UNIVERSITY OF IOANNINA
SCHOOL OF HEALTH SCIENCES
FACULTY OF MEDICINE

SECTOR OF SOCIAL MEDICINE AND MENTAL HEALTH
DEPARTMENT OF HYGIENE AND EPIDEMIOLOGY

MULTIVARIATE ANALYSIS OF GENETIC DATA ON COMPLEX DISEASES AND PHENOTYPES

FOTIOS KOSKERIDIS

Thesis submitted for the degree of
DOCTOR OF PHILOSOPHY

IOANNINA 2024

ΠΑΝΕΠΙΣΤΗΜΙΟ ΙΩΑΝΝΙΝΩΝ
ΣΧΟΛΗ ΕΠΙΣΤΗΜΩΝ ΥΓΕΙΑΣ
ΤΜΗΜΑ ΙΑΤΡΙΚΗΣ

ΤΟΜΕΑΣ ΚΟΙΝΩΝΙΚΗΣ ΙΑΤΡΙΚΗΣ ΚΑΙ ΨΥΧΙΚΗΣ ΥΓΕΙΑΣ
ΕΡΓΑΣΤΗΡΙΟ ΥΓΙΕΙΝΗΣ ΚΑΙ ΕΠΙΔΗΜΙΟΛΟΓΙΑΣ

ΠΟΛΥΜΕΤΑΒΛΗΤΗ ΑΝΑΛΥΣΗ ΓΕΝΕΤΙΚΩΝ ΔΕΔΟΜΕΝΩΝ ΣΕ ΣΥΝΘΕΤΑ ΝΟΣΗΜΑΤΑ ΚΑΙ ΦΑΙΝΟΤΥΠΟΥΣ

ΦΩΤΙΟΣ ΚΟΣΚΕΡΙΔΗΣ

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

ΙΩΑΝΝΙΝΑ 2024

Η έγκριση της διδακτορικής διατριβής από το Τμήμα Ιατρικής του Πανεπιστημίου Ιωαννίνων δεν υποδηλώνει αποδοχή των γνωμών του συγγραφέα Ν. 5343/32, άρθρο 202, παράγραφος 2 (νομική κατοχύρωση του Ιατρικού Τμήματος).

Ημερομηνία αίτησης του κ. Κοσκερίδη Φώτιου: 17-10-2018

Ημερομηνία ορισμού Τριμελούς Συμβουλευτικής Επιτροπής: Γ.Σ. αριθμ. 876α/18-12-2018

Μέλη Τριμελούς Συμβουλευτικής Επιτροπής:
Επιβλέπων:
Ευαγγέλου Ευάγγελος, Επίκουρος Καθηγητής Υγιεινής με έμφαση στην Κλινική και Μοριακή Επιδημολογία του Τμήματος Ιατρικής του Πανεπιστημίου Ιωαννίνων
Μέλη:
Ντζάνη Ευαγγελία, Αναπληρώτρια Καθηγήτρια Υγιεινής με ιδιαίτερη έμφαση στην Επιδημιολογία του Τμήματος Ιατρικής του Πανεπιστημίου Ιωαννίνων
Τσιλίδης Κωνσταντίνος, Επίκουρος Καθηγητής Υγιεινής με έμφαση στην Επιδημιολογία του Τμήματος Ιατρικής του Πανεπιστημίου Ιωαννίνων
Ημερομηνία ορισμού θέματος: 25-01-2019

«Πολυμεταβλητή ανάλυση γενετικών δεδομένων σε σύνθετα νοσήματα»

«Multivariate analysis of genetic data on complex diseases»

Ανασυγκρότηση Τριμελούς Συμβουλευτικής Επιτροπής: Γ.Σ. αριθμ. 1075α/26-09-2023
Επιβλέπων:

Τσιλίδης Κωνσταντίνος, Αναπληρωτής Καθηγητής Υγιεινής με έμφαση στην Επιδημιολογία του Τμήματος Ιατρικής του Πανεπιστημίου Ιωαννίνων

Μέλη:

Ντζάνη Ευαγγελία, Καθηγήτρια Υγιεινής με ιδιαίτερη έμφαση στην Επιδημιολογία του Τμήματος Ιατρικής του Πανεπιστημίου Ιωαννίνων

Τζουλάκη Ιωάννα, Διευθύντρια Ερευνών Ιδρύματος Ιατροβιολογικών Ερευνών Ακαδημίας Αθηνών

Μερική τροποποίηση τίτλου θέματος διδακτορικής διατριβής: Γ.Σ. αριθμ. 1123α/26-09-2023

«Πολυμεταβλητή ανάλυση γενετικών δεδομένων σε σύνθετα νοσήματα και φαινότυπους»

«Multivariate analysis of genetic data on complex diseases and phenotypes»

ΟΡΙΣΜΟΣ ΕΠΤΑΜΕΛΟΥΣ ΕΞΕΤΑΣΤΙΚΗΣ ΕΠΙΤΡΟΠΗΣ: 1125α/03-12-2024

1. Τσιλίδης Κωνσταντίνος, Καθηγητής Υγιεινής με έμφαση στην Επιδημιολογία του Τμήματος Ιατρικής του Πανεπιστημίου Ιωαννίνων
2. Ντζάνη Ευαγγελία, Καθηγήτρια Υγιεινής με ιδιαίτερη έμφαση στην Επιδημιολογία του Τμήματος Ιατρικής του Πανεπιστημίου Ιωαννίνων
3. Τζουλάκη Ιωάννα, Διευθύντρια Ερευνών, Ίδρυμα Ιατροβιολογικών Ερευνών Ακαδημίας Αθηνών
4. Χριστοφορίδης Σάββας, Καθηγητής Βιολογικής Χημείας του Τμήματος Ιατρικής του Πανεπιστημίου Ιωαννίνων
5. Μαυρίδης Δημήτριος, Αναπληρωτής Καθηγητής Στατιστικής στην Εκπαίδευση του Παιδαγωγικού Τμήματος Δημοτικής Εκπαίδευσης του Πανεπιστημίου Ιωαννίνων
6. Θεοδωράτου Ευρώπη, Professor of Cancer Epidemiology and Global Health, The University of Edinburgh
7. Μπάγκος Παντελής, Καθηγητής Βιοπληροφορικής και Βιοστατιστικής του Τμήματος Πληροφορικής με Εφαρμογές στη Βιοϊατρική του Πανεπιστημίου Θεσσαλίας

Έγκριση Διδακτορικής Διατριβής με βαθμό **«ΑΡΙΣΤΑ»** στις **17-12-2024**

Ιωάννινα 09-05-2025
ΠΡΟΕΔΡΟΣ ΤΟΥ ΤΜΗΜΑΤΟΣ ΙΑΤΡΙΚΗΣ
Σπυρίδων Κονιτσιώτης
Καθηγητής Νευρολογίας

ΙΑΤΡΙΚΗ ΣΧΟΛΗ ΠΑΝΕΠΙΣΤΗΜΙΟΥ ΙΩΑΝΝΙΝΩΝ 1977

Η Γραμματέας του Τμήματος
Αικατερίνη Λαγού

To the memory of

Dr Evangelos Evangelou

ACKNOWLEDGEMENTS

First and foremost, I wish to honour the memory of my late supervisor, Dr Evangelos Evangelou, whose wisdom and guidance have been instrumental in shaping this work and my growth as a researcher.

I am also deeply grateful to Dr Konstantinos Tsilidis, who graciously accepted to substitute for my late supervisor during a challenging time.

Sincere thanks to the members of my Three-member Advisory Committee, Dr Evangelia Ntzani, and Dr Ioanna Tzoulaki, for their valuable insights, critical evaluations, and continuous encouragement throughout this process. I am especially grateful to Dr Tzoulaki for a constructive collaboration on several projects, many of which are part of this thesis, for co-supervising my thesis and for her motivating support during its preparation.

I also extend my thanks to the members of the Seven-member Examining Committee, Dr Savvas Christoforidis, Dr Dimitrios Mavridis, Dr Evropi Theodoratou, and Dr Pantelis Bagos, for their insightful comments and valuable feedback.

To my colleagues and lab mates, their camaraderie and knowledge-sharing have been invaluable.

To my family and friends, their unwavering support and belief in me have been a constant source of strength and motivation.

PREFACE

Understanding the genetics behind complex diseases is fundamental in modern research. It offers a chance to untangle how genetics and environment interact to shape disease risk. Most conditions usually arise from the interplay between genetic predispositions and environmental or lifestyle factors. Traditional epidemiology has provided key insights into disease risk, but genetic epidemiology has taken this further by revealing the underlying biological mechanisms. Genetic epidemiological studies have identified thousands of genes for complex traits. However, translating these discoveries to clinical benefits is not always an easy task. The fact that these traits are highly polygenic and are affected by several sources of small genetic effects which can interact with external factors makes the discovery of the involved pathways and the understanding of their role very challenging.

Many biological pathways of multiple conditions are often interconnected. However, most genetic studies focus on single traits. This one-dimensional (univariate) approach might neglect these interconnections between phenotypes and overlook critical insights into their shared underlying mechanisms. Therefore, more innovative approaches are needed that combine genetic information from multiple traits into a single framework.

Multivariate methods can explore the shared genetic architecture between multiple traits, offering a more comprehensive view of their connections. These techniques allow to identify new genetic risk factors and shared pathways, improving the understanding of disease pathophysiologies and genetic links between phenotypes.

This thesis investigates multivariate analytical methods to explore the genetic structure of complex phenotypes. It begins with an introduction to the genetic epidemiology of complex traits and the rationale behind multivariate approaches. It then reviews existing multi-trait methods and applies the most suitable of them to real data on various diseases and other phenotypes to investigate genetic effects. The aim is to deepen the understanding of the genetic basis for these conditions and contribute to more effective preventive and therapeutic strategies.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ACAT	Aggregated Cauchy Association Test
AD	Alzheimer's Disease
ADHD	Attention deficit hyperactivity disorder
AF	Atrial Fibrillation
AIC	Akaike Information Criterion
ANOVA	Analysis of Variance
ASD	Autism spectrum disorder
BAS	Basophil counts
BD	Bipolar disease
BF	Bayes Factor
BMI	Body Mass Index
BP	Blood Pressure
BrCa	Breast cancer
CAD	Coronary Artery Disease
CCA	Canonical Correlation Analysis
cIMT	Carotid Intima-Media Thickness
CPD	Cigarettes per Day
CRP	C-Reactive Protein
CV	Cardiovascular
CVD	Cardiovascular Diseases
DBP	Diastolic Blood Pressure
DR	Dimension Reduction
EHR	Electronic Health Records
EOS	Eosinophil count
eQTL	Expression Quantitative Trait Loci
FDR	False Discovery Rate
FUMA	Functional Mapping and Annotation of GWAS
GEE	Generalised Estimating Equations
GLMM	Generalised Linear Mixed-effects Models
Gluc	Fasting glucose
GTEx	Genotype-Tissue Expression
GWAS	Genome-Wide Association Study
GWS	Genome-Wide Significance/Significant
HCT	Haematocrit
HDL	High-Density Lipoproteins
HEIDI	Heterogeneity In Dependent Instruments
HES	Hospital Episode Statistics
HGB	Haemoglobin
HWE	Hardy-Weinberg Equilibrium
IBD	Identity By Descent
IBS	Identity By State
ICD	International Classification of Diseases
Insl	Fasting insulin
IVW	Inverse Variance Weighted
LD	Linkage Disequilibrium
LDL	Low-Density Lipoproteins
LDSC	Linkage Disequilibrium Score regression
LMM	Linear Mixed-effects Models
LYM	Lymphocyte count
MAF	Minor Allele Frequency
MANOVA	Multivariate Analysis of Variance
MCH	Mean Haematocrit
MCHC	Mean Corpuscular Haemoglobin Concentration
MCV	Mean Corpuscular Volume
MDD	Major depressive disorder
MI	Myocardial infarction

MLE	Maximum Likelihood Estimate
MON	Monocyte count
MPV	Mean Platelet Volume
MR	Mendelian Randomisation
MR2	Multi-Response Mendelian Randomisation
MTAG	Multi-Trait Analysis of GWAS
MVMR	Multi-Variable Mendelian Randomisation
NEU	Neutrophil count
OR	Odds Ratio
PCA	Principal Component Analysis
PheWAS	Phenome-Wide Association Study
PLT	Platelet count
PP	Posterior Probability
PrCa	Prostate Cancer
PRS	Polygenic Risk Score
RBC	Red Blood Cell count
RDW	Red Distribution Width
SBP	Systolic Blood Pressure
Schz	Schizophrenia
SEM	Structural Equation Modelling
SMR	Summary-data-based Mendelian Randomisation
SNP	Single-nucleotide polymorphism
SSU	Sum of Squared score
T2D	Type 2 diabetes
TDT	Transmission Disequilibrium Test
TG	Triglycerides
UKB	UK Biobank
UTSC	Univariate Test Statistics Combination
WBC	White Blooc Cell count
WC	Waist circumference
WHO	World Health Organisation
WHR	Waist-to-hip ratio
WM	Weighted Median

GENERAL PART – INTRODUCTION

CHAPTER 1

INTRODUCTION TO GENETIC EPIDEMIOLOGY

Genetic epidemiology is the scientific field that combines the principles of genetics and epidemiology aiming to study the genetic determinants of human health. In contrast with traditional epidemiology, which studies non-genetic factors, such as environmental and lifestyle, it focuses on genetic variations and their role in the pathophysiology of diseases (Duncan, 2004).

1.1 Overview of Genetic Epidemiology

Genetic epidemiology arose as a distinct field in the 1980s. It brought together methodologies from genetics, epidemiology, biostatistics, and bioinformatics under the need to investigate the human genetic basis and its impact on the overall health (Morton and Chung, 1978, Evangelou, 2018).

Its initial focused was restricted to the study of Mendelian (monogenic) diseases, where a single gene is the main causal factor of the disease (Zheng et al., 2012). Therefore, family-based studies were designed (see section 3.1.1) to track the inheritance through family trees in order to identify the causal genes that affect the diseases (Duncan, 2004).

After the completion of the Human Genome Project along with other advancements in genomic technologies, genetic epidemiology was able to expand its focus to complex phenotypes, which are influenced by multiple factors (Lander et al., 2001, Zheng et al., 2012). As a result, new methodologies were developed leading to the discovery of thousands of novel genes. Nevertheless, the study of complex traits is still a very challenging task considering the influence of multiple genetic variants of smaller effect, which are not easy to discover, that interact each other and with external factors (Balding et al., 2007).

