8091226
433. Hellwege JN, Keaton JM, Giri A, Gao X, Velez Edwards DR, Edwards TL. Population Stratification in Genetic Association Studies. *Curr Protoc Hum Genet*. 2017;95(1):1.22.1. doi:10.1002/CPHG.48
434. Vilhjálmsson BJ, Nordborg M. The nature of confounding in genome-wide association studies. *Nature Reviews Genetics 2012 14:1*. 2012;14(1):1-2. doi:10.1038/nrg3382
435. Loh PR, Tucker G, Bulik-Sullivan BK, et al. Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat Genet*. 2015;47(3):284-290. doi:10.1038/ng.3190
436. Zaidi AA, Mathieson I. Demographic history mediates the effect of stratification on polygenic scores. *Elife*. 2020;9:1-30. doi:10.7554/ELIFE.61548
437. Novembre J, Johnson T, Bryc K, et al. Genes mirror geography within Europe. *Nature*. 2008;456(7218):98. doi:10.1038/NATURE07331
438. Berg JJ, Harpak A, Sinnott-Armstrong N, et al. Reduced signal for polygenic adaptation of height in UK biobank. *Elife*. 2019;8. doi:10.7554/eLife.39725
439. Sohail M, Maier RM, Ganna A, et al. Polygenic adaptation on height is overestimated due to uncorrected stratification in genome-wide association studies. *Elife*. 2019;8. doi:10.7554/eLife.39702
440. Wood AR, Esko T, Yang J, et al. Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat Genet*. 2014;46(11):1173. doi:10.1038/NG.3097
441. Diaz-Papkovich A, Anderson-Trocmé L, Ben-Eghan C, Gravel S. UMAP reveals cryptic population structure and phenotype heterogeneity in large genomic cohorts. *PLoS Genet*. 2019;15(11). doi:10.1371/journal.pgen.1008432
442. Flanagin A, Frey T, Christiansen SL. Updated Guidance on the Reporting of Race and Ethnicity in Medical and Science Journals. *JAMA*. 2021;326(7):621-627. doi:10.1001/JAMA.2021.13304

443. Kidd JM, Gravel S, Byrnes J, et al. Population Genetic Inference from Personal Genome Data: Impact of Ancestry and Admixture on Human Genomic Variation. *Am J Hum Genet*. 2012;91(4):660. doi:10.1016/J.AJHG.2012.08.025
444. Homburger JR, Moreno-Estrada A, Gignoux CR, et al. Genomic Insights into the Ancestry and Demographic History of South America. *PLoS Genet*. 2015;11(12). doi:10.1371/JOURNAL.PGEN.1005602
445. Moreno-Estrada A, Gravel S, Zakharia F, et al. Reconstructing the Population Genetic History of the Caribbean. *PLoS Genet*. 2013;9(11). doi:10.1371/JOURNAL.PGEN.1003925
446. Montinaro F, Busby GBJ, Pascali VL, Myers S, Hellenthal G, Capelli C. Unravelling the hidden ancestry of American admixed populations. *Nat Commun*. 2015;6. doi:10.1038/NCOMMS7596
447. Geibel J, Reimer C, Weigend S, Weigend A, Pook T, Simianer H. How array design creates SNP ascertainment bias. *PLoS One*. 2021;16(3 March):e0245178-e0245178. doi:10.1371/journal.pone.0245178
448. Lachance J, Tishkoff SA. SNP ascertainment bias in population genetic analyses: Why it is important, and how to correct it. *BioEssays*. 2013;35(9):780-786. doi:10.1002/bies.201300014
449. Albrechtsen A, Nielsen FC, Nielsen R. Ascertainment biases in SNP chips affect measures of population divergence. *Mol Biol Evol*. 2010;27(11):2534-2547. doi:10.1093/molbev/msq148
450. Lu Y, Patterson N, Zhan Y, Mallick S, Reich D. Technical design document for a SNP array that is optimized for population genetics.
451. Price RN, Commons RJ, Battle KE, Thriemer K, Mendis K. Plasmodium vivax in the Era of the Shrinking P. falciparum Map. *Trends Parasitol*. 2020;36(6):560-570. doi:10.1016/j.pt.2020.03.009
452. Kariuki SN, Williams TN. *Human Genetics and Malaria Resistance*. Vol 139. Springer; 2020:801-811. doi:10.1007/S00439-020-02142-6
453. Liu W, Li Y, Learn GH, et al. Origin of the human malaria parasite Plasmodium falciparum in gorillas. *Nature*. 2010;467(7314):420. doi:10.1038/NATURE09442
454. LEVINE ND, CORLISS JO, COX FEG, et al. A Newly Revised Classification of the Protozoa. *J Protozool*. 1980;27(1):37-58. doi:10.1111/J.1550-7408.1980.TB04228.X
455. Wu R, Trubl G, Taş N, Jansson JK. Permafrost as a potential pathogen reservoir. *One Earth*. 2022;5(4):351-360. doi:10.1016/J.ONEEAR.2022.03.010
456. Xu R, Zhang M, Lin H, et al. Response of soil protozoa to acid mine drainage in a contaminated terrace. *J Hazard Mater*. 2022;421:126790. doi:10.1016/J.JHAZMAT.2021.126790

457. Peacock CS, Seeger K, Harris D, et al. Comparative genomic analysis of three Leishmania species that cause diverse human disease. *Nature Genetics 2007 39:7*. 2007;39(7):839-847. doi:10.1038/ng2053
458. Combs TP, Nagajyothi, Mukherjee S, et al. The adipocyte as an important target cell for Trypanosoma cruzi infection. *Journal of Biological Chemistry*. 2005;280(25):24085-24094. doi:10.1074/jbc.M412802200
459. Sasai M, Yamamoto M. Innate, adaptive, and cell-autonomous immunity against Toxoplasma gondii infection. *Experimental & Molecular Medicine 2019 51:12*. 2019;51(12):1-10. doi:10.1038/s12276-019-0353-9
460. Beri D, Ramdani G, Balan B, et al. Insights into physiological roles of unique metabolites released from Plasmodium-infected RBCs and their potential as clinical biomarkers for malaria. *Scientific Reports 2019 9:1*. 2019;9(1):1-11. doi:10.1038/s41598-018-37816-9
461. Matuschewski K. Getting infectious: formation and maturation of Plasmodium sporozoites in the Anopheles vector. *Cell Microbiol*. 2006;8(10):1547-1556. doi:10.1111/J.1462-5822.2006.00778.X
462. Faust C, Dobson AP. Primate malarias: Diversity, distribution and insights for zoonotic Plasmodium. *One Health*. 2015;1:66-75. doi:10.1016/J.ONEHLT.2015.10.001
463. Sato S. Plasmodium—a brief introduction to the parasites causing human malaria and their basic biology. *J Physiol Anthropol*. 2021;40(1):1-13. doi:10.1186/S40101-020-00251-9/TABLES/1
464. Kotepui M, Kotepui KU, Milanez GD, Masangkay FR. Global prevalence and mortality of severe Plasmodium malariae infection: A systematic review and meta-analysis. *Malar J*. 2020;19(1):1-13. doi:10.1186/S12936-020-03344-Z/TABLES/3
465. Kotepui M, Kotepui KU, Milanez GD, Masangkay FR. Severity and mortality of severe Plasmodium ovale infection: A systematic review and meta-analysis. *PLoS One*. 2020;15(6):e0235014. doi:10.1371/JOURNAL.PONE.0235014
466. Mahittikorn A, Masangkay FR, Kotepui KU, Milanez GDJ, Kotepui M. Comparison of Plasmodium ovale curtisi and Plasmodium ovale wallikeri infections by a meta-analysis approach. *Scientific Reports 2021 11:1*. 2021;11(1):1-15. doi:10.1038/s41598-021-85398-w
467. Kotepui M, Kotepui KU, Milanez GD, Masangkay FR. Prevalence of severe Plasmodium knowlesi infection and risk factors related to severe complications compared with non-severe P. knowlesi and severe P. falciparum malaria: a systematic review and meta-analysis. *Infect Dis Poverty*. 2020;9(1):1-14. doi:10.1186/S40249-020-00727-X/FIGURES/9

468. Howes RE, Battle KE, Mendis KN, et al. Global Epidemiology of Plasmodium vivax. *Am J Trop Med Hyg*. 2016;95(6 Suppl):15. doi:10.4269/AJTMH.16-0141
469. Phyo AP, Dahal P, Mayxay M, Ashley EA. Clinical impact of vivax malaria: A collection review. *PLoS Med*. 2022;19(1):e1003890. doi:10.1371/JOURNAL.PMED.1003890
470. Snow RW. Global malaria eradication and the importance of Plasmodium falciparum epidemiology in Africa. *BMC Med*. 2015;13(1):1-3. doi:10.1186/S12916-014-0254-7/FIGURES/1
471. Shaw WR, Marcenac P, Catteruccia F. Plasmodium development in Anopheles: a tale of shared resources. *Trends Parasitol*. 2022;38(2):124-135. doi:10.1016/J.PT.2021.08.009
472. Amino R, Thiberge S, Martin B, et al. Quantitative imaging of Plasmodium transmission from mosquito to mammal. *Nature Medicine 2006 12:2*. 2006;12(2):220-224. doi:10.1038/nm1350
473. Sturm A, Graewe S, Franke-Fayard B, et al. Alteration of the Parasite Plasma Membrane and the Parasitophorous Vacuole Membrane during Exo-Erythrocytic Development of Malaria Parasites. *Protist*. 2009;160(1):51-63. doi:10.1016/J.PROTIS.2008.08.002
474. Sturm A, Amino R, Van De Sand C, et al. Manipulation of host hepatocytes by the malaria parasite for delivery into liver sinusoids. *Science (1979)*. 2006;313(5791):1287-1290. doi:10.1126/SCIENCE.1129720/SUPPL_FILE/STURM.SOM.PDF
475. Cowman AF, Crabb BS. Invasion of red blood cells by malaria parasites. *Cell*. 2006;124(4):755-766. doi:10.1016/j.cell.2006.02.006
476. Ndila CM, Uyoga S, Macharia AW, et al. Human candidate gene polymorphisms and risk of severe malaria in children in Kilifi, Kenya: a case-control association study. *Lancet Haematol*. 2018;5(8):e333-e345. doi:10.1016/S2352-3026(18)30107-8
477. Jaskiewicz E, Jodłowska M, Kaczmarek R, Zerka A. Erythrocyte glycophorins as receptors for Plasmodium merozoites. *Parasit Vectors*. 2019;12(1):1-11. doi:10.1186/S13071-019-3575-8/FIGURES/3
478. Miller LH, Mason SJ, Clyde DF, McGinniss MH. The Resistance Factor to Plasmodium vivax in Blacks. *https://doi.org/101056/NEJM197608052950602*. 1976;295(6):302-304. doi:10.1056/NEJM197608052950602
479. Blackman MJ. Malarial proteases and host cell egress: an 'emerging' cascade. *Cell Microbiol*. 2008;10(10):1925-1934. doi:10.1111/J.1462-5822.2008.01176.X

480. Bancells C, Llorà-Batlle O, Poran A, et al. Revisiting the initial steps of sexual development in the malaria parasite Plasmodium falciparum. *Nature Microbiology 2018 4:1*. 2018;4(1):144-154. doi:10.1038/s41564-018-0291-7
481. Boyle MJ, Wilson DW, Beeson JG. New approaches to studying Plasmodium falciparum merozoite invasion and insights into invasion biology. *Int J Parasitol*. 2013;43(1):1-10. doi:10.1016/J.IJPARA.2012.11.002
482. Boddey J. Plasmodium Nesting: Remaking the Erythrocyte from the Inside Out Structural studies of invasion processes during malaria infection View project malaria View project. *Article in Annual Review of Microbiology*. Published online 2013. doi:10.1146/annurev-micro-092412-155730
483. Smith LM, Motta FC, Chopra G, et al. An intrinsic oscillator drives the blood stage cycle of the malaria parasite Plasmodium falciparum. *Science (1979)*. 2020;368(6492):754-759. doi:10.1126/SCIENCE.ABA4357/SUPPL_FILE/ABA4357_SMITH_SM.PDF
484. Sinden RE, Strong K. An ultrastructural study of the sporogonic development of Plasmodium falciparum in Anopheles gambiae. *Trans R Soc Trop Med Hyg*. 1978;72(5):477-491. doi:10.1016/0035-9203(78)90167-0
485. Robert Shaw WI, Holmdahl ID IE, Itoe MA, et al. Multiple blood feeding in mosquitoes shortens the Plasmodium falciparum incubation period and increases malaria transmission potential. Published online 2020. doi:10.1371/journal.ppat.1009131
486. Kappe S, Bruderer T, Gantt S, Fujioka H, Nussenzweig V, Ménard R. Conservation of a Gliding Motility and Cell Invasion Machinery in Apicomplexan Parasites. *J Cell Biol*. 1999;147(5):937. doi:10.1083/JCB.147.5.937
487. Sultan AA, Thathy V, Frevert U, et al. TRAP Is Necessary for Gliding Motility and Infectivity of Plasmodium Sporozoites. *Cell*. 1997;90(3):511-522. doi:10.1016/S0092-8674(00)80511-5
488. Moxon CA, Gibbins MP, McGuinness D, Milner DA, Marti M. New Insights into Malaria Pathogenesis. *Annual Review of Pathology: Mechanisms of Disease*. 2020;15(1):315-343. doi:10.1146/annurev-pathmechdis-012419-032640
489. Knackstedt SL, Georgiadou A, Apel F, et al. Neutrophil extracellular traps drive inflammatory pathogenesis in malaria. 2019;4(40):336. doi:10.1126/SCIIMMUNOL.AAW0336
490. Sierro F, Grau GER. The ins and outs of cerebral malaria pathogenesis: Immunopathology, extracellular vesicles, immunometabolism, and trained immunity. *Front Immunol*. 2019;10(MAR):830. doi:10.3389/fimmu.2019.00830

491. Cela D, Knackstedt SL, Groves S, et al. PAD4 controls chemoattractant production and neutrophil trafficking in malaria. *J Leukoc Biol*. Published online 2021. doi:10.1002/JLB.4AB1120-780R
492. WHO. *Global Technical Strategy for Malaria 2016-2030, 2021 Update*.; 2021.
493. Cohen JM, Okumu F, Moonen B. The fight against malaria: Diminishing gains and growing challenges. *Sci Transl Med*. 2022;14(651):3256. doi:10.1126/SCITRANSLMED.ABN3256
494. Allison AC. Protection afforded by sickle-cell trait against subtertian malarial infection. *Br Med J*. 1954;1(4857):290-294. doi:10.1136/bmj.1.4857.290
495. Kwiatkowski DP. How Malaria Has Affected the Human Genome and What Human Genetics Can Teach Us about Malaria. *The American Journal of Human Genetics*. 2005;77(2):171-192. doi:10.1086/432519
496. Mackinnon MJ, Mwangi TW, Snow RW, Marsh K, Williams TN. Heritability of Malaria in Africa. Foote S, ed. *PLoS Med*. 2005;2(12):e340. doi:10.1371/journal.pmed.0020340
497. Sakuntabhai A, Ndiaye R, Casadémont I, et al. Genetic Determination and Linkage Mapping of Plasmodium falciparum Malaria Related Traits in Senegal. Gwinn K, ed. *PLoS One*. 2008;3(4):e2000. doi:10.1371/journal.pone.0002000
498. Koboldt DC, Steinberg KM, Larson DE, Wilson RK, Mardis ER. The Next-Generation Sequencing Revolution and Its Impact on Genomics. *Cell*. 2013;155(1):27-38. doi:10.1016/J.CELL.2013.09.006
499. Ahouidi A, Ali M, Almagro-Garcia J, et al. An open dataset of Plasmodium falciparum genome variation in 7,000 worldwide samples. *Wellcome Open Res*. 2021;6:61. doi:10.12688/WELLCOMEOPENRES.16168.1
500. Militello KT, Dodge M, Bethke L, Wirth DF. Identification of regulatory elements in the Plasmodium falciparum genome. *Mol Biochem Parasitol*. 2004;134(1):75-88. doi:10.1016/J.MOLBIOPARA.2003.11.004
501. Mu J, Awadalla P, Duan J, et al. Genome-wide variation and identification of vaccine targets in the Plasmodium falciparum genome. *Nature Genetics 2006 39:1*. 2006;39(1):126-130. doi:10.1038/ng1924
502. Verra F, Mangano VD, Modiano D. Genetics of susceptibility to Plasmodium falciparum: from classical malaria resistance genes towards genome-wide association studies. *Parasite Immunol*. 2009;31(5):234-253. doi:10.1111/J.1365-3024.2009.01106.X
503. Atallah-Yunes SA, Ready A, Newburger PE. Benign ethnic neutropenia. *Blood Rev*. 2019;37:100586. doi:10.1016/j.blre.2019.06.003
504. Rippey JJ. LEUCOPENIA IN WEST INDIANS AND AFRICANS. *The Lancet*. 1967;290(7505):44. doi:https://doi.org/10.1016/S0140-6736(67)90086-4

505. Denic S, Showqi S, Klein C, Takala M, Nagelkerke N, Agarwal MM. Prevalence, phenotype and inheritance of benign neutropenia in Arabs. *BMC Blood Disord*. 2009;9:3. doi:10.1186/1471-2326-9-3
506. Hsieh MM, Everhart JE, Byrd-Holt DD, Tisdale JF, Rodgers GP. Prevalence of Neutropenia in the U.S. Population: Age, Sex, Smoking Status, and Ethnic Differences. *Ann Intern Med*. 2007;146(7):486. doi:10.7326/0003-4819-146-7-200704030-00004
507. Amulic B, Cazalet C, Hayes GL, Metzler KD, Zychlinsky A. Neutrophil Function: From Mechanisms to Disease. *https://doi.org/101146/annurev-immunol-020711-074942*. 2012;30:459-489. doi:10.1146/ANNUREV-IMMUNOL-020711-074942
508. Rappoport N, Simon AJ, Amariglio N, Rechavi G. The Duffy antigen receptor for chemokines, ACKR1,– 'Jeanne DARC' of benign neutropenia. *Br J Haematol*. 2019;184(4):497-507. doi:10.1111/bjh.15730
509. Palmblad J, Höglund P. Ethnic benign neutropenia: A phenomenon finds an explanation. *Pediatr Blood Cancer*. 2018;65(12):e27361. doi:10.1002/pbc.27361
510. Kho S, Minigo G, Andries B, et al. Circulating Neutrophil Extracellular Traps and Neutrophil Activation Are Increased in Proportion to Disease Severity in Human Malaria. *J Infect Dis*. 2019;219(12):1994-2004. doi:10.1093/INFDIS/JIY661
511. Van Wolfswinkel ME, Vliegenthart-Jongbloed K, De Mendonça Melo M, et al. Predictive value of lymphocytopenia and the neutrophil-lymphocyte count ratio for severe imported malaria. *Malar J*. 2013;12(1):1-8. doi:10.1186/1475-2875-12-101/TABLES/3
512. Berens-Riha N, Kroidl I, Schunk M, et al. Evidence for significant influence of host immunity on changes in differential blood count during malaria. *Malar J*. 2014;13(1):1-9. doi:10.1186/1475-2875-13-155/TABLES/3
513. Amulic B, Moxon C, Cunnington A. *A More Granular View of Neutrophils in Malaria 1 2*.
514. Aitken EH, Alemu A, Rogerson SJ. Neutrophils and Malaria. *Front Immunol*. 2018;9:3005. doi:10.3389/fimmu.2018.03005
515. Lee HJ, Georgiadou A, Walther M, et al. Integrated pathogen load and dual transcriptome analysis of systemic host-pathogen interactions in severe malaria. *Sci Transl Med*. 2018;10(447). doi:10.1126/SCITRANSLMED.AAR3619
516. Munde EO, Okeyo WA, Raballah E, et al. Association between Fcγ receptor IIA, IIIA and IIIB genetic polymorphisms and susceptibility to severe malaria anemia in children in western Kenya. *BMC Infect Dis*. 2017;17(1). doi:10.1186/S12879-017-2390-0

517. Faik I, Van Tong H, Lell B, Meyer CG, Kremsner PG, Velavan TP. Pyruvate Kinase and Fcγ Receptor Gene Copy Numbers Associated With Malaria Phenotypes. *J Infect Dis*. 2017;216(2):276-282. doi:10.1093/INFDIS/JIX284
518. Zelter T, Strahilevitz J, Simantov K, et al. Neutrophils impose strong selective pressure against PfEMP1 variants implicated in cerebral malaria. *bioRxiv*. Published online May 9, 2021:2021.05.09.443317. doi:10.1101/2021.05.09.443317
519. Smith GD, Ebrahim S. Mendelian randomization: prospects, potentials, and limitations. *Int J Epidemiol*. 2004;33(1):30-42. doi:10.1093/ije/dyh132
520. Marchini J, Howie B, Myers S, McVean G, Donnelly P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature Genetics 2007 39:7*. 2007;39(7):906-913. doi:10.1038/ng2088
521. Burton PR, Clayton DG, Cardon LR, et al. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*. 2007;447(7145):661. doi:10.1038/NATURE05911
522. de Bakker PIW, Ferreira MAR, Jia X, Neale BM, Raychaudhuri S, Voight BF. Practical aspects of imputation-driven meta-analysis of genome-wide association studies. *Hum Mol Genet*. 2008;17(R2):R122. doi:10.1093/HMG/DDN288
523. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet*. 2011;88(1):76-82. doi:10.1016/j.ajhg.2010.11.011
524. Yang J, Ferreira T, Morris AP, et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat Genet*. 2012;44(4):369-S3. doi:10.1038/ng.2213
525. Marees AT, de Kluiver H, Stringer S, et al. A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. *Int J Methods Psychiatr Res*. 2018;27(2):1-10. doi:10.1002/mpr.1608
526. Pruim RJ, Welch RP, Sanna S, et al. LocusZoom: regional visualization of genome-wide association scan results. *BIOINFORMATICS APPLICATIONS NOTE*. 2010;26(18):2336-2337. doi:10.1093/bioinformatics/btq419
527. Boughton AP, Welch RP, Taliun D, et al. Interactive, shareable plots of GWAS data with LocusZoom.
528. Major T, Takei R. LocusZoom-like Plots for GWAS Results. Published online August 3, 2021. doi:10.5281/ZENODO.5154379
529. Yang J, Weedon MN, Purcell S, et al. Genomic inflation factors under polygenic inheritance. *European Journal of Human Genetics*. 2011;19(7):807. doi:10.1038/EJHG.2011.39

530. Ehret GB. Genome-Wide Association Studies: Contribution of Genomics to Understanding Blood Pressure and Essential Hypertension. *Curr Hypertens Rep*. 2010;12(1):17. doi:10.1007/S11906-009-0086-6
531. Watanabe K, Taskesen E, Van Bochoven A, Posthuma D. Functional mapping and annotation of genetic associations with FUMA. *Nature Communications 2017 8:1*. 2017;8(1):1-11. doi:10.1038/s41467-017-01261-5
532. Choi KW, Stein MB, Nishimi KM, et al. An exposure-wide and mendelian randomization approach to identifying modifiable factors for the prevention of depression. *American Journal of Psychiatry*. 2020;177(10):944-954. doi:10.1176/APPI.AJP.2020.19111158/ASSET/IMAGES/LARGE/APPI.AJP.2020.19111158F4.JPEG
533. Noyce AJ, Bandres-Ciga S, Kim J, et al. The Parkinson's Disease Mendelian Randomization Research Portal. *Movement Disorders*. 2019;34(12):1864. doi:10.1002/MDS.27873
534. Wheeler DL, Barrett T, Benson DA, et al. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*. 2007;35(Database issue):D5. doi:10.1093/NAR/GKL1031
535. McLaren W, Gil L, Hunt SE, et al. The Ensembl Variant Effect Predictor. *Genome Biol*. 2016;17(1):1-14. doi:10.1186/S13059-016-0974-4/TABLES/8
536. Cunningham F, Allen JE, Allen J, et al. Ensembl 2022. *Nucleic Acids Res*. 2022;50(D1):D988-D995. doi:10.1093/NAR/GKAB1049
537. Safran M, Rosen N, Twik M, et al. The GeneCards Suite. doi:10.1007/978-981-16-5812-9_2
538. Uhlén M, Fagerberg L, Hallström BM, et al. Tissue-based map of the human proteome. *Science (1979)*. 2015;347(6220). doi:10.1126/SCIENCE.1260419
539. Yang J, Benyamin B, McEvoy BP, et al. Common SNPs explain a large proportion of heritability for human height. *Nat Genet*. 2010;42(7):565. doi:10.1038/NG.608
540. Visscher PM, Hemani G, Vinkhuyzen AAE, et al. Statistical Power to Detect Genetic (Co)Variance of Complex Traits Using SNP Data in Unrelated Samples. *PLoS Genet*. 2014;10(4):e1004269. doi:10.1371/JOURNAL.PGEN.1004269
541. Yang J, Bakshi A, Zhu Z, et al. Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nature Genetics 2015 47:10*. 2015;47(10):1114-1120. doi:10.1038/ng.3390
542. Evans LM, Tahmasbi R, Vrieze SI, et al. Comparison of methods that use whole genome data to estimate the heritability and genetic architecture of complex traits. *Nature Genetics 2018 50:5*. 2018;50(5):737-745. doi:10.1038/s41588-018-0108-x

543. Sanna S, Jackson AU, Nagaraja R, et al. Common variants in the GDF5-UQCC region are associated with variation in human height. *Nat Genet*. 2008;40(2):198-203. doi:10.1038/NG.74
544. Willer CJ, Sanna S, Jackson AU, et al. Newly identified loci that influence lipid concentrations and risk of coronary artery disease. *Nat Genet*. 2008;40(2):161-169. doi:10.1038/NG.76
545. Bulik-Sullivan B, Loh PR, Finucane HK, et al. LD Score Regression Distinguishes Confounding from Polygenicity in Genome-Wide Association Studies. *Nat Genet*. 2015;47(3):291. doi:10.1038/NG.3211
546. The Python Language Reference — Python 3.7.13 documentation. Accessed August 5, 2022. https://docs.python.org/3.7/reference/
547. Age groups - GOV.UK Ethnicity facts and figures. Accessed August 17, 2022. https://www.ethnicity-facts-figures.service.gov.uk/uk-population-by-ethnicity/demographics/age-groups/latest
548. Health Survey for England: Weight. Accessed August 17, 2022. http://healthsurvey.hscic.gov.uk/data-visualisation/data-visualisation/explore-the-trends/weight.aspx
549. Goh L, Yap VB. Effects of normalization on quantitative traits in association test. *BMC Bioinformatics*. 2009;10:415. doi:10.1186/1471-2105-10-415
550. SHAPIRO SS, WILK MB. An analysis of variance test for normality (complete samples). *Biometrika*. 1965;52(3-4):591-611. doi:10.1093/BIOMET/52.3-4.591
551. Rahman MM, Govindarajulu Z. A modification of the test of Shapiro and Wilk for normality. *http://dx.doi.org/101080/02664769723828*. 2010;24(2):219-236. doi:10.1080/02664769723828
552. Sul JH, Martin LS, Eskin E. Population structure in genetic studies: Confounding factors and mixed models. *PLoS Genet*. 2018;14(12). doi:10.1371/JOURNAL.PGEN.1007309
553. Mathieson I, McVean G. Differential confounding of rare and common variants in spatially structured populations. *Nature Genetics 2012 44:3*. 2012;44(3):243-246. doi:10.1038/ng.1074
554. Hong EP, Park JW. Sample Size and Statistical Power Calculation in Genetic Association Studies. *Genomics Inform*. 2012;10(2):117. doi:10.5808/GI.2012.10.2.117
555. Retshabile G, Mlotshwa BC, Williams L, et al. Whole-Exome Sequencing Reveals Uncaptured Variation and Distinct Ancestry in the Southern African Population of Botswana. *Am J Hum Genet*. 2018;102(5):731-743. doi:10.1016/j.ajhg.2018.03.010

556. Klein RJ. Power analysis for genome-wide association studies. *BMC Genet*. 2007;8:58. doi:10.1186/1471-2156-8-58
557. Pierce BL, Burgess S. Efficient Design for Mendelian Randomization Studies: Subsample and 2-Sample Instrumental Variable Estimators. *Am J Epidemiol*. 2013;178(7):1177. doi:10.1093/AJE/KWT084
558. Weissbrod O, Kanai M, Shi H, et al. Leveraging fine-mapping and multi-population training data to improve cross-population polygenic risk scores. *Nat Genet*. 2022;54(4):450. doi:10.1038/S41588-022-01036-9
559. Coleman JRI, Euesden J, Patel H, Folarin AA, Newhouse S, Breen G. Quality control, imputation and analysis of genome-wide genotyping data from the Illumina HumanCoreExome microarray. *Brief Funct Genomics*. 2016;15(4). doi:10.1093/BFGP/ELV037
560. Kang HM, Sul JH, Service SK, et al. Variance component model to account for sample structure in genome-wide association studies. *Nat Genet*. 2010;42(4):348. doi:10.1038/NG.548
561. Polushina T, Giddaluru S, Bettella F, et al. Analysis of the joint effect of SNPs to identify independent loci and allelic heterogeneity in schizophrenia GWAS data. *Transl Psychiatry*. 2017;7(12):1289. doi:10.1038/s41398-017-0033-2
562. Lichou F, Trynka G. Functional studies of GWAS variants are gaining momentum. *Nature Communications 2020 11:1*. 2020;11(1):1-4. doi:10.1038/s41467-020-20188-y
563. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res*. 2010;38(16):e164-e164. doi:10.1093/NAR/GKQ603
564. He B, Shi J, Wang X, Jiang H, Zhu HJ. Genome-wide pQTL analysis of protein expression regulatory networks in the human liver. *BMC Biol*. 2020;18(1). doi:10.1186/S12915-020-00830-3
565. Machiela MJ, Chanock SJ. LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. *Bioinformatics*. 2015;31(21):3555. doi:10.1093/BIOINFORMATICS/BTV402
566. Nica AC, Dermitzakis ET. Expression quantitative trait loci: present and future. *Philosophical Transactions of the Royal Society B: Biological Sciences*. 2013;368(1620). doi:10.1098/RSTB.2012.0362
567. Xiong X, Lai X, Li A, Liu Z, Ma N. Diversity roles of CHD1L in normal cell function and tumorigenesis. *Biomark Res*. 2021;9(1). doi:10.1186/S40364-021-00269-W
568. Nilsen TW, Graveley BR. Expansion of the eukaryotic proteome by alternative splicing. *Nature*. 2010;463(7280):457. doi:10.1038/NATURE08909

569. Garrido-Martín D, Borsari B, Calvo M, Reverter F, Guigó R. Identification and analysis of splicing quantitative trait loci across multiple tissues in the human genome. *Nature Communications 2021 12:1*. 2021;12(1):1-16. doi:10.1038/s41467-020-20578-2
570. McCartney DL, Min JL, Richmond RC, et al. Genome-wide association studies identify 137 genetic loci for DNA methylation biomarkers of aging. *Genome Biol*. 2021;22(1):25. doi:10.1186/S13059-021-02398-9
571. Zhou F, Xing Y, Xu X, et al. NBPF is a potential DNA-binding transcription factor that is directly regulated by NF-κB. *Int J Biochem Cell Biol*. 2013;45(11):2479-2490. doi:10.1016/J.BIOCEL.2013.07.022
572. Moore CB, Verma A, Pendergrass S, et al. Phenome-wide Association Study Relating Pretreatment Laboratory Parameters With Human Genetic Variants in AIDS Clinical Trials Group Protocols. *Open Forum Infect Dis*. 2015;2(1). doi:10.1093/OFID/OFU113
573. Gurdasani D, Carstensen T, Fatumo S, et al. Uganda Genome Resource Enables Insights into Population History and Genomic Discovery in Africa. *Cell*. 2019;179(4). doi:10.1016/J.CELL.2019.10.004
574. Reiner AP, Lettre G, Nalls MA, et al. Genome-Wide Association Study of White Blood Cell Count in 16,388 African Americans: the Continental Origins and Genetic Epidemiology Network (COGENT). Abecasis GR, ed. *PLoS Genet*. 2011;7(6):e1002108. doi:10.1371/journal.pgen.1002108
575. Soremekun O, Soremekun C, Machipisa T, et al. Genome-Wide Association and Mendelian Randomization Analysis Reveal the Causal Relationship Between White Blood Cell Subtypes and Asthma in Africans. *Front Genet*. 2021;12:749415. doi:10.3389/FGENE.2021.749415/FULL
576. Jain D, Hodonsky CJ, Schick UM, et al. Genome-wide association of white blood cell counts in Hispanic/Latino Americans: the Hispanic Community Health Study/Study of Latinos. *Hum Mol Genet*. 2017;26(6). doi:10.1093/HMG/DDX024
577. Legge SE, Pardiñas AF, Helthuis M, et al. A genome-wide association study in individuals of African ancestry reveals the importance of the Duffy-null genotype in the assessment of clozapine-related neutropenia. *Molecular Psychiatry 2019 24:3*. 2019;24(3):328-337. doi:10.1038/s41380-018-0335-7
578. Kachuri L, Jeon S, DeWan AT, et al. Genetic determinants of blood-cell traits influence susceptibility to childhood acute lymphoblastic leukemia. *Am J Hum Genet*. 2021;108(10):1823-1835. doi:10.1016/J.AJHG.2021.08.004
579. Höglund J, Hadizadeh F, Ek WE, Karlsson T, Johansson Å. Gene-Based Variant Analysis of Whole-Exome Sequencing in Relation to Eosinophil Count. *Front Immunol*. 2022;13. doi:10.3389/FIMMU.2022.862255/FULL

580. Sakaue S, Kanai M, Tanigawa Y, et al. A cross-population atlas of genetic associations for 220 human phenotypes. *Nature Genetics 2021 53:10*. 2021;53(10):1415-1424. doi:10.1038/s41588-021-00931-x
581. Kim JH, Youn Y, Hwang JH. NCAPH Stabilizes GEN1 in Chromatin to Resolve Ultra-Fine DNA Bridges and Maintain Chromosome Stability. *Mol Cells*. 2022;45(11):792. doi:10.14348/MOLCELLS.2022.0048
582. Li C, Meng J, Zhang T. NCAPH is a prognostic biomarker and associated with immune infiltrates in lung adenocarcinoma. *Sci Rep*. 2022;12(1). doi:10.1038/S41598-022-12862-6
583. Fernandez-Rozadilla C, Timofeeva M, Chen Z, et al. Deciphering colorectal cancer genetics through multi-omic analysis of 100,204 cases and 154,587 controls of European and east Asian ancestries. 2022;55(1). doi:10.1038/s41588-022-01222-9
584. Kanai M, Akiyama M, Takahashi A, et al. Genetic analysis of quantitative traits in the Japanese population links cell types to complex human diseases. *Nature Genetics 2018 50:3*. 2018;50(3):390-400. doi:10.1038/s41588-018-0047-6
585. Kim MH, Kim S. Structures and functions of multi-tRNA synthetase complexes. *Enzymes (Essen)*. 2020;48:149-173. doi:10.1016/BS.ENZ.2020.06.008
586. De Filippo K, Rankin SM. CXCR4, the master regulator of neutrophil trafficking in homeostasis and disease. *Eur J Clin Invest*. 2018;48(Suppl Suppl 2). doi:10.1111/ECI.12949
587. Eash KJ, Means JM, White DW, Link DC. CXCR4 is a key regulator of neutrophil release from the bone marrow under basal and stress granulopoiesis conditions. *Blood*. 2009;113(19):4711. doi:10.1182/BLOOD-2008-09-177287
588. Weber C, Kraemer S, Drechsler M, et al. Structural determinants of MIF functions in CXCR2-mediated inflammatory and atherogenic leukocyte recruitment. *Proc Natl Acad Sci U S A*. 2008;105(42):16278. doi:10.1073/PNAS.0804017105
589. Ghosh S, Jiang N, Farr L, Ngobeni R, Moonah S. Parasite-produced MIF cytokine: Role in immune evasion, invasion, and pathogenesis. *Front Immunol*. 2019;10(AUG):1995. doi:10.3389/FIMMU.2019.01995/BIBTEX
590. Bando H, Pradipta A, Iwanaga S, et al. CXCR4 regulates Plasmodium development in mouse and human hepatocytes. *J Exp Med*. 2019;216(8):1733.
591. Rodrigues DAS, Prestes EB, Gama AMS, et al. CXCR4 and MIF are required for neutrophil extracellular trap release triggered by Plasmodium-infected erythrocytes. *PLoS Pathog*. 2020;16(8):e1008230. doi:10.1371/JOURNAL.PPAT.1008230

592. Schipper S, Springer E, Hahn J, et al. Characterization of Plasmodium falciparum macrophage migration inhibitory factor homologue and its cysteine deficient mutants. *Parasitol Int*. 2022;87:102513. doi:10.1016/J.PARINT.2021.102513
593. Park S, Song J, Baek IJ, et al. Loss of Acot12 contributes to NAFLD independent of lipolysis of adipose tissue. *Experimental & Molecular Medicine 2021 53:7*. 2021;53(7):1159-1169. doi:10.1038/s12276-021-00648-1
594. Lu M, Zhu WW, Wang X, et al. ACOT12-Dependent Alteration of Acetyl-CoA Drives Hepatocellular Carcinoma Metastasis by Epigenetic Induction of Epithelial-Mesenchymal Transition. *Cell Metab*. 2019;29(4):886-900.e5. doi:10.1016/j.cmet.2018.12.019
595. Kichaev G, Bhatia G, Loh PR, et al. Leveraging Polygenic Functional Enrichment to Improve GWAS Power. *Am J Hum Genet*. 2019;104(1):65. doi:10.1016/J.AJHG.2018.11.008
596. Xing J, Zhao X, Li X, et al. Variation at ACOT12 and CT62 locus represents susceptibility to psoriasis in Han population. *Mol Genet Genomic Med*. 2020;8(2). doi:10.1002/MGG3.1098
597. Pérez MM, Pimentel VE, Fuzo CA, et al. Acetylcholine, Fatty Acids, and Lipid Mediators Are Linked to COVID-19 Severity. *The Journal of Immunology*. 2022;209(2):250-261. doi:10.4049/JIMMUNOL.2200079
598. Sutter ML, Console L, Fahner AF, et al. The role of cholesterol recognition (CARC/CRAC) mirror codes in the allosterism of the human organic cation transporter 2 (OCT2, SLC22A2). *Biochem Pharmacol*. 2021;194:114840. doi:10.1016/J.BCP.2021.114840
599. Yee SW, Giacomini KM. Special Section on New Era of Transporter Science-Minireview Emerging Roles of the Human Solute Carrier 22 Family S. Published online 2022. doi:10.1124/dmd.121.000702
600. Sinnott-Armstrong N, Tanigawa Y, Amar D, et al. Genetics of 35 blood and urine biomarkers in the UK Biobank. *Nat Genet*. 2021;53(2):185. doi:10.1038/S41588-020-00757-Z
601. Kjaerulff O, Brodin L, Jung A. The Structure and Function of Endophilin Proteins. *Cell Biochemistry and Biophysics 2010 60:3*. 2010;60(3):137-154. doi:10.1007/S12013-010-9137-5
602. Oskarsson GR, Oddsson A, Magnusson MK, et al. Predicted loss and gain of function mutations in ACO1 are associated with erythropoiesis. *Commun Biol*. 2020;3(1). doi:10.1038/S42003-020-0921-5
603. Yang L, Gong X, Wang J, et al. Functional mechanisms of TRPS1 in disease progression and its potential role in personalized medicine. *Pathol Res Pract*. 2022;237:154022. doi:10.1016/J.PRP.2022.154022

604. Elster D, Tollot M, Schlegelmilch K, et al. TRPS1 shapes YAP/TEAD-dependent transcription in breast cancer cells. *Nat Commun*. 2018;9(1). doi:10.1038/S41467-018-05370-7
605. Tängdén T, Gustafsson S, Rao AS, Ingelsson E. A genome-wide association study in a large community-based cohort identifies multiple loci associated with susceptibility to bacterial and viral infections. *Scientific Reports 2022 12:1*. 2022;12(1):1-14. doi:10.1038/s41598-022-05838-z
606. Yosef N, Shalek AK, Gaublomme JT, et al. Dynamic regulatory network controlling TH17 cell differentiation. *Nature 2013 496:7446*. 2013;496(7446):461-468. doi:10.1038/nature11981
607. Lopez M, Choin J, Sikora M, et al. Genomic Evidence for Local Adaptation of Hunter-Gatherers to the African Rainforest. *Current Biology*. 2019;29(17):2926-2935.e4. doi:10.1016/J.CUB.2019.07.013
608. Yang L, Wu Y, Xu H, et al. Identification and Validation of a Novel Six-lncRNA-Based Prognostic Model for Lung Adenocarcinoma. *Front Oncol*. 2022;11:775583. doi:10.3389/FONC.2021.775583/FULL
609. Cano-Gamez E, Trynka G. From GWAS to Function: Using Functional Genomics to Identify the Mechanisms Underlying Complex Diseases. *Front Genet*. 2020;11:424. doi:10.3389/FGENE.2020.00424/BIBTEX
610. Thaler RH. Anomalies: The Winner's Curse. *Journal of Economic Perspectives*. 1988;2(1):191-202. doi:10.1257/JEP.2.1.191
611. Panagiotou OA, Ioannidis JPA, Hirschhorn JN, et al. What should the genome-wide significance threshold be? Empirical replication of borderline genetic associations. *Int J Epidemiol*. 2012;41(1):273-286. doi:10.1093/IJE/DYR178
612. Chen Z, Boehnke M, Wen X, Mukherjee B. Revisiting the genome-wide significance threshold for common variant GWAS. *G3 Genes|Genomes|Genetics*. 2021;11(2). doi:10.1093/G3JOURNAL/JKAA056
613. Risch N, Merikangas K. The Future of Genetic Studies of Complex Human Diseases. *Science (1979)*. 1996;273(5281):1516-1517. doi:10.1126/SCIENCE.273.5281.1516
614. Kraft P. Curses - Winner's and otherwise - In genetic epidemiology. *Epidemiology*. 2008;19(5):649-651. doi:10.1097/EDE.0B013E318181B865
615. Ioannidis JPA. Why most discovered true associations are inflated. *Epidemiology*. 2008;19(5):640-648. doi:10.1097/EDE.0B013E31818131E7
616. Schwartz E, Sadetzki S, Murad H, Raveh D. Age as a risk factor for severe Plasmodium falciparum malaria in nonimmune patients. *Clinical Infectious Diseases*. 2001;33(10):1774-1777. doi:10.1086/322522/2/33-10-1774-TBL003.GIF

617. Carneiro I, Roca-Feltrer A, Griffin JT, et al. Age-Patterns of Malaria Vary with Severity, Transmission Intensity and Seasonality in Sub-Saharan Africa: A Systematic Review and Pooled Analysis. *PLoS One*. 2010;5(2). doi:10.1371/JOURNAL.PONE.0008988
618. Ortmann W, Kolaczkowska E. Age is the work of art? Impact of neutrophil and organism age on neutrophil extracellular trap formation. *Cell and Tissue Research 2017 371:3*. 2017;371(3):473-488. doi:10.1007/S00441-017-2751-4
619. World Population Prospects - Population Division - United Nations. Published 2019. Accessed January 14, 2023. https://population.un.org/wpp/
620. Brækkan SK, Mathiesen EB, NjøLstad I, Wilsgaard T, StøRmer J, Hansen JB. Mean platelet volume is a risk factor for venous thromboembolism: the Tromsø study. *Journal of Thrombosis and Haemostasis*. 2010;8(1):157-162. doi:10.1111/J.1538-7836.2009.03498.X
621. Wendelboe AM, Raskob GE. Global Burden of Thrombosis. *Circ Res*. 2016;118(9):1340-1347. doi:10.1161/CIRCRESAHA.115.306841
622. Baaten CCFMJ, ten Cate H, van der Meijden PEJ, Heemskerk JWM. Platelet populations and priming in hematological diseases. *Blood Rev*. 2017;31(6):389-399. doi:https://doi.org/10.1016/j.blre.2017.07.004
623. Mackman N. New insights into the mechanisms of venous thrombosis. *J Clin Invest*. 2012;122(7):2331-2336. doi:10.1172/JCI60229
624. Stone J, Hangge P, Albadawi H, et al. Deep vein thrombosis: pathogenesis, diagnosis, and medical management. *Cardiovasc Diagn Ther*. 2017;7(Suppl 3):S276-S284. doi:10.21037/cdt.2017.09.01
625. *Heart Disease and Stroke Statistics-2021 Update A Report from the American Heart Association*. Lippincott Williams and Wilkins; 2021.
626. ONS. Mortality statistics. Official Labour Market Statistics.
627. Silverstein MD, Heit JA, Mohr DN, Petterson TM, O'Fallon WM, Melton LJ. Trends in the Incidence of Deep Vein Thrombosis and Pulmonary Embolism: A 25-Year Population-Based Study. *Arch Intern Med*. 1998;158(6):585-593. doi:10.1001/ARCHINTE.158.6.585
628. What is Venous Thromboembolism? | CDC. Accessed September 23, 2021. https://www.cdc.gov/ncbddd/dvt/facts.html
629. Giustozzi M, Franco L, Vedovati MC, Becattini C, Agnelli G. Safety of direct oral anticoagulants versus traditional anticoagulants in venous thromboembolism. *J Thromb Thrombolysis*. 2019;48(3):439-453. doi:10.1007/S11239-019-01878-X/TABLES/4
630. Samuelson Bannow BT, Konkle BA. Laboratory biomarkers for venous thromboembolism risk in patients with hematologic malignancies: A review.

Thromb Res. 2018;163:138-145. doi:https://doi.org/10.1016/j.thromres.2018.01.037

631. Klovaite J, Benn M, Nordestgaard BG. Obesity as a causal risk factor for deep venous thrombosis: a Mendelian randomization study. *J Intern Med*. 2014;277(5):573-584. doi:10.1111/joim.12299
632. Memon Ashfaque A, Sundquist K, PirouziFard M, et al. Identification of novel diagnostic biomarkers for deep venous thrombosis. *Br J Haematol*. 2018;181(3):378-385. doi:10.1111/bjh.15206
633. Wijten P, Holten T van, Woo LL, et al. High Precision Platelet Releasate Definition by Quantitative Reversed Protein Profiling—Brief Report. *Arterioscler Thromb Vasc Biol*. 2013;33(7):1635-1638. doi:10.1161/ATVBAHA.113.301147
634. Gold L, Walker JJ, Wilcox SK, Williams S. Advances in human proteomics at high scale with the SOMAscan proteomics platform. *N Biotechnol*. 2012;29(5):543-549. doi:10.1016/J.NBT.2011.11.016
635. Imming P, Sinning C, Meyer A. Drugs, their targets and the nature and number of drug targets. *Nat Rev Drug Discov*. 2006;5(10):821-834. doi:10.1038/nrd2132
636. Panova-Noeva M, Wagner B, Nagler M, et al. Comprehensive platelet phenotyping supports the role of platelets in the pathogenesis of acute venous thromboembolism – results from clinical observation studies. *EBioMedicine*. 2020;60:102978. doi:10.1016/j.ebiom.2020.102978
637. Xiong X, Li T, Yu S, Cheng B. Association Between Platelet Indices and Preoperative Deep Vein Thrombosis in Elderly Patients Undergoing Total Joint Arthroplasty: A Retrospective Study. *Clinical and Applied Thrombosis/Hemostasis*. 2023;29. doi:10.1177/10760296221149699/ASSET/IMAGES/LARGE/10.1177_10760296221149699-FIG2.JPEG
638. Edvardsen MS, Hansen ES, Hindberg K, et al. Combined effects of plasma von Willebrand factor and platelet measures on the risk of incident venous thromboembolism. *Blood*. 2021;138(22):2269-2277. doi:10.1182/BLOOD.2021011494
639. Goudswaard LJ, Bell JA, Hughes DA, et al. Effects of adiposity on the human plasma proteome: observational and Mendelian randomisation estimates. *International Journal of Obesity 2021*. 2021;6(2):1-9. doi:10.1038/s41366-021-00896-1
640. Zaghlool SB, Sharma S, Molnar M, et al. Revealing the role of the human blood plasma proteome in obesity using genetic drivers. *Nature Communications 2021 12:1*. 2021;12(1):1-13. doi:10.1038/s41467-021-21542-4

641. Millard LAC, Davies NM, Gaunt TR, Davey Smith G, Tilling K. Software Application Profile: PHESANT: a tool for performing automated phenome scans in UK Biobank. *Int J Epidemiol*. 2018;47(1):29-35. doi:10.1093/ije/dyx204
642. Battle A, Khan Z, Wang SH, et al. Impact of Regulatory Variation from RNA to Protein. *Science*. 2015;347(6222):664. doi:10.1126/SCIENCE.1260793
643. Sun BB, Maranville JC, Peters JE, et al. Genomic atlas of the human plasma proteome. *Nature*. 2018;558(7708):73-79. doi:10.1038/s41586-018-0175-2
644. Folkersen L, Fauman E, Sabater-Lleal M, et al. Mapping of 79 loci for 83 plasma protein biomarkers in cardiovascular disease. *PLoS Genet*. 2017;13(4):e1006706. doi:10.1371/journal.pgen.1006706
645. Suhre K, Arnold M, Bhagwat AM, et al. Connecting genetic risk to disease end points through the human blood plasma proteome. *Nat Commun*. 2017;8:14357. doi:10.1038/ncomms14357
https://www.nature.com/articles/ncomms14357#supplementary-information
646. Yao C, Chen G, Song C, et al. Genome-wide mapping of plasma protein QTLs identifies putatively causal genes and pathways for cardiovascular disease. *Nat Commun*. 2018;9(1):3268. doi:10.1038/s41467-018-05512-x
647. Hemani G, Zheng J, Wade KH, et al. MR-Base: a platform for systematic causal inference across the phenome using billions of genetic associations. *bioRxiv*. Published online 2016.
648. Bowden J, Spiller W, Del Greco M F, et al. Improving the visualization, interpretation and analysis of two-sample summary data Mendelian randomization via the Radial plot and Radial regression. *Int J Epidemiol*. 2018;47(4):1264-1278. doi:10.1093/ije/dyy101
649. Zheng J, Richardson T, Millard L, et al. PhenoSpD: an atlas of phenotypic correlations and a multiple testing correction for the human phenome. *bioRxiv*. Published online 2017. doi:10.1101/148627
650. Nyholt DR. A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. *Am J Hum Genet*. 2004;74(4):765-769. doi:10.1086/383251
651. Li J, Ji L. Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity (Edinb)*. 2005;95(3):221-227. doi:10.1038/sj.hdy.6800717
652. Lloyd-Jones LR, Robinson MR, Yang J, Visscher PM. Transformation of Summary Statistics from Linear Mixed Model Association on All-or-None Traits to Odds Ratio. *Genetics*. 2018;208(4):1397.