Finally, genetic epidemiology plays a crucial role in precision (personalised) medicine. In this innovative field, based on individual's genetic and molecular profiles, therapy

and prevention strategies can be adjusted to each patient's needs optimising the safest treatment scheme for the patient (Dawn Teare, 2011).

1.2 Population Genetics and its role in Genetic Epidemiology

Population genetics consists of the basis of genetic epidemiology providing the essential theoretical foundation to understand how genetic variation affects phenotypes. This field investigates how evolutionary forces, like natural selection and mutations, cause changes to the distribution and frequency of genetic variants within a population over time (Kar et al., 2022). Below are discussed a few basic principles of population genetics.

1.2.1 Allele Frequency

Allele frequency reflects the genetic diversity within populations and refers to the proportion of an allele at a genetic locus within a population. The most frequent allele at a locus is called major allele, while the least frequent is called minor allele and its frequency minor allele frequency (MAF). The study of allele frequency is vital in genetic association studies. Common diseases are assumed to be influenced mainly by common genetic variants (MAF > 0.05), otherwise, they wouldn't be common in a population. Nevertheless, the role or rare variants (MAF < 0.01) in common diseases cannot be excluded either (Evangelou, 2018, Zheng et al., 2012). Allele frequencies can be affected due to genetic drift, migration, selection, and mutation. They can be used to understand population structure or minimise potential confounding factors such as population stratification. For instance, certain alleles may be rare in the general population but common in specific ethnic groups or isolated populations (Evangelou, 2018).

1.2.2 Hardy-Weinberg Equilibrium

Hardy-Weinberg equilibrium (HWE) examines weather allele and genotype frequencies in a population remain constant from generation to generation under the absence of evolutionary influences, such as selection, mutation, or migration. Deviations from HWE can indicate selection bias, non-random mating, or population stratification (Balding et al., 2007, Evangelou, 2018).

For a biallelic variant with alleles A and a and frequencies in the population p and q, respectively, there are three possible genotypes: AA, Aa and aa. Then, the genotype equilibrium frequencies for the variant can be calculated by the algebraic equation:

$$(p+q)^2 = p^2 + 2pq + q^2 \qquad (1.1)$$

The HWE equation (1.1) provides the expected frequencies of the alleles for the next generation. **Figure 1.1** is a schematic overview of the above equation. It shows how the alleles of a genetic variant are inherited from one generation to the next one. A deviation from HWE test is performed using a Pearson's chi-squared hypothesis test comparing the expected allele frequencies with the observed ones calculated from data.

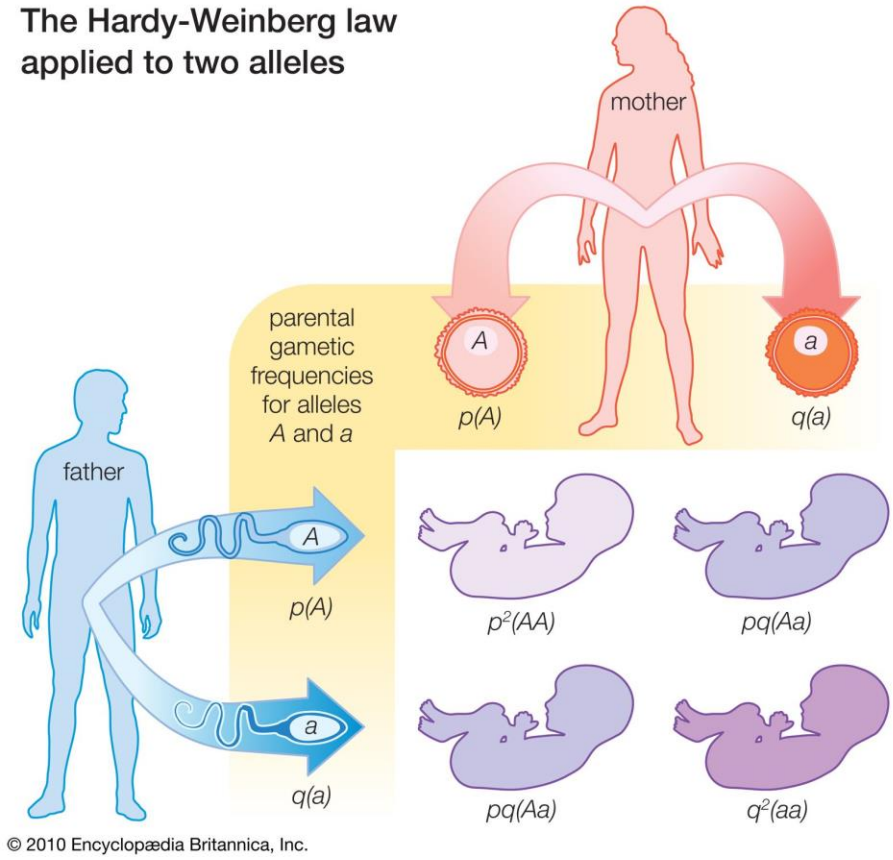


Figure 1.1 The Hardy-Weinberg law for a variant of two alleles

Source: https://www.britannica.com/science/Hardy-Weinberg-law

1.2.3 Linkage Disequilibrium

Linkage disequilibrium (LD) refers to the non-random allocation of alleles at different genetic loci within a population. It occurs when alleles within the same or neighbouring

regions are inherited together more frequently than expected by chance (haplotypes or LD blocks), often due to limited recombination events that prevent these alleles from segregating independently (Dawn Teare, 2011). Recombination events within families gradually disintegrate chromosomal segments across generations and in a population basis, their effect is amplified tending to break apart the segments until all alleles become independent (linkage equilibrium) (**Figure 1.2**) (Bush and Moore, 2012). However, this is not possible to happen due to natural selection mechanisms and LD blocks can be formed eventually even under a uniform recombination rate (Wang et al., 2002).

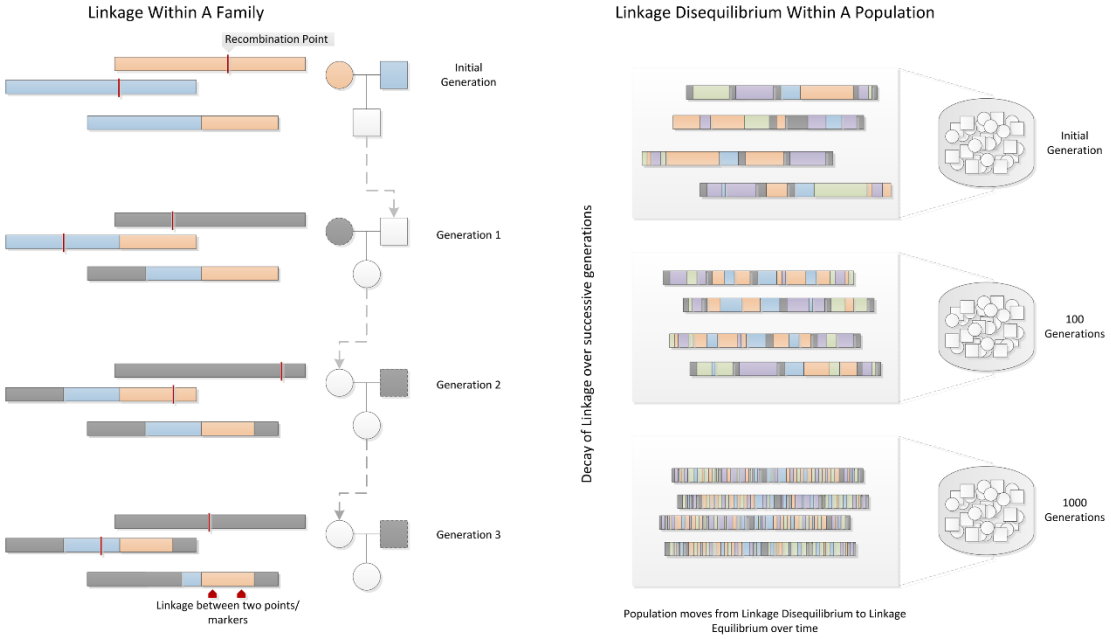


Figure 1.2 Linkage within a family (left) and linkage disequilibrium within a population (right)

Source: (Bush and Moore, 2012)

Several measures of pairwise LD have been suggested, but all of them are related to the LD coefficient D:

$$D = p_{AB} - p_A p_B \tag{1.2}$$

where p_{AB} is the frequency of haplotype AB (a genotype with alleles A and B at two loci), while p_A and p_B are the frequencies of alleles A and alleles B, respectively. When D = 0 means that there is a linkage equilibrium, which is similar to HWE as it indicates

that the alleles are randomly associated (Slatkin, 2008). The magnitude of D represents the degree of LD, where positive values mean that the haplotype is more frequent than expected and negative values that it is less frequent than expected.

However, D is not often the optimal measurement because of its infinite range, which is determined by the allele frequencies. Hence, other measurements are preferred instead, such as D′ and r^2, which help quantify the degree of association between alleles within a more limited and comparable range of values. D′ reflects the proportion of the maximum possible disequilibrium observed:

$$D' = \frac{D}{D_{max}} \tag{1.3}$$

where $$D_{max} = \begin{cases} min\left(p_A p_B, (1-p_A)(1-p_B)\right), & when\ D < 0 \\ min\left(p_A(1-p_B), p_B(1-p_A)\right), & when\ D > 0 \end{cases}$$

and it ranges $-1 \leq D' \leq 1$ with the zero indicating independence of the alleles, positive values that the alleles are observed more frequently than expected and negative values that the alleles are observed less frequently than expected (Lewontin, 1964).

The other commonly used LD measurement, r^2, provides a correlation coefficient that is more interpretable for association studies and can be calculated by the equation:

$$r^2 = \frac{D^2}{p_A(1-p_A)p_B(1-p_B)} \tag{1.4}$$

It ranges from 0 to 1 with values close to 0 indicating independence while values close to 1 high LD. Thus, information for one variant's allele can accurately indicate the allele for other variants in high LD, enabling more cost-efficient genotyping using imputation techniques. However, it might complicate the identification of true causal variants leading to multiple false positive associations (Evangelou, 2018, Balding et al., 2007).

LD allows to map genetic markers to close genes and infer their associations with phenotypic outcomes. It also allows to test fewer markers while capturing broader genomic information, but in return, it requires careful interpretation to distinguish true signals from spurious associations caused by LD with nearby variants (Zheng et al., 2012). LD patterns vary across populations and can be influenced by several factors such as genetic drift, mutation, and recombination rates (Duncan, 2004).

1.2.4 Identity by Descent and Identity by State

Identity by descent (IBD) refers to alleles inherited from a common ancestor, while **identity by state (IBS)** to identical in allele sequences, which are not necessarily inherited by the same ancestor. IBD and IBS can be used to detect genetic relationships between individuals, such as kinship coefficients, which measure the probability that two individuals share alleles IBD, and therefore, can assess the genetic relatedness within populations (Evangelou, 2018, Duncan, 2004).

1.3 Importance of Genetic Epidemiology in Public Health

Genetic epidemiology has evolved into an irreplaceable pillar for public health. The field has witnessed a significant growth over the past years allowing to conduct studies of bigger and bigger sample sizes, where millions of genetic variants are examined with a plethora of phenotypes, which in turn results in an increasing publication rate (**Figure 1.3**).

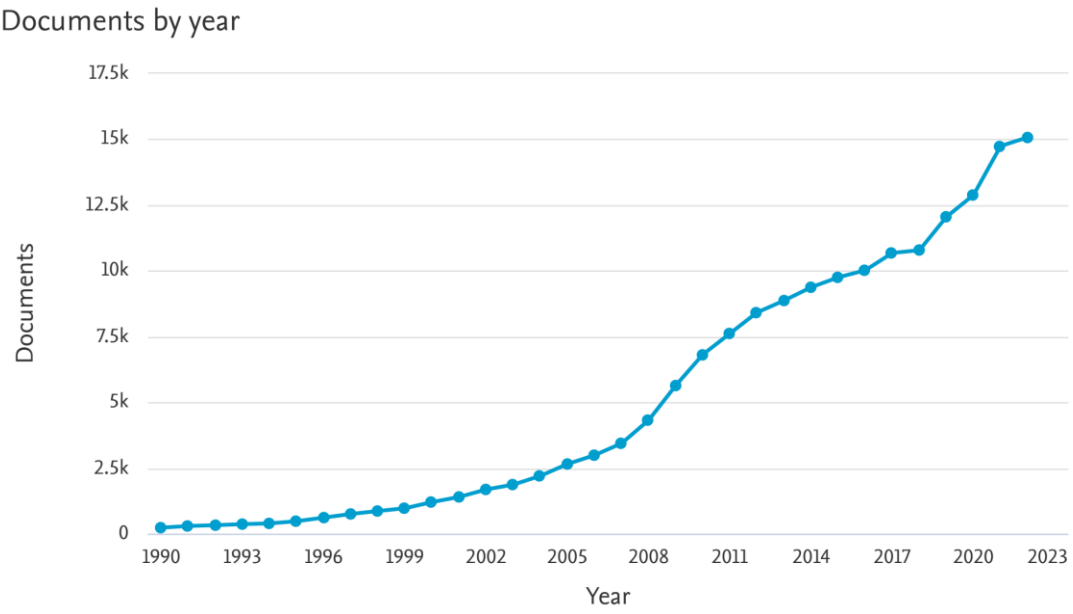


Figure 1.3 Publications related to genetic association studies over time

Source: www.scopus.com

Genetic predisposition holds a significant role for various diseases. A detrimental genetic effect can increase the disease burden, worsen quality of life and eventually

reduce healthy lifespan (**Figure 1.4**). Thus, genetic epidemiology is, beyond doubt, a crucial field to public health contributing to better understand disease aetiologies, pushing forward the development of targeted interventions and improved treatment strategies (Manolio et al., 2009, Visscher et al., 2017). Additionally, it is also determining to the development of precision medicine. Individuals with high genetic predisposition for a disease can benefit in a way that limit the disease risk after implementing more accurate and early interventions (Collins et al., 2003, Khera and Kathiresan, 2017, Torkamani et al., 2018). Finally, it informs public health policy to develop more targeted interventions based on population-level risk factors (Khoury et al., 2016, Manolio and Collins, 2009).