653. Giambartolomei C, Vukcevic D, Schadt EE, et al. Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. *PLoS Genet*. 2014;10(5):e1004383. doi:10.1371/journal.pgen.1004383
654. Lyon MS, Andrews SJ, Elsworth B, Gaunt TR, Hemani G, Marcora E. The variant call format provides efficient and robust storage of GWAS summary statistics. *Genome Biology 2021 22:1*. 2021;22(1):1-10. doi:10.1186/S13059-020-02248-0
655. Boyd A, Golding J, Macleod J, et al. Cohort Profile: The 'Children of the 90s'—the index offspring of the Avon Longitudinal Study of Parents and Children. *Int J Epidemiol*. 2013;42(1):111. doi:10.1093/IJE/DYS064
656. Fraser A, Macdonald-wallis C, Tilling K, et al. Cohort Profile: The Avon Longitudinal Study of Parents and Children: ALSPAC mothers cohort. *Int J Epidemiol*. 2013;42(1):97. doi:10.1093/IJE/DYS066
657. Gregson J, Kaptoge S, Bolton T, et al. Cardiovascular Risk Factors Associated With Venous Thromboembolism. *JAMA Cardiol*. 2019;4(2):163-173. doi:10.1001/JAMACARDIO.2018.4537
658. Zeng H, Lin C, Wang S, Zheng Y, Gao X. Genetically predicted body composition in relation to cardiometabolic traits: a Mendelian randomization study. *Eur J Epidemiol*. 2021;36(11):1157-1168. doi:10.1007/S10654-021-00779-9/FIGURES/6
659. Roetker NS, Armasu SM, Pankow JS, et al. Taller height as a risk factor for venous thromboembolism: a Mendelian randomization meta-analysis. *Journal of Thrombosis and Haemostasis*. 2017;15(7):1334-1343. doi:10.1111/JTH.13719
660. Cushman M, O'Meara ES, Heckbert SR, Zakai NA, Rosamond W, Folsom AR. Body size measures, hemostatic and inflammatory markers and risk of venous thrombosis: The Longitudinal Investigation of Thromboembolism Etiology. *Thromb Res*. 2016;144:127-132. doi:10.1016/j.thromres.2016.06.012
661. Samama MM, Group for the SS. An Epidemiologic Study of Risk Factors for Deep Vein Thrombosis in Medical Outpatients: The Sirius Study. *Arch Intern Med*. 2000;160(22):3415-3420. doi:10.1001/ARCHINTE.160.22.3415
662. Srisawat S, Sitasuwan T, Ungprasert P. Increased risk of venous thromboembolism among patients with hyperthyroidism: a systematic review and meta-analysis of cohort studies. *Eur J Intern Med*. 2019;67:65-69. doi:10.1016/J.EJIM.2019.06.012
663. Homoncik M, Gessl A, Ferlitsch A, Jilma B, Vierhapper H. Altered Platelet Plug Formation in Hyperthyroidism and Hypothyroidism. *J Clin Endocrinol Metab*. 2007;92(8):3006-3012. doi:10.1210/JC.2006-2644

664. Horacek J, Maly J, Svilias I, et al. Prothrombotic changes due to an increase in thyroid hormone levels. *Eur J Endocrinol*. 2015;172(5):537-542. doi:10.1530/EJE-14-0801
665. Mousa SS, Davis FB, Davis PJ, Mousa SA. Human Platelet Aggregation and Degranulation Is Induced In Vitro by L-Thyroxine, but Not by 3,5,3′-Triiodo-L-Thyronine or Diiodothyropropionic Acid (DITPA): *http://dx.doi.org/101177/1076029609348315*. 2009;16(3):288-293. doi:10.1177/1076029609348315
666. Chang SLSW, Hu S, Huang YL, et al. Treatment of Varicose Veins Affects the Incidences of Venous Thromboembolism and Peripheral Artery Disease. *Circ Cardiovasc Interv*. Published online 2021.
667. Müller B, Leutgeb, Engeser, Achankeng N, Szecsenyi, Laux. Varicose veins are a risk factor for deep venous thrombosis in general practice patients. *Vasa*. 2012;41(5):360-365. doi:10.1024/0301-1526/a000222
668. BERTOLETTI L, COUTURAUD F. COPD is not only one of the several VTE risk factors. *Eur J Intern Med*. 2021;84:14-15. doi:10.1016/J.EJIM.2020.12.013
669. Lankeit M, Held M. Incidence of venous thromboembolism in COPD: linking inflammation and thrombosis? doi:10.1183/13993003.01679-2015
670. Kaze AD, Bigna JJ, Nansseu JR, Noubiap JJ. Body size measures and risk of venous thromboembolism: protocol for a systematic review and meta-analysis. *BMJ Open*. 2018;8(3):e018958-e018958. doi:10.1136/bmjopen-2017-018958
671. Thaler E, Lechner K. Antithrombin III Deficiency and Thromboembolism. *Clin Haematol*. 1981;10(2):369-390. doi:10.1016/S0308-2261(21)00229-0
672. Tang J, Zhu W, Mei X, Zhang Z. Plasminogen activator inhibitor-1: A risk factor for deep vein thrombosis after total hip arthroplasty. *J Orthop Surg Res*. 2018;13(1):1-5. doi:10.1186/S13018-018-0716-2/TABLES/3
673. Maki RG. Small Is Beautiful: Insulin-Like Growth Factors and Their Role in Growth, Development, and Cancer. *Journal of Clinical Oncology*. 2010;28(33):4985. doi:10.1200/JCO.2009.27.5040
674. Fashanu OE, Heckbert SR, Aguilar D, et al. Galectin-3 and venous thromboembolism incidence: the Atherosclerosis Risk in Communities (ARIC) Study. *Res Pract Thromb Haemost*. 2017;1(2):223-230. doi:10.1002/RTH2.12038
675. Huebner BR, Moore EE, Moore HB, et al. Thrombin Provokes Degranulation of Platelet α-Granules Leading to the Release of Active Plasminogen Activator Inhibitor-1 (PAI-1). *Shock*. 2018;50(6):671. doi:10.1097/SHK.0000000000001089
676. Klarin D, Busenkell E, Judy R, et al. Genome-wide association analysis of venous thromboembolism identifies new risk loci and genetic overlap with arterial vascular

disease. *Nature Genetics 2019 51:11*. 2019;51(11):1574-1579. doi:10.1038/s41588-019-0519-3

677. Klarin D, Emdin CA, Natarajan P, et al. Genetic Analysis of Venous Thromboembolism in UK Biobank Identifies the ZFPM2 Locus and Implicates Obesity as a Causal Risk Factor. *Circ Cardiovasc Genet*. 2017;10(2):e001643. doi:10.1161/CIRCGENETICS.116.001643
678. Isogai C, Laug WE, Shimada H, et al. Plasminogen activator inhibitor-1 promotes angiogenesis by stimulating endothelial cell migration toward fibronectin. *Cancer Res*. 2001;61(14):5587-5594.
679. Hjortland GO, Lillehammer T, Somme S, et al. Plasminogen activator inhibitor-1 increases the expression of VEGF in human glioma cells. *Exp Cell Res*. 2004;294(1):130-139. doi:10.1016/J.YEXCR.2003.10.013
680. ZHANG Q, LEI L, JING D. Knockdown of SERPINE1 reverses resistance of triple-negative breast cancer to paclitaxel via suppression of VEGFA. *Oncol Rep*. 2020;44(5):1875. doi:10.3892/OR.2020.7770
681. Zhang Q, Zhang X, Zhang J, et al. Vascular endothelial growth factor and the risk of venous thromboembolism: a genetic correlation and two-sample Mendelian randomization study. *Thromb J*. 2022;20(1):1-11. doi:10.1186/S12959-022-00427-6/TABLES/2
682. Frischmuth T, Hindberg K, Aukrust P, et al. Elevated plasma levels of plasminogen activator inhibitor-1 are associated with risk of future incident venous thromboembolism. *Journal of Thrombosis and Haemostasis*. 2022;20(7):1618-1626. doi:10.1111/JTH.15701
683. Mo JW, Zhang DF, Ji GL, Liu XZ, Fan B. TGF-β1 and serpine 1 expression changes in traumatic deep vein thrombosis. *Genetics and Molecular Research*. 2015;14(4):13835-13842. doi:10.4238/2015.October.29.3
684. Shimomura I, Funahashi T, Takahashi M, et al. Enhanced expression of PAI–1 in visceral fat: Possible contributor to vascular disease in obeisty. *Nature Medicine 1996 2:7*. 1996;2(7):800-803. doi:10.1038/nm0796-800
685. Gadekar T, Dudeja P, Basu I, Vashisht S, Mukherji S. Correlation of visceral body fat with waist–hip ratio, waist circumference and body mass index in healthy adults: A cross sectional study. *Med J Armed Forces India*. 2020;76(1):41-46. doi:10.1016/J.MJAFI.2017.12.001
686. Sillen M, Declerck PJ. Targeting PAI-1 in Cardiovascular Disease: Structural Insights Into PAI-1 Functionality and Inhibition. *Front Cardiovasc Med*. 2020;7:364. doi:10.3389/FCVM.2020.622473/BIBTEX
687. Laumen H, Skurk T, Hauner H. The HMG-CoA reductase inhibitor rosuvastatin inhibits plasminogen activator inhibitor-1 expression and secretion in human

adipocytes. *Atherosclerosis*. 2008;196(2):565-573. doi:10.1016/J.ATHEROSCLEROSIS.2007.06.005

688. Glynn RJ, Danielson E, Fonseca FA, et al. A Randomized Trial of Rosuvastatin in the Prevention of Venous Thromboembolism: the JUPITER Trial. *N Engl J Med*. 2009;360(18):1851. doi:10.1056/NEJMOA0900241
689. Schol-Gelok S, de Maat MPM, Biedermann JS, et al. Rosuvastatin use increases plasma fibrinolytic potential: a randomised clinical trial. *Br J Haematol*. 2020;190(6):916-922. doi:10.1111/BJH.16648
690. Jückstock J, Kimmich T, Mylonas I, Friese K, Dian D. The inhibin-βC subunit is down-regulated, while inhibin-βE is up-regulated by interferon-β1a in Ishikawa carcinoma cell line. *Archives of Gynecology and Obstetrics 2013 288:4*. 2013;288(4):883-888. doi:10.1007/S00404-013-2848-2
691. Thomas TZ, Chapman SM, Hong W, et al. Inhibins, Activins, and Follistatins: Expression of mRNAs and Cellular Localization in Tissues From Men With Benign Prostatic Hyperplasia. *Prostate*. 1998;34:34-43. doi:10.1002/(SICI)1097-0045(19980101)34:1
692. Détriché G, Gendron N, Philippe A, et al. Gonadotropins as novel active partners in vascular diseases: Insight from angiogenic properties and thrombotic potential of endothelial colony-forming cells. *Journal of Thrombosis and Haemostasis*. 2022;20(1):230-237. doi:10.1111/JTH.15549
693. LaFoya B, Munroe JA, Mia MM, et al. Notch: A multi-functional integrating system of microenvironmental signals. *Dev Biol*. 2016;418(2):227-241. doi:10.1016/J.YDBIO.2016.08.023
694. Chaurasia SN, Ekhlak M, Kushwaha G, Singh V, Mallick RL, Dash D. Notch signaling functions in noncanonical juxtacrine manner in platelets to amplify thrombogenicity. *Elife*. 2022;11:1-25. doi:10.7554/eLife.79590
695. Mancarella S, Serino G, Dituri F, et al. Crenigacestat, a selective NOTCH1 inhibitor, reduces intrahepatic cholangiocarcinoma progression by blocking VEGFA/DLL4/MMP13 axis. *Cell Death & Differentiation 2020 27:8*. 2020;27(8):2330-2343. doi:10.1038/s41418-020-0505-4
696. Casulo C, Ruan J, Dang NH, et al. Safety and Preliminary Efficacy Results of a Phase I First-in-Human Study of the Novel Notch-1 Targeting Antibody Brontictuzumab (OMP-52M51) Administered Intravenously to Patients with Hematologic Malignancies. *Blood*. 2016;128(22):5108. doi:10.1182/BLOOD.V128.22.5108.5108
697. Zhou Y, Zhang Y, Lian X, et al. Therapeutic target database update 2022: facilitating drug discovery with enriched comparative data of targeted agents. *Nucleic Acids Res*. 2022;50(D1):D1398. doi:10.1093/NAR/GKAB953

698. Konstantinides S V., Torbicki A, Agnelli G, et al. 2014 ESC Guidelines on the diagnosis and management of acute pulmonary embolismThe Task Force for the Diagnosis and Management of Acute Pulmonary Embolism of the European Society of Cardiology (ESC)Endorsed by the European Respiratory Society (ERS). *Eur Heart J*. 2014;35(43):3033-3080. doi:10.1093/EURHEARTJ/EHU283
699. Yu N, Chen FC, Ota S, et al. *Larger Genetic Differences Within Africans Than Between Africans and Eurasians*.
700. Thibord F, Klarin D, Brody JA, et al. Cross-Ancestry Investigation of Venous Thromboembolism Genomic Predictors. *Circulation*. 2022;146(16):1225-1242. doi:10.1161/CIRCULATIONAHA.122.059675
701. Fatumo S, Mugisha J, Soremekun OS, et al. Uganda Genome Resource: A rich research database for genomic studies of communicable and non-communicable diseases in Africa. *Cell Genomics*. 2022;2(11):100209. doi:10.1016/J.XGEN.2022.100209
702. Kanduri C, Sandve GK, Hovig E, De S, Layer RM. Editorial: Genomic Colocalization and Enrichment Analyses. *Front Genet*. 2021;11:1621. doi:10.3389/FGENE.2020.617876/BIBTEX
703. Võsa U, Claringbould A, Westra HJ, et al. Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. *Nature Genetics 2021 53:9*. 2021;53(9):1300-1310. doi:10.1038/s41588-021-00913-z
704. Grant AJ, Gill D, Kirk PDW, Burgess S. Noise-augmented directional clustering of genetic association data identifies distinct mechanisms underlying obesity. 2022;18(1):e1009975. doi:10.1371/JOURNAL.PGEN.1009975
705. Burgess S, Mason AM, Grant AJ, et al. Using genetic association data to guide drug discovery and development: Review of methods and applications. *The American Journal of Human Genetics*. 2023;110(2):195-214. doi:10.1016/J.AJHG.2022.12.017
706. Elhaik E. Principal Component Analyses (PCA)-based findings in population genetic studies are highly biased and must be reevaluated. *Scientific Reports 2022 12:1*. 2022;12(1):1-35. doi:10.1038/s41598-022-14395-4
707. Babatunde KA, Adenuga OF. Neutrophils in malaria: A double-edged sword role. *Front Immunol*. 2022;13:3993.
708. Kenny EE, Timpson NJ, Sikora M, et al. Melanesian blond hair is caused by an amino acid change in TYRP1. *Science (1979)*. 2012;336(6081):554. doi:10.1126/SCIENCE.1217849/SUPPL_FILE/KENNY.SM.PDF

709. Campbell MC, Tishkoff SA. African Genetic Diversity: Implications for Human Demographic History, Modern Human Origins, and Complex Disease Mapping. doi:10.1146/annurev.genom.9.081307.164258
710. Collins FS, Doudna JA, Lander ES, Rotimi CN. Human Molecular Genetics and Genomics — Important Advances and Exciting Possibilities. *New England Journal of Medicine*. 2021;384(1):1-4. doi:10.1056/NEJMP2030694/SUPPL_FILE/NEJMP2030694_DISCLOSURES.PDF
711. Our Future Health. Accessed March 2, 2023. https://ourfuturehealth.org.uk/
712. Seven things you need to know about Our Future Health – Our Future Health. Accessed March 2, 2023. https://ourfuturehealth.org.uk/news/seven-things-you-need-to-know-about-the-uks-largest-ever-health-research-programme/
713. Fry A, Littlejohns TJ, Sudlow C, et al. Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. *Am J Epidemiol*. 2017;186(9):1026-1034. doi:10.1093/aje/kwx246
714. Squire P. WHY THE 1936 LITERARY DIGEST POLL FAILED. *Public Opin Q*. 1988;52(1):125-133. doi:10.1086/269085
715. Caelers D. Plan for network of Genomics Centres of Excellence across Africa. *Nature Africa*. Published online February 26, 2023. doi:10.1038/D44148-023-00052-Z
716. Burgess S, Timpson NJ, Ebrahim S, Smith GD. Mendelian randomization: Where are we now and where are we going? *Int J Epidemiol*. 2015;44(2):379-388. doi:10.1093/ije/dyv108
717. Liu J, Chou EL, Lau KK, et al. A Mendelian randomization-based exploration of red blood cell distribution width and mean corpuscular volume with risk of hemorrhagic strokes. *Human Genetics and Genomics Advances*. 2022;3(4):100135. doi:10.1016/J.XHGG.2022.100135
718. Pongdee T, Bielinski SJ, Decker PA, Kita H, Larson NB. White blood cells and chronic rhinosinusitis: a Mendelian randomization study. *Allergy, Asthma & Clinical Immunology 2022 18:1*. 2022;18(1):1-8. doi:10.1186/S13223-022-00739-2
719. Guyatt A, John C, Williams AT, et al. Mendelian randomisation of eosinophils and other cell types in relation to lung function and disease. *Thorax*. Published online May 10, 2022:thoraxjnl-2021-217993. doi:10.1136/THORAXJNL-2021-217993
720. Ulrich A, Wharton J, Thayer TE, et al. Mendelian randomisation analysis of red cell distribution width in pulmonary arterial hypertension. *European Respiratory Journal*. 2020;55(2). doi:10.1183/13993003.01486-2019

APPENDICES

Appendices for the thesis are found below and are linked throughout the text.

Appendix 1. *Description of meta-analysed studies for WBC count.*

Study	Age	% Female	BASO (10^9/L)	EOS (10^9/L)	LYM (10^9/L)	MONO (10^9/L)	NEU (10^9/L)	WBC (10^9/L)	Full study name
Airwave	40.88 (9.03)	37.40	0.06 (0.04)	0.20 (0.12)	1.81 (0.62)	0.40 (0.17)	3.95 (1.36)	6.53 (1.72)	Airwave Health Monitoring Study
n	13113	4910	13104	13105	13104	13105	13105	13105	
BioME (EUR)	69.76 (9.74)	43.1	0.02 (0.04)	0.15 (0.13)	1.69 (1.00)	0.56 (0.20)	4.65 (2.45)	7.20 (2.46)	BioMe™ BioBank Program
n	1861	802	460	460	460	460	460	789	
CaPS	65.46 (4.45)	0	0.08 (0.03)	0.22 (0.17)	2.75 (0.72)	0.75 (0.18)	5.98 (0.83)	7.09 (2.03)	Caerphilly Prospective Study
n	1181	0	1173	1176	1173	1139	1177	1179	
CHS (EUR)	72.33 (5.38)	60.79	-	-	-	-	-	6.26 (1.94)	Cardiovascular Health Study
n	3249	1975	-	-	-	-	-	3249	
Estonia_chip	39.0 (15.8)	51.24	0.03 (0.03)	0.16 (0.13)	1.95 (0.58)	0.52 (0.18)	3.67 (1.41)	6.34 (1.88)	Estonia SNP Chip
n	1085	556	1084	1081	1079	1079	1081	1084	
Estonia_WGS	50.5 (15.7)	49.65	0.03 (0.02)	0.15 (0.12)	1.97 (0.64)	0.51 (0.18)	3.68 (1.39)	6.36 (1.92)	Estonia Whole Genome Sequencing
n	1009	501	1009	1009	1007	1009	1009	1008	
FHS	55.86 (16.27)	52.78	0.04 (0.02)	0.19 (0.12)	1.63 (0.58)	0.52 (0.15)	3.63 (1.21)	6.15 (1.65)	Framingham Heart Study
n	6458	3409	4293	4293	4293	4293	4293	6131	
FINCAVAS	53.3 (13.9)	42	0.04 (0.03)	0.23 (0.20)	2.18 (1.05)	0.52 (0.22)	4.13 (2.15)	7.26 (2.45)	The Finnish Cardiovascular Study
n	911	383	396	436	396	396	396	910	
GERA (EUR)	63.19 (12.95)	61.15	-	-	1.89 (0.73)	0.58 (0.22)	4.06 (1.92)	6.67 (2.27)	Genetic Epidemiology Research on Adult Health and Aging
n	53822	32912	-	-	43479	43475	43479	53822	
GERA (EUR_LATchip)	60.87 (13.35)	70.48	-	-	2.01 (0.68)	0.6 (0.23)	4.3 (2.16)	7.06 (2.79)	Genetic Epidemiology Research on Adult Health and Aging

Study	Age	% Female	BASO (10^9/L)	EOS (10^9/L)	LYM (10^9/L)	MONO (10^9/L)	NEU (10^9/L)	WBC (10^9/L)	Full study name
n	1504	1060	-	-	1204	1204	1204	1504	
INTERVAL	44.15 (13.88)	49.8	0.04 (0.04)	0.17 (0.13)	1.95 (0.53)	0.53 (0.15)	3.75 (1.25)	6.47 (1.95)	INTERVAL Study
n	42524	21177	39192	38883	36693	36693	36571	39260	
MESA (EUR)	69.85 (9.347)	51.02	0.03 (0.03)	0.17 (0.13)	1.69 (1.20)	0.49 (0.18)	3.81 (1.34)	6.20 (1.96)	The Multi-Ethnic Study of Atherosclerosis
n	1172	598	1172	1172	1172	1172	1172	1172	
MHIphase1	66.2 (9.26)	26.2	0.05 (0.05)	0.18 (0.13)	1.84 (0.92)	0.63 (0.22)	4.95 (2.37)	7.65 (2.65)	Montreal Heart Institute Biobank phase1
n	1417	371	1417	1417	1417	1417	1417	1417	
MHIphase2	65.5 (9.11)	24.9	0.05 (0.06)	0.18 (0.13)	1.77 (1.05)	0.6 (0.22)	4.74 (2.32)	7.33 (2.67)	Montreal Heart Institute Biobank phase2
n	1879	468	1879	1879	1879	1879	1879	1879	
RS-I	79.5 (4.8)	59.2	-	-	2.21 (0.91)	0.46 (0.17)	-	7.08 (1.79)	Rotterdam Study I
n	1455	862	-	-	1455	1455	-	1455	
RS-II	72.4 (5.2)	54.7	-	-	2.37 (0.78)	0.48 (0.19)	-	7.08 (1.79)	Rotterdam Study II
n	1269	694	-	-	1269	1269	-	1269	
RS-III	62.4 (5.8)	56.6	-	-	2.42 (0.72)	0.46 (0.16)	-	7.07 (1.85)	Rotterdam Study III
n	2378	1345	-	-	2378	2378	-	2378	
SHIP	49.19 (16.09)	50.9	-	-	-	-	-	6.74 (2.02)	Study of Helath in Pomerania
n	3164	1610	-	-	-	-	-	3159	
SHIP-TREND	51.82 (15.37)	51.3	0.03 (0.02)	0.15 (0.10)	1.68 (0.47)	0.50 (0.15)	3.34 (1.18)	5.70 (1.47)	Study of Health in Pomerania Trend
n	4099	2103	938	939	939	939	939	940	
UKBB_EUR	57.04 (8.10)	54	0.039 (0.044)	0.17 (0.13)	1.96 (1.17)	0.47 (0.28)	4.23 (1.40)	6.88 (2.10)	UK Biobank European-ancestry
n	463523	250302	453395	455195	455895	452885	455735	456785	

Study	Age	% Female	BASO (10^9/L)	EOS (10^9/L)	LYM (10^9/L)	MONO (10^9/L)	NEU (10^9/L)	WBC (10^9/L)	Full study name
WHI	66.7 (6.7)	100	0.04 (0.04)	0.21 (0.14)	1.8 (3.6)	0.62 (0.40)	3.8 (1.4)	6.5 (15.2)	Womens' Health Initiative
n	17,682	17,682	3168	3175	3193	3193	3193	17672	
YFS	41.9 (5.0)	55.3	-	-	-	-	-	5.64 (1.63)	The Cardiovascular Risk in Young Finns Study
n	1889	1044	-	-	-	-	-	1888	
Total	**626644**	**344764**	**522680**	**524220**	**572485**	**569440**	**567110**	**612055**	

Derived from Chen et al.
https://doi.org/10.1016/j.cell.2020.06.045

Appendix 2. *Description of meta-analysed studies for CRC risk.*

Meta-analysis Stage	Study Acronym	Study Name	Country	N total	N Cases (Ad)	N Controls
1	ASTERISK	Association STudy Evaluating RISK for sporadic colorectal cancer	France	1839	892 (0)	947
1	ATBC	Alpha-Tocopherol, Beta Carotene Cancer Prevention Study	Finland	177	147 (0)	30
1	CCFR_1	Colon Cancer Family Registry	USA, Canada, Australia	2014	1036 (0)	978
1	CCFR_2	Colon Cancer Family Registry	USA, Canada, Australia	716	331 (0)	385
1	CCFR_3	Colon Cancer Family Registry	USA, Canada, Australia	1851	1190 (0)	661
1	CCFR_4	Colon Cancer Family Registry	USA, Canada, Australia	2124	1590 (0)	534
1	Colo2&3	Hawai'i Colorectal Cancer Studies 2&3	USA	211	87 (0)	124
1	ColoCare_Heidelberg	ColoCare Consortium	Germany	223	187 (0)	36
1	ColoCare_Seattle	ColoCare Consortium	USA	169	169 (0)	0
1	CPSII_1	American Cancer Society Cancer Prevention Study II nested case-control study	USA	1076	540 (0)	536
1	CRCGEN	Colorectal Cancer Genetics & Genomics, Spanish study	Spain	1546	760 (0)	786
1	DACHS_1	Darmkrebs: Chancen der Verhütung durch Screening	Germany	3409	1707 (0)	1702
1	DACHS_2	Darmkrebs: Chancen der Verhütung durch Screening	Germany	1164	666 (0)	498
1	DALS_1	Diet, Activity and Lifestyle Study	USA	1411	702 (0)	709
1	DALS_2	Diet, Activity and Lifestyle Study	USA	863	402 (0)	461

1	ESTHER_VERDI	Epidemiologische Studie zu Chancen der Verhütung, Früherkennung und optimierten Therapie chronischer Erkrankungen in der älteren Bevölkerung; Verlauf der diagnotischen Abklärung bei Krebspatienten	Germany	817	397 (0)	420
1	HCES-CRC	The Hwasun Cancer Epidemiology Study-Colon and Rectum Cancer	Korea	5294	3026 (0)	2268
1	HPFS_1	Health Professionals Follow-Up Study	USA	455	227 (0)	228
1	HPFS_2	Health Professionals Follow-Up Study	USA	348	176 (0)	172
1	HPFS_3_AD	Health Professionals Follow-Up Study	USA	655	312 (312)	343
1	Kentucky	Kentucky Case-Control Study	USA	2167	1035 (0)	1132
1	MCCS	Melbourne Collaborative Cohort Study	Australia	1343	709 (0)	634
1	MEC_1	Multiethnic Cohort Study	USA	816	389 (0)	427
1	MECC_1	Molecular Epidemiology of Colorectal Cancer Study	Israel	978	483 (0)	495
1	MECC_2	Molecular Epidemiology of Colorectal Cancer Study	Israel	1901	1093 (0)	808
1	MECC_3	Molecular Epidemiology of Colorectal Cancer Study	Israel	4380	2570 (0)	1810
1	MSKCC	Memorial Sloan Kettering Cancer Center Cohort	USA	68	68 (0)	0
1	NFCCR	Newfoundland Case-Control Study	Canada	660	193 (0)	467
1	NGCCS	PopGen Biobank	Germany	1103	1103 (0)	0
1	NHS_1	Nurses' Health Study	USA	1165	391 (0)	774
1	NHS_2	Nurses' Health Study	USA	339	158 (0)	181
1	NHS_3_AD	Nurses' Health Study	USA	1090	513 (513)	577

1	NHSII	Nurses' Health Study	USA	167	87 (0)	80
1	OFCCR	Ontario Familial Colorectal Cancer Registry	Canada	1116	594 (0)	522
1	PHS	Physicians' Health Study	USA	764	375 (0)	389
1	PLCO_1	Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial	USA	2496	524 (0)	1972
1	PLCO_2	Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial	USA	889	475 (0)	414
1	PMH-CCFR	Postmenopausal Hormones Supplementary Study to the Colon Cancer Family Registry	USA	398	276 (0)	122
1	SEARCH	Studies of Epidemiology and Risk Factors in Cancer Heredity	UK	4288	4173 (0)	115
1	SLRCCS	Swedish Low-Risk Colorectal Cancer Study	Sweden	4785	2504 (0)	2281
1	SMC_COSM	Swedish Mammography Cohort and Cohort of Swedish Men	Sweden	1397	566 (0)	831
1	USC_HRT_CRC	Los Angeles County Cancer Surveillance Program	USA	708	321 (0)	387
1	VITAL	VITamins And Lifestyle	USA	565	279 (0)	286
1	WHI_1	Women's Health Initiative Study	USA	1991	468 (0)	1523
1	WHI_2	Women's Health Initiative Study	USA	1984	978 (0)	1006
2	CLUEII	Campaign against Cancer and Heart Disease II	USA	518	258 (0)	260

2	COLON	Colorectal Cancer: Longitudinal Observational study on Nutritional and lifestyle factors that influence colorectal tumor recurrence, survival and quality of life	Netherlands	1335	643 (0)	692
2	CORSA_1	Colorectal Cancer Study of Austria	Austria	2234	1460 (519)	774
2	CORSA_2	Colorectal Cancer Study of Austria	Austria	2483	1210 (687)	1273
2	CPSII_2	American Cancer Society Cancer Prevention Study II nested case-control study	USA	688	339 (0)	349
2	Czech	Czech Republic CCS	Czech Republic	3293	1675 (0)	1618
2	DACHS_3	Darmkrebs: Chancen der Verhütung durch Screening Study	Germany	1827	1210 (0)	617
2	EDRN	Early Detection Research Network	USA	589	273 (6)	316
2	EPIC	European Prospective Investigation into Cancer and Nutrition	Europe	4401	2095 (0)	2306
2	EPICOLON	EPICOLON	Spain	609	267 (0)	342
2	HawaiiCCS_AD	Hawaii Adenoma Study	USA	628	85 (85)	543
2	HPFS_4	Health Professionals Follow-Up Study	USA	380	183 (0)	197
2	HPFS_5_AD	Health Professionals Follow-Up Study	USA	260	155 (155)	105
2	LCCS	Leeds Colorectal Cancer Study	UK	2183	1482 (0)	701
2	NCCCSI	North Carolina Colon Cancer Study, I	USA	720	251 (0)	469
2	NCCCSII	North Carolina Colon Cancer Study, II	USA	1281	595 (0)	686
2	NHS_4	Nurses' Health Study	USA	611	308 (0)	303
2	NHS_5_AD	Nurses' Health Study	USA	477	251 (251)	226
2	NSHDS	The Northern Sweden Health and Disease Study	Sweden	829	416 (0)	413

2	OSUMC	Columbus-area HNPCC study, Ohio Colorectal Cancer Prevention Initiative, Ohio State University Medical Center	USA	5527	3094 (0)	2433
2	PLCO_4_AD	Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial	USA	2105	797 (794)	1308
2	SELECT	Selenium and Vitamin E Prevention Trial	USA	533	264 (0)	269
2	SMS_AD	Screening Markers for Colorectal Cancer Study (advanced adenomas)	USA	171	41 (0)	130
2	WHI_3	Women's Health Initiative Study	USA	1113	554 (0)	559
Total				98715	52775	45940

Ad, advanced adenoma

Derived from Huyghe et al. https://doi.org/10.1038/s41588-018-0286-6

Appendix 3. *Description of meta-analysed studies for allergic disease.*

Study	Study-type	GWA Covariates	Original case ascertainment	N total	N controls	Number of cases	Age (mean, range)	% Females
UKBiobank	Population-based	BOLT-LMM with age, sex and SNP chip	NA	138354	96108	42246	56.7 (39-73)	53
23andMe	Population-based	Logistic regression with age, sex and PCs 1:5	NA	118269	34934	83335	49.5 (1-114)	48
GERA	Population-based	PLINK 1.9 with age, sex and PCs 1:10	NA	51218	15999	35219	62.3 (18-90)	59
CATSS	Population-based	RAREMETALWORKER with age and PCs 1:4	NA	11068	7488	3580	9.8 (9-23)	49
NTR	Population-based	GCTA --mlma-loco with age, sex and PCs 1:20	NA	10242	7919	2323	40.1 (4-94)	64
LifeLines	Population-based	PLINK 1.9 with age and sex	NA	8560	4837	3723	46.2 (18-88)	58
TWINGENE	Population-based	PLINK 1.9 with sex and PCs 1:4	NA	5517	3762	1755	58.3 (41-93)	51
ALSPAC$	Population-based	SNPTEST with sex	NA	4964	2330	2634	A/E: 10.8 (10-13); H: 13.9 (13-16)	49
SALTY	Population-based	RAREMETALWORKER with age and PCs 1:4	NA	4062	2761	1301	49.8 (41-72)	49
AAGC	Selected case-control	SNPTEST with age and sex	Asthma	2435	460	1975	35.1 (3-89)	56
GENEVA	Selected case-control	SNPTEST with sex	Eczema	2633	1274	1359	43.9 (0-85)	56
GENUFAD-SHIP-1	Selected case-control	Mach2dat with sex and PCs 1:2	Eczema	1781	1364	417	Cases: 3.9 (1-34); Controls: 50.0 (20-81)	Cases: 39; Controls: 50
GENUFAD-SHIP-2	Selected case-control	Mach2dat with sex and PCs 1:2	Eczema	1735	1473	262	Cases: 8.3 (1-26); Controls: 50.0 (20-81)	Cases: 54; Controls: 50
Total				**360838**	**180709**	**180129**		

Derived from Ferreira et al. https://doi.org/10.1038/ng.3985

$ For ALSPAC, information from different surveys was used to define asthma (A) and eczema (E) when compared to those used to define hay fever (H). For this reason, age of participants is reported separately for A/E and H.

Appendix 4. *STROBE-MR checklist of recommended items to address in reports of Mendelian randomization studies.*

Item No.	Section	Checklist item	Page No.	Relevant text from manuscript
1	**TITLE and ABSTRACT**	Indicate Mendelian randomization (MR) as the study's design in the title and/or the abstract if that is a main purpose of the study	1-3	
	INTRODUCTION			
2	**Background**	Explain the scientific background and rationale for the reported study. What is the exposure? Is a potential causal relationship between exposure and outcome plausible? Justify why MR is a helpful method to address the study question	3-5	
3	**Objectives**	State specific objectives clearly, including pre-specified causal hypotheses (if any). State that MR is a method that, under specific assumptions, intends to estimate causal effects	6	Introduction, paragraphs 4,5
	METHODS			
4	**Study design and data sources**	Present key elements of the study design early in the article. Consider including a table listing sources of data for all phases of the study. For each data source contributing to the analysis, describe the following:		
	a)	Setting: Describe the study design and the underlying population, if possible. Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection, when available.	6-8	Methods, paragraphs 1-4
	b)	Participants: Give the eligibility criteria, and the sources and methods of selection of participants. Report the sample size, and whether any power or sample size calculations were carried out prior to the main analysis	7-8	Methods, paragraphs 2-4. Supplementary Tables 1-4
	c)	Describe measurement, quality control and selection of genetic variants	8	Methods, paragraphs 2-4. More in the manuscript associated with each study.

Item No.	Section	Checklist item	Page No.	Relevant text from manuscript
	d)	For each exposure, outcome, and other relevant variables, describe methods of assessment and diagnostic criteria for diseases	7-8	Methods, paragraphs 2-4. Supplementary Tables 1-4
	e)	Provide details of ethics committee approval and participant informed consent, if relevant	22	Ethics declaration
5	**Assumptions**	Explicitly state the three core IV assumptions for the main analysis (relevance, independence and exclusion restriction) as well assumptions for any additional or sensitivity analysis	5	Introduction, paragraph 4
6	**Statistical methods: main analysis**	Describe statistical methods and statistics used	8-11	Methods, paragraphs 6,7,11
	a)	Describe how quantitative variables were handled in the analyses (i.e., scale, units, model)	6	Methods, paragraph 1
	b)	Describe how genetic variants were handled in the analyses and, if applicable, how their weights were selected	9	Methods, paragraph 5
	c)	Describe the MR estimator (e.g. two-stage least squares, Wald ratio) and related statistics. Detail the included covariates and, in case of two-sample MR, whether the same covariate set was used for adjustment in the two samples	9,10	Methods, paragraphs 8,9
	d)	Explain how missing data were addressed	N/A	
	e)	If applicable, indicate how multiple testing was addressed	N/A	
7	**Assessment of assumptions**	Describe any methods or prior knowledge used to assess the assumptions or justify their validity	9	Methods, paragraph 7
8	**Sensitivity analyses and**	Describe any sensitivity analyses or additional analyses performed (e.g. comparison of effect estimates from different approaches, independent	10	Methods, paragraphs 8,9

Item No.	Section	Checklist item	Page No.	Relevant text from manuscript
	additional analyses	replication, bias analytic techniques, validation of instruments, simulations)		
9	**Software and pre-registration**			
	a)	Name statistical software and package(s), including version and settings used	10	Methods, paragraph 10
	b)	State whether the study protocol and details were pre-registered (as well as when and where)	N/A	
	RESULTS			
10	**Descriptive data**			
	a)	Report the numbers of individuals at each stage of included studies and reasons for exclusion. Consider use of a flow diagram	11	Supplementary Figure 2
	b)	Report summary statistics for phenotypic exposure(s), outcome(s), and other relevant variables (e.g. means, SDs, proportions)	11	Table 1, Supplementary
	c)	If the data sources include meta-analyses of previous studies, provide the assessments of heterogeneity across these studies	8	Methods, paragraph 5
	d)	For two-sample MR: i. Provide justification of the similarity of the genetic variant-exposure associations between the exposure and outcome samples ii. Provide information on the number of individuals who overlap between the exposure and outcome studies	N/A	0% overlap between samples
11	**Main results**			
	a)	Report the associations between genetic variant and exposure, and between genetic variant and outcome, preferably on an interpretable scale	N/A	Summary statistics for CRC are not currently publicly available.

Item No.	Section		Checklist item	Page No.	Relevant text from manuscript
		b)	Report MR estimates of the relationship between exposure and outcome, and the measures of uncertainty from the MR analysis, on an interpretable scale, such as odds ratio or relative risk per SD difference	13-16	Results, paragraphs 3,4,6,7,10,11. Supplementary Tables 10,14
		c)	If relevant, consider translating estimates of relative risk into absolute risk for a meaningful time period	N/A	
		d)	Consider plots to visualize results (e.g. forest plot, scatterplot of associations between genetic variants and outcome versus between genetic variants and exposure)		Figure 2,3
12	**Assessment of assumptions**				
		a)	Report the assessment of the validity of the assumptions	14,16	Results, paragraphs 8,12
		b)	Report any additional statistics (e.g., assessments of heterogeneity across genetic variants, such as I^2, Q statistic or E-value)	14,16	Results, paragraphs 8,12
13	**Sensitivity analyses and additional analyses**				
		a)	Report any sensitivity analyses to assess the robustness of the main results to violations of the assumptions	14,16	Results, paragraphs 8,12
		b)	Report results from other sensitivity analyses or additional analyses	14,16	Results, paragraphs 8,12
		c)	Report any assessment of direction of causal relationship (e.g., bidirectional MR)	14	Results, paragraph 8
		d)	When relevant, report and compare with estimates from non-MR analyses	16-18	Discussion

Item No.	Section	Checklist item	Page No.	Relevant text from manuscript
	e)	Consider additional plots to visualize results (e.g., leave-one-out analyses)	N/A	
	DISCUSSION			
14	**Key results**	Summarize key results with reference to study objectives	16,19	Discussion, paragraphs 1 & 7
15	**Limitations**	Discuss limitations of the study, taking into account the validity of the IV assumptions, other sources of potential bias, and imprecision. Discuss both direction and magnitude of any potential bias and any efforts to address them	18,19	Discussion, paragraph 6
16	**Interpretation**			
	a)	Meaning: Give a cautious overall interpretation of results in the context of their limitations and in comparison with other studies	16-18	
	b)	Mechanism: Discuss underlying biological mechanisms that could drive a potential causal relationship between the investigated exposure and the outcome, and whether the gene-environment equivalence assumption is reasonable. Use causal language carefully, clarifying that IV estimates may provide causal effects only under certain assumptions	16-18	
	c)	Clinical relevance: Discuss whether the results have clinical or public policy relevance, and to what extent they inform effect sizes of possible interventions		
17	**Generalizability**	Discuss the generalizability of the study results (a) to other populations, (b) across other exposure periods/timings, and (c) across other levels of exposure	19	Methods, paragraphs 5,6. Discussion, paragraph 6
	OTHER INFORMATION			

Item No.	Section	Checklist item	Page No.	Relevant text from manuscript
18	**Funding**	Describe sources of funding and the role of funders in the present study and, if applicable, sources of funding for the databases and original study or studies on which the present study is based	21-22	
19	**Data and data sharing**	Provide the data used to perform all analyses or report where and how the data can be accessed, and reference these sources in the article. Provide the statistical code needed to reproduce the results in the article, or report whether the code is publicly accessible and if so, where	19	
20	**Conflicts of Interest**	All authors should declare all potential conflicts of interest	22	

Appendix 5. *STROBE Statement—Checklist of items that should be included in reports of cohort studies.*

	Item No	Recommendation	Page No
Title and abstract	1	(*a*) Indicate the study's design with a commonly used term in the title or the abstract	1
		(*b*) Provide in the abstract an informative and balanced summary of what was done and what was found	3-4
Introduction			
Background/rationale	2	Explain the scientific background and rationale for the investigation being reported	4-5
Objectives	3	State specific objectives, including any prespecified hypotheses	5
Methods			
Study design	4	Present key elements of study design early in the paper	6
Setting	5	Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection	9-10
Participants	6	(*a*) Give the eligibility criteria, and the sources and methods of selection of participants. Describe methods of follow-up	9-10
		(*b*) For matched studies, give matching criteria and number of exposed and unexposed	N/A
Variables	7	Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable	9-11
Data sources/ measurement	8*	For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group	9-10
Bias	9	Describe any efforts to address potential sources of bias	11
Study size	10	Explain how the study size was arrived at	9,10,15
Quantitative variables	11	Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen and why	9-10
Statistical methods	12	(*a*) Describe all statistical methods, including those used to control for confounding	10-11

		(*b*) Describe any methods used to examine subgroups and interactions	N/A
		(*c*) Explain how missing data were addressed	10
		(*d*) If applicable, explain how loss to follow-up was addressed	N/A
		(*e*) Describe any sensitivity analyses	10-11
Results			
Participants	13*	(a) Report numbers of individuals at each stage of study—eg numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analysed	15
		(b) Give reasons for non-participation at each stage	9-10
		(c) Consider use of a flow diagram	15
Descriptive data	14*	(a) Give characteristics of study participants (eg demographic, clinical, social) and information on exposures and potential confounders	15
		(b) Indicate number of participants with missing data for each variable of interest	15
		(c) Summarise follow-up time (eg, average and total amount)	15
Outcome data	15*	Report numbers of outcome events or summary measures over time	15
Main results	16	(a) Give unadjusted estimates and, if applicable, confounder-adjusted estimates and their precision (eg, 95% confidence interval). Make clear which confounders were adjusted for and why they were included	16
		(b) Report category boundaries when continuous variables were categorized	N/A
		(c) If relevant, consider translating estimates of relative risk into absolute risk for a meaningful time period	N/A
Other analyses	17	Report other analyses done—eg analyses of subgroups and interactions, and sensitivity analyses	16-17
Discussion			
Key results	18	Summarise key results with reference to study objectives	18
Limitations	19	Discuss limitations of the study, taking into account sources of potential bias or imprecision. Discuss both direction and magnitude of any potential bias	20

Interpretation	20	Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence	18-19
Generalisability	21	Discuss the generalisability (external validity) of the study results	18-19
Other information			
Funding	22	Give the source of funding and the role of the funders for the present study and, if applicable, for the original study on which the present article is based	21-22

*Give information separately for exposed and unexposed groups.

Note: An Explanation and Elaboration article discusses each checklist item and gives methodological background and published examples of transparent reporting. The STROBE checklist is best used in conjunction with this article (freely available on the Web sites of PLoS Medicine at http://www.plosmedicine.org/, Annals of Internal Medicine at http://www.annals.org/, and Epidemiology at http://www.epidem.com/). Information on the STROBE Initiative is available at http://www.strobe-statement.org.

Appendix 6. *Univariable MR analysis of WBC count on CRC with sensitivity methods.*

Exposure	Outcome	Method	No SNPs	OR	LCI	UCI	P-value
Basophil	Colon	Inverse variance weighted	173	0.85	0.74	0.98	0.022
Basophil	Colon	MR Egger	173	0.83	0.58	1.19	0.311
Basophil	Colon	MR PRESSO	173	0.87	0.78	0.98	0.019
Basophil	Colon	Weighted median	173	0.83	0.7	0.99	0.033
Basophil	Colon	Weighted mode	173	0.83	0.64	1.08	0.163
Basophil	Distal	Inverse variance weighted	174	0.82	0.7	0.97	0.019
Basophil	Distal	MR Egger	174	0.8	0.53	1.22	0.302
Basophil	Distal	MR PRESSO	174	0.86	0.75	0.99	0.034
Basophil	Distal	Weighted median	174	0.82	0.66	1.04	0.098
Basophil	Distal	Weighted mode	174	0.81	0.56	1.17	0.269
Basophil	Female	Inverse variance weighted	174	0.92	0.79	1.07	0.271
Basophil	Female	MR Egger	174	1.03	0.7	1.52	0.883
Basophil	Female	MR PRESSO	174	0.97	0.84	1.11	0.636
Basophil	Female	Weighted median	174	0.96	0.78	1.19	0.710
Basophil	Female	Weighted mode	174	1.03	0.75	1.42	0.854
Basophil	Male	Inverse variance weighted	173	0.89	0.78	1.01	0.066
Basophil	Male	MR Egger	173	0.82	0.59	1.14	0.239
Basophil	Male	MR PRESSO	173				NA
Basophil	Male	Weighted median	173	0.81	0.67	0.99	0.037
Basophil	Male	Weighted mode	173	0.82	0.61	1.09	0.172
Basophil	Overall	Inverse variance weighted	171	0.88	0.78	0.99	0.037
Basophil	Overall	MR Egger	171	0.89	0.64	1.23	0.485
Basophil	Overall	MR PRESSO	171	0.9	0.81	0.99	0.039
Basophil	Overall	Weighted median	171	0.92	0.8	1.06	0.248
Basophil	Overall	Weighted mode	171	0.94	0.75	1.19	0.628
Basophil	Proximal	Inverse variance weighted	174	0.87	0.74	1.03	0.103
Basophil	Proximal	MR Egger	174	0.75	0.49	1.15	0.191
Basophil	Proximal	MR PRESSO	174	0.91	0.79	1.06	0.226
Basophil	Proximal	Weighted median	174	0.86	0.69	1.08	0.205
Basophil	Proximal	Weighted mode	174	0.9	0.64	1.26	0.543
Basophil	Rectal	Inverse variance weighted	176	0.93	0.79	1.09	0.355

Exposure	Outcome	Method	No SNPs	OR	LCI	UCI	P-value
Basophil	Rectal	MR Egger	176	0.92	0.61	1.38	0.680
Basophil	Rectal	MR PRESSO	176	0.91	0.78	1.07	0.252
Basophil	Rectal	Weighted median	176	1.01	0.81	1.25	0.930
Basophil	Rectal	Weighted mode	176	1.02	0.69	1.49	0.933
Eosinophil	Colon	Inverse variance weighted	397	0.9	0.84	0.96	0.001
Eosinophil	Colon	MR Egger	397	0.78	0.68	0.89	0.000
Eosinophil	Colon	MR PRESSO	397	0.9	0.84	0.96	0.001
Eosinophil	Colon	Weighted median	397	0.87	0.79	0.95	0.002
Eosinophil	Colon	Weighted mode	397	0.87	0.76	1	0.049
Eosinophil	Distal	Inverse variance weighted	397	0.89	0.82	0.97	0.007
Eosinophil	Distal	MR Egger	397	0.75	0.63	0.88	0.001
Eosinophil	Distal	MR PRESSO	397	0.9	0.83	0.97	0.009
Eosinophil	Distal	Weighted median	397	0.86	0.77	0.97	0.012
Eosinophil	Distal	Weighted mode	397	0.83	0.68	1.01	0.064
Eosinophil	Female	Inverse variance weighted	393	0.91	0.85	0.99	0.021
Eosinophil	Female	MR Egger	393	0.8	0.68	0.93	0.004
Eosinophil	Female	MR PRESSO	393	0.93	0.87	1.01	0.076
Eosinophil	Female	Weighted median	393	0.91	0.82	1.02	0.095
Eosinophil	Female	Weighted mode	393	0.93	0.76	1.13	0.473
Eosinophil	Male	Inverse variance weighted	398	0.94	0.88	1.01	0.119
Eosinophil	Male	MR Egger	398	0.83	0.72	0.96	0.011
Eosinophil	Male	MR PRESSO	398	0.93	0.87	1	0.045
Eosinophil	Male	Weighted median	398	0.95	0.86	1.05	0.336
Eosinophil	Male	Weighted mode	398	0.94	0.8	1.12	0.494
Eosinophil	Overall	Inverse variance weighted	396	0.93	0.88	0.98	0.012
Eosinophil	Overall	MR Egger	396	0.82	0.73	0.92	0.001
Eosinophil	Overall	MR PRESSO	396	0.94	0.89	0.99	0.016
Eosinophil	Overall	Weighted median	396	0.9	0.83	0.98	0.010
Eosinophil	Overall	Weighted mode	396	0.91	0.81	1.03	0.130
Eosinophil	Proximal	Inverse variance weighted	392	0.89	0.82	0.96	0.003
Eosinophil	Proximal	MR Egger	392	0.81	0.69	0.96	0.014
Eosinophil	Proximal	MR PRESSO	392	0.91	0.84	0.98	0.019
Eosinophil	Proximal	Weighted median	392	0.91	0.81	1.02	0.090

Exposure	Outcome	Method	No SNPs	OR	LCI	UCI	P-value
Eosinophil	Proximal	Weighted mode	392	0.92	0.74	1.13	0.419
Eosinophil	Rectal	Inverse variance weighted	393	0.96	0.89	1.04	0.305
Eosinophil	Rectal	MR Egger	393	0.88	0.75	1.02	0.086
Eosinophil	Rectal	MR PRESSO	393				NA
Eosinophil	Rectal	Weighted median	393	0.94	0.83	1.05	0.268
Eosinophil	Rectal	Weighted mode	393	0.98	0.81	1.19	0.846
Lymphocyte	Colon	Inverse variance weighted	453	0.96	0.89	1.03	0.215
Lymphocyte	Colon	MR Egger	453	0.91	0.78	1.05	0.186
Lymphocyte	Colon	MR PRESSO	453	0.97	0.91	1.04	0.393
Lymphocyte	Colon	Weighted median	453	0.99	0.9	1.08	0.763
Lymphocyte	Colon	Weighted mode	453	1.03	0.83	1.27	0.802
Lymphocyte	Distal	Inverse variance weighted	449	0.92	0.85	1	0.056
Lymphocyte	Distal	MR Egger	449	0.89	0.74	1.06	0.194
Lymphocyte	Distal	MR PRESSO	449	0.95	0.87	1.03	0.180
Lymphocyte	Distal	Weighted median	449	0.96	0.86	1.08	0.499
Lymphocyte	Distal	Weighted mode	449	0.97	0.75	1.26	0.845
Lymphocyte	Female	Inverse variance weighted	443	0.96	0.88	1.04	0.273
Lymphocyte	Female	MR Egger	443	0.79	0.66	0.94	0.008
Lymphocyte	Female	MR PRESSO	443	0.98	0.91	1.06	0.672
Lymphocyte	Female	Weighted median	443	0.91	0.81	1.03	0.126
Lymphocyte	Female	Weighted mode	443	0.86	0.62	1.21	0.391
Lymphocyte	Male	Inverse variance weighted	444	0.97	0.9	1.04	0.395
Lymphocyte	Male	MR Egger	444	0.99	0.85	1.17	0.945
Lymphocyte	Male	MR PRESSO	444	0.98	0.91	1.06	0.637
Lymphocyte	Male	Weighted median	444	1.03	0.92	1.15	0.622
Lymphocyte	Male	Weighted mode	444	1.08	0.87	1.36	0.483
Lymphocyte	Overall	Inverse variance weighted	444	0.94	0.89	1	0.057
Lymphocyte	Overall	MR Egger	444	0.87	0.76	0.99	0.040
Lymphocyte	Overall	MR PRESSO	444	0.95	0.9	1.01	0.099
Lymphocyte	Overall	Weighted median	444	0.94	0.87	1.02	0.134
Lymphocyte	Overall	Weighted mode	444	0.9	0.74	1.08	0.255
Lymphocyte	Proximal	Inverse variance weighted	455	0.96	0.88	1.04	0.335

Exposure	Outcome	Method	No SNPs	OR	LCI	UCI	P-value
Lymphocyte	Proximal	MR Egger	455	0.86	0.72	1.02	0.092
Lymphocyte	Proximal	MR PRESSO	455	0.99	0.91	1.07	0.806
Lymphocyte	Proximal	Weighted median	455	1.03	0.91	1.16	0.641
Lymphocyte	Proximal	Weighted mode	455	1.06	0.81	1.37	0.685
Lymphocyte	Rectal	Inverse variance weighted	452	0.93	0.85	1	0.063
Lymphocyte	Rectal	MR Egger	452	0.84	0.7	1	0.047
Lymphocyte	Rectal	MR PRESSO	452	0.92	0.85	1	0.049
Lymphocyte	Rectal	Weighted median	452	0.93	0.82	1.05	0.235
Lymphocyte	Rectal	Weighted mode	452	0.93	0.73	1.17	0.518
Monocyte	Colon	Inverse variance weighted	484	0.97	0.91	1.02	0.242
Monocyte	Colon	MR Egger	484	0.97	0.87	1.07	0.543
Monocyte	Colon	MR PRESSO	484	0.98	0.93	1.04	0.574
Monocyte	Colon	Weighted median	484	1	0.92	1.09	0.993
Monocyte	Colon	Weighted mode	484	1	0.91	1.09	0.942
Monocyte	Distal	Inverse variance weighted	479	0.95	0.89	1.03	0.204
Monocyte	Distal	MR Egger	479	0.98	0.86	1.12	0.749
Monocyte	Distal	MR PRESSO	479	0.99	0.92	1.06	0.690
Monocyte	Distal	Weighted median	479	1	0.91	1.11	0.948
Monocyte	Distal	Weighted mode	479	1.02	0.9	1.14	0.783
Monocyte	Female	Inverse variance weighted	477	0.98	0.92	1.04	0.467
Monocyte	Female	MR Egger	477	0.96	0.85	1.07	0.443
Monocyte	Female	MR PRESSO	477	0.98	0.92	1.04	0.459
Monocyte	Female	Weighted median	477	0.94	0.85	1.05	0.286
Monocyte	Female	Weighted mode	477	0.93	0.83	1.05	0.246
Monocyte	Male	Inverse variance weighted	480	0.95	0.89	1.01	0.089
Monocyte	Male	MR Egger	480	1	0.9	1.12	0.982
Monocyte	Male	MR PRESSO	480	0.95	0.89	1.01	0.086
Monocyte	Male	Weighted median	480	1.03	0.94	1.13	0.544
Monocyte	Male	Weighted mode	480	1.02	0.91	1.13	0.749
Monocyte	Overall	Inverse variance weighted	477	0.95	0.9	1	0.054
Monocyte	Overall	MR Egger	477	0.95	0.87	1.04	0.306
Monocyte	Overall	MR PRESSO	477	0.96	0.92	1.01	0.140
Monocyte	Overall	Weighted median	477	0.97	0.91	1.04	0.458