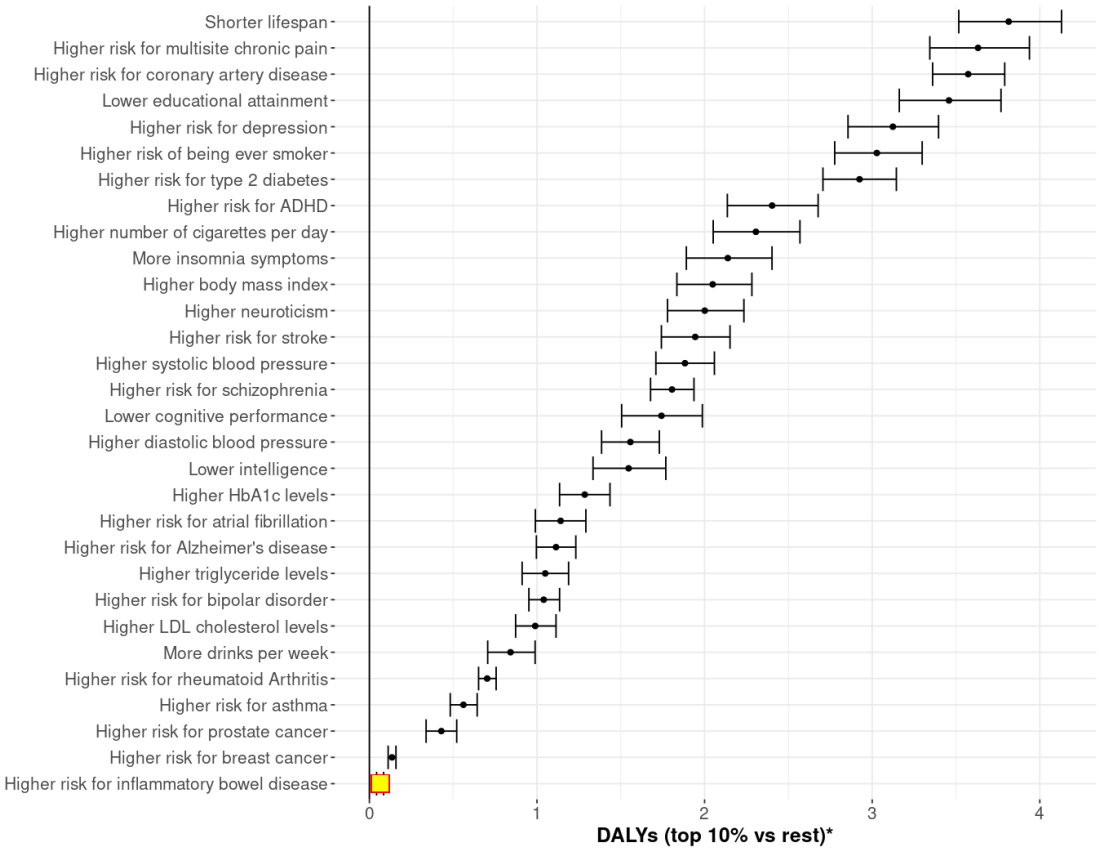


Figure 1.4 The impact of different polygenic scores on public health

*DALYs Disability-Adjusted Life Years. Expected loss of healthy life years for individuals in the top 10% of the score compared to the rest 90%.

Source: Data Science – Genetic Epidemiology Lab, Institute for Molecular Medicine Finland, HiLIFE, University of Helsinki and Massachusetts General Hospital, Harvard Medical School (https://www.dsgelab.org)

CHAPTER 2

INTRODUCTION TO COMPLEX PHENOTYPES

Complex traits are phenotypes that correspond to biological features or health conditions. They are influenced by the combined effects of multiple genetic and environmental factors. Thus, it is very challenging to predict their development and progression. Yet, they are often responsible for a significant amount of the disease burden in populations.

Their development involves various biological mechanisms, including metabolic pathways, regulatory systems, and cellular signalling, but it is also influenced by non-genetic factors like age, sex, diet, lifestyle, and other environmental exposures. Each causal factor may effect on the outcome either directly or through interactions with other factors (Rowe and Tenesa, 2012).

2.1 Genetic Architecture of Complex Phenotypes

The genetic structure of complex traits is very challenging and cannot effectively be addressed by traditional epidemiological approaches. It requires the consideration of the combined effects of many genetic variants, each contributing a small amount to the overall phenotype (Lander, 2011, McCarthy et al., 2008). This section summarises some basic parameters of the genetic architecture of complex traits.

2.1.1 Polygenicity

During the last years, previous genetic epidemiological studies have discovered thousands of genes affecting complex traits. However, there is still a significant proportion of phenotypic variance that remains unexplained. Complex traits are characterised by **polygenicity**, a phenomenon where the final shape of a phenotype is controlled by the effect of many genetic variants (Visscher et al., 2017). These variants might be located in non-coding regions of the genome regulating gene expressions rather than directly coding for proteins (Maurano et al., 2012). The polygenic nature of complex traits makes their prediction and study very challenging (Plomin et al., 2016).

The higher the polygenicity of a phenotype, the higher the number of the susceptibility genetic variants that are associated with this phenotype and at the same time, the lower the effect size of these variants. This practically means the more polygenic a trait is, the higher sample sizes are needed (**Figure 2.1**) (Matoba et al., 2022).

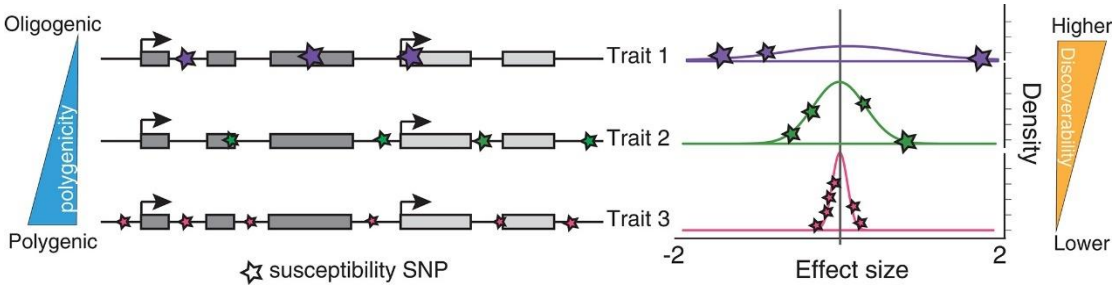


Figure 2.1 Connection of polygenicity and discoverability of genetic variants

SNP Single-Nucleotide Polymorphism

Source: (Matoba et al., 2022)

2.1.2 Inheritance, Penetrance and Genetic Effect Models

Inheritance refers to the process where the genetic information is transmitted from parents to their offspring following the principles of Mendelian genetics (Griffiths, 2000).

Penetrance concerns the probability that a person carrying a specific genetic variant will express the associated phenotype. It provides information about the phenotypic expression, as even individuals with the same genetic variant might not always exhibit the phenotype (Dawn Teare, 2011).

Genetic effect models refer to the modes of inheritance that describe how different alleles at a genetic locus can contribute to the overall phenotype (Zheng et al., 2012). The most common models include:

1. **Additive**: the effects of the two alleles are summed.
2. **Dominant**: the presence of one allele (dominant) is sufficient to express the trait, regardless of the second allele.
3. **Recessive**: two copies of the allele (recessive) are required for the trait to be expressed.

4. **Multiplicative**: the two alleles interact each other and their effect is multiplied (Evangelou, 2018).

2.1.3 Heritability

Heritability measures the proportion of phenotypic variation in a population that is attributed to genetic differences among individuals. In other words, it quantifies the genetic contribution to the phenotypes. It can be distinguished to broad-sense heritability (H^2), which measures the overall genetic effect, and narrow-sense heritability (h^2), which focuses on additive genetic effects only (Evangelou, 2018, Dawn Teare, 2011, Griffiths, 2000).

Heritability is not a fixed characteristic and varies across populations (Balding et al., 2007). It is calculated as the ratio of genetic to total phenotypic variance for continuous traits, while for binary traits (e.g., diseases), its calculation is more challenging due to more complex modelling and requires careful interpretation (Griffiths, 2000).

2.1.4 Gene-Gene and Gene-Environment Interaction

Gene-gene interactions (epistasis) occur when the effect of one gene on a phenotype is influenced by the presence of another (Evangelou, 2018). On the other hand, **gene-environment interactions** refer to the phenomenon where the genetic effect on a phenotype is influenced by environmental factors. In other words, the phenotypic expression of genetic predispositions can be modified, either enhanced or suppressed by external influences (Hunter, 2005). The study of these interactions provides insights into understanding the role of genetic mechanisms on the phenotypic expression under the presence or absence of other genetic or environmental exposures (Evangelou, 2018).

2.2 Genetic Pleiotropy and its Impact on Complex Traits

Genetic pleiotropy refers to the phenomenon where a single genetic locus influences multiple phenotypic traits. The effects of a single genetic variant may manifest across different traits, making it difficult to isolate its specific contribution to a particular trait. Pleiotropic genes may be involved in multiple pathways, increasing the complexity of their roles in disease aetiology (Solovieff et al., 2013).

Pleiotropy can be observed at allelic or gene level and can be categorised into three main types based on how a genetic variant affects multiple traits (**Figure 2.2**):

1. **Biological Pleiotropy (horizontal)**: the genetic variant directly influences more than one phenotypic trait.
2. **Mediated Pleiotropy (vertical)**: the genetic variant influences one phenotype, which affects another phenotype. This type of pleiotropy suggests a causal relationships between traits.
3. **Spurious Pleiotropy**: a genetic variant appears to be associated with multiple traits due to confounding factors rather than a true biological effect. Spurious pleiotropy can arise from issues like population stratification or LD (Solovieff et al., 2013).

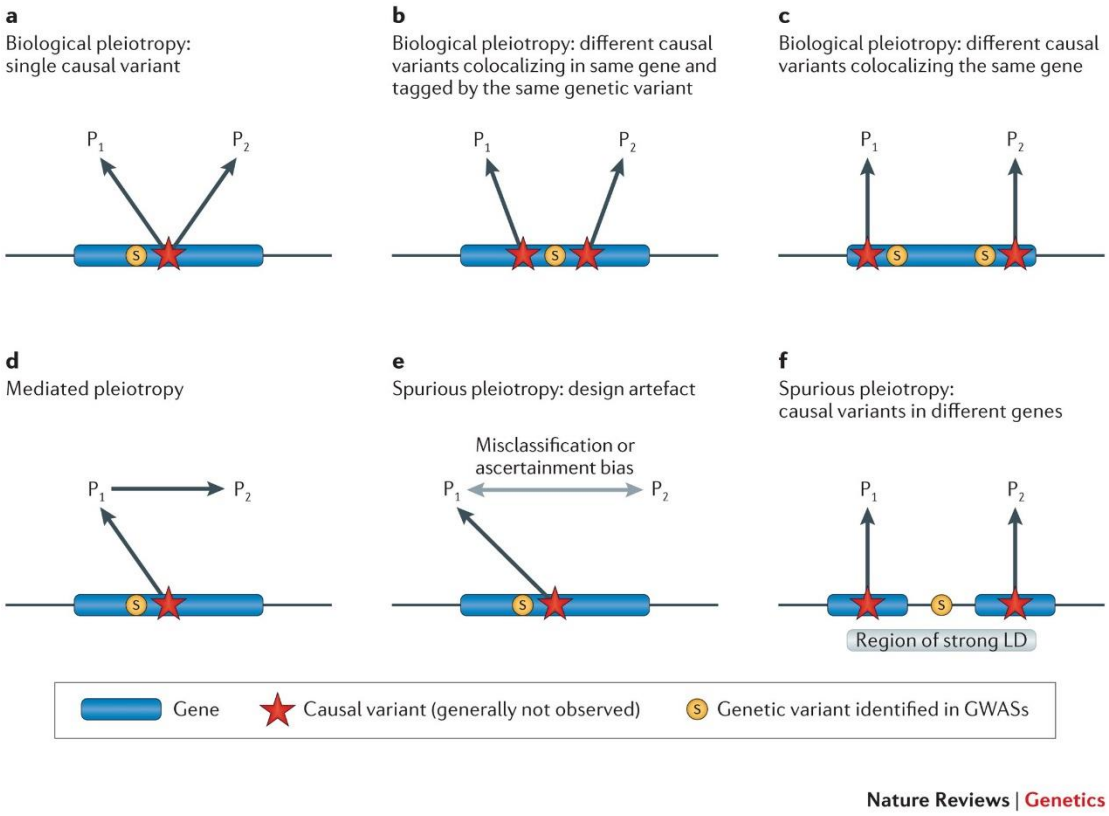


Figure 2.2 Types of pleiotropy

Source: (Solovieff et al., 2013)

Pleiotropy is intrinsically connected to comorbidity (see section 2.3) as pleiotropic genes might contribute to shared genetic pathways underlying different diseases (Sivakumaran et al., 2011, Solovieff et al., 2013). The identification of pleiotropic

effects can provide insights into shared genetic mechanisms across different traits (Pickrell et al., 2016).

2.3 Comorbidity in Complex Traits

Comorbidity refers to the co-occurrence of two or more diseases or conditions in an individual. Four different models have been proposed to explain the aetiological relationships that lead to comorbidity.: i) direct causation, ii) associated risk factors, iii) heterogeneity, and iv) independence (**Figure 2.3**) (Valderas et al., 2009).