Exposure	Outcome	Method	No SNPs	OR	LCI	UCI	P-value
Monocyte	Overall	Weighted mode	477	0.96	0.88	1.05	0.378
Monocyte	Proximal	Inverse variance weighted	487	0.96	0.89	1.03	0.216
Monocyte	Proximal	MR Egger	487	0.98	0.86	1.11	0.751
Monocyte	Proximal	MR PRESSO	487	0.96	0.9	1.02	0.200
Monocyte	Proximal	Weighted median	487	0.99	0.89	1.11	0.885
Monocyte	Proximal	Weighted mode	487	0.98	0.86	1.12	0.805
Monocyte	Rectal	Inverse variance weighted	484	0.95	0.89	1.03	0.200
Monocyte	Rectal	MR Egger	484	0.98	0.86	1.12	0.791
Monocyte	Rectal	MR PRESSO	484	0.97	0.91	1.05	0.479
Monocyte	Rectal	Weighted median	484	0.97	0.86	1.1	0.665
Monocyte	Rectal	Weighted mode	484	1.03	0.9	1.17	0.692
Neutrophil	Colon	Inverse variance weighted	390	0.95	0.87	1.03	0.225
Neutrophil	Colon	MR Egger	390	0.96	0.8	1.15	0.640
Neutrophil	Colon	MR PRESSO	390	0.95	0.88	1.02	0.168
Neutrophil	Colon	Weighted median	390	1.04	0.93	1.15	0.521
Neutrophil	Colon	Weighted mode	390	1.13	0.93	1.37	0.233
Neutrophil	Distal	Inverse variance weighted	398	0.96	0.87	1.06	0.440
Neutrophil	Distal	MR Egger	398	1.06	0.85	1.32	0.603
Neutrophil	Distal	MR PRESSO	398	0.94	0.86	1.03	0.191
Neutrophil	Distal	Weighted median	398	0.98	0.86	1.12	0.796
Neutrophil	Distal	Weighted mode	398	0.97	0.76	1.22	0.768
Neutrophil	Female	Inverse variance weighted	390	0.97	0.89	1.06	0.465
Neutrophil	Female	MR Egger	390	0.98	0.81	1.19	0.852
Neutrophil	Female	MR PRESSO	390	0.97	0.9	1.06	0.525
Neutrophil	Female	Weighted median	390	0.98	0.87	1.11	0.769
Neutrophil	Female	Weighted mode	390	1	0.81	1.25	0.967
Neutrophil	Male	Inverse variance weighted	391	1.01	0.92	1.1	0.851
Neutrophil	Male	MR Egger	391	1.15	0.95	1.39	0.142
Neutrophil	Male	MR PRESSO	391	0.98	0.9	1.07	0.681
Neutrophil	Male	Weighted median	391	1.05	0.93	1.19	0.442
Neutrophil	Male	Weighted mode	391	1.09	0.88	1.35	0.447
Neutrophil	Overall	Inverse variance weighted	387	0.97	0.9	1.05	0.500

Exposure	Outcome	Method	No SNPs	OR	LCI	UCI	P-value
Neutrophil	Overall	MR Egger	387	1.03	0.88	1.21	0.671
Neutrophil	Overall	MR PRESSO	387	0.99	0.93	1.05	0.656
Neutrophil	Overall	Weighted median	387	1	0.92	1.09	0.970
Neutrophil	Overall	Weighted mode	387	1	0.87	1.17	0.952
Neutrophil	Proximal	Inverse variance weighted	382	0.93	0.84	1.02	0.116
Neutrophil	Proximal	MR Egger	382	0.82	0.67	1	0.051
Neutrophil	Proximal	MR PRESSO	382	0.96	0.88	1.05	0.353
Neutrophil	Proximal	Weighted median	382	0.99	0.86	1.13	0.849
Neutrophil	Proximal	Weighted mode	382	1.02	0.82	1.26	0.880
Neutrophil	Rectal	Inverse variance weighted	396	1.03	0.93	1.13	0.607
Neutrophil	Rectal	MR Egger	396	1.07	0.88	1.31	0.496
Neutrophil	Rectal	MR PRESSO	396	1.02	0.93	1.12	0.609
Neutrophil	Rectal	Weighted median	396	1.03	0.9	1.18	0.637
Neutrophil	Rectal	Weighted mode	396	1.07	0.83	1.39	0.580

Appendix 7. *MVMR analyses summary.*

Exposure	Outcome	Method	OR	OR LCI	OR UCI	P-value	BH P-v
Basophil	Colon	IVW MVMR (direct)	1.14	0.84	1.55	0.398245294	2.26E-0
Basophil	Distal	IVW MVMR (direct)	1.29	0.90	1.86	0.165464984	2.38E-0
Basophil	Female	IVW MVMR (direct)	1.27	0.90	1.79	0.167114023	5.10E-0
Basophil	Male	IVW MVMR (direct)	0.99	0.71	1.36	0.926907635	5.86E-0
Basophil	Overall	IVW MVMR (direct)	1.19	0.92	1.54	0.196510154	1.35E-0
Basophil	Proximal	IVW MVMR (direct)	1.13	0.78	1.63	0.513197046	2.27E-0
Basophil	Rectal	IVW MVMR (direct)	1.13	0.80	1.59	0.502402269	3.40E-0
Eosinophil	Colon	IVW MVMR (direct)	0.84	0.75	0.94	0.001929124	4.79E-0
Eosinophil	Distal	IVW MVMR (direct)	0.88	0.77	1.00	0.048927969	8.74E-0
Eosinophil	Female	IVW MVMR (direct)	0.83	0.73	0.94	0.003895395	1.71E-0
Eosinophil	Male	IVW MVMR (direct)	0.92	0.81	1.03	0.142140639	2.80E-0
Eosinophil	Overall	IVW MVMR (direct)	0.88	0.80	0.97	0.0109405	3.30E-0
Eosinophil	Proximal	IVW MVMR (direct)	0.79	0.69	0.90	0.000437523	3.83E-0
Eosinophil	Rectal	IVW MVMR (direct)	0.95	0.83	1.07	0.387621542	3.90E-0
Lymphocyte	Colon	IVW MVMR (direct)	0.85	0.76	0.96	0.006790761	4.05E-0
Lymphocyte	Distal	IVW MVMR (direct)	0.77	0.67	0.88	0.00013601	4.14E-0
Lymphocyte	Female	IVW MVMR (direct)	0.76	0.67	0.87	6.46E-05	4.21E-0
Lymphocyte	Male	IVW MVMR (direct)	0.94	0.83	1.06	0.305639437	4.24E-0
Lymphocyte	Overall	IVW MVMR (direct)	0.84	0.76	0.93	0.000670022	5.63E-0
Lymphocyte	Proximal	IVW MVMR (direct)	0.89	0.77	1.02	0.08805435	5.89E-0
Lymphocyte	Rectal	IVW MVMR (direct)	0.86	0.75	0.98	0.02246793	6.09E-0
Monocyte	Colon	IVW MVMR (direct)	1.02	0.93	1.11	0.687426188	6.12E-0
Monocyte	Distal	IVW MVMR (direct)	0.99	0.88	1.10	0.822614976	6.14E-0
Monocyte	Female	IVW MVMR (direct)	1.00	0.91	1.11	0.941591766	6.34E-0
Monocyte	Male	IVW MVMR (direct)	0.94	0.86	1.04	0.217817562	6.41E-0
Monocyte	Overall	IVW MVMR (direct)	0.97	0.90	1.05	0.506276497	6.46E-0
Monocyte	Proximal	IVW MVMR (direct)	1.05	0.94	1.17	0.420872341	6.56E-0
Monocyte	Rectal	IVW MVMR (direct)	0.92	0.83	1.02	0.113170722	6.76E-0
Neutrophil	Colon	IVW MVMR (direct)	0.94	0.82	1.08	0.403461408	8.30E-0
Neutrophil	Distal	IVW MVMR (direct)	1.00	0.85	1.17	0.987074608	8.42E-0
Neutrophil	Female	IVW MVMR (direct)	1.01	0.87	1.17	0.91773417	9.29E-0
Neutrophil	Male	IVW MVMR (direct)	0.94	0.82	1.08	0.419566677	9.69E-0
Neutrophil	Overall	IVW MVMR (direct)	0.95	0.85	1.06	0.348251285	9.83E-0
Neutrophil	Proximal	IVW MVMR (direct)	0.90	0.77	1.05	0.19262925	9.87E-0
Neutrophil	Rectal	IVW MVMR (direct)	0.97	0.84	1.13	0.721558819	1.00E+

[1]Benjamini-Hochberg (FDR) multiple testing correction adjusted P-values for 35 independent tests.

Appendix 8. *Univariable observational analysis of WBC count on overall, and by genetic sex CRC.*

Exposure	Outcome	Analysis type	Model	OR	LCI	UCI	P-value
Basophil	Overall	Univariable	Model 1	1.06	1.02	1.09	0.0004455
Basophil	Male	Univariable	Model 1	1.02	0.98	1.07	0.240227
Basophil	Female	Univariable	Model 1	1.03	0.98	1.08	0.252134
Eosinophil	Overall	Univariable	Model 1	0.97	0.94	1	0.0966543
Eosinophil	Male	Univariable	Model 1	0.96	0.92	1	0.0485831
Eosinophil	Female	Univariable	Model 1	1	0.95	1.05	0.9492865
Lymphocyte	Overall	Univariable	Model 1	1	0.97	1.04	0.7633771
Lymphocyte	Male	Univariable	Model 1	1	0.96	1.04	0.9784208
Lymphocyte	Female	Univariable	Model 1	1.03	0.98	1.08	0.1986498
Monocyte	Overall	Univariable	Model 1	1.05	1.02	1.08	0.0032073
Monocyte	Male	Univariable	Model 1	1.02	0.98	1.06	0.3028392
Monocyte	Female	Univariable	Model 1	1.06	1.01	1.11	0.0118104
Neutrophil	Overall	Univariable	Model 1	1.09	1.06	1.13	1.94E-08
Neutrophil	Male	Univariable	Model 1	1.08	1.04	1.13	0.0001752
Neutrophil	Female	Univariable	Model 1	1.07	1.02	1.12	0.0060642
Basophil	Overall	Univariable	Model 2	1.04	1.01	1.07	0.0139836
Basophil	Male	Univariable	Model 2	1.02	0.98	1.06	0.4221856
Basophil	Female	Univariable	Model 2	1.01	0.96	1.06	0.6618079
Eosinophil	Overall	Univariable	Model 2	0.96	0.93	0.99	0.0215153
Eosinophil	Male	Univariable	Model 2	0.95	0.91	0.99	0.0201037
Eosinophil	Female	Univariable	Model 2	0.99	0.94	1.03	0.5515385
Lymphocyte	Overall	Univariable	Model 2	0.98	0.95	1.02	0.3392471
Lymphocyte	Male	Univariable	Model 2	0.97	0.94	1.02	0.2176043
Lymphocyte	Female	Univariable	Model 2	1.01	0.97	1.06	0.5394318
Monocyte	Overall	Univariable	Model 2	1.03	1	1.07	0.0331332
Monocyte	Male	Univariable	Model 2	1	0.96	1.04	0.9797341
Monocyte	Female	Univariable	Model 2	1.06	1.01	1.11	0.0266179
Neutrophil	Overall	Univariable	Model 2	1.08	1.05	1.11	1.51E-06
Neutrophil	Male	Univariable	Model 2	1.07	1.02	1.11	0.0020307
Neutrophil	Female	Univariable	Model 2	1.06	1.01	1.11	0.0263914

Appendix 9. *Observational analysis of WBC traits, adjusted between each other on overall, and by genetic sex CRC.*

Exposure	Outcome	Analysis type	Model	OR	LCI	UCI	P-value
Basophil	Overall	Multivariable	Model 1	1.04	1.01	1.08	0.008168
Basophil	Male	Multivariable	Model 1	1.02	0.97	1.06	0.4469032
Basophil	Female	Multivariable	Model 1	1.01	0.96	1.06	0.6789682
Eosinophil	Overall	Multivariable	Model 1	0.96	0.93	0.99	0.0092955
Eosinophil	Male	Multivariable	Model 1	0.95	0.91	0.99	0.018828
Eosinophil	Female	Multivariable	Model 1	0.98	0.93	1.03	0.3967718
Lymphocyte	Overall	Multivariable	Model 1	0.98	0.94	1.01	0.1903065
Lymphocyte	Male	Multivariable	Model 1	0.99	0.95	1.04	0.6876938
Lymphocyte	Female	Multivariable	Model 1	1	0.95	1.06	0.869115
Monocyte	Overall	Multivariable	Model 1	1.03	1	1.07	0.084775
Monocyte	Male	Multivariable	Model 1	1.01	0.96	1.06	0.7300268
Monocyte	Female	Multivariable	Model 1	1.05	0.99	1.1	0.0874193
Neutrophil	Overall	Multivariable	Model 1	1.08	1.05	1.12	1.92E-06
Neutrophil	Male	Multivariable	Model 1	1.08	1.04	1.13	0.0003537
Neutrophil	Female	Multivariable	Model 1	1.05	1	1.11	0.0463839
Basophil	Overall	Multivariable	Model 2	1.03	1	1.07	0.0475975
Basophil	Male	Multivariable	Model 2	1.02	0.97	1.06	0.4290144
Basophil	Female	Multivariable	Model 2	1	0.95	1.05	0.9332657
Eosinophil	Overall	Multivariable	Model 2	0.96	0.93	0.99	0.0058224
Eosinophil	Male	Multivariable	Model 2	0.95	0.91	0.99	0.0250332
Eosinophil	Female	Multivariable	Model 2	0.97	0.93	1.02	0.2623026
Lymphocyte	Overall	Multivariable	Model 2	0.97	0.94	1	0.0552767
Lymphocyte	Male	Multivariable	Model 2	0.97	0.93	1.02	0.2245981
Lymphocyte	Female	Multivariable	Model 2	1	0.95	1.05	0.8695197
Monocyte	Overall	Multivariable	Model 2	1.03	0.99	1.06	0.1205044
Monocyte	Male	Multivariable	Model 2	1	0.95	1.04	0.8797589
Monocyte	Female	Multivariable	Model 2	1.05	1	1.11	0.0687917
Neutrophil	Overall	Multivariable	Model 2	1.07	1.04	1.11	1.57E-05
Neutrophil	Male	Multivariable	Model 2	1.07	1.03	1.12	0.0013675
Neutrophil	Female	Multivariable	Model 2	1.04	0.99	1.1	0.0865551

Appendix 10. *UVMR analysis between allergic disease and CRC.*

Exposure	Outcome	Method	No SNPs	OR	LCI	UCI	P-value
Allergic disease	Colon	Inverse variance weighted	79	0.88	0.8	0.97	0.00787327
Allergic disease	Colon	MR Egger	79	0.81	0.63	1.03	0.09398785
Allergic disease	Colon	MR PRESSO	79	0.88	0.82	0.94	0.00036776
Allergic disease	Colon	Weighted median	79	0.92	0.83	1.01	0.0750271
Allergic disease	Colon	Weighted mode	79	0.93	0.79	1.09	0.37493087
Allergic disease	Distal	Inverse variance weighted	78	0.91	0.82	1.01	0.06698952
Allergic disease	Distal	MR Egger	78	0.87	0.66	1.15	0.33837906
Allergic disease	Distal	MR PRESSO	78	0.89	0.82	0.98	0.01487843
Allergic disease	Distal	Weighted median	78	0.88	0.78	1	0.0506209
Allergic disease	Distal	Weighted mode	78	0.83	0.68	1.02	0.07936828
Allergic disease	Female	Inverse variance weighted	77	0.87	0.78	0.96	0.00541118
Allergic disease	Female	MR Egger	77	0.68	0.52	0.89	0.00703118
Allergic disease	Female	MR PRESSO	77	0.87	0.8	0.96	0.00399098
Allergic disease	Female	Weighted median	77	0.85	0.75	0.95	0.00459797
Allergic disease	Female	Weighted mode	77	0.82	0.69	0.98	0.02914724
Allergic disease	Male	Inverse variance weighted	80	0.92	0.85	0.99	0.03389726
Allergic disease	Male	MR Egger	80	1.15	0.93	1.42	0.21443635
Allergic disease	Male	MR PRESSO	80				NA
Allergic disease	Male	Weighted median	80	0.93	0.83	1.04	0.18496623
Allergic disease	Male	Weighted mode	80	0.99	0.84	1.15	0.8738974
Allergic disease	Overall	Inverse variance weighted	81	0.89	0.82	0.96	0.00319495
Allergic disease	Overall	MR Egger	81	0.84	0.68	1.05	0.12746211
Allergic disease	Overall	MR PRESSO	81	0.88	0.83	0.94	0.00026143
Allergic disease	Overall	Weighted median	81	0.93	0.85	1.01	0.08495569
Allergic disease	Overall	Weighted mode	81	0.94	0.83	1.07	0.37462223
Allergic disease	Proximal	Inverse variance weighted	79	0.87	0.78	0.97	0.01100713
Allergic disease	Proximal	MR Egger	79	0.77	0.57	1.03	0.08357488
Allergic disease	Proximal	MR PRESSO	79	0.86	0.79	0.94	0.00145658
Allergic disease	Proximal	Weighted median	79	0.83	0.73	0.94	0.00313491
Allergic disease	Proximal	Weighted mode	79	0.8	0.63	1.01	0.06986964
Allergic disease	Rectal	Inverse variance weighted	80	0.97	0.88	1.07	0.5052078
Allergic disease	Rectal	MR Egger	80	0.97	0.74	1.27	0.84002279
Allergic disease	Rectal	MR PRESSO	80	0.94	0.86	1.03	0.1902516
Allergic disease	Rectal	Weighted median	80	1	0.89	1.12	0.94142787
Allergic disease	Rectal	Weighted mode	80	0.99	0.85	1.17	0.94798691

Appendix 11. *Steps undertaken in the PCA analysis.*

Continental ancestry group (CAG)	Nr. individuals with 80% ancestry assigned to CAG	Step 1 - Unrelated individuals	Step 2 - LD independent SNPs	Step 3 - LD independent PLINK files	Step 4 - smartrel Related individuals projected	Step 5 - smartpca Generating PCs	Step 6 - Outlier removal
European	50,685	39006	40095	values from column B and D	12182 (derived from step1, as smartrel failed with the EUR sample)	Smartpca to estimate PCs only on unrelated followed by projection of those related + 1KG corresponding Superpopulation	NA (ran in fast mode with no exclusions)
African	6,653	5306	48818	values from column B and D	541	Smartpca to estimate PCs only on unrelated followed by projection of those related + 1KG corresponding Superpopulation	148
East Asian	2868	1692	47113	values from column B and D	29	Smartpca to estimate PCs only on unrelated followed by projection of those related + 1KG corresponding Superpopulation	89

South Asian	3271	1919	43915	values from column B and D	208	Smartpca to estimate PCs only on unrelated followed by projection of those related + 1KG corresponding Superpopulation	92

Appendix 12. *Continents used in the study along with UN regions and countries associated with each region.*

Continent	Geoscheme region	Country
Africa	Northern Africa	Algeria
Africa	Northern Africa	Egypt
Africa	Northern Africa	Libya
Africa	Northern Africa	Morocco
Africa	Northern Africa	Sudan
Africa	Northern Africa	Tunisia
Africa	Northern Africa	Western Sahara
Africa	Eastern Africa	British Indian Ocean Territory
Africa	Eastern Africa	Burundi
Africa	Eastern Africa	Comoros
Africa	Eastern Africa	Djibouti
Africa	Eastern Africa	Eritrea
Africa	Eastern Africa	Ethiopia
Africa	Eastern Africa	French Southern Territories
Africa	Eastern Africa	Kenya
Africa	Eastern Africa	Madagascar
Africa	Eastern Africa	Malawi
Africa	Eastern Africa	Mauritius
Africa	Eastern Africa	Mayotte
Africa	Eastern Africa	Mozambique
Africa	Eastern Africa	Reunion
Africa	Eastern Africa	Rwanda
Africa	Eastern Africa	Saychelles
Africa	Eastern Africa	Somalia
Africa	Eastern Africa	South Sudan
Africa	Eastern Africa	Uganda
Africa	Eastern Africa	United Republic of Tanzania
Africa	Eastern Africa	Zambia
Africa	Eastern Africa	Zimbabwe
Africa	Central/Middle Africa	Angola
Africa	Central/Middle Africa	Cameroon
Africa	Central/Middle Africa	Central African Republic
Africa	Central/Middle Africa	Chad
Africa	Central/Middle Africa	Congo
Africa	Central/Middle Africa	Democratic Republic of the Congo
Africa	Central/Middle Africa	Equatorial Guinea
Africa	Central/Middle Africa	Gabon
Africa	Central/Middle Africa	Sao Tome and Principe

Continent	Geoscheme region	Country
Africa	Southern Africa	Botswana
Africa	Southern Africa	Eswatini
Africa	Southern Africa	Lesotho
Africa	Southern Africa	Namibia
Africa	Southern Africa	South Africa
Africa	Western Africa	Benin
Africa	Western Africa	Burkina Faso
Africa	Western Africa	Cabo Verde
Africa	Western Africa	Cote d'Ivoire
Africa	Western Africa	Gambia
Africa	Western Africa	Ghana
Africa	Western Africa	Guinea
Africa	Western Africa	Guinea-Bissau
Africa	Western Africa	Liberia
Africa	Western Africa	Mali
Africa	Western Africa	Mauritania
Africa	Western Africa	Niger
Africa	Western Africa	Nigeria
Africa	Western Africa	Saint Helena
Africa	Western Africa	Senegal
Africa	Western Africa	Sierra Leone
Africa	Western Africa	Togo
Asia	Central Asia	Kazakhstan
Asia	Central Asia	Kyrgystan
Asia	Central Asia	Tajikistan
Asia	Central Asia	Turkmenistan
Asia	Central Asia	Uzbekistan
Asia	Eastern Asia	China
Asia	Eastern Asia	China, Hong Kong Special Administrative Region
Asia	Eastern Asia	China, Macao Special Administrative Region
Asia	Eastern Asia	Democratic People's Republic of Korea
Asia	Eastern Asia	Japan
Asia	Eastern Asia	Mongolia
Asia	Eastern Asia	Republic of Korea
Asia	South-eastern Asia	Brunei Darussalam
Asia	South-eastern Asia	Cambodia
Asia	South-eastern Asia	Indonesia
Asia	South-eastern Asia	Lao People's Democratic Republic
Asia	South-eastern Asia	Malaysia
Asia	South-eastern Asia	Myanmar

Continent	Geoscheme region	Country
Asia	South-eastern Asia	Philippines
Asia	South-eastern Asia	Singapore
Asia	South-eastern Asia	Thailand
Asia	South-eastern Asia	Timor-Leste
Asia	South-eastern Asia	Vietnam
Asia	Southern Asia	Afghanistan
Asia	Southern Asia	Bangladesh
Asia	Southern Asia	Bhutan
Asia	Southern Asia	India
Asia	Southern Asia	Islamic Republic of Iran
Asia	Southern Asia	Maldives
Asia	Southern Asia	Nepal
Asia	Southern Asia	Pakistan
Asia	Southern Asia	Sri Lanka
Asia	Western Asia	Armenia
Asia	Western Asia	Azerbaijan
Asia	Western Asia	Bahrain
Asia	Western Asia	Cyprus
Asia	Western Asia	Georgia
Asia	Western Asia	Iraq
Asia	Western Asia	Israel
Asia	Western Asia	Jordan
Asia	Western Asia	Kuwait
Asia	Western Asia	Lebanon
Asia	Western Asia	Oman
Asia	Western Asia	Qatar
Asia	Western Asia	Saudi Arabia
Asia	Western Asia	State of Palestine
Asia	Western Asia	Syrian Arab Republic
Asia	Western Asia	Turkey
Asia	Western Asia	United Arab Emirates
Asia	Western Asia	Yemen
Europe	Eastern Europe	Belarus
Europe	Eastern Europe	Bulgaria
Europe	Eastern Europe	Czechia
Europe	Eastern Europe	Hungary
Europe	Eastern Europe	Poland
Europe	Eastern Europe	Republic of Moldova
Europe	Eastern Europe	Romania
Europe	Eastern Europe	Russian Federation

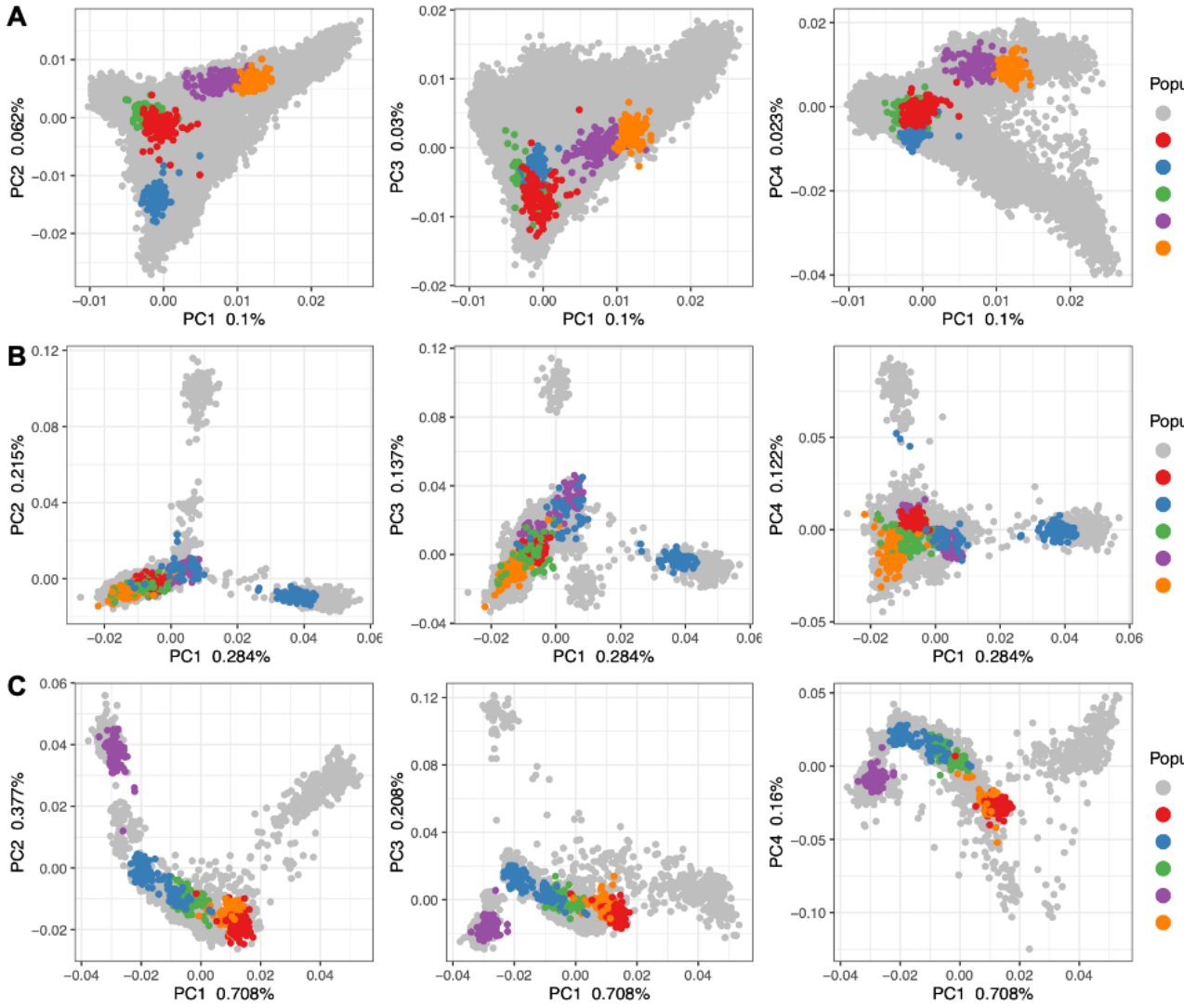
Continent	Geoscheme region	Country
Europe	Eastern Europe	Slovakia
Europe	Eastern Europe	Ukraine
Europe	Northern Europe	Aland Islands
Europe	Northern Europe	Channel Islands
Europe	Northern Europe	Denmark
Europe	Northern Europe	Estonia
Europe	Northern Europe	Faroe Islands
Europe	Northern Europe	Finland
Europe	Northern Europe	Iceland
Europe	Northern Europe	Ireland
Europe	Northern Europe	Isle of Man
Europe	Northern Europe	Latvia
Europe	Northern Europe	Lithuania
Europe	Northern Europe	Norway
Europe	Northern Europe	Svalbard and Jan Mayen Islands
Europe	Northern Europe	Sweden
Europe	Northern Europe	United Kingdom of Great Britain and Northern Ireland
Europe	Southern Europe	Albania
Europe	Southern Europe	Andorra
Europe	Southern Europe	Bosnia and Herzegovina
Europe	Southern Europe	Croatia
Europe	Southern Europe	Gibraltar
Europe	Southern Europe	Greece
Europe	Southern Europe	Holy See
Europe	Southern Europe	Italy
Europe	Southern Europe	Malta
Europe	Southern Europe	Montenegro
Europe	Southern Europe	North Macedonia
Europe	Southern Europe	Portugal
Europe	Southern Europe	San Marino
Europe	Southern Europe	Serbia
Europe	Southern Europe	Slovenia
Europe	Southern Europe	Spain
Europe	Western Europe	Austria
Europe	Western Europe	Belgium
Europe	Western Europe	France
Europe	Western Europe	Germany
Europe	Western Europe	Lichtenstein
Europe	Western Europe	Luxembourg
Europe	Western Europe	Monaco

Continent	Geoscheme region	Country
Europe	Western Europe	Netherlands
Europe	Western Europe	Switzerland

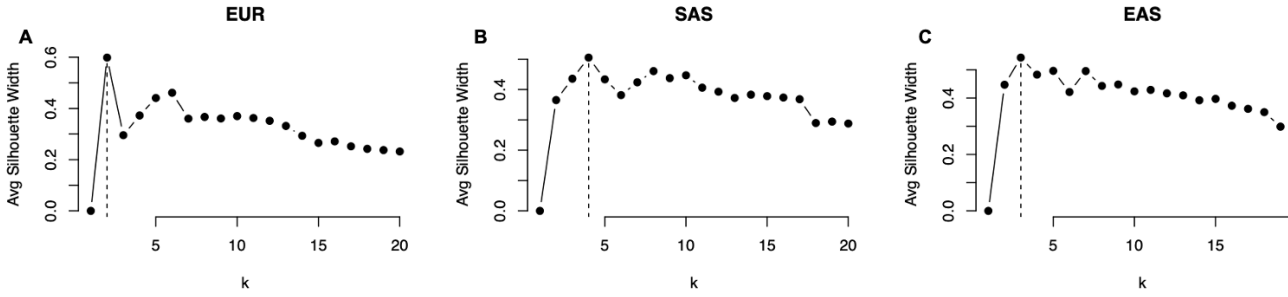
Appendix 13. *PCs for each CAG and their respective eigenvalues.*

Continental ancestry group	Principal component	Eigenvalue (%)
African	PC1	10.82
African	PC2	10.00
African	PC3	4.95
African	PC4	2.82
South Asian	PC1	7.11
South Asian	PC2	5.38
South Asian	PC3	3.42
South Asian	PC4	3.06
South Asian	PC5	2.66
East Asian	PC1	15.93
East Asian	PC2	8.48
East Asian	PC3	4.69
East Asian	PC4	3.60
European	PC1	38.30

Appendix 14. *PC1 on PCs 2-7 (**A-F**) with variance explained on the axis labels. Information on each 1000 Genomes sub-population is available at https://catalog.coriell.org/0/Sections/Collections/NHGRI/1000genome.aspx.*



Appendix 15. *Silhouette analysis for optimal k K-cluster identification. Average silhouette width was calculated for k = 2-20. The x-axis represents the K-cluster number, while the y-axis is the average silhouette width, a larger value indicating a better fit.*

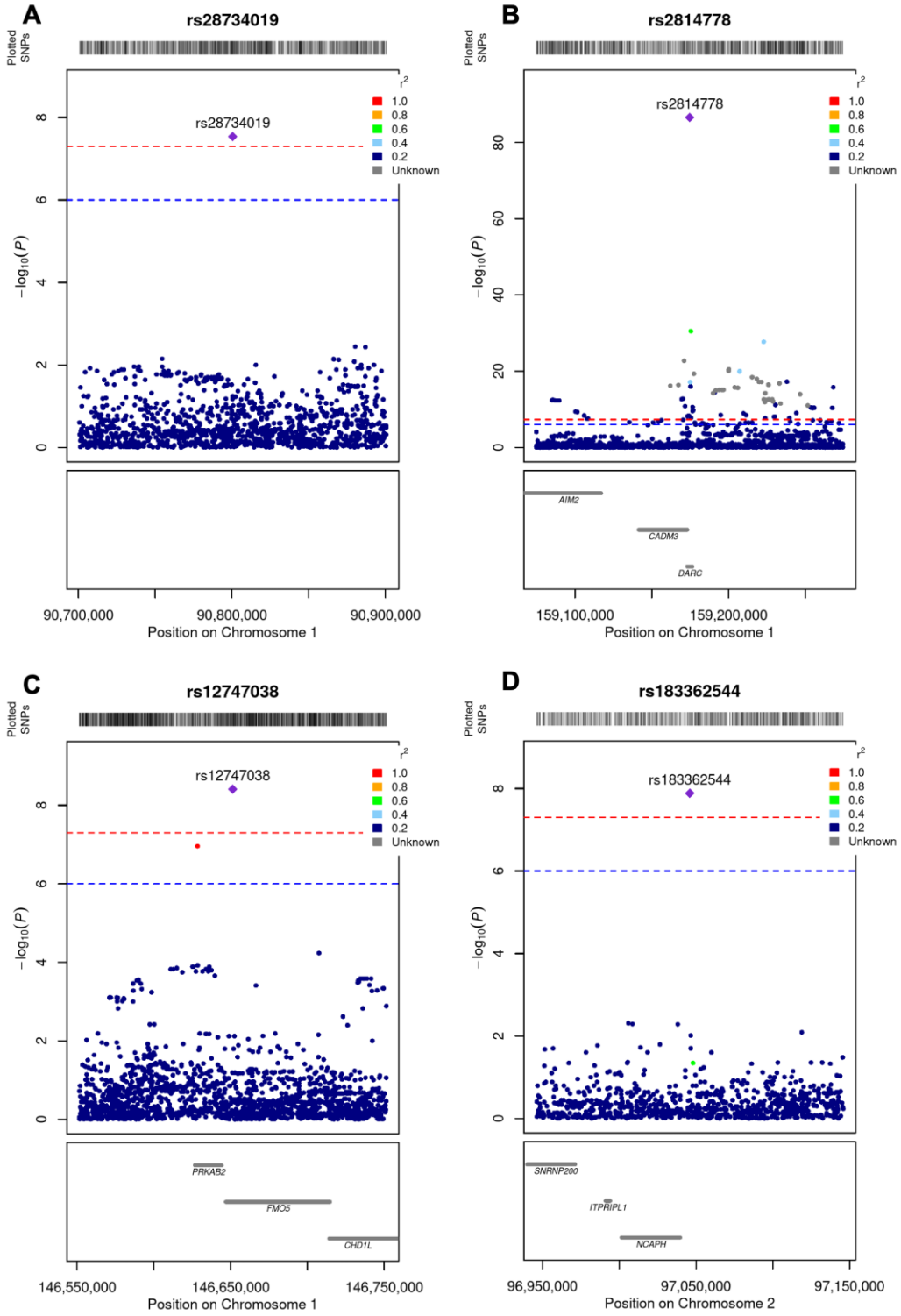


Appendix 16. *Correspondence analyses for EUR, SAS and EAS CAGs.*

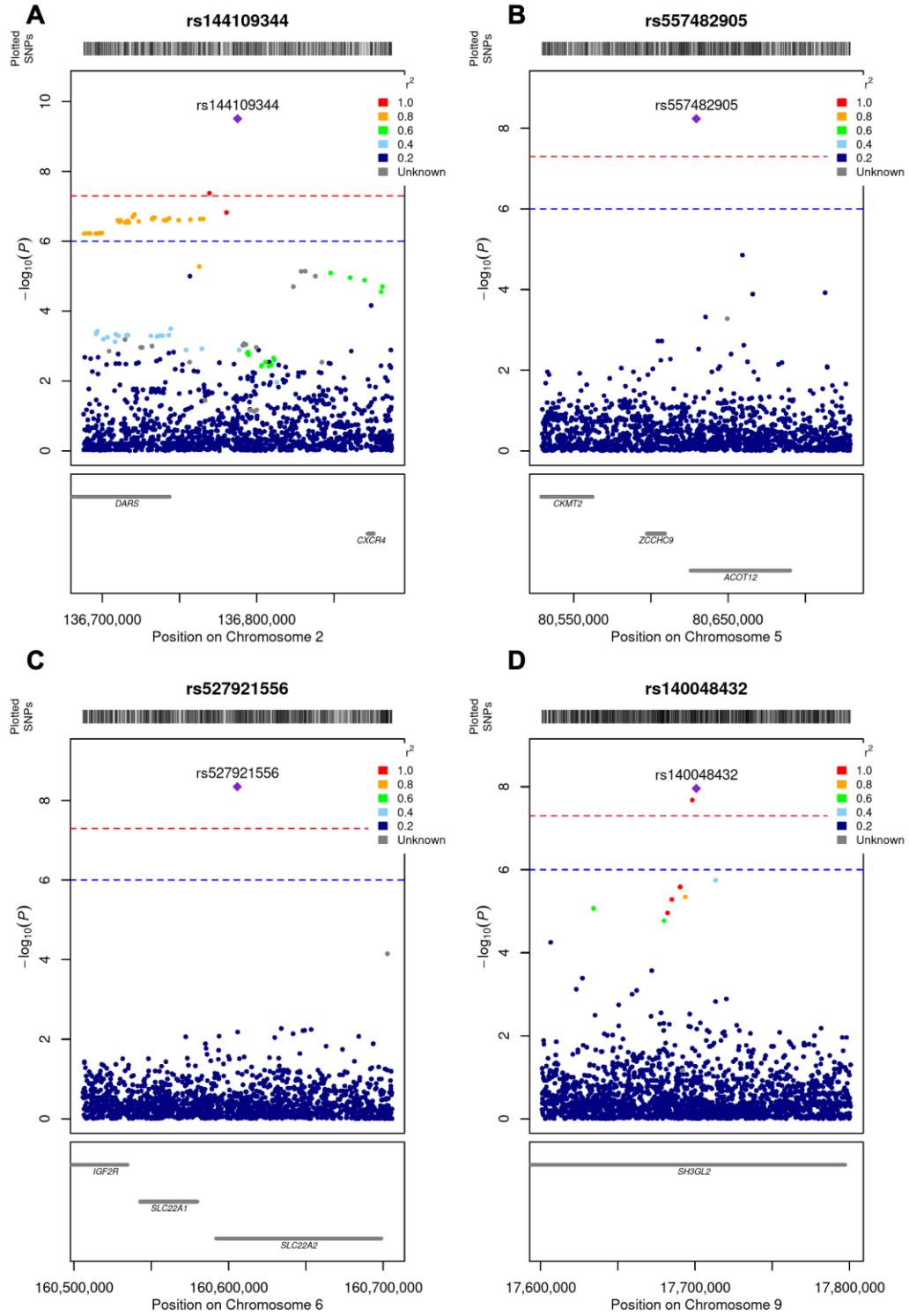
Continental ancestry group	Type[1]	Dimension	Variance explained (%)
South Asian	ROB	Dimension 1	80.00
South Asian	ROB	Dimension 2	18.20
South Asian	COB	Dimension 1	61.60
South Asian	COB	Dimension 2	25.31
South Asian	COB	Dimension 3	13.09
East Asian	ROB	Dimension 1	92.11
East Asian	ROB	Dimension 2	7.89
East Asian	COB	Dimension 1	50.49
East Asian	COB	Dimension 2	49.51
European	ROB	Dimension 1	58.25
European	ROB	Dimension 2	28.67
European	COB	Dimension 1	40.43
European	COB	Dimension 2	31.89
European	COB	Dimension 3	22.09
European	COB	Dimension 4	4.53

[1]COB = country of birth; ROB = UN region based on COB data

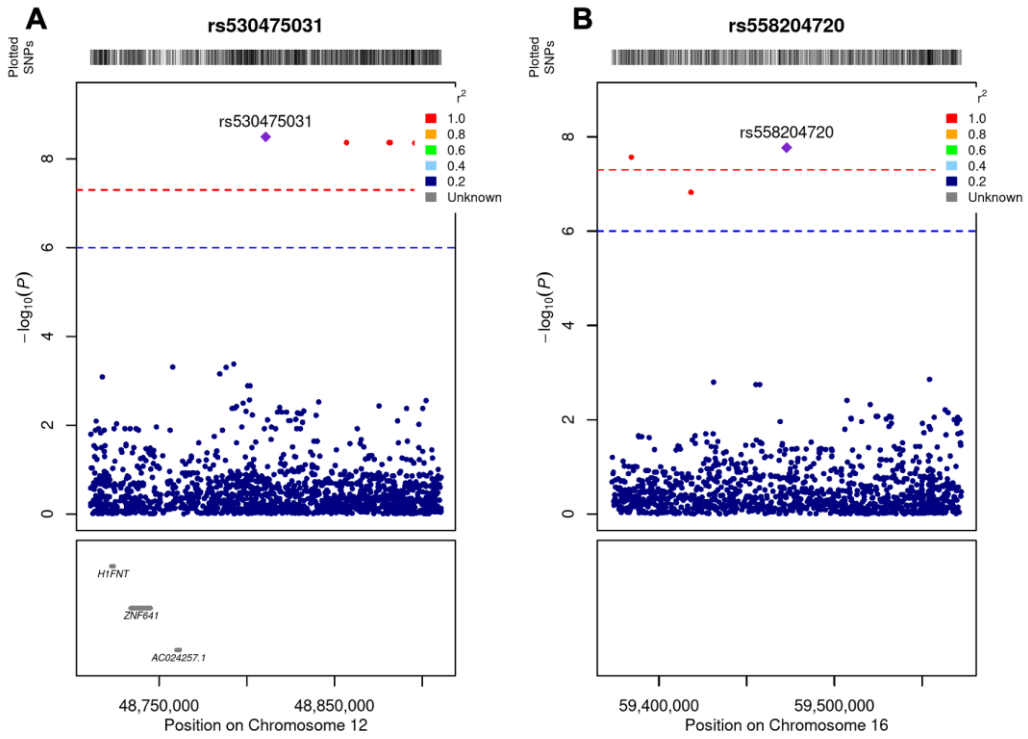
Appendix 17. *Regional plots of index SNPs (1).*



Appendix 18. *Regional plots of index SNPs (2).*



Appendix 19. *Regional plots of index SNPs (3).*



Appendix 20. *Summary statistics for neutrophil count (clumping r^2 = 0.1).*

N.META = sample-size in the SNPTEST/META run; K.META = number of K-means clusters used in the SNPTEST/META run.

SNP	CHR	BP (GRCh37)	EA	NEA	EAF	INFO	BETA.BOLT	SE.BOLT	P.BOLT	N.META	K.META
rs28734019	1	90800573	C	T	0.998	0.35	-0.65	0.12	2.90E-08	4124	4
rs12747038	1	146651428	T	G	0.990	0.93	-0.22	0.04	3.90E-09	5976	7
rs112483667	1	151651180	A	AT	0.974	0.97	-0.13	0.02	1.70E-08	5976	7
rs11581802	1	152936164	G	A	0.976	0.91	-0.17	0.02	1.10E-12	5976	7
rs9330298	1	153590254	C	A	0.942	0.80	-0.11	0.02	4.70E-10	5976	7
rs72696290	1	154345686	T	C	0.944	0.98	-0.09	0.02	5.60E-09	5976	7
rs61811432	1	154684462	C	T	0.975	0.71	-0.16	0.03	5.00E-09	5976	7
rs4845401	1	154941593	C	G	0.979	1.00	-0.14	0.03	1.40E-08	5976	7
rs61811895	1	154976137	G	T	0.995	1.00	-0.31	0.05	7.50E-10	4124	4
rs11582072	1	155477570	T	C	0.971	1.00	-0.21	0.02	1.00E-19	5976	7
rs670523	1	155878732	A	G	0.959	0.91	-0.15	0.02	2.10E-14	5976	7
rs3768276	1	156198366	G	A	0.966	0.84	-0.18	0.02	1.40E-15	5976	7
rs10908505	1	156468243	T	A	0.968	0.91	-0.17	0.02	2.20E-14	5976	7
rs11264504	1	156560624	C	T	0.848	0.99	-0.06	0.01	1.10E-10	5976	7
rs12566986	1	156728317	G	A	0.977	0.93	-0.14	0.02	2.70E-08	5976	7
rs2768759	1	156852463	A	C	0.931	1.00	-0.10	0.01	2.00E-13	5976	7
rs17404670	1	157066893	G	A	0.971	0.77	-0.15	0.02	2.20E-10	5976	7
rs10908530	1	157076715	C	T	0.966	0.70	-0.13	0.02	3.10E-08	5976	7
rs12138690	1	157415618	T	C	0.985	0.77	-0.21	0.03	1.80E-10	5976	7
rs3811035	1	157485561	G	A	0.910	0.86	-0.11	0.01	1.10E-15	5976	7
rs12406899	1	157540651	T	G	0.911	0.95	-0.07	0.01	1.90E-08	5975	7
rs7535596	1	157673356	A	G	0.125	0.94	0.07	0.01	1.20E-10	5975	7
rs2210914	1	157673628	T	C	0.011	0.88	0.34	0.04	8.80E-22	4643	5

rs4272616	1	157865663	T	C	0.041	0.85	0.11	0.02	2.40E-08	5976	7
rs927698	1	157926455	G	A	0.972	0.82	-0.19	0.02	9.80E-16	5976	7
rs1888821	1	157977753	C	A	0.995	1.00	-0.29	0.05	1.50E-08	3859	5
rs6427419	1	158058109	C	A	0.962	1.00	-0.16	0.02	2.90E-17	5976	7
rs74781198	1	158075448	C	T	0.998	1.00	-0.46	0.08	3.00E-09	3157	3
rs371178711	1	158186653	C	T	0.969	0.90	-0.17	0.02	6.90E-15	5976	7
rs74802440	1	158490785	A	T	0.994	0.87	-0.30	0.05	1.30E-10	3859	5
rs12044097	1	158597309	A	G	0.992	1.00	-0.22	0.04	2.60E-08	3157	3
rs34542525	1	158664483	A	G	0.998	1.00	-0.48	0.08	1.00E-09	3157	3
rs539456851	1	158731459	T	TTA	0.982	0.73	-0.27	0.03	2.20E-17	5976	7
1:158777618_CT_C	1	158777618	CT	C	0.050	0.76	0.12	0.02	1.40E-10	5976	7
rs34167592	1	158883393	C	A	0.960	0.78	-0.12	0.02	4.10E-09	5976	7
rs1103805	1	158924741	C	A	0.929	0.91	-0.08	0.01	3.20E-08	5976	7
rs146677619	1	158995984	A	AT	0.991	0.66	-0.35	0.05	3.70E-13	5793	6
rs703153	1	159105879	C	G	0.007	0.86	0.27	0.05	4.40E-09	5793	6
rs2814778	1	159174683	T	C	0.036	1.00	0.43	0.02	2.70E-87	5793	6
rs56921594	1	159314302	A	G	0.993	0.97	-0.26	0.04	4.10E-11	5976	7
rs77238873	1	159369847	T	C	0.997	0.91	-0.37	0.06	2.60E-09	4826	5
rs1446968	1	159535940	G	C	0.994	1.00	-0.30	0.05	3.30E-10	3157	3
rs61823703	1	159542164	T	C	0.987	0.82	-0.31	0.04	6.00E-19	5793	6
rs61380083	1	159604428	C	A	0.995	1.00	-0.29	0.05	3.40E-08	3859	5
rs12760041	1	159714035	C	T	0.979	0.94	-0.19	0.03	2.10E-14	5793	6
rs6677719	1	159723120	C	T	0.027	0.86	0.14	0.02	3.10E-09	5976	7
rs11591079	1	159796302	G	T	0.990	0.71	-0.23	0.04	4.00E-08	5975	7
rs11422063	1	159799599	C	CA	0.022	0.50	0.19	0.03	1.20E-08	5976	7
rs4656856	1	159859392	C	T	0.987	1.00	-0.31	0.03	1.50E-22	3340	4

rs2789423	1	159885500	G	A	0.946	1.00	-0.11	0.02	1.80E-13	5976	7
rs1320568	1	159912346	A	G	0.958	1.00	-0.11	0.02	1.30E-09	5976	7
rs16831234	1	159977732	T	A	0.981	0.91	-0.15	0.03	4.30E-08	5976	7
rs1186685	1	160029212	A	G	0.984	0.85	-0.20	0.03	2.70E-11	5976	7
rs12402888	1	160059748	C	A	0.995	0.82	-0.34	0.05	3.40E-10	4307	4
rs2369725	1	160587202	A	T	0.979	0.93	-0.18	0.03	2.00E-12	5976	7
rs535622	1	160710745	C	T	0.982	0.87	-0.21	0.03	4.50E-13	5976	7
rs11576058	1	161111446	C	A	0.979	0.98	-0.16	0.02	9.80E-11	5976	7
rs78603008	1	161559720	G	A	0.984	1.00	-0.16	0.03	1.70E-09	5793	6
rs6695760	1	161885545	G	C	0.966	0.95	-0.20	0.02	1.90E-20	5976	7
rs12737539	1	162198429	G	A	0.942	0.84	-0.11	0.02	1.60E-11	5976	7
rs4657188	1	162347877	G	C	0.972	1.00	-0.12	0.02	1.20E-08	5976	7
rs12730805	1	162873220	C	A	0.993	1.00	-0.26	0.04	1.00E-09	4124	4
rs10733036	1	162918475	G	A	0.015	0.90	0.17	0.03	1.50E-08	5976	7
rs12037463	1	164512386	C	T	0.985	0.93	-0.16	0.03	3.80E-08	5976	7
rs183362544	2	97045902	C	T	0.998	0.67	0.61	0.11	1.30E-08	2717	4
rs144109344	2	136787730	C	T	0.964	0.93	-0.12	0.02	3.10E-10	5976	7
rs557482905	5	80629499	C	T	0.998	0.75	0.55	0.10	5.80E-09	3778	4
rs527921556	6	160605701	T	C	0.996	0.69	0.40	0.07	4.50E-09	5793	6
rs10096834	8	116281087	T	C	0.573	0.98	0.04	0.01	2.30E-08	5976	7
rs140048432	9	17700893	T	C	0.996	0.86	-0.33	0.06	1.10E-08	5976	7
rs530475031	12	48810860	G	T	0.998	0.37	0.73	0.12	3.20E-09	4952	5
rs558204720	16	59472815	T	C	0.998	0.84	0.52	0.09	1.70E-08	1486	2
rs138163369	18	6492075	T	C	0.998	0.58	0.53	0.10	4.90E-08	4619	5

Appendix 21. *Independent SNPs from the main GWAS inside GWAS Catalog.*

SNP	CHR	BP (GRCh37)	PMID	First.Author	Date	Trait
rs11581802	1	152887412	34187551	McCartney DL	29/06/2021	DNA methylation-estimated granulocyte proportions
rs11584009	1	153081204	32888493	Chen MH	01/09/2020	Monocyte count
rs9330298	1	153628645	32888493	Chen MH	01/09/2020	Monocyte count
rs11582072	1	155273869	34187551	McCartney DL	29/06/2021	DNA methylation-estimated granulocyte proportions
rs12134456	1	155722506	31152163	Wuttke M	31/05/2019	Blood urea nitrogen levels
rs12134456	1	155722506	31676860	Zhao B	01/11/2019	Brain region volumes
rs12134456	1	155722506	32888493	Chen MH	01/09/2020	Mean corpuscular hemoglobin concentration
rs12134456	1	155722506	31578528	Tin A	02/10/2019	Urate levels
rs12134456	1	155722506	31578528	Tin A	02/10/2019	Urate levels
rs12134456	1	155722506	31578528	Tin A	02/10/2019	Urate levels
rs3856261	1	155876613	26198764	Goes FS	21/07/2015	Schizophrenia
rs3856261	1	155876613	30595370	Kichaev G	27/12/2018	Eczema
rs11582072	1	155878732	23128233	Jostins L	01/11/2012	Inflammatory bowel disease
rs11582072	1	155878732	26192919	Liu JZ	20/07/2015	Crohn's disease
rs11582072	1	155878732	31043758	Warrington NM	01/05/2019	Birth weight
rs11582072	1	155907823	31217584	Wojcik GL	19/06/2019	White blood cell count
rs10908505	1	156406381	30595370	Kichaev G	27/12/2018	Body mass index
rs10908505	1	156468243	30595370	Kichaev G	27/12/2018	Height
rs10908505	1	156468243	32888494	Vuckovic D	01/09/2020	Plateletcrit
rs849830	1	157527250	34187551	McCartney DL	29/06/2021	DNA methylation-estimated granulocyte proportions

SNP	CHR	BP (GRCh37)	PMID	First.Author	Date	Trait
rs11582663	1	157559122	29875488	Sun BB	06/06/2018	Blood protein levels
rs11582663	1	157561420	30072576	Emilsson V	02/08/2018	Blood protein levels
rs11582663	1	157561420	30072576	Emilsson V	02/08/2018	Blood protein levels
rs11264798	1	157668993	27723758	Bronson PG	10/10/2016	Selective IgA deficiency
rs11264798	1	157670816	21829393	Plagnol V	04/08/2011	Insulinoma-associated antigen 2 autoantibody levels in type 1 diabetes
rs11264798	1	157670816	29875488	Sun BB	06/06/2018	Blood protein levels
rs6427401	1	157674997	24390342	Okada Y	25/12/2013	Rheumatoid arthritis
rs6427401	1	157674997	30423114	Laufer VA	13/11/2018	Rheumatoid arthritis
rs6427401	1	157674997	32514122	Ishigaki K	08/06/2020	Graves' disease
rs6427401	1	157674997	33272962	Yin X	03/12/2020	Systemic lupus erythematosus
rs6427401	1	157748564	30072576	Emilsson V	02/08/2018	Blood protein levels
rs6427401	1	157779182	29875488	Sun BB	06/06/2018	Blood protein levels
rs6427401	1	157798000	30895295	Jonnalagadda M	21/03/2019	Iris heterochromicity
rs7534518	1	157798923	30072576	Emilsson V	02/08/2018	Blood protein levels
rs7534518	1	157798923	30072576	Emilsson V	02/08/2018	Blood protein levels
rs7534518	1	157798923	30072576	Emilsson V	02/08/2018	Blood protein levels
rs7534518	1	157798923	30072576	Emilsson V	02/08/2018	Blood protein levels
rs6427419	1	158058109	23251661	Comuzzie AG	04/12/2012	Obesity-related traits