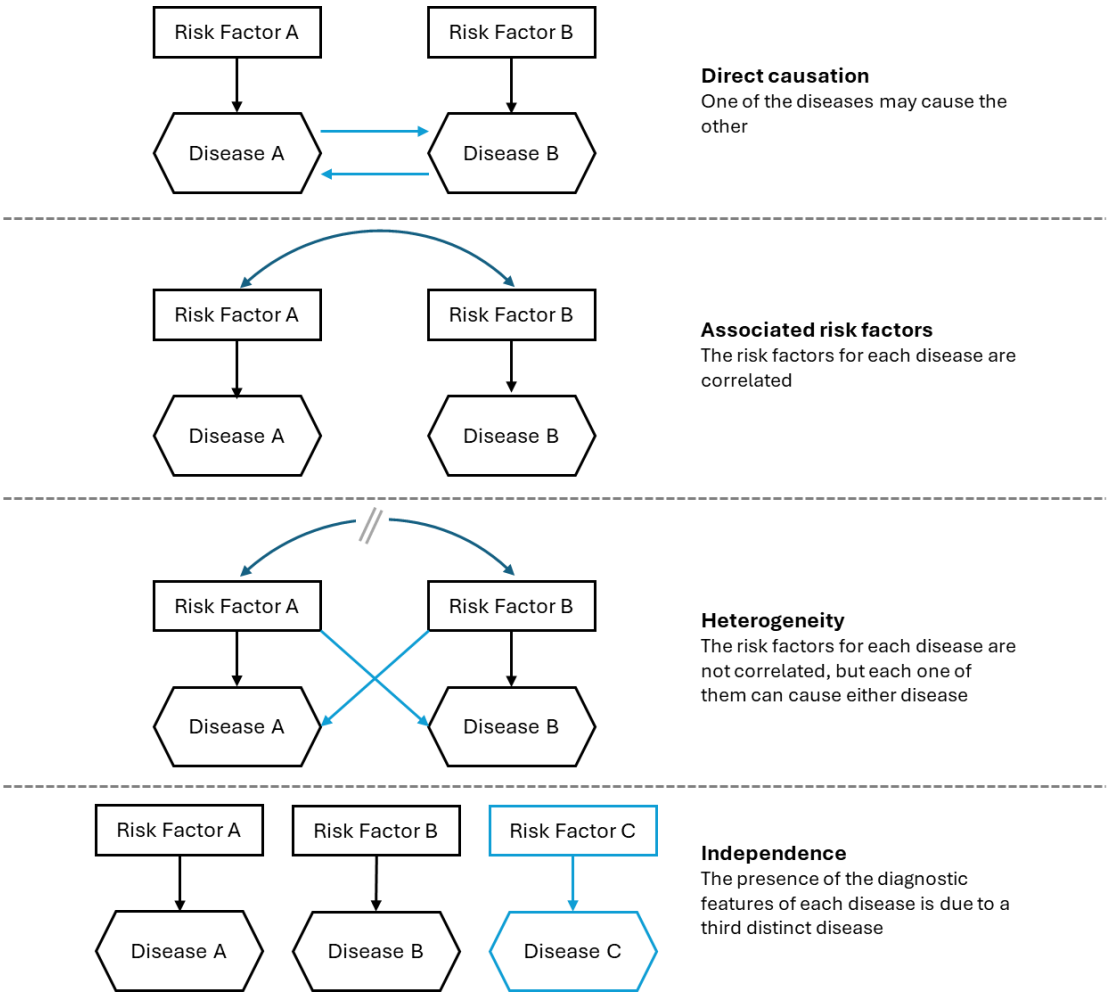


Figure 2.3 Aetiological models of comorbid diseases

Source: (Valderas et al., 2009)

Integrative approaches, combining genetic, transcriptomic, and proteomic data, help in elucidating the molecular underpinnings of comorbid conditions. For instance, gene

expression data can capture both the cause and consequence of diseases and in combination with genetic data can improve the understanding of comorbidity relationships (Sanchez-Valle and Valencia, 2023).

Genetic studies of comorbidity also have implications for public health. By identifying individuals at risk for multiple conditions, healthcare systems can better allocate resources and develop targeted interventions. Understanding the genetic links between comorbid conditions can lead to the discovery of new therapeutic targets that may benefit patients with multiple related diseases.

CHAPTER 3

METHODOLOGIES IN GENETIC EPIDEMIOLOGY

Understanding the genetic basis of complex traits requires a robust methodological framework that integrates both classical and advanced statistical techniques. In this chapter, the key methodologies that underpin genetic epidemiology are discussed, focusing on most important study designs in the field and analytical techniques used for the score of the present thesis. The methodologies discussed here provide the foundation for understanding how genetic variation influences phenotypes, guiding the identification of genetic associations. This chapter sets the stage for the application of these methods in the subsequent analysis of genetic data, highlighting their significance and relevance to the field of genetic epidemiology.

3.1 Study Designs in Genetic Epidemiology

The study designs in genetic epidemiology often adapt traditional epidemiological approaches. They can be broadly categorised into **family-based** and **population-based** studies (Dawn Teare, 2011). Family-based designs focus on the genetic relationships within families to identify heritable traits, while population-based designs examine genetic variation across broader populations to uncover associations with disease and traits (Ott et al., 2011).

3.1.1 Family-Based Studies

Family-based studies have been widely utilised to investigate the genetic basis of complex traits. These studies leverage the genetic relationships within families, making them particularly effective in controlling for confounding factors such as population stratification, which can significantly bias results in population-based studies (Dawn Teare, 2011, Balding et al., 2007).

One of the most used methods in family-based studies is the Transmission Disequilibrium Test (TDT). This test compares the transmission of alleles from parents to affected offspring (Spielman et al., 1993). It can eliminate the effects of population

stratification because the non-transmitted alleles serve as internal controls. This method is particularly useful in case-parent trio designs, where genotypes from both parents and an affected child are available (Zheng et al., 2012). The TDT has been extended to various scenarios, including situations where only one parent is available or when analysing haplotypes rather than single markers (Balding et al., 2007, Dawn Teare, 2011).

Family-based designs are not without limitations. One of them is the difficulty in recruiting suitable family members, especially for late-onset diseases where parents may no longer be available (Dawn Teare, 2011). Additionally, the shared genetic background of family members can lead to reduced statistical power (Zheng et al., 2012).

3.1.2 Population-Based Studies

Population-based studies in genetic epidemiology often utilise similar frameworks to traditional epidemiology. The most basic study designs include case-control, cohorts, and cross-sectional studies (Rothman et al., 2008).

Two primary methodologies have emerged in population-based genetic studies: **Candidate Gene Studies** and **Genome-Wide Association Studies (GWAS)**. The candidate gene studies (3.1.2.1) focus on investigating specific genes offering a targeted analysis but may miss associations in other, untested regions of the genome (Tabor et al., 2002). On the other hand, GWAS (3.1.2.2) provides a more comprehensive approach by scanning the entire genome, without requiring prior hypotheses about specific genes (Hirschhorn and Daly, 2005, McCarthy et al., 2008).

3.1.2.1 Candidate Gene Studies

Candidate gene studies investigate specific genes hypothesised to be associated with the phenotype of interest. This approach is grounded in prior biological knowledge, often derived from functional studies or the known physiological role of the gene in disease pathways (Tabor et al., 2002). Thus, genes are selected based on their suspected involvement in a biological process relevant to the disease. These genes are then examined for genetic variation to determine whether certain variants are more frequently observed in cases compared to controls (Risch and Merikangas, 1996). This approach is very powerful when the gene's role in the disease is known, allowing for a more directed investigation compared to hypothesis-free methods like GWAS

(Hirschhorn and Daly, 2005). They have been used extensively to validate findings from other types of studies, such as GWAS. However, one of their limitations is its reliance on existing knowledge means that it may miss important associations outside the studied genes (Tabor et al., 2002).

3.1.2.2 Genome-Wide Association Studies

Genome-Wide Association Studies (GWAS) is a key approach in genetic epidemiology. It is an agnostic approach scanning the entire genome to detect associations between genetic variants and traits (Uffelmann et al., 2021).

GWAS scan the whole genome aiming to identify genetic variants that occur more frequently in cases compared to controls. The process begins with the collection of DNA samples from a large cohort of participants, which are then genotyped to decode the information of target SNPs across the genome. These genetic data are quality controlled and used to impute the information of millions of untyped variants based on LD reference panels. The genotyped and imputed variants are analysed using statistical methods to identify associations between SNPs and the trait of interest (**Figure 3.1**) (Hirschhorn and Daly, 2005, Uffelmann et al., 2021).

A typical GWAS analysis generally includes linear or logistic regression models, depending on the phenotype nature (continuous or binary, respectively), including covariates, such as age and sex, to adjust for potential confounders.

GWAS have identified thousands of genetic associations with phenotypes by analysing huge numbers of variants across the genome (Visscher et al., 2017). However, they also face several limitations, such as multiple testing burden, which increases dramatically the likelihood of false positives (Pe'er et al., 2008). To control for Type I error, a genome-wide significance threshold of $P < 5\times10^{-8}$ has been established, but there are GWAS using an even more conservative threshold of $P < 1\times10^{-8}$ or $P < 5\times10^{-9}$. Additionally, although GWAS is effective at identifying genetic loci, it is often challenging to detect the exact causal variants, mostly due to LD (Bush and Moore, 2012). Another challenge is the fact that GWAS results can only explain a small fraction of the overall heritability of the phenotype. Rare variants, gene-gene interactions, and gene-environment interactions possibly play a significant role in the phenotypic variation, which is not fully captured by GWAS (Manolio et al., 2009). Population stratification, a phenomenon observed in GWAS where allele frequencies differ

between cases and controls due to different LD structure across population ancestries, is another challenge as such underlying genetic differences can cause confounding issues resulting in spurious associations (Marchini et al., 2004). This issue can be addressed using suitable methods, such as principal component analysis, limiting the probability of false-positive associations (Price et al., 2006).

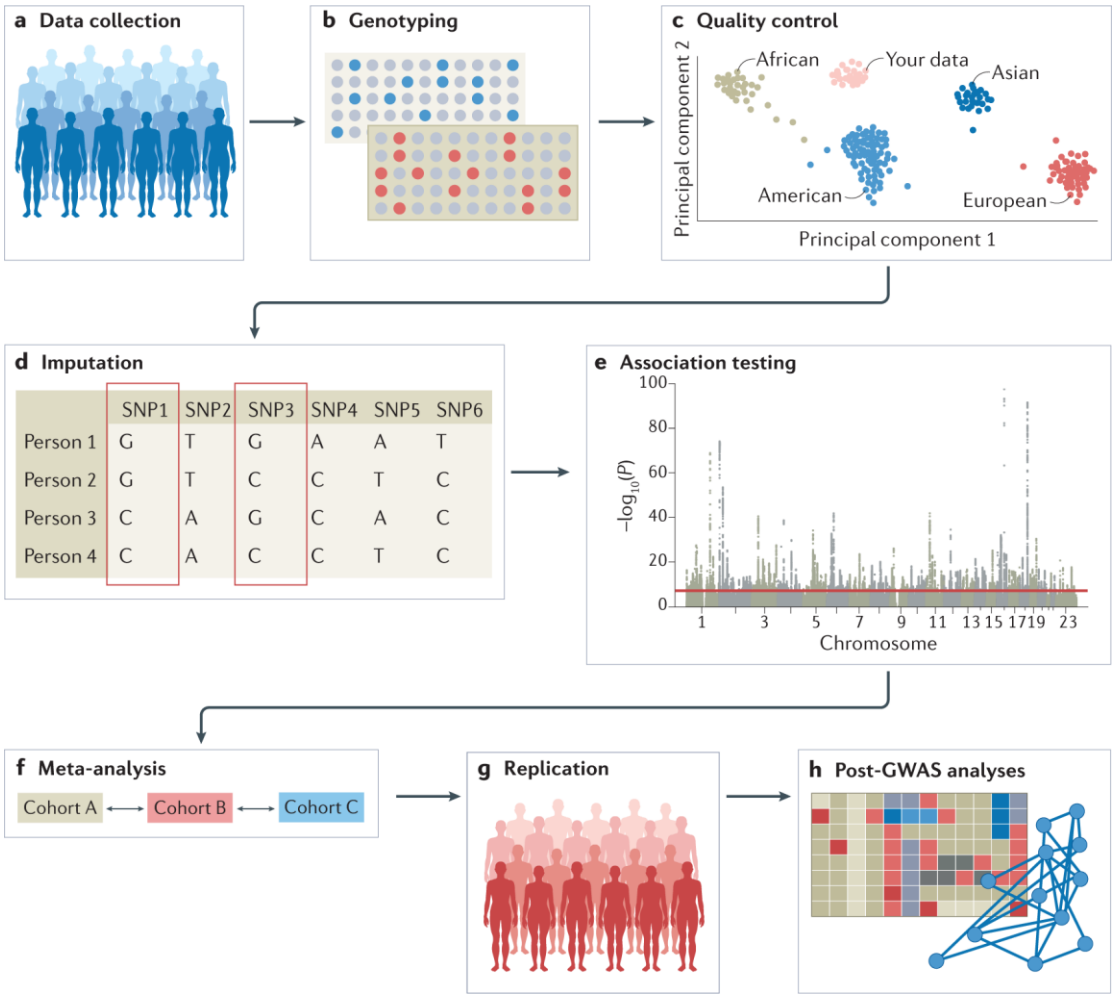


Figure 3.1 Overview of a genome-wide association study workflow

Source: (Uffelmann et al., 2021)

3.2 Advanced Analytical Techniques in Genetic Epidemiology

In addition to the multivariate methods that form the core analytical framework of this thesis, several other key techniques have been employed. These analytical techniques are crucial for validating and refining the findings derived from multivariate approaches. These methodologies enhance the precision of genetic associations and

provide deeper insights into the underlying biological mechanisms. The following sections provide detail on techniques applied for the scope of this research.

3.2.1 Fine-Mapping and Functional Annotation

Fine-mapping pinpoints specific genetic variants within a locus that are most likely responsible for the observed associations. It narrows down the broad regions identified in GWAS by using additional statistical methods and incorporating biological data, such as LD patterns and functional genomics data. The goal is to identify causal variants that directly influence the phenotype (Benner et al., 2016, Schaid et al., 2018).

Functional annotation complements fine-mapping by integrating various biological datasets to predict the functional impact of the identified variants. This process involves mapping the variants to known regulatory elements, protein-coding regions, or non-coding RNAs, and assessing their potential role in gene regulation or protein function (Roadmap Epigenomics Consortium et al., 2015). Together, fine-mapping and functional annotation help to translate statistical associations into biological insights.

3.2.2 Colocalisation

Colocalisation analysis is a method to assess whether two or more traits share a common causal variant at a specific locus. Therefore, it is particularly useful in investigating pleiotropy elucidating whether genetic associations between traits arise from the same causal variant or occur through different variants in LD (Solovieff et al., 2013, Giambartolomei et al., 2014).

Several methods are widely used for colocalisation analysis. COLOC, a popular Bayesian approach, estimates the probability that two traits share a causal variant within a specific region. It tests five hypotheses where multiple traits may share common genetic loci (Giambartolomei et al., 2014, Wallace, 2021):

H_0: neither trait has a causal variant in the locus
H_1: only the first trait has a causal variant in the locus
H_2: only the second trait has a causal variant in the locus
H_3: both traits have causal variants in the locus, but not the same one (**Figure 3.2A**)
H_4: both traits share the same causal variant in the locus (**Figure 3.2B and C**)

However, colocalisation cannot distinguish between horizontal (**Figure 3.2B**) and vertical pleiotropy (**Figure 3.2C**) and therefore infer whether the two traits are causally

related. Additional analyses are needed for that including Mendelian randomisation (see section 3.2.4).

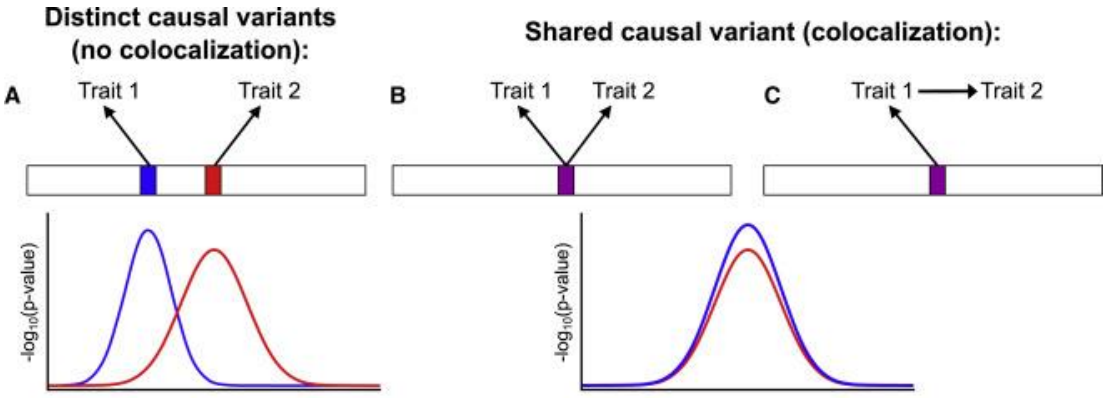


Figure 3.2 Schematic illustration in three colocalisation scenarios

A) No colocalisation, B) colocalisation due to horizontal pleiotropy, C) colocalisation due to vertical pleiotropy

Source: (Zuber et al., 2022)

An extension of the previous method, HyPrColoc, can handle multiple traits simultaneously. It also uses a Bayesian algorithm to build a hierarchical framework and identify shared causal variants across multiple phenotypes (Foley et al., 2021).

Colocalisation analysis has been extensively used to understand the genetic architecture of complex diseases. By integrating GWAS with expression quantitative trait loci (eQTL) data, colocalisation methods help to identify genetic loci that influence disease risk and also regulate gene expression in disease-relevant tissues. This can provide insights into the biological pathways affected by genetic variants (Zhu et al., 2016).

3.2.3 Phenome-Wide Association Study

The **Phenome-Wide Association Study (PheWAS)** is a method that explores the association between a single genetic variant and a wide array of phenotypes. It can use electronic health records (EHRs) or other phenotypic measures, linking them with genetic information. Unlike GWAS, which examines many genetic variants with a single trait, PheWAS scans multiple phenotypes with specific genetic variants (Denny et al., 2010).

It can provide a broad view of the potential genetic impact on different aspects of human health and thus, it can reveal pleiotropic effects, which might remain hidden in a GWAS approach. It can also contribute to the identification of new disease associations or common pathways between seemingly unrelated diseases (Tyler et al., 2016).

PheWAS can be combined with other genetic approaches, such as Polygenic Risk Score (PRS), a score that quantifies the cumulative effect of multiple genetic variants on an individual's risk for a given trait. PRS is typically used in a disease-specific context to predict the genetic predisposition to a single phenotype (Chatterjee et al., 2016). However, using PheWAS, PRS can be applied across a broad range of phenotypes allowing the exploration of pleiotropy and comorbidity (Wang et al., 2021).

3.2.4 Mendelian Randomisation for Causal Inference

Mendelian randomisation (MR) uses genetic variants as instrumental variables to infer causal relationships between risk factors (exposures) and health outcomes. The underlying principle of MR is based on Mendel's law of inheritance, which ensures the random allocation of alleles during meiosis. This natural mechanism mimics the randomisation process in controlled trials, thereby reducing the risk of confounding and reverse causation, creating a natural experiment for causal inference (Smith and Ebrahim, 2003, Davies et al., 2018).

In MR, genetic variants (instruments) are chosen as proxies for modifiable exposures. These genetic instruments are assumed to influence the outcome only through their effect on the exposure, satisfying three key assumptions (**Figure 3.3**) (Burgess et al., 2013):

1. **Relevance:** The genetic instruments are associated with the exposure.
2. **Independence:** The instruments are independent (not associated) of other factors (confounders) which affect the outcome.
3. **Exclusion Restriction (no pleiotropy):** The instrument affects the outcome exclusively through the exposure, not via any alternative pathways

There are several approaches to MR analysis, each addressing different scenarios and data types. **One-sample MR,** where both the genetic instruments and the exposure/outcome data come from the same sample of individuals (Burgess et al., 2015).

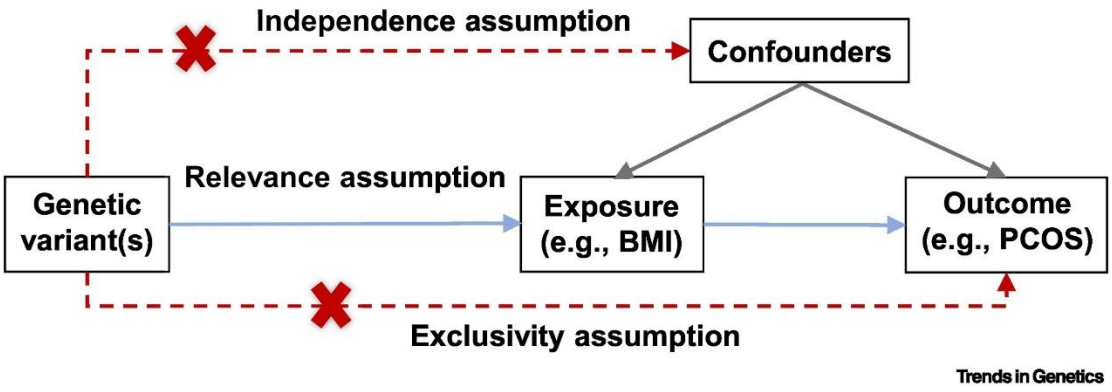


Figure 3.3 Mendelian randomisation overview and assumptions

Source: (Wang et al., 2022)

Two-sample MR utilises genetic instruments for the exposure and outcome derived from separate populations. Common methods used in two-sample MR include inverse variance weighted (IVW) analysis, MR-Egger, and weighted median (WM) approaches (Burgess et al., 2015, Hemani et al., 2018). IVW-MR is one of the most widely used approaches for MR. It combines the effect estimates from each genetic variant weighted by the inverse of their variance, providing a meta-analytic estimate of the causal effect. It assumes that all genetic instruments are valid without horizontal pleiotropy. MR-Egger allows for the possibility of horizontal pleiotropy, where genetic variants influence the outcome through pathways other than the exposure. It performs a weighted linear regression of the outcome-exposure associations and if the intercept is significantly different from zero, it indicates the presence of horizontal pleiotropy. WM-MR provides a causal estimate by taking the median of the weighted distribution of the genetic instruments. This approach is robust to invalid instruments, provided that at least 50% of the weight in the analysis comes from valid instruments (Bowden et al., 2015, Burgess et al., 2015).

Bidirectional MR allows the investigation of potential reverse causation by examining the effect in both directions between two traits (Richmond and Davey Smith, 2022). **Multivariable MR (MVMR)** extends standard MR by incorporating multiple exposures, adjusting the estimation of direct effects of each exposure on the outcome (Sanderson, 2021). **Multi-response MR (MR2)**, a recently suggested method, is a joint multivariable (multiple exposures) and multivariate (multiple outcomes) MR model,

designed to explore the unmeasured pleiotropic pathways of multiple related exposures on multiple outcomes (Zuber et al., 2023).

CHAPTER 4

MULTIVARIATE METHODS IN GENETIC EPIDEMIOLOGY

Multivariate analysis is an advanced statistical approach that examines multiple variables simultaneously. It can be used in genetic epidemiology to uncover genetic relationships with complex traits by enabling the study of multiple phenotypes together.

Most genetic epidemiological studies follow univariate approaches focusing on single traits that examine associations between one independent variable (e.g., a single SNP) and one outcome variable (e.g., a specific phenotype). However, many correlated traits are one multivariate phenotype in nature with several components. Thus, univariate methods might fall short in capturing the complex structure of interconnected traits, which may correspond to linked biological pathways, or the co-occurrence of symptoms. Multivariate approaches allow the joint analysis of multiple phenotypes and can offer essential insights into pleiotropy and shared genetic architecture of phenotypes (Yang and Wang, 2012, Vroom et al., 2019).

Additionally, many univariate genetic studies may be underpowered to detect small-to-moderate genetic effects. Hence, the rationale for employing multivariate analysis is based on its ability to improve statistical power. This can be accomplished through the exploitation of the covariance among phenotypes that can lead to increased probability to detect genetic loci associated with phenotypes of interest, especially when phenotypes are correlated. Furthermore, these methods can limit multiple testing issues that arise from testing multiple phenotypes separately (Yang and Wang, 2012, Vroom et al., 2019).

With a few exceptions, multivariate methods generally test a null hypothesis of no association with any trait. However, they present more differences in terms of the exact alternative hypothesis they test. Some of them test the hypothesis that the joint effect for a genetic variant on two or more traits deviates from zero, while others that the genetic variant is associated with at least one trait. Also, the choice of multivariate methods can be influenced by the type of data they are able to analyse. Some of them can only incorporate continuous phenotypes only, others dichotomous, categorical or

mixed. All methods suitable for continuous traits assume a multivariate norma distribution for all traits (Vroom et al., 2019).

Methods within multivariate analysis vary in their objectives and methodologies, from regression-based models that accommodate correlated outcomes to dimension reduction techniques that simplify high-dimensional data by identifying key underlying factors. This section provides an overview of the key multivariate methods of all continuous and dichotomous outcomes following Yang and Wang's conceptual classification (Yang and Wang, 2012).

4.1 Univariate Test Statistics Combination Methods

Univariate test statistics combination methods combine P-values or test statistics derived from univariate analyses to test a multivariate hypothesis. The simplicity is the main benefit of such approaches as using univariate methods is generally easier and conceptually simpler than multivariate methods. These methods are usually applicable to any type of phenotypes including both continuous and dichotomous outcomes. The test statistics that can be combined include beta effect estimate coefficients, P-values or test statistics values. However, the most challenging part of these methods is to handle effectively the statistics correlations, which come from the correlations of the phenotypes. (Yang and Wang, 2012, Vroom et al., 2019).

The alternative hypothesis (H_1) that most of these methods test is whether the genetic marker is associated with at least one of the examined traits against the null hypothesis (H_0) of no associations to any trait. These methods can assume homogeneous or heterogeneous genetic effects. The models assuming homogeneous genetic effects constraint all genetic variants effects to be the same and thus they are most powerful under this scenario while the heterogeneous genetic effects models are expected to be more powerful when the genetic effects differ in size or sign across traits (Yang and Wang, 2012, Vroom et al., 2019).

4.2 Dimension Reduction Methods

Dimension reduction methods aim to reduce the number of variables. With a few exceptions (e.g. factor analysis), these methods are applicable only to continuous phenotypes that follow an approximate normal distribution. As a principle, these

methods initially create new variates that are linear combinations of the original phenotypic variation:

$$\tilde{y} = b_0 + b_1 y_1 + b_2 y_2 + \cdots + b_m y_m \quad (4.1)$$

where $b_1, \ldots, b_m$ are coefficients weighting the contribution of each original trait to the new variates. Afterwards, the generated variable is tested with the genetic variants for any potential associations (Yang and Wang, 2012, Vroom et al., 2019). This section discusses three key methods performing dimension reduction.

4.2.1 Principal Component Analysis

Principal Component Analysis (PCA) transforms a set of correlated and standardised traits into uncorrelated linear combinations that account for the most variability in the original variables. The new variates are called Principal Components (PC) and their number cannot be bigger than the number of the original traits. For each PC, weights are selected in such a way to maximise its variance. The first PC (PC1) explains the largest part of the overall multidimensional variability of the full set of the original traits, followed by the second PC (PC2), the third (PC3) and so forth (Vroom et al., 2019, Yang et al., 2016).

4.2.2 Factor Analysis

Factor Analysis creates a common factor from a set of correlated traits, calculates a factor score and then uses it as a dependent variable in regression models. The method determines the weights $b_1, \ldots, b_m$ in equation (4.1) so that the factor accounts for the maximum possible variance common to the traits (Vroom et al., 2019). Although factor analysis and PCA show many similarities, they show some differences. First, PCA aims to explain the total variance of the traits, while factor analysis focuses on the shared covariance between the traits (Lawley and Maxwell, 1962). Second, the two methods differ in their concept as factors can represent an underlying theoretical construct (e.g. a health condition that cannot measured directly), while PCs are just statistical combinations. Third, unlike PCA, factor analysis assumes that residuals are uncorrelated (Vroom et al., 2019).

4.2.3 Canonical Correlation Analysis

Canonical Correlation Analysis (CCA) extracts a linear combination (canonical variable) of a set of traits that has the highest possible correlation (canonical correlation)

with the genetic variant. Thus, for each genetic variant, the weights of the canonical variable differ across traits indicating their strength of association with the genetic variant. CCA is the only dimension reduction method that incorporates information from the genetic variants to the new variate (Ferreira and Purcell, 2009). The covariance matrix of a multivariate outcome Y and a genetic marker X is estimated as follows:

$$cov\begin{bmatrix} Y \\ X \end{bmatrix} = \begin{bmatrix} \Sigma_{YY} & \Sigma_{YX} \\ \Sigma_{XY} & \Sigma_{XX} \end{bmatrix} \tag{4.2}$$

where Σ_{YY} is the variance matrix of Y, Σ_{YX} and its transpose Σ_{XY} are the covariance matrices between Y and X and Σ_{XX} is the variance matrix of X. Under an additive genetic model of codominance, CCA's performance is identically with MANOVA (see section 4.3.1) (Yang and Wang, 2012).

4.3 Multivariate Regression Modelling Methods

Multivariate regression modelling is another common approach to analyse associations between a genetic variant and continuous or categorical multivariate traits. As a rule, most multivariate regression-based models assume a multivariate normal distribution of the traits. In general, these models follow the equation:

$$Y = BX + E \tag{4.3}$$

where Y is a matrix containing the phenotypic values, X is a matrix of predictors with genotypic information, B is a matrix of the trait-specific regression weights and E is a matrix of trait-specific zero-mean residuals (error or disturbance terms). The residuals are assumed to have the same variance (homoscedasticity) across traits. Including E in the model allows to account for all sources of residual covariance (Vroom et al., 2019).