SNP	CHR	BP (GRCh37)	PMID	First.Author	Date	Trait
rs74802440	1	158518050	33462484	Sinnott-Armstrong N	18/01/2021	Glycated hemoglobin levels
rs34542525	1	158664483	33462484	Sinnott-Armstrong N	18/01/2021	Glycated hemoglobin levels
rs10489844	1	158728389	34187551	McCartney DL	29/06/2021	DNA methylation-estimated granulocyte proportions
rs4657616	1	158971086	23263863	Li J	20/12/2012	Hematology traits
rs4657616	1	158971086	31217584	Wojcik GL	19/06/2019	White blood cell count
rs856046	1	158983593	34187551	McCartney DL	29/06/2021	DNA methylation-estimated granulocyte proportions
rs856046	1	158987941	34187551	McCartney DL	29/06/2021	DNA methylation-estimated granulocyte proportions
rs856046	1	159012646	25884002	Moore CB	09/01/2015	Neutrophil count in HIV-infection
rs856046	1	159062436	25096241	Keller MF	05/08/2014	White blood cell count
rs856046	1	159062436	25096241	Keller MF	05/08/2014	White blood cell count
rs856046	1	159062436	31217584	Wojcik GL	19/06/2019	White blood cell count
rs4656165	1	159167290	30038396	Lee JJ	23/07/2018	Educational attainment (years of education)
rs4656165	1	159167290	30038396	Lee JJ	23/07/2018	Educational attainment (MTAG)
rs4656165	1	159169463	32296059	Han Y	15/04/2020	Asthma
rs4656165	1	159169463	31374203	Lam M	01/08/2019	Cognitive ability, years of educational attainment or schizophrenia (pleiotropy)
rs2814778	1	159174683	21507922	Ramsuran V	01/05/2011	Neutrophil count
rs2814778	1	159174683	22037903	Crosslin DR	30/10/2011	White blood cell count
rs2814778	1	159174683	31869403	Kowalski MH	23/12/2019	White blood cell count

SNP	CHR	BP (GRCh37)	PMID	First.Author	Date	Trait
rs2814778	1	159174683	31675503	Gurdasani D	01/10/2019	White blood cell count
rs2814778	1	159174683	31675503	Gurdasani D	01/10/2019	Monocyte count
rs2814778	1	159174683	31675503	Gurdasani D	01/10/2019	Neutrophil count
rs2814778	1	159174683	25884002	Moore CB	09/01/2015	Neutrophil count in HIV-infection
rs2814778	1	159174683	28158719	Jain D	01/02/2017	White blood cell count
rs2814778	1	159174683	28158719	Jain D	01/02/2017	White blood cell count (monocyte)
rs2814778	1	159174683	28158719	Jain D	01/02/2017	White blood cell count (neutrophil)
rs2814778	1	159174683	31708768	Liu C	25/10/2019	Cerebrospinal fluid sTREM-2 levels
rs2814778	1	159174683	27863252	Astle WJ	17/11/2016	Granulocyte count
rs2814778	1	159174683	27863252	Astle WJ	17/11/2016	Sum neutrophil eosinophil counts
rs2814778	1	159174683	27863252	Astle WJ	17/11/2016	Myeloid white cell count
rs2814778	1	159174683	27863252	Astle WJ	17/11/2016	Neutrophil count
rs2814778	1	159174683	27863252	Astle WJ	17/11/2016	Sum basophil neutrophil counts
rs2814778	1	159174683	27863252	Astle WJ	17/11/2016	Lymphocyte percentage of white cells
rs2814778	1	159174683	27863252	Astle WJ	17/11/2016	Neutrophil percentage of white cells
rs2814778	1	159174683	32888493	Chen MH	01/09/2020	Monocyte count
rs2814778	1	159174683	32888493	Chen MH	01/09/2020	Monocyte count
rs2814778	1	159174683	32888493	Chen MH	01/09/2020	Neutrophil count
rs2814778	1	159174683	32888493	Chen MH	01/09/2020	Neutrophil count
rs2814778	1	159174683	32888493	Chen MH	01/09/2020	Neutrophil count
rs2814778	1	159174683	32888494	Vuckovic D	01/09/2020	Lymphocyte percentage of white cells

SNP	CHR	BP (GRCh37)	PMID	First.Author	Date	Trait
rs2814778	1	159174683	32888493	Chen MH	01/09/2020	White blood cell count
rs2814778	1	159174683	31217584	Wojcik GL	19/06/2019	White blood cell count
rs2814778	1	159174683	32888494	Vuckovic D	01/09/2020	Monocyte count
rs2814778	1	159174683	32888493	Chen MH	01/09/2020	White blood cell count
rs2814778	1	159174683	32888493	Chen MH	01/09/2020	White blood cell count
rs2814778	1	159174683	30647433	Legge SE	15/01/2019	Neutrophil level response to clozapine in treatment-resistant schizophrenia
rs2814778	1	159174683	32888494	Vuckovic D	01/09/2020	Neutrophil count
rs2814778	1	159174683	32888494	Vuckovic D	01/09/2020	Neutrophil percentage of white cells
rs2814778	1	159174683	32888494	Vuckovic D	01/09/2020	White blood cell count
rs2814778	1	159174683	29596498	Charles BA	29/03/2018	Low white blood cell count
rs2814778	1	159174683	34187551	McCartney DL	29/06/2021	DNA methylation-estimated granulocyte proportions
rs2814778	1	159174683	34187551	McCartney DL	29/06/2021	DNA methylation PhenoAge acceleration
rs12075	1	159175354	23017229	Voruganti VS	24/09/2012	Monocyte chemoattractant protein-1 levels
rs12075	1	159175354	23251661	Comuzzie AG	04/12/2012	Obesity-related traits
rs12075	1	159175354	22037903	Crosslin DR	30/10/2011	White blood cell count
rs12075	1	159175354	22291609	Naitza S	26/01/2012	Inflammatory biomarkers
rs12075	1	159175354	25201988	Rietveld CA	08/09/2014	Educational attainment
rs12075	1	159175354	29875488	Sun BB	06/06/2018	Blood protein levels
rs12075	1	159175354	29875488	Sun BB	06/06/2018	Blood protein levels
rs12075	1	159175354	27989323	Ahola-Olli AV	13/12/2016	Growth-regulated protein alpha levels

SNP	CHR	BP (GRCh37)	PMID	First.Author	Date	Trait
rs12075	1	159175354	27989323	Ahola-Olli AV	13/12/2016	Eotaxin levels
rs12075	1	159175354	27989323	Ahola-Olli AV	13/12/2016	Interleukin-8 levels
rs12075	1	159175354	27989323	Ahola-Olli AV	13/12/2016	Monocyte chemoattractant protein-1 levels
rs12075	1	159175354	27863252	Astle WJ	17/11/2016	Myeloid white cell count
rs12075	1	159175354	27863252	Astle WJ	17/11/2016	White blood cell count (basophil)
rs12075	1	159175354	27863252	Astle WJ	17/11/2016	Monocyte count
rs12075	1	159175354	27863252	Astle WJ	17/11/2016	Basophil percentage of white cells
rs12075	1	159175354	27863252	Astle WJ	17/11/2016	Basophil percentage of granulocytes
rs12075	1	159175354	32641083	Hillary RF	08/07/2020	Monocyte chemoattractant protein-4 levels
rs12075	1	159175354	32888493	Chen MH	01/09/2020	Monocyte count
rs12075	1	159175354	27532455	Sun W	17/08/2016	Blood protein levels
rs12075	1	159175354	32888493	Chen MH	01/09/2020	Basophil count
rs12075	1	159175354	32888494	Vuckovic D	01/09/2020	Basophil count
rs12075	1	159175354	32888494	Vuckovic D	01/09/2020	Basophil percentage of white cells
rs12075	1	159175354	32888493	Chen MH	01/09/2020	Basophil count
rs12075	1	159175354	32888494	Vuckovic D	01/09/2020	Lymphocyte percentage of white cells
rs12075	1	159175354	31217584	Wojcik GL	19/06/2019	White blood cell count
rs12075	1	159175354	32888494	Vuckovic D	01/09/2020	Monocyte count
rs12075	1	159175354	32888494	Vuckovic D	01/09/2020	Neutrophil count
rs12075	1	159175354	32888494	Vuckovic D	01/09/2020	Monocyte percentage of white cells
rs12075	1	159175354	33067605	Folkersen L	16/10/2020	C-X-C motif chemokine 6 levels

SNP	CHR	BP (GRCh37)	PMID	First.Author	Date	Trait
rs12075	1	159175354	32888494	Vuckovic D	01/09/2020	White blood cell count
rs12075	1	159175354	33067605	Folkersen L	16/10/2020	Monocyte chemoattractant protein-1 levels
rs12075	1	159175354	33227023	Wang Y	23/11/2020	Monocyte chemoattractant protein-1 levels
rs12075	1	159175354	31217265	Sliz E	19/06/2019	Monocyte chemoattractant protein-1 levels
rs13962	1	159175527	22075330	Granada M	08/11/2011	IgE levels
rs13962	1	159175527	29875488	Sun BB	06/06/2018	Blood protein levels
rs3845622	1	159176490	33462484	Sinnott-Armstrong N	18/01/2021	C-reactive protein levels
rs11265155	1	159218266	22291609	Naitza S	26/01/2012	Inflammatory biomarkers
rs11265155	1	159218266	31217584	Wojcik GL	19/06/2019	C-reactive protein levels
rs863016	1	159223787	34187551	McCartney DL	29/06/2021	DNA methylation-estimated granulocyte proportions
rs78478121	1	159237950	34187551	McCartney DL	29/06/2021	DNA methylation-estimated granulocyte proportions
rs77383163	1	159282664	33462484	Sinnott-Armstrong N	18/01/2021	C-reactive protein levels
rs4656236	1	159326880	22075330	Granada M	08/11/2011	IgE levels
rs9427014	1	159342439	20237162	Bozaoglu K	17/03/2010	Chemerin levels
rs9427014	1	159357684	20237162	Bozaoglu K	17/03/2010	Chemerin levels
rs4656236	1	159410975	27989323	Ahola-Olli AV	13/12/2016	Growth-regulated protein alpha levels
rs4656236	1	159410975	27989323	Ahola-Olli AV	13/12/2016	Monocyte chemoattractant protein-1 levels
rs1446968	1	159535940	33462484	Sinnott-Armstrong N	18/01/2021	Serum phosphate levels

SNP	CHR	BP (GRCh37)	PMID	First.Author	Date	Trait
rs17457976	1	159580873	33462484	Sinnott-Armstrong N	18/01/2021	C-reactive protein levels
rs61380083	1	159608855	34187551	McCartney DL	29/06/2021	DNA methylation-estimated granulocyte proportions
rs12760041	1	159670145	23696881	Wood AR	16/05/2013	Blood protein levels
rs12760041	1	159670336	27286809	Ligthart S	10/06/2016	C-reactive protein levels or triglyceride levels (pleiotropy)
rs12760041	1	159722056	30388399	Ligthart S	01/11/2018	C-reactive protein levels
rs12760041	1	159722056	33462484	Sinnott-Armstrong N	18/01/2021	C-reactive protein levels
1:159891160_TAAC_T	1	159892088	27863252	Astle WJ	17/11/2016	Mean platelet volume
1:159891160_TAAC_T	1	159892088	32888493	Chen MH	01/09/2020	Mean platelet volume
1:159891160_TAAC_T	1	159892088	32888493	Chen MH	01/09/2020	Mean platelet volume
1:159891160_TAAC_T	1	159892088	32888494	Vuckovic D	01/09/2020	Mean platelet volume
rs1934073	1	159936733	32939015	Zhu Z	16/09/2020	Ebbinghaus illusion (overestimation)
rs1934073	1	159936733	34187551	McCartney DL	29/06/2021	DNA methylation-estimated granulocyte proportions
rs4656342	1	161945419	31217584	Wojcik GL	19/06/2019	White blood cell count
rs10918701	1	162090536	31689377	Wootton RE	06/11/2019	Lifetime smoking index

SNP	CHR	BP (GRCh37)	PMID	First.Author	Date	Trait
rs4424487	1	162198429	32527199	van Duijvenboden S	11/06/2020	QT dynamics during recovery from exercise
rs115653138	2	136769426	32888493	Chen MH	01/09/2020	White blood cell count
rs144109344	2	136787730	32888493	Chen MH	01/09/2020	Neutrophil count

Appendix 22. *Single SNP MR analysis of neutrophil count on severe malaria.*

Exposure	Outcome	SNP	CHR	BP	EA	NEA	EAF	b.MR	se.MR	p.MR	LD Proxy
Neutrophil count	Overall severe malaria	rs144109344	2	136787730	C	T	0.96	0.01	0.07	0.83	rs144109344
Neutrophil count	Cerebral malaria	rs144109344	2	136787730	C	T	0.96	-0.02	0.10	0.82	rs144109344
Neutrophil count	Severe malaria anaemia	rs144109344	2	136787730	C	T	0.96	0.19	0.16	0.23	rs144109344
Neutrophil count	Other severe malaria	rs144109344	2	136787730	C	T	0.96	-0.06	0.09	0.50	rs144109344
Neutrophil count	Overall severe malaria	rs2325919	1	159222811	G	T	0.98	-0.14	0.25	0.57	rs2814778
Neutrophil count	Cerebral malaria	rs2325919	1	159222811	G	T	0.98	-0.29	0.36	0.42	rs2814778
Neutrophil count	Severe malaria anaemia	rs2325919	1	159222811	G	T	0.98	-0.18	0.63	0.78	rs2814778
Neutrophil count	Other severe malaria	rs2325919	1	159222811	G	T	0.98	0.09	0.38	0.82	rs2814778
Neutrophil count	Overall severe malaria	rs7460611	8	116272546	C	T	0.57	0.03	0.02	0.22	rs10096834
Neutrophil count	Cerebral malaria	rs7460611	8	116272546	C	T	0.57	0.00	0.03	0.90	rs10096834
Neutrophil count	Severe malaria anaemia	rs7460611	8	116272546	C	T	0.57	0.07	0.05	0.14	rs10096834
Neutrophil count	Other severe malaria	rs7460611	8	116272546	C	T	0.57	0.04	0.03	0.16	rs10096834

Appendix 23. *Single SNP MR analysis of severe malaria on neutrophil count.*

Exposure	Outcome	SNP	CHR	BP	EA	NEA	EAF	b.MR	se.MR	p.MR
Overall severe malaria	Neutrophil count	rs113892119	11	5273865	C	G	0.05	0.93	0.98	0.34
Overall severe malaria	Neutrophil count	rs116423146	3	160396863	T	C	0.09	-4.16	4.15	0.32
Overall severe malaria	Neutrophil count	rs1419114	1	203652444	A	G	0.31	-7.01	6.98	0.32
Overall severe malaria	Neutrophil count	rs553707144	4	144988500	A	T	0.06	2.32	2.44	0.34
Overall severe malaria	Neutrophil count	rs557568961	11	5497277	C	G	0.04	0.70	0.70	0.31
Overall severe malaria	Neutrophil count	rs57032711	9	129250119	A	G	0.13	4.86	4.85	0.32
Overall severe malaria	Neutrophil count	rs8176751	9	136131022	T	C	0.19	-4.87	4.92	0.32
Cerebral malaria	Neutrophil count	rs113892119	11	5273865	C	G	0.06	0.77	0.81	0.34
Cerebral malaria	Neutrophil count	rs543034558	11	4986130	T	C	0.06	0.75	0.81	0.35
Other severe malaria	Neutrophil count	rs113892119	11	5273865	C	G	0.05	1.16	1.22	0.34
Other severe malaria	Neutrophil count	rs116423146	3	160396863	T	C	0.09	-3.32	3.31	0.32
Other severe malaria	Neutrophil count	rs557568961	11	5497277	C	G	0.04	0.78	0.78	0.31

Appendix 24. *STROBE-MR checklist of recommended items to address in reports of Mendelian randomization studies.*

Item No.	Section	Checklist item	Page No.	Relevant text from manuscript
1	**TITLE and ABSTRACT**	Indicate Mendelian randomization (MR) as the study's design in the title and/or the abstract if that is a main purpose of the study	1-3	
	INTRODUCTION			
2	**Background**	Explain the scientific background and rationale for the reported study. What is the exposure? Is a potential causal relationship between exposure and outcome plausible? Justify why MR is a helpful method to address the study question	3-4	Introduction, paragraphs 1-4.
3	**Objectives**	State specific objectives clearly, including pre-specified causal hypotheses (if any). State that MR is a method that, under specific assumptions, intends to estimate causal effects	4-5	Introduction, paragraphs 5-6.
	METHODS			
4	**Study design and data sources**	Present key elements of the study design early in the article. Consider including a table listing sources of data for all phases of the study. For each data source contributing to the analysis, describe the following:		
	a)	Setting: Describe the study design and the underlying population, if possible. Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection, when available.	5-6	Methods, paragraphs 1-5. Referenced UK Biobank paper by Bycroft et al.
	b)	Participants: Give the eligibility criteria, and the sources and methods of selection of participants. Report the sample size, and whether any power or sample size calculations were carried out prior to the main analysis	5-6	Supplementary Table 2. Referenced UK Biobank paper by Bycroft et al.
	c)	Describe measurement, quality control and selection of genetic variants	5-7	Supplementary Table 2. Referenced UK Biobank paper by Bycroft et al.
	d)	For each exposure, outcome, and other relevant variables, describe methods of assessment and diagnostic criteria for diseases	5-6	Referenced UK Biobank paper by Bycroft et al and Neale Lab study.

Item No.	Section	Checklist item	Page No.	Relevant text from manuscript
	e)	Provide details of ethics committee approval and participant informed consent, if relevant	8,16,17	
5	**Assumptions**	Explicitly state the three core IV assumptions for the main analysis (relevance, independence and exclusion restriction) as well assumptions for any additional or sensitivity analysis	4,7	Supplementary Figure 1, Supplementary Methods.
6	**Statistical methods: main analysis**	Describe statistical methods and statistics used		
	a)	Describe how quantitative variables were handled in the analyses (i.e., scale, units, model)	6	Methods, Supplementary Table 2.
	b)	Describe how genetic variants were handled in the analyses and, if applicable, how their weights were selected	6,7	Methods, paragraph 6.
	c)	Describe the MR estimator (e.g. two-stage least squares, Wald ratio) and related statistics. Detail the included covariates and, in case of two-sample MR, whether the same covariate set was used for adjustment in the two samples	7	Methods, paragraph 7.
	d)	Explain how missing data were addressed	7	
	e)	If applicable, indicate how multiple testing was addressed	8	Methods, paragraph 9. Supplementary Methods.
7	**Assessment of assumptions**	Describe any methods or prior knowledge used to assess the assumptions or justify their validity	8-9	Methods. Supplementary Methods.
8	**Sensitivity analyses and additional analyses**	Describe any sensitivity analyses or additional analyses performed (e.g. comparison of effect estimates from different approaches, independent replication, bias analytic techniques, validation of instruments, simulations)	7	Methods. Supplementary Methods.

Item No.	Section	Checklist item	Page No.	Relevant text from manuscript
9	**Software and pre-registration**			
	a)	Name statistical software and package(s), including version and settings used	5-7,9	Methods, paragraphs 2-3,6-9. Supplementary Methods.
	b)	State whether the study protocol and details were pre-registered (as well as when and where)	N/A	
	RESULTS			
10	**Descriptive data**			
	a)	Report the numbers of individuals at each stage of included studies and reasons for exclusion. Consider use of a flow diagram	5,10	Methods, paragraph 1-3. Results, paragraph 1. Figure 1.
	b)	Report summary statistics for phenotypic exposure(s), outcome(s), and other relevant variables (e.g. means, SDs, proportions)		Supplementary Tables 1 and 2.
	c)	If the data sources include meta-analyses of previous studies, provide the assessments of heterogeneity across these studies	N/A	
	d)	For two-sample MR: i. Provide justification of the similarity of the genetic variant-exposure associations between the exposure and outcome samples ii. Provide information on the number of individuals who overlap between the exposure and outcome studies	5-6	Methods, paragraphs 2-3.
11	**Main results**			
	a)	Report the associations between genetic variant and exposure, and between genetic variant and outcome, preferably on an interpretable scale	8-11	

Item No.	Section		Checklist item	Page No.	Relevant text from manuscript
		b)	Report MR estimates of the relationship between exposure and outcome, and the measures of uncertainty from the MR analysis, on an interpretable scale, such as odds ratio or relative risk per SD difference	8-11	
		c)	If relevant, consider translating estimates of relative risk into absolute risk for a meaningful time period	N/A	
		d)	Consider plots to visualize results (e.g. forest plot, scatterplot of associations between genetic variants and outcome versus between genetic variants and exposure)		Figures 2-3
12	**Assessment of assumptions**				
		a)	Report the assessment of the validity of the assumptions	9-10	Table 1
		b)	Report any additional statistics (e.g., assessments of heterogeneity across genetic variants, such as I^2, Q statistic or E-value)	10-12	Table 1
13	**Sensitivity analyses and additional analyses**				
		a)	Report any sensitivity analyses to assess the robustness of the main results to violations of the assumptions	9-10	
		b)	Report results from other sensitivity analyses or additional analyses		Supplementary table 4
		c)	Report any assessment of direction of causal relationship (e.g., bidirectional MR)	10	
		d)	When relevant, report and compare with estimates from non-MR analyses	9-13	Methods & Discussion.
		e)	Consider additional plots to visualize results (e.g., leave-one-out analyses)		

Item No.	Section	Checklist item	Page No.	Relevant text from manuscript
	DISCUSSION			
14	**Key results**	Summarize key results with reference to study objectives	11	
15	**Limitations**	Discuss limitations of the study, taking into account the validity of the IV assumptions, other sources of potential bias, and imprecision. Discuss both direction and magnitude of any potential bias and any efforts to address them	6,7	
16	**Interpretation**			
	a)	Meaning: Give a cautious overall interpretation of results in the context of their limitations and in comparison with other studies	11-13	
	b)	Mechanism: Discuss underlying biological mechanisms that could drive a potential causal relationship between the investigated exposure and the outcome, and whether the gene-environment equivalence assumption is reasonable. Use causal language carefully, clarifying that IV estimates may provide causal effects only under certain assumptions	11-14	
	c)	Clinical relevance: Discuss whether the results have clinical or public policy relevance, and to what extent they inform effect sizes of possible interventions	3,13-14	
17	**Generalizability**	Discuss the generalizability of the study results (a) to other populations, (b) across other exposure periods/timings, and (c) across other levels of exposure		
	OTHER INFORMATION			
18	**Funding**	Describe sources of funding and the role of funders in the present study and, if applicable, sources of funding for the databases and original study or studies on which the present study is based	15-16	

Item No.	Section	Checklist item	Page No.	Relevant text from manuscript
19	**Data and data sharing**	Provide the data used to perform all analyses or report where and how the data can be accessed, and reference these sources in the article. Provide the statistical code needed to reproduce the results in the article, or report whether the code is publicly accessible and if so, where	15	
20	**Conflicts of Interest**	All authors should declare all potential conflicts of interest	17	

Appendix 25. *Protein quantitative trait loci data used in mediation analysis.*

Protein full name	Author	Year	PMID	UniProtID	SomaID	Protein abr.	N	Beta*	SE*	P_val*
Leptin	Goudswaard LJ	2021	34226637	P41159	SL000498	LEP	2728	0.68	0.08	3.88E-17
Leptin	Goudswaard LJ	2021	34226637	P41159	SL000498	LEP	2728	0.63	0.08	1.55E-15
Fatty acid-binding protein	Goudswaard LJ	2021	34226637	P15090	SL005086	Adipocyte	2728	0.65	0.09	6.69E-12
Leptin	Goudswaard LJ	2021	34226637	P41159	SL000498	LEP	2728	0.61	0.09	2.81E-11
Fumarylacetoacetase	Goudswaard LJ	2021	34226637	P16930	SL008049	FAAA	2728	0.51	0.11	2.15E-06
Receptor-type tyrosine-protein phosphatase delta	Goudswaard LJ	2021	34226637	P23468	SL008499	PTPRD	2728	-0.49	0.11	4.28E-06
Inhibin beta C chain	Goudswaard LJ	2021	34226637	P55103	SL007288	INHBC	2728	0.45	0.10	1.08E-05
Complement C5	Goudswaard LJ	2021	34226637	P01031	SL000319	C5	2728	0.50	0.11	1.10E-05
Sex hormone-binding globulin	Goudswaard LJ	2021	34226637	P04278	SL005102	SHBG	2728	-0.45	0.10	1.21E-05
PILR alpha-associated neural protein	Goudswaard LJ	2021	34226637	Q8IYJ0	SL019019	PIANP	2728	-0.49	0.11	1.35E-05
Leptin	Zaghlool SB	2021	33627659	P41159	2575-5_5	LEP	992	0.27	0.04	1.32E-12

Protein full name	Author	Year	PMID	UniProtID	SomaID	Protein abr.	N	Beta*	SE*	P_val*
Insulin-like growth factor-binding protein 1	Zaghlool SB	2021	33627659	P08833	2771-35_2	IGFBP1	992	-0.21	0.03	4.72E-10
Insulin-like growth factor-binding protein 2	Zaghlool SB	2021	33627659	P18065	2570-72_5	IGFBP2	992	-0.19	0.03	3.60E-09
Plasminogen activator inhibitor 1	Zaghlool SB	2021	33627659	P05121	2925-9_1	SERPINE1	992	0.17	0.03	2.84E-08
WAP, Kazal, immunoglobulin, Kunitz and NTR domain-containing protein 2	Zaghlool SB	2021	33627659	Q8TEU8	3235-50_2	WFIKKN2	992	-0.17	0.03	5.83E-08
Dickkopf-related protein 3	Zaghlool SB	2021	33627659	Q9UBP4	3607-71_1	DKK3	992	-0.16	0.03	7.85E-07
Galectin-3-binding protein	Zaghlool SB	2021	33627659	Q08380	5000-52_1	LGALS3BP	992	0.15	0.03	1.18E-06
Sex hormone-binding globulin	Zaghlool SB	2021	33627659	P04278	4929-55_1	SHBG	992	-0.16	0.03	1.88E-06
Growth hormone receptor	Zaghlool SB	2021	33627659	P10912	2948-58_2	GHR	992	0.15	0.03	3.38E-06
Growth/differentiation factor 2	Zaghlool SB	2021	33627659	Q9UK05	4880-21_1	GDF2	992	-0.15	0.03	4.35E-06
Netrin receptor UNC5D	Zaghlool SB	2021	33627659	Q6UXZ4	5140-56_3	UNC5D	992	-0.15	0.03	4.63E-06
Neurogenic locus notch homolog protein 1	Zaghlool SB	2021	33627659	P46531	5107-7_2	NOTCH1	992	-0.15	0.03	5.08E-06
Hepatocyte growth factor receptor	Zaghlool SB	2021	33627659	P08581	2837-3_2	MET	992	-0.14	0.03	6.68E-06
Antithrombin-III	Zaghlool SB	2021	33627659	P01008	3344-60_4	SERPINC1	992	-0.15	0.03	6.82E-06