Next, some of the most common multivariate regression-based methods are presented. Of them, all except for the reversed ordinal multiple regression constitute specific instances of the equation (4.3).

4.3.1 Multivariate Analysis of Variance

Multivariate Analysis of Variance (MANOVA) extends the traditional analysis of variance (ANOVA) allowing to assess the joint effects of genetic variants on multiple traits simultaneously. Standard MANOVA uses an unconstrained covariance matrix testing the null hypothesis of all zero regression coefficients. It compares variance-

covariance matrices within and between genotypes (Yang et al., 2016). Some methods assume equal regression weights of the genetic variants (homogenous effect) while others allow for different regression weights (heterogenous effect) across traits (Vroom et al., 2019). This method assumes a multivariate normal distribution for the traits, which can offer many statistical advantages in hypothesis tests, but it comes with strong assumptions, which might be violated in many cases (Yang et al., 2016).

4.3.2 Mixed Effects Models

Mixed effects models are multivariate regression models, where fixed effects are assumed to estimate the genetic effects, while additional random effects are used to account for the correlations between traits. In general, these models can address various types of potential confounding, such as familial relatedness or population stratification. Depending on the type of outcomes, two different models are mainly used. **Linear Mixed-Effects Models (LMM)** are used to model continuous outcomes following the form:

$$Y_k = \beta_0 + \beta_k X + \eta_k + e_k \quad (4.4)$$

where β_0 is the intercept capturing the non-genetic fixed effect, β_k is the genetic fixed effect size of the genetic variant on the k^{th} phenotype, η_k corresponds to the random effects correlation between traits assuming to follow a normal distribution and e_k represent the random errors (residuals) also following a normal distribution. In models using individual level data, η_k is expected to be correlated between traits within an individual, but independent between different individuals.

When the outcomes are categorical, the **Generalised Linear Mixed-Effects Models (GLMM)** are used as follows:

$$E(Y_k|\eta_k) = \mu^{-1}(\beta_0 + \beta_k X + \eta_k) \quad (4.5)$$

where μ is the link function. For normally distributed traits, the μ corresponds to the identity link and GLMM equation (4.5) becomes identical to LMM (4.4), while for binary outcomes, μ corresponds to the logit link.

The null hypothesis for both LMM and GLMM assumes a zero genetic effect on any trait ($H_0: \beta_1 = ... = \beta_k = 0$) and can be tested using the likelihood ratio or a Wald chi-squared test (Yang and Wang, 2012, Vroom et al., 2019).

4.3.3 Generalised Estimating Equations

Generalised Estimating Equations (GEE) belong to the class of marginal models. Unlike mixed effects model, which use η_k and residuals as separate terms to model phenotypic correlation, GEE collapse random effects and random residual errors in the model. The structure of the residual correlation matrix can be assumed to correspond to conditionally uncorrelated traits, conditionally correlated traits or freely estimated conditional correlations. The null hypothesis of no association between the genetic variant and any phenotype can be tested using a Wald test statistic (Yang and Wang, 2012, Vroom et al., 2019).

4.3.4 Structural Equation Modelling

Structural equation modelling (SEM) is a multivariate statistical approach modelling directly and indirectly observed (latent) variables. SEM is a hypothesis-driven approach (not agnostic) that uses a system of linear equations combining techniques such as regression models, factor analysis and path analysis. It comprises two sub-models: the measurement model, which estimates relationships between the observed (indicators) and unobserved (latent) variables using a factor analysis framework and the structural model, which develops relationships between the latent variables. The measurement model consists multiple equations (one for each indicator) following the form:

$$X_i = \Lambda_X \xi + \delta_i \tag{4.6}$$

$$Y_i = \Lambda_Y \eta + \varepsilon_i \tag{4.7}$$

where X_i and Y_i are the indicators, ξ and η are the latent variables, Λ_X and Λ_Y are the factor loadings, δ_i and ε_i are the residuals. The structural model consists of the equation:

$$\eta = \alpha + B\xi + \zeta \tag{4.8}$$

where η is a vector of the latent variables, α is a vector of intercepts, B is a matrix of the genetic coefficients on η and ζ is a vector of disturbances. The model assumes that δ, ε and ζ are mutually uncorrelated (Stein et al., 2012).

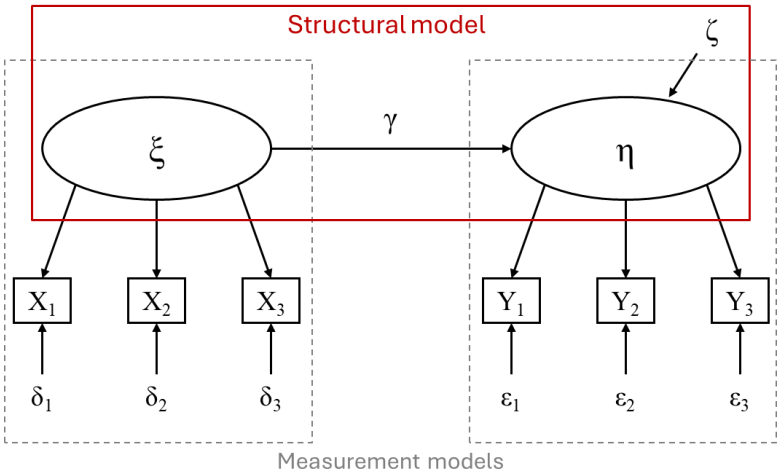


Figure 4.1 Structural Equations Modelling workflow under a simple scenario

4.3.5 Reversed Ordinal Multiple Regression

This method, named MultiPhen, reverses the regression model, considering genetic variants as an ordinal outcome [0, 1, 2] and traits as predictors. Thus, the distributional assumptions of the phenotypes (e.g., conditional multivariate normality) are not necessary anymore. The phenotypes can be any combination of continuous and categorical (as dummy-coded) variables. The null hypothesis is that all coefficients in the regression model are zero (O'Reilly et al., 2012).

4.4 Multivariate Meta-Analytical Methods

Multivariate meta-analysis allows the joint synthesis of multiple outcomes accounting for their correlation. Most meta-analyses aim to estimate multiple summary results jointly from the same meta-analysis model (van Houwelingen et al., 2002). Most meta-analytical methods are based on one or more of the previous categories to produce summary statistics using various approaches and, on this aspect, they do not strictly form a distinct class of multivariate methods. They allow a larger number of studies to contribute into the meta-analysis results for each outcome, which can improve efficiency and reduce risk of bias compared to performing separate univariate meta-analyses for each outcome (Hattle et al., 2022).

4.5 Bayesian Multivariate Modelling Methods

There have been suggested some multivariate methods using a **Bayesian** approach. Many of these methods use multivariate regression models, but they infer based on a Bayes Factor (BF) instead of P-value. They use various forms of prior information including priors about the genetic effect model (additive, dominant, recessive) and priors about the genetic effect size (Marchini et al., 2007, Stephens, 2013).

CHAPTER 5

OBJECTIVES OF THE THESIS

Genetic association studies typically investigate associations between single genetic variants and single phenotypes. Yet, multivariate approaches, which allow the joint analysis of genetic effects on multiple phenotypes, have the potential to capture the shared genetic architecture between correlated traits. These multi-trait methods exhibit many advantages; they can increase the statistical power to detect novel genetic loci and they allow for a more holistic investigation of genetic pleiotropy.

The present thesis explores the multivariate methods that utilise data from GWAS, makes a comparison and prioritises the most effective methods to discover novel genetic loci and investigate pleiotropy. Specific objectives:

1. To systematically review and compare the existing multivariate methods that combine summary statistics from GWAS on human complex traits aiming to prioritise the most efficient multivariate methods for the discovery of novel loci and for the investigation of pleiotropy.
2. To discover novel genetic loci of blood cell traits using suitable multivariate method to GWAS summary statistics on blood cell traits and prioritise genetic variants with global effect.
3. To investigate the shared genetic architecture between inflammation and cardiometabolism. More specifically, using multi-trait methods on genomic data, this thesis aimed to identify genetic loci with pleiotropic effect between C-reactive protein, a marker of chronic inflammation, and several cardiometabolic risk factors and also, to discover shared biological mechanisms.
4. To understand the comorbidity of Alzheimer's and cardiovascular diseases attributed to genetic predisposition through the application of multi-trait GWAS, to discover novel genetic loci for Alzheimer's disease, identify pleiotropic loci and shared biological pathways aiming to define common therapeutic targets.

SPECIAL PART - APPLICATIONS AND FINDINGS

CHAPTER 6

MULTIVARIATE METHODS IN GENOME-WIDE ASSOCIATION STUDIES: A SYSTEMATIC REVIEW

6.1 Background

GWAS typically examine associations in a univariate approach between single genetic variants and single traits. Although univariate GWAS successfully identified thousands of novel genetic loci, yet they might not be able to explain only a sufficient proportion of heritability of complex phenotypes (Manolio et al., 2009, Eichler et al., 2010). However, multivariate methods, which allow the joint analysis of multiple phenotypes, can be more efficient in capturing the complex genetic architecture of multiple correlated traits offering several advantages (Galesloot et al., 2014, Zhu et al., 2015a). They can improve the statistical power by leveraging the genetic correlation of the examined traits, and thus identify more genetic variants of small-to-moderate effects. They can also reduce the number of tests limiting the inflation of type I error due to multiple testing (Allison et al., 1998, Yang and Wang, 2012, Galesloot et al., 2014). Under the presence of pleiotropy, multivariate modelling is more consistent with biology (Chavali et al., 2010).

Several multivariate methodologies have been suggested varying in their objectives and concepts. These methods can be classified into five main categories: methods that combine test statistics from univariate models, dimension reduction approaches, multivariate regression models, multivariate meta-analysis and Bayesian multivariate methods. Each category shows unique advantages and limitations (Yang and Wang, 2012, Vroom et al., 2019). While most multivariate methods were initially applicable to individual-level data only, recently, there has been a recent increasing number of multivariate methods that can utilise summary statistics from GWAS. Yet, there has not been a systematic review assessing the existing multivariate methods that incorporate summary statistics.

This study aims to systematically review all the existing multivariate methods that perform a single SNP – multiple phenotypes GWAS using exclusively summary statistics from univariate GWAS. It also aims to compare and understand their varying approaches and prioritise the most widely used methods with practical applications.

6.2 Methods

6.2.1 Eligibility Criteria

All studies suggesting new multivariate methods to perform joint analysis of single SNPs - multiple phenotypes genome-wide associations using GWAS summary statistics were included. The studies that met any of the following criteria were excluded: not introducing a novel method to jointly analyse multiple phenotypes, not applicable to human populations, not performing genome-wide association analysis between single SNPs and multiple complex traits, not incorporating exclusively summary statistics, extensions or implementation of existing multivariate methods, and absence of software with practical application.

6.2.2 Information Sources and Search Strategy

A systematic search of the literature was conducted to detect multivariate methods applicable to GWAS summary statistics. We searched PubMed on 12th December 2023 using the algorithm: “(multivariate OR "multivariate analysis" OR multitrait* OR multi-trait* OR "multiple trait" OR "multiple traits" OR multiphenotype* OR multi-phenotype* OR "cross phenotype" OR "cross phenotypes" OR crossphenotype* OR cross-phenotype*) AND (GWA OR GWAS OR genome wide OR genome scan)”.

To identify any studies that were not included in PubMed database or were missed by the search algorithm, an additional manual search for eligible articles was performed on the references of the included articles.

6.2.3 Study Selection and Data Extraction

The retrieved citations from PubMed were screened at two stages: first, at a title and abstract level and then at a full text level. The articles were assessed against the predefined eligibility criteria. The references of the eligible articles were manually reviewed, and the articles meeting the inclusion criteria that were not included in the

list of citations from PubMed were examined using the same screening approach and eligibility criteria as described earlier.

Data extraction included information on the first author's name, publication year, journal, multivariate name, type and classification of multivariate method, data type that the method uses (summary statistics or individual level data), software environment (e.g. Python, R), number of phenotypes the method can incorporate, type of input phenotypes (continuous, categorical or both), type of output phenotype (combined or trait-specific), type of input summary statistics (beta/se, Z, P-value, etc), type of output summary statistics, feasibility of joint analysis on multiple variants, and suitability for rare variants analysis.

6.2.4 Citation Analysis

A citation analysis of the included studies was conducted using the Scopus database to prioritise the most cited methods. The extracted information included the total number of citations and the average annual number of citations up to 31 December 2023.

6.3 Results

6.3.1 Evidence Base Overview

The search in PubMed identified 3827 citations, which corresponded to 26 eligible articles while one additional article was identified from their references (**Figure 6.1**). The 27 included articles (26 from PubMed and 1 from the reference search) referred to 24 distinct multivariate methods (**Table 6.1**).

The 24 methods, published between 2013 and 2023, perform a multi-trait genome-wide association analysis using exclusively summary statistics from univariate GWAS. In summary, 8 of them can be classified as pure univariate test statistics combination (UTSC) methods, 2 as pure dimension reduction (DR) methods and another 2 as a combination of UTSC and DR methods. Moreover, 5 methods follow a pure meta-analytical approach while 3 other methods are based on combinations between meta-analysis and techniques from other categories. Five methods incorporate linear modelling approaches including multinomial logistic regression or linear combination models. Finally, 2 pure Bayesian methods were found and another 2 combining Bayesian with approaches from other categories.

Table 6.1 Characteristics of the 24 included multivariate methods

Name (Publication)	Trait Type	Method	Method Classification Category	Input Stats	Output Stats	Environment
aMAT * (Wu, 2020)	Cont.	minP	UTSC	Z	P	R
aSPU ‡ (Kim et al., 2015)	Both	minP and meta-analysis	UTSC	Z	P	R
CGWAS (Xiong et al., 2022)	Both		Meta-analysis	Beta, P	P	R
CONFIT (Gai and Eskin, 2018)	Cont.	Likelihood ratio	Meta-analysis	Beta, SE	P	Python
CPASSOC (Zhu et al., 2015b)	Both	IVW meta-analysis and max weighted sum Z	Meta-analysis and UTSC	Z	P	R
EBMMT (Liu et al., 2022)	Both	ACAT (PCA and minP)	DR and UTSC	Z	P	R
GenomicSEM‡ (Grotzinger et al., 2019)	Both	SEM	Linear combination and meta-analysis	Beta, P	Beta, SE, P	R
GWAMA § (Baselmans et al., 2019)	Both	Weighted sum Z	Meta-analysis (UTSC)	Beta, SE or Z, P	Beta, SE	R
HIPO (Qi and Chatterjee, 2018)	Both	Orthogonal linear combinations	Linear combination	Z, P	Z, P	R
HSVS-M ‡ (Yang et al., 2022)	Both	Bayesian Lasso	Bayesian	Z	PP	R
iMAP (Zeng et al., 2018)	Both	Penalised Gaussian mixture models	Multivariate regression model and Bayesian	Z	PP	R
JASS (Julienne et al., 2020)	Both	Weighted sum Z	UTSC	Z	P	Python/Java
metaCCA ‡ (Cichonska et al., 2016)	Cont.	CCA	DR	Beta, SE	r, P	R
metaUSAT (Ray and Boehnke, 2018)	Both	minP weighted (MANOVA and SSU)	Meta-analysis (UTSC and linear modelling)	Z	P	R

MPAT (Liu and Lin, 2019)	Cont.	PCA	DR	Z	P	R
MRP *‡ (Venkataraman et al., 2021)	Both	BF	Bayesian	Beta, SE	BF	Python
MTAFS (Deng et al., 2024)	Cont.	Adaptive Fisher	UTSC	Z	P	R
MTAG § (Turley et al., 2018)	Both	Generalised Method of Moments	Linear combination and meta-analysis	Beta, SE or Z	Beta, SE, P	Python
MTAR (Guo and Wu, 2019a)	Cont.	PCA and weighted sum Z and minP	DR and UTSC	Z	P	R
MultiMeta § (Vuckovic et al., 2015)	Cont.	IVW meta-analysis	Meta-analysis	Beta, Var	Beta, P	R
PAT (Taraszka et al., 2022)	Cont.	Likelihood ratio	Bayesian meta-analysis	Z	Z, P	Python
PLACO ¥ (Ray and Chatterjee, 2020)	Both	MLE	UTSC	Z	P	R
PolarMorphism (von Berg et al., 2022)	Both	Euclidean trigonometric transformation of Z	UTSC	Beta, SE	r, P and θ, P	R
TATES (van der Sluis et al., 2013)	Both	min weighted P	UTSC	P	P	FORTRAN

* Can efficiently analyse rare variants (MAF ≤ 0.01)

§ Generates trait-specific summary statistics

‡ Can also perform a multivariate SNP-set analysis

¥ Can perform a bivariate analysis of two phenotypes only

minP: minimum P-value; IVW: inverse-variance weighted; ACAT: aggregated Cauchy association test; PCA: principal component analysis; SEM: structural equation modelling; CCA: canonical correlation analysis; SSU: sum of squared score; BF: Bayes factor; MLE: maximum likelihood estimate; UTSC: univariate test statistics combination; DR: dimension reduction; Z: Z statistic; SE: standard error; P: P-value; Var: variance; r: correlation value measuring overall effect; θ: angle parameter measuring the sharedness between the traits

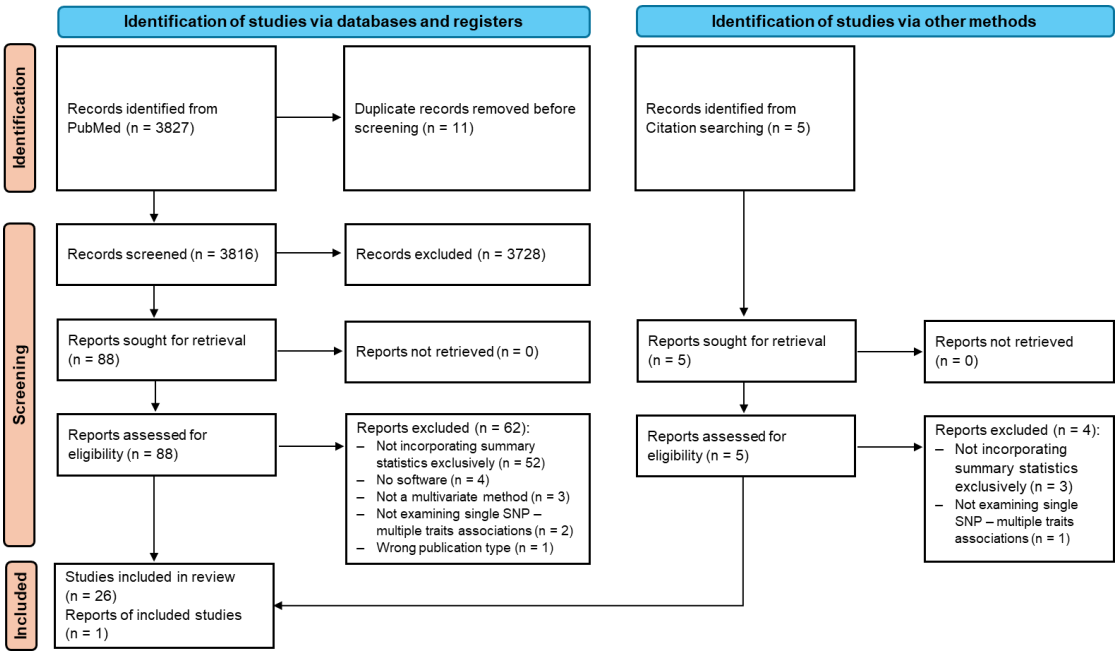


Figure 6.1 Flow chart for the selection of eligible studies

Source: Page MJ, et al. BMJ 2021;372:n71. doi: 10.1136/bmj.n71.

The null hypothesis in most methods assumes that the SNP has no genetic effect on any of the phenotypes, while the alternative hypothesis assumes the existence of genetic effect between the SNP and at least one phenotype. In contrast, PLACO assumes that the SNP effects zero or one trait under the null hypothesis and that it effects both traits under the alternative hypothesis. A Bayesian method, iMAP, in a simple scenario of two traits, it tests four hypotheses: the SNP effects none of the traits (H_0), the first trait only (H_1), the second trait only (H_2) and both traits (H_3).

Three methods (MultiMeta, MTAG and GWAMA) generate trait-specific summary statistics per SNP, while the rest methods produce single summary statistics per SNP summarising the joint genetic effect on the set of the examined traits. PLACO can only perform a bivariate analysis while the rest methods are able to incorporate two or more phenotypes. The methods aMAT and MRP can efficiently analyse rare variants (MAF < 0.01) while the methods aSPU, HSVS-M, metaCCA and MRP can perform a multivariate set-SNP analysis.

A software or code is provided for all methods, with eighteen available in the R environment, five in Python, and one in FORTRAN.

6.3.2 Citation Metrics

The citation analysis on the included studies indicated MTAG as the method with the highest impact, followed by Genomic SEM and CPASSOC (**Table 6.2**). The number of citations for these three methods corresponded to the 85% of the overall number of the citations for all methods, while the number of citations for the three methods separately corresponded to the 40% for MTAG, 26.7% for Genomic SEM and 18.6% for CPASSOC. **Figures 6.2** and **Figure 6.3** highlight changes in citation counts and cumulative average rate of citations across years for the nine most cited studies.

Table 6.2 Number of citations per year for the included studies

Method	Publ Year	≥2016	2017	2018	2019	2020	2021	2022	2023	Tot Cit	Avg Cit
MTAG	2018	0	0	21	63	56	88	99	117	**444**	**74.0**
GenomicSEM	2019	0	0	0	12	27	76	94	88	**297**	**59.4**
GWAMA	2019	0	0	0	7	19	39	44	35	**144**	**28.8**
CPASSOC	2015	20	26	25	26	20	32	31	27	**207**	**23.0**
TATES	2013	50	20	14	16	16	14	13	10	**153**	**13.9**
metaCCA	2016	0	9	11	10	16	13	8	9	**76**	**9.5**
PLACO	2020	0	0	0	0	0	6	7	19	**32**	**8.0**
metaUSAT	2018	0	0	1	6	10	8	5	7	**37**	**6.2**
aSPU	2015	2	6	8	4	11	7	4	8	**50**	**5.6**
MPAT	2019	0	0	0	0	4	4	6	4	**18**	**3.6**
iMAP	2018	0	0	0	2	2	7	6	3	**20**	**3.3**
HIPO	2018	0	0	0	1	4	6	5	3	**19**	**3.2**
MTAR	2019	0	0	0	0	4	3	2	7	**16**	**3.2**
CGWAS	2022	0	0	0	0	0	0	0	5	**5**	**2.5**
aMAT	2020	0	0	0	0	2	2	2	2	**8**	**2.0**
EBMMT	2022	0	0	0	0	0	0	1	2	**3**	**1.5**
MultiMeta	2015	1	2	0	2	3	0	2	2	**12**	**1.3**
CONFIT	2018	0	0	0	0	2	1	1	2	**6**	**1.0**
JASS	2020	0	0	0	0	0	1	1	2	**4**	**1.0**
MRP	2021	0	0	0	0	0	0	2	0	**2**	**0.7**
PolarMorphism	2022	0	0	0	0	0	0	0	1	**1**	**0.5**
PAT	2022	0	0	0	0	0	0	0	1	**1**	**0.5**
HSVS-M	2022	0	0	0	0	0	0	0	0	**0**	**0.0**
MTAFS	2023	0	0	0	0	0	0	0	0	**0**	**0.0**

Publ Year: Publication Year; Tot Cit: Total number of citations; Avg Cit: Average number of Citations from publication year

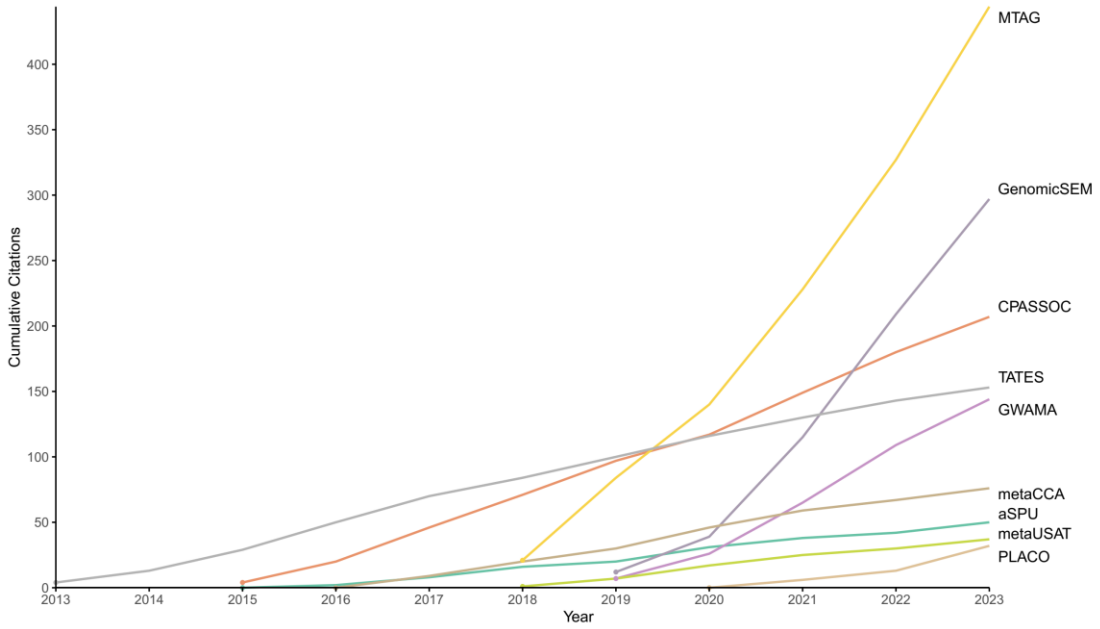


Figure 6.2 Trend of the total citations over time for the most cited studies

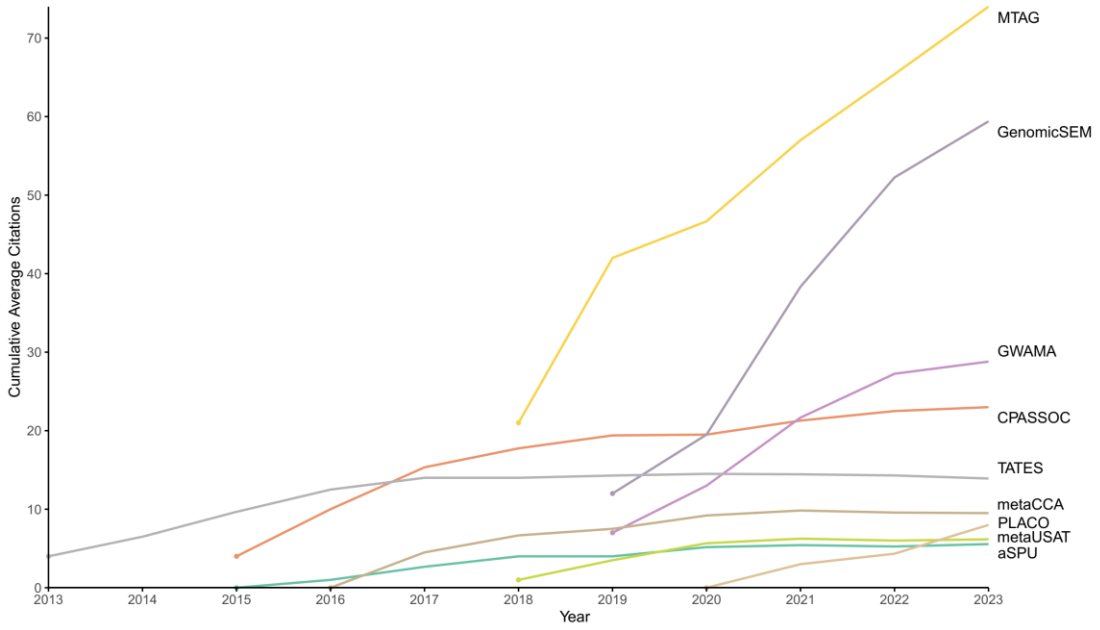


Figure 6.3 Trend of the cumulative average number of citations over time for the most cited studies

6.3.3 Overview of MTAG

Multi-Trait Analysis of GWAS (MTAG) (Turley et al., 2018) can boost statistical power combining information from the correlation of the variant's effects and from the correlation of the estimation error of the variant's effect to account for sample overlap and biases such as population stratification or cryptic relatedness. This method models the shared genetic structure between traits in three steps. First, it estimates the variance-covariance matrix of the estimation errors of genetic effects using an LD score regression approach (LDSC) (Bulik-Sullivan et al., 2015). Second, it estimates the variance-covariance matrix of the genetic effects using a generalised method of moments. Third, it generates the MTAG estimates based on a weighted sum of the original GWAS estimates incorporating the information from the two estimated variance-covariance matrices.