Protein full name	Author	Year	PMID	UniProtID	SomaID	Protein abr.	N	Beta*	SE*	P_val*
C-reactive protein	Zaghlool SB	2021	33627659	P02741	4337-49_2	CRP	992	0.13	0.03	2.99E-05
Neural cell adhesion molecule 1, 120 kDa isoform	Zaghlool SB	2021	33627659	P13591	4498-62_2	NCAM1	992	-0.13	0.03	3.05E-05
Protein jagged-1	Zaghlool SB	2021	33627659	P78504	5092-51_3	JAG1	992	-0.13	0.03	3.45E-05
Cystatin-M	Zaghlool SB	2021	33627659	Q15828	3303-23_2	CST6	992	-0.14	0.03	3.56E-05
Endothelial cell-specific molecule 1	Zaghlool SB	2021	33627659	Q9NQ30	3805-16_2	ESM1	992	-0.13	0.03	4.33E-05

Goudswaard et al (https://doi.org/10.1038/s41366-021-00896-1)
Zaghlool et al (https://doi.org/10.1038/s41467-021-21542-4)
*Effect estimate of BMI on the specified protein
*Duplicate values in the first column indicate different heptamers of the same protein

Appendix 26. *MR results for the proteins with pQTL SNPs in the DVT GWAS.*

Exposure	Gene symbol	Author	No. SNP	MR_method	Log Risk Ratio*	CI (95%)*		SE*	P-value*	Beta - BMI to protein estimate*	Proportion mediated (%)
Neurogenic locus notch homolog protein 1	NOTCH1	Sun BB	1	Wald ratio	0.57	0.45	0.68	0.057	1.12E-23	-0.15	20.778
Plasminogen activator inhibitor 1	SERPINE1	Sun BB	1	Wald ratio	0.42	0.30	0.54	0.061	4.27E-12	0.17	18.557
Inhibin beta C chain	INHBC	Sun BB	1	Wald ratio	-1.18	-2.18	-0.69	0.380	1.96E-03	0.45	133.350
Growth hormone receptor	GHR	Sun BB	1	Wald ratio	0.19	-0.01	0.35	0.092	4.17E-02	0.15	6.922
Endothelial cell-specific molecule 1	ESM1	Sun BB	2	Inverse variance weighted	0.13	-0.03	0.26	0.073	8.00E-02	-0.13	4.175
Antithrombin-III	SERPINC1	Sun BB	1	Wald ratio	0.17	-0.05	0.34	0.101	1.04E-01	-0.15	6.054
Fumarylacetoacetase	FAAA	Sun BB	1	Wald ratio	0.05	-0.01	0.11	0.031	1.19E-01	0.51	6.138
PILR alpha-associated neural protein	PIANP	Sun BB	1	Wald ratio	0.16	-0.07	0.35	0.107	1.31E-01	-0.49	19.775
Dickkopf-related protein 3	DKK3	Sun BB	1	Wald ratio	0.08	-0.03	0.17	0.051	1.32E-01	-0.16	3.078
WAP, Kazal, immunoglobulin, Kunitz and NTR	WFIKKN2	Sun BB	1	Wald ratio	0.03	-0.02	0.08	0.027	2.51E-01	-0.17	1.346

Exposure	Gene symbol	Author	No. SNP	MR_method	Log Risk Ratio*	CI (95%)*		SE*	P-value*	Beta - BMI to protein estimate*	Proportion mediated (%)
domain-containing protein 2											
Cystatin-M	CST6	Sun BB	2	Inverse variance weighted	-0.09	-0.26	0.05	0.081	2.51E-01	-0.14	3.216
Fatty acid binding protein 4	FABP4	Folkersen L	3	Simple mode	0.04	-0.04	0.12	0.040	3.11E-01	0.65	1.292
Neural cell adhesion molecule 1, 120 kDa isoform	NCAM1	Sun BB	3	Simple mode	0.06	-0.07	0.18	0.064	3.25E-01	-0.13	0.912
Neural cell adhesion molecule 1, 120 kDa isoform	NCAM1	Sun BB	3	Weighted mode	0.06	-0.08	0.19	0.070	3.76E-01	-0.13	1.911
Neural cell adhesion molecule 1, 120 kDa isoform	NCAM1	Sun BB	3	Weighted median	0.05	-0.07	0.16	0.057	3.77E-01	-0.13	1.673
Cystatin-M	CST6	Sun BB	1	Wald ratio	-0.08	-0.31	0.10	0.105	4.18E-01	-0.14	2.937
Neural cell adhesion molecule 1, 120 kDa isoform	NCAM1	Sun BB	3	Inverse variance weighted	0.04	-0.06	0.13	0.049	4.29E-01	-0.13	1.275

Exposure	Gene symbol	Author	No. SNP	MR_method	Log Risk Ratio*	CI (95%)*		SE*	P-value*	Beta - BMI to protein estimate*	Proportion mediated (%)
Receptor-type tyrosine-protein phosphatase delta	PTPRD	Sun BB	1	Wald ratio	0.08	-0.17	0.28	0.114	4.94E-01	-0.49	9.568
Fatty acid binding protein 4	FABP4	Folkersen L	3	Inverse variance weighted	0.01	-0.04	0.06	0.026	6.44E-01	0.65	1.918
Fatty acid binding protein 4	FABP4	Folkersen L	3	Weighted median	0.01	-0.05	0.07	0.030	6.59E-01	0.65	2.108
Fatty acid binding protein 4	FABP4	Folkersen L	3	Weighted mode	-0.01	-0.08	0.06	0.033	8.10E-01	0.65	0.742
C-reactive protein	CRP	Sun BB	3	Weighted mode	-0.02	-0.23	0.15	0.096	8.18E-01	0.13	1.301
C-reactive protein	CRP	Sun BB	3	Inverse variance weighted	0.01	-0.12	0.13	0.063	8.34E-01	0.13	0.441
C-reactive protein	CRP	Sun BB	3	Simple mode	-0.01	-0.23	0.17	0.100	9.14E-01	0.13	0.183
C-reactive protein	CRP	Sun BB	3	Weighted median	0.00	-0.15	0.13	0.073	9.96E-01	0.13	0.013

*Log risk ratios per SD increase in circulating protein level

Exposure	Gene symbol	Author	No. SNP	MR_method	Log Risk Ratio*	CI (95%)*	SE*	P-value*	Beta - BMI to protein estimate*	Proportion mediated (%)

*BMI-Protein MR effect estimates from Goudswaard et al (https://doi.org/10.1038/s41366-021-00896-1) and Zaghlool et al (https://doi.org/10.1038/s41467-021-21542-4)

*Multiple-testing corrected P-value threshold: 0.003

Appendix 27. *Reverse MR analysis of DVT on traits which had evidence of an effect on DVT in the MR-PheWAS analysis.*

Outcome	No. SNP	MR method	Beta	SE	P-value	$P_{Het\ (ML)}$	P_{Plt}
Treatment/medication code: warfarin	9	IVW	0.29	0.02	3.81E-32	9.63E-02	5.11E-01
Stearidonate (18:4n3)	5	IVW	1.35	0.50	6.78E-03	9.11E-01	8.41E-01
Leg predicted mass (left)	9	IVW	0.51	0.23	2.73E-02	4.19E-04	6.13E-01
Leg fat-free mass (left)	9	IVW	0.50	0.23	2.86E-02	5.20E-04	6.10E-01
Leg predicted mass (right)	9	IVW	0.47	0.23	4.12E-02	3.95E-04	6.03E-01
Long-standing illness disability or infirmity	9	IVW	0.19	0.10	4.69E-02	2.25E-01	9.50E-01
Leg fat-free mass (right)	9	IVW	0.47	0.23	4.71E-02	3.47E-04	6.09E-01
Taking other prescription medications	9	IVW	0.16	0.10	8.84E-02	7.97E-01	2.34E-01
Varicose veins	9	IVW	0.02	0.01	9.89E-02	7.21E-01	3.19E-01
Eicosapentaenoate (EPA; 20:5n3)	5	IVW	0.58	0.44	1.87E-01	5.75E-01	8.16E-01
Leg fat percentage (left)	9	IVW	-0.35	0.27	1.91E-01	8.77E-07	9.80E-01
Qualifications: None of the above	9	IVW	-0.11	0.09	2.06E-01	1.02E-01	8.73E-01
Varicose veins of lower extremities	9	IVW	0.04	0.03	2.38E-01	2.08E-01	5.93E-01
Weight	9	IVW	0.22	0.25	3.62E-01	1.95E-02	5.78E-01
Leg fat percentage (right)	9	IVW	-0.25	0.28	3.80E-01	3.18E-07	9.90E-01
Hyperthyroidism/thyrotoxicosis	9	IVW	-0.01	0.02	3.89E-01	9.48E-01	6.27E-01
Arm fat percentage (left)	9	IVW	0.35	0.42	3.96E-01	6.65E-12	8.82E-01
Arm fat percentage (right)	9	IVW	0.34	0.41	4.15E-01	1.70E-11	8.55E-01

Outcome	No. SNP	MR method	Beta	SE	P-value	$P_{Het\ (ML)}$	P_{Plt}
Arm fat mass (left)	9	IVW	0.31	0.39	4.21E-01	1.84E-05	7.77E-01
Arachidonate (20:4n6)	5	IVW	0.24	0.30	4.28E-01	8.53E-01	8.94E-01
Hip circumference	9	IVW	0.22	0.29	4.37E-01	8.58E-03	9.32E-01
Basal metabolic rate	9	IVW	0.21	0.28	4.54E-01	1.46E-06	6.68E-01
Whole body water mass	9	IVW	0.20	0.32	5.31E-01	1.38E-09	7.29E-01
Whole body fat-free mass	9	IVW	0.20	0.32	5.42E-01	1.22E-09	7.16E-01
Waist circumference	9	IVW	0.12	0.25	6.24E-01	1.70E-02	9.81E-01
Obesity class 2	5	IVW	1.20	2.53	6.33E-01	7.08E-01	5.15E-01
Arm predicted mass (right)	9	IVW	-0.13	0.32	6.91E-01	1.30E-10	8.85E-01
Overweight	5	IVW	-0.46	1.17	6.95E-01	6.26E-01	8.70E-01
Trunk fat percentage	9	IVW	0.18	0.45	6.98E-01	3.16E-09	9.39E-01
Whole body fat mass	9	IVW	0.13	0.36	7.18E-01	1.43E-04	6.77E-01
Arm fat-free mass (right)	9	IVW	-0.11	0.33	7.30E-01	9.99E-11	7.45E-01
Arm predicted mass (left)	9	IVW	-0.11	0.33	7.47E-01	1.21E-10	8.46E-01
Comparative height size at age 10	9	IVW	0.07	0.25	7.70E-01	2.37E-04	6.67E-01
Treatment/medication code: carbimazole	9	IVW	0.00	0.01	7.83E-01	3.46E-01	9.43E-01
Arm fat-free mass (left)	9	IVW	-0.09	0.32	7.92E-01	5.50E-10	8.24E-01
Mania/bipolar/manic depression	9	IVW	0.00	0.01	8.09E-01	4.14E-01	8.69E-01
Trunk predicted mass	9	IVW	0.07	0.39	8.67E-01	1.33E-15	7.36E-01

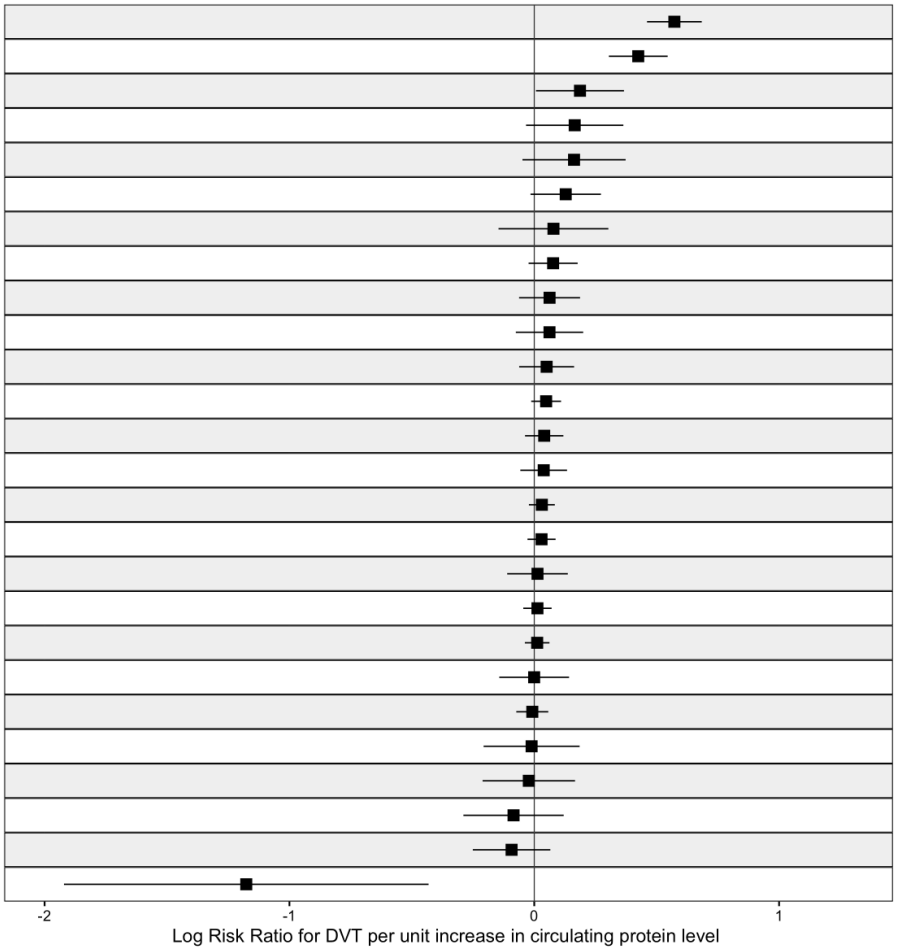
Outcome	No. SNP	MR method	Beta	SE	P-value	$P_{Het\ (ML)}$	P_{Plt}
Leg fat mass (right)	9	IVW	-0.04	0.29	8.84E-01	1.75E-04	8.88E-01
Trunk fat-free mass	9	IVW	0.05	0.39	9.03E-01	1.59E-15	7.25E-01
Body fat percentage	9	IVW	0.04	0.37	9.06E-01	5.73E-09	8.91E-01

*Method: Inverse variance weighted (IVW).

*Beta column represents the effect estimate from the MR analysis of DVT on trait risk

Appendix 28. *Forest plot of MR results for the proteins with pQTL SNPs in the DVT GWAS.*

Protein	GWAS Author	MR method	No.SNP	P-val
NOTCH1	Sun BB	Wald ratio	1	1.12e-23
SERPINE1	Sun BB	Wald ratio	1	4.27e-12
GHR	Sun BB	Wald ratio	1	0.04175
SERPINC1	Sun BB	Wald ratio	1	0.104
PIANP	Sun BB	Wald ratio	1	0.131
ESM1	Sun BB	Inverse variance weighted	2	0.08
PTPRD	Sun BB	Wald ratio	1	0.494
DKK3	Sun BB	Wald ratio	1	0.132
NCAM1	Sun BB	Simple mode	3	0.325
NCAM1	Sun BB	Weighted mode	3	0.376
NCAM1	Sun BB	Weighted median	3	0.377
FAAA	Sun BB	Wald ratio	1	0.119
FABP4	Folkersen L	Simple mode	3	0.311
NCAM1	Sun BB	Inverse variance weighted	3	0.429
WFIKKN2	Sun BB	Wald ratio	1	0.251
WFIKKN2	Sun BB	Wald ratio	1	0.311
CRP	Sun BB	Inverse variance weighted	3	0.834
FABP4	Folkersen L	Weighted median	3	0.659
FABP4	Folkersen L	Inverse variance weighted	3	0.644
CRP	Sun BB	Weighted median	3	0.996
FABP4	Folkersen L	Weighted mode	3	0.81
CRP	Sun BB	Simple mode	3	0.914
CRP	Sun BB	Weighted mode	3	0.818
CST6	Sun BB	Wald ratio	1	0.418
CST6	Sun BB	Inverse variance weighted	2	0.251
INHBC	Sun BB	Wald ratio	1	0.00196



Appendix 29. *Colocalization analysis results for traits with only one SNP as instrument for the MR analysis.*

Trait	nr SNP	PP.H0	PP.H1	PP.H1	PP.H3	PP.H4	PP.S*
Neurogenic locus notch homolog protein 1	3856	1.0694E-79	4.778E-73	2.2382E-07	0.99999972	6.0801E-08	0.00%
Inhibin beta C chain	4079	1.1109E-29	2.6137E-23	4.2502E-07	0.99999948	9.3591E-08	0.00%
Lysine	547	2.4588E-11	0.98338278	3.2772E-13	0.01310352	0.0035137	0.35%
Bipolar disorder / mania	3533	0.47264348	0.43702738	0.03965284	0.03665077	0.01402554	1.40%
Chronic obstructive pulmonary disorder	4229	0.0766326	0.83975957	0.00333097	0.03645779	0.04381907	4.38%
X-14473	655	6.292E-07	0.83967623	6.959E-08	0.09280245	0.06752062	6.75%
Docosapentaenoate	614	1.9077E-08	0.62830917	1.1181E-09	0.03649044	0.33520037	33.50%
Adrenate	626	1.8886E-18	0.5838098	1.1167E-19	0.03413747	0.38205274	38.20%
Stearidonate	674	5.34E-11	0.50441818	3.2335E-12	0.03007888	0.46550294	46.60%
Eicosapentanoate	633	2.8064E-17	0.22721212	1.6606E-18	0.01268473	0.76010315	76.00%
Arachidonate	626	4.9721E-77	0.17796851	2.9399E-78	0.00971061	0.81232088	81.20%
Plasminogen activator inhibitor 1	2604	3.0254E-13	1.9614E-06	3.9637E-09	0.02472248	0.97527556	97.50%

Posterior probabilities for: H0 (no causal variant), H1 (causal variant for trait 1 only), H2 (causal variant for trait 2 only), H3 (two distinct causal variants) and H4 (one common causal variant).

*PP.S is the posterior probability of the genetic variant being causal for the shared signal if H4 is true.

Appendix 30. *The count for each category in the initial dataset of 973 traits.*

Category	N (Count)
Lipid	140
Unknown metabolite	83
Medication	80
Health	77
Behavioural	66
Amino acid	58
Fatty acid	50
Psychiatric / neurological	42
Anthropometric	38
Cardiovascular	33
Autoimmune / inflammatory	29
Cancer	23
Immune cell subset frequency	19
Peptide	15
Education	14
Lung	14
Immune cell-surface protein expression levels	12
Aging	9
Bone	9
Other	9
Diabetes	8
Glycemic	8
Reproductive aging	8
Sleeping	8
Supplement	8
Anthropometric-fat-free	7
Carbohydrate	7
Cofactors and vitamins	7
Hormone	7
Kidney	7
Metal	7
Protein	7
Energy	6
Infection	6
Nucleotide	6
Personality	6
Anthropometric-height	5
Anthropometric-impedance	5
Benign	5
Intelligence	4
Bladder	3
Haematological	3

Geographical	2
Haemotological	2
Immune system	2
Metabolite	2
Metabolites ratio	2
Eye	1
Keto acid	1
Liver	1
Skin	1
Xenobiotics	1

Appendix 31. *First 50 rows (ordered by Benjamini-Hochberg P-value) of the thyrotoxicosis MR-PheWAS on the circulating proteome.*

Exposure	Outcome	Method	No. SNPs	BETA	SE	P-value	Protein ID	P_BH
Thyrotoxicosis	Membrane protein FAM174A || id:prot-a-1034	IVW	13	-13.96	3.91	0.0004	prot-a-1034	0.55
Thyrotoxicosis	Interleukin-21 || id:prot-a-1506	IVW	13	-21.53	6.11	0.0004	prot-a-1506	0.60
Thyrotoxicosis	Secreted frizzled-related protein 3 || id:prot-a-1140	IVW	13	-15.04	4.40	0.0006	prot-a-1140	0.61
Thyrotoxicosis	Chondroadherin || id:prot-a-533	IVW	13	-13.36	4.02	0.0009	prot-a-533	0.66
Thyrotoxicosis	Fibrinogen C domain-containing protein 1 || id:prot-a-1108	IVW	13	-11.25	3.54	0.0015	prot-a-1108	0.81
Thyrotoxicosis	Signal transducer and activator of transcription 3 || id:prot-a-2869	IVW	13	11.31	3.54	0.0014	prot-a-2869	0.81
Thyrotoxicosis	Wnt inhibitory factor 1 || id:prot-a-3230	IVW	13	-11.41	3.61	0.0016	prot-a-3230	0.81
Thyrotoxicosis	Transmembrane protein 87B || id:prot-a-3016	IVW	13	12.32	4.00	0.0021	prot-a-3016	0.96
Thyrotoxicosis	Glycoproteins || id:met-c-862	IVW	10	-5.94	2.04	0.0036	met-c-862	0.99
Thyrotoxicosis	Prothrombin || id:prot-a-1007	IVW	13	-10.39	3.54	0.0034	prot-a-1007	0.99
Thyrotoxicosis	Apolipoprotein E (isoform E2) || id:prot-a-132	IVW	13	-10.21	3.54	0.0039	prot-a-132	0.99
Thyrotoxicosis	Insulin-like growth factor-binding protein 7 || id:prot-a-1451	IVW	13	-10.51	3.54	0.0030	prot-a-1451	0.99
Thyrotoxicosis	Interleukin-22 receptor subunit alpha-1 || id:prot-a-1509	IVW	13	10.88	3.62	0.0027	prot-a-1509	0.99

Thyrotoxicosis	Laminin subunit alpha-4 || id:prot-a-1696	IVW	13	10.86	3.71	0.0034	prot-a-1696	0.99
Thyrotoxicosis	Protocadherin alpha-7 || id:prot-a-2200	IVW	13	10.26	3.54	0.0038	prot-a-2200	0.99
Thyrotoxicosis	Periostin || id:prot-a-2332	IVW	13	-10.22	3.54	0.0039	prot-a-2332	0.99
Thyrotoxicosis	Spondin-1 || id:prot-a-2829	IVW	13	-11.14	3.76	0.0030	prot-a-2829	0.99
Thyrotoxicosis	Angiopoietin-2 || id:prot-a-94	IVW	13	-12.31	4.15	0.0030	prot-a-94	0.99
Thyrotoxicosis	Leukocyte immunoglobulin-like receptor subfamily B member 4 || id:prot-a-1746	IVW	13	10.14	3.54	0.0042	prot-a-1746	1.00
Thyrotoxicosis	Albumin || id:met-c-841	IVW	10	0.36	2.15	0.8666	met-c-841	1.00
Thyrotoxicosis	Apolipoprotein A-I || id:met-c-842	IVW	10	0.68	2.19	0.7559	met-c-842	1.00
Thyrotoxicosis	Apolipoprotein B || id:met-c-843	IVW	10	2.57	2.56	0.3148	met-c-843	1.00
Thyrotoxicosis	Glycoprotein acetyls || id:met-c-863	IVW	10	1.80	2.06	0.3830	met-c-863	1.00
Thyrotoxicosis	APOBEC1 complementation factor || id:prot-a-1	IVW	13	3.57	3.54	0.3133	prot-a-1	1.00
Thyrotoxicosis	Histo-blood group ABO system transferase || id:prot-a-10	IVW	13	-2.24	3.54	0.5278	prot-a-10	1.00
Thyrotoxicosis	Angiopoietin-related protein 7 || id:prot-a-100	IVW	13	4.97	3.54	0.1606	prot-a-100	1.00
Thyrotoxicosis	RNA-binding protein EWS || id:prot-a-1000	IVW	13	1.76	3.54	0.6195	prot-a-1000	1.00
Thyrotoxicosis	Exosome complex component CSL4 || id:prot-a-1001	IVW	13	2.23	3.54	0.5292	prot-a-1001	1.00

Thyrotoxicosis	Exosome complex component RRP40 || id:prot-a-1002	IVW	13	-1.13	3.54	0.7495	prot-a-1002	1.00
Thyrotoxicosis	Exostosin-like 2 || id:prot-a-1003	IVW	13	-1.47	3.80	0.6995	prot-a-1003	1.00
Thyrotoxicosis	Ezrin || id:prot-a-1004	IVW	13	4.30	3.63	0.2365	prot-a-1004	1.00
Thyrotoxicosis	Coagulation factor Xa || id:prot-a-1005	IVW	13	4.25	4.25	0.3175	prot-a-1005	1.00
Thyrotoxicosis	Coagulation Factor X || id:prot-a-1006	IVW	13	5.49	4.05	0.1745	prot-a-1006	1.00
Thyrotoxicosis	Tissue Factor || id:prot-a-1008	IVW	13	-5.40	3.54	0.1275	prot-a-1008	1.00
Thyrotoxicosis	Coagulation Factor VIII || id:prot-a-1009	IVW	13	3.78	4.14	0.3608	prot-a-1009	1.00
Thyrotoxicosis	Ankyrin-2 || id:prot-a-101	IVW	13	4.10	4.49	0.3614	prot-a-101	1.00
Thyrotoxicosis	Fatty-acid amide hydrolase 2 || id:prot-a-1010	IVW	13	0.83	3.54	0.8152	prot-a-1010	1.00
Thyrotoxicosis	Fatty acid-binding protein, liver || id:prot-a-1011	IVW	13	-1.07	3.54	0.7625	prot-a-1011	1.00
Thyrotoxicosis	Fatty acid-binding protein, heart || id:prot-a-1012	IVW	13	2.97	3.54	0.4019	prot-a-1012	1.00
Thyrotoxicosis	Fatty acid-binding protein, adipocyte || id:prot-a-1013	IVW	13	4.40	3.54	0.2146	prot-a-1013	1.00
Thyrotoxicosis	Fatty acid-binding protein, epidermal || id:prot-a-1014	IVW	13	-1.33	3.54	0.7075	prot-a-1014	1.00
Thyrotoxicosis	FAS-associated factor 2 || id:prot-a-1015	IVW	13	-0.45	4.13	0.9122	prot-a-1015	1.00
Thyrotoxicosis	FAS-associated factor 2 || id:prot-a-1016	IVW	13	4.36	3.54	0.2185	prot-a-1016	1.00

Thyrotoxicosis	FAS-associated factor 2 || id:prot-a-1017	IVW	13	7.40	3.54	0.0367	prot-a-1017	1.00
Thyrotoxicosis	Fumarylacetoacetase || id:prot-a-1018	IVW	13	-3.36	3.54	0.3433	prot-a-1018	1.00
Thyrotoxicosis	Fas apoptotic inhibitory molecule 3 || id:prot-a-1019	IVW	13	-7.36	4.01	0.0669	prot-a-1019	1.00
Thyrotoxicosis	Ankyrin repeat domain-containing protein 27 || id:prot-a-102	IVW	13	0.84	4.66	0.8571	prot-a-102	1.00
Thyrotoxicosis	Protein FAM107A || id:prot-a-1020	IVW	13	-9.65	3.54	0.0064	prot-a-1020	1.00
Thyrotoxicosis	Protein FAM107B || id:prot-a-1021	IVW	13	1.37	4.61	0.7666	prot-a-1021	1.00
Thyrotoxicosis	Protein FAM107B || id:prot-a-1022	IVW	13	4.45	3.54	0.2093	prot-a-1022	1.00