MTAG can be considered a generalisation of inverse-variance-weighted meta-analysis that incorporates summary statistics from k number of single-trait GWAS and returns a k number of MTAG-generated trait-specific summary statistics. Thus, the MTAG statistics can be interpreted similarly with those from a univariate GWAS and can be used in subsequent post-GWAS analyses.

MTAG assumes homoscedasticity for the genetic effects across traits under a random effect model and homogeneity for the effect's estimation error across all SNPs. These are two strong assumptions, which might be violated under many scenarios, especially for variants with truly null effects on some traits and truly non-null effects on other traits, which might increase the probability of false positive associations.

6.3.4 Overview of Genomic SEM

Genomic Structural Equation Modelling (Genomic Sem) (Grotzinger et al., 2019) is a SEM approach performing in two steps. The first step includes the estimation of the empirical genetic covariance matrix and its associated sampling covariance matrix. These two covariance matrices are modified to include information from the SNPs effects.

The second step comprises a multivariate system of regression to specify a SEM and covariance associations to estimate a model-implied genetic covariance matrix. A set of parameters for these associations are selected such that the discrepancy between the model-implied and the empirical covariance matrices is minimised. In the measurement

model of SEM, the latent genetic factors are defined as linear functions based on the SNP effects on each phenotype, trying to fit the model to data. In the structural model of SEM, regression coefficients are estimated that relate latent variables to each other using another linear model.

Genomic SEM assumes that the SNPs affect phenotypes through the common factors exclusively. However, the method calculates a heterogeneity statistic of effect sizes Q_{snp}, similar to a heterogeneity Q statistic in meta-analysis, which can be used as an index measuring the violation of this assumption.

6.3.5 Overview of GWAMA

Genome-Wide Association Meta-Analysis (GWAMA) (Baselmans et al., 2019) includes two different meta-analytical approaches that can be used complementary to each other: N-weighted GWAMA (N-GWAMA) and Model-Averaging GWAMA (MA-GWAMA). Both methods can account for sample overlap, population stratification and cryptic relatedness using LDSC (Bulik-Sullivan et al., 2015).

N-GWAMA assumes a homogenous genetic effect across traits. It is a UTSC method where the multivariate statistic is generated as a sample size based weighted sum of Z that follows a standard normal distribution under the null hypothesis.

MA-GWAMA relaxes the assumption of homogenous genetic effects across traits. It is a model-averaging approach that combines results from multiple models under different assumptions about the genetic structure of the traits weighting the models based on the Akaike information criterion (AIC) (Akaike, 1979). Then, it generates trait-specific effect estimates for each SNP, which include the model averaged effect sizes for the effect of a particular variant on the trait and follow a multivariate normal distribution.

6.4 Discussion

Multivariate methods generally can leverage the genetic correlation between traits and boost the statistical power for associations between genetic variants and phenotypes. This systematic review provides a comprehensive evaluation of multivariate methods in GWAS that utilise summary statistics. Twenty-four methods were identified and classified under various multivariate categories, including UTSC, DR, linear modelling, multivariate meta-analysis and Bayesian approaches. Each of these categories exhibits unique features and follows different frameworks to generate

multivariate test statistics (Yang and Wang, 2012, Vroom et al., 2019). A citation analysis on the 24 multivariate methods indicated that MTAG, Genomic SEM and GWAMA are the methods with the highest impact.

The first multivariate methods with applications in GWAS summary statistics were mainly UTSC methods following a simpler approach. Although these methods perform different calculations to produce the combined test statistic, most of them share the same rationale of minimising the generated P-value or maximising a Z-based test statistic. However, after the development of techniques that allow the estimation of the genetic covariance between traits using GWAS summary statistics (Bulik-Sullivan et al., 2015, Yang et al., 2011), several multivariate methods were suggested that follow more complex and advanced frameworks such as MTAG and Genomic SEM.

Many of the reviewed methods show significant advantages in the investigation and interpretation of pleiotropy. They enable the discovery of shared genetic factors across traits and contribute to a deeper understanding of the genetic pleiotropic architecture of complex traits. Some multivariate methods such as MTAG and GWAMA can generate trait-specific summary statistics allowing for the incorporation of the multivariate statistics into subsequent post-GWAS analyses. Therefore, they can be more efficient in investigating pleiotropy of shared mechanisms underlying two or more conditions, providing deep insights into shared genetic mechanisms. Multivariate methods can also increase the statistical power and detect more associations compared to univariate methods. However, these benefits are fruitful under certain circumstances such as high genetic correlation between the examined traits and other assumptions related to each method separately and therefore, they should be used after careful consideration of the assumptions and complementary to univariate GWAS.

Despite their benefits, multivariate methods face several limitations. Many rely on assumptions of homoscedasticity or homogeneity of genetic effects across traits, which can be violated under certain scenarios leading to inflated false-positive rates. Additionally, these methods incorporate summary statistics only and thus, any biases of the original univariate GWAS are inherited to the multivariate methods. For the same reason, these methods do not allow to address any potential confounding factors that have not been considered in the original GWAS. Furthermore, the interpretation of results from multivariate methods is more challenging compared to univariate methods and should be done with caution. Also, multivariate methods usually show higher

probability of false positive findings and therefore, additional methods and analyses should be performed for validation.

Another limitation of this study is that it did not consider multivariate methods that use individual-level data. However, such methods have been studied previously and discussed in detail elsewhere (Vroom et al., 2019). This systematic review focused only methods that perform multivariate analysis between single SNPs and multiple phenotypes. However, there are methods that perform multivariate analysis between sets of SNPs and multiple phenotypes such as MSKAT (Guo and Wu, 2019b) and MTAR (Luo et al., 2020). Finally, the search strategy of this study included citations only from only from PubMed, thus methods not included in PubMed database might not have been considered. However, a manual search on the references of the included studies was conducted to limit the probability of non-identified studies.

This systematic review identified multivariate methods designed for single-variant and multi-trait genome-wide association analysis using GWAS summary statistics. It indicated MTAG as the most influential method within the scientific community and presented its statistical framework in detail. Additionally, it highlighted its advantages such as the generation of trait-specific summary statistics that can be used in post-GWAS analyses and pointed out its limitations that need carefully consideration in practical applications.

CHAPTER 7

MULTI-TRAIT GWAS DISCOVERS NOVEL GENETIC LOCI FOR BLOOD CELL TRAITS

7.1 Background

Haematopoiesis is a highly regulated and hierarchical biological process that involves the differentiation of primitive hematopoietic stem cells to mature blood cells. Blood cell traits are complex and highly polygenic, with strong heritability. Their study is valuable to understand the genetic architecture of haematopoiesis (Bao et al., 2019). Although GWAS have previously identified thousands of genes linked to blood traits, the genetic loci discovered so far accounts for only a small proportion of the phenotypic variance (e.g., up to 28% in European population) revealing the highly polygenic nature of these traits (Visscher et al., 2017, Timpson et al., 2018).

The shared origin of blood cells through hematopoietic process explains the strong genetic correlation between blood traits and the numerous shared mechanisms underlying them. Many of the identified genes have been associated with multiple blood cell phenotypes indicating pleiotropic effects (Vuckovic et al., 2020). While most of the previous GWAS studied multiple blood traits, they mainly employed single-trait approaches that might fail to capture the correlation between traits (Vuckovic et al., 2020, Chen et al., 2020, Astle et al., 2016). In contrast, multi-trait approaches can account for these correlations offering several advantages such as increased statistical power (Turley et al., 2018, Zhou and Stephens, 2014, Grotzinger et al., 2019, Baselmans et al., 2019). These methods have the potential to complement and extend the findings from previous single-trait GWAS when applied to complex traits with shared genetic architecture such as blood cell phenotypes.

This study sought to identify novel genetic loci associated with blood cell traits and further explore their genetic architecture. It included multi-trait GWAS on fifteen blood cell traits, gene, pathway, gene expression and colocalisation analyses. The findings

extend the current knowledge and highlight target variants with broad effects across multiple phenotypes.

7.2 Methods

7.2.1 Study Design & Population Sample

The summary statistics from 15 GWAS on blood cell traits were used including 6 GWAS on traits related to white blood cells: basophil count (BAS), eosinophil count (EOS), neutrophil count (NEU), lymphocyte count (LYM), monocyte count (MON), white blood cell count (WBC); 7 GWAS on traits related to red blood cells: haematocrit (HCT), haemoglobin (HGB), mean corpuscular haemoglobin (MCH), mean corpuscular haemoglobin concentration (MCHC), mean corpuscular volume (MCV), red blood cell count (RBC), red distribution width (RDW); and 2 GWAS related to platelets: platelet count (PLT), mean platelet volume (MPV). Details on the population characteristics of the included GWAS are described elsewhere (Vuckovic et al., 2020).

7.2.2 Genotypic Quality Control

Initially, approximately 47m genotyped and imputed SNPs were included for each trait. Insertions, deletions, rare variants (MAF < 0.01), variants in sex chromosomes, variants with sample size less than the 2/3 of the 90th percentile, ambiguous/palindromic SNPs and non-overlapping SNPs across the analysed traits were excluded from the analysis. After exclusions, approximately 8.6m SNPs were retained for the analysis.

7.2.3 Multi-Trait Genome-Wide Association Analysis

The 15 blood traits GWAS were categorised into three sets: 1) red blood cell-related traits, 2) white blood cell-related traits and 3) platelet-related traits. The traits of each set were jointly analysed using MTAG (Turley et al., 2018) resulting in three separate MTAG analyses. For each analysis, the genetic correlation between the respective traits was calculated and bivariate LD score regression was applied to account for sample overlap. Fifteen GWAS datasets were generated in total by MTAG, with each dataset containing the trait-specific summary statistics, allowing for an interpretation in a similar way with those from a univariate single-trait GWAS (Turley et al., 2018).

7.2.4 Functional Mapping and Annotation

The Functional Mapping and Annotation of GWAS (FUMA) platform (Watanabe et al., 2017) was used to functionally analyse the 15 trait-specific summary statistics from MTAG. For each trait, the genome-wide significance (GWS) threshold was set at $P = 5\times10^{-9}$ and clumping was performed twice using different r^2 thresholds: initially, at $r^2 < 0.6$ to determine the borders of the LD blocks and subsequently, at $r^2 < 0.1$ to identify independent signals. The GWS SNPs with $r^2 < 0.6$ were included for further annotation and gene prioritisation, while those at $0.1 \leq r^2 < 0.6$ were assigned to the same LD block. LD blocks closer than 250kb were merged into one genomic region. The lead variant of each genomic locus was the GWS independent SNP with the smallest MTAG P-value. The European sample from 1000 Genome Project Phase 3 was used as a reference panel to calculate pairwise LD between variants using PLINK v1.96. ANOVAR (Wang et al., 2010) was used to positionally map the SNPs to their nearest protein coding genes, based on Ensembl build v92, with a maximum distance of 10kb.

7.2.5 Novel Genetic Risk Regions Identification

As the main aim was to identify novel loci, genomic regions previously identified to be associated with the trait of interest were excluded from the annotation analysis. Specifically, variants within 500kb from the top signals of previous GWAS (Vuckovic et al., 2020) for the respective traits were excluded from further consideration.

7.2.6 Gene-Based, Gene-Set and Tissue Expression Analysis

We performed gene, gene-set and gene-property analysis on 15 trait specific summary results from MTAG using Multi-marker Analysis of GenoMic Annotation software (MAGMA v1.08)9. In the gene analysis, we used the 1000 Genomes phase3 to prioritise the genes which the GWA significant SNPs were located in. We implemented a SNP-wide mean model for gene tests following the default parameters of the FUMA pipeline. In the gene-set analysis, we tested more than 15,000 different gene-sets derived from MsigDB v7.010.

We also performed a tissue expression analysis to test possible associations between prioritised genes and tissue-specific gene expression profiles using the Genotype-Tissue Expression version 8 (GTEx v8) for 30 general tissue types. The gene, gene-set and gene-property analyses were all corrected using a Bonferroni significance threshold.

7.2.7 Colocalisation Analysis

A multi-trait colocalisation analysis was conducted on the loci identified by MTAG to prioritise potential shared causal variants across blood traits. A Bayesian divisive clustering algorithm was applied, as implemented in HyPrColoc v.1.0.0R8. The analysis investigated regions ±200kb from top signals using the MTAG summary statistics. Variant-specific priors were applied under the assumption that the probability of a variant colocalising with a set of traits decreases as the number of traits increases. A prior probability for an association between a variant and a single trait was set at $P = 1\times10^{-4}$ while a conditional prior probability for an association with an additional trait at $P = 0.02$. Only regions with strong evidence of colocalisation (Posterior Probability, PP > 0.75) where the shared association was sufficiently explained by one single causal variant (proportion of PP explained by the SNP, %PP ≥ 80%) were considered.

For colocalised loci with a candidate shared causal variant, an additional colocalisation analysis was performed using expression quantitative trait loci (eQTL) in the most significantly associated tissues identified in tissue expression analysis in order to detect shared genes across blood traits. The eQTL information was derived from Genotype Tissue Expression version 7 (GTEx v7). The same parameters with multi-trait colocalisation analysis were applied as described earlier.

7.3 Results

7.3.1 Multi-Trait GWAS and Novel Loci of Blood Cell Traits

Three distinct MTAG analyses were conducted using: 1) seven red blood cell-related traits (RBC-MTAG), 2) six white blood cell-related traits (WBC-MTAG), and 3) two platelet-related traits (PLT-MTAG). The three analyses identified 3,100 novel conditionally independent trait-variant associations ($P < 5\times10^{-9}$) involving 2,317 unique SNPs, assigned to 711 genomic loci (**Figure 7.1**). The number of loci per blood cell phenotype ranged from 47 to 141. Specifically, 1,447 associations (302 loci) concerned RBC-related traits, 1,188 (374 loci) WBC-related traits, and 465 (137 loci) PLT-related traits. Out of the 711 identified loci, 29 (corresponding to 53 trait-locus associations) had not been previously linked to any blood trait in populations of European ancestry (**Table 7.1**).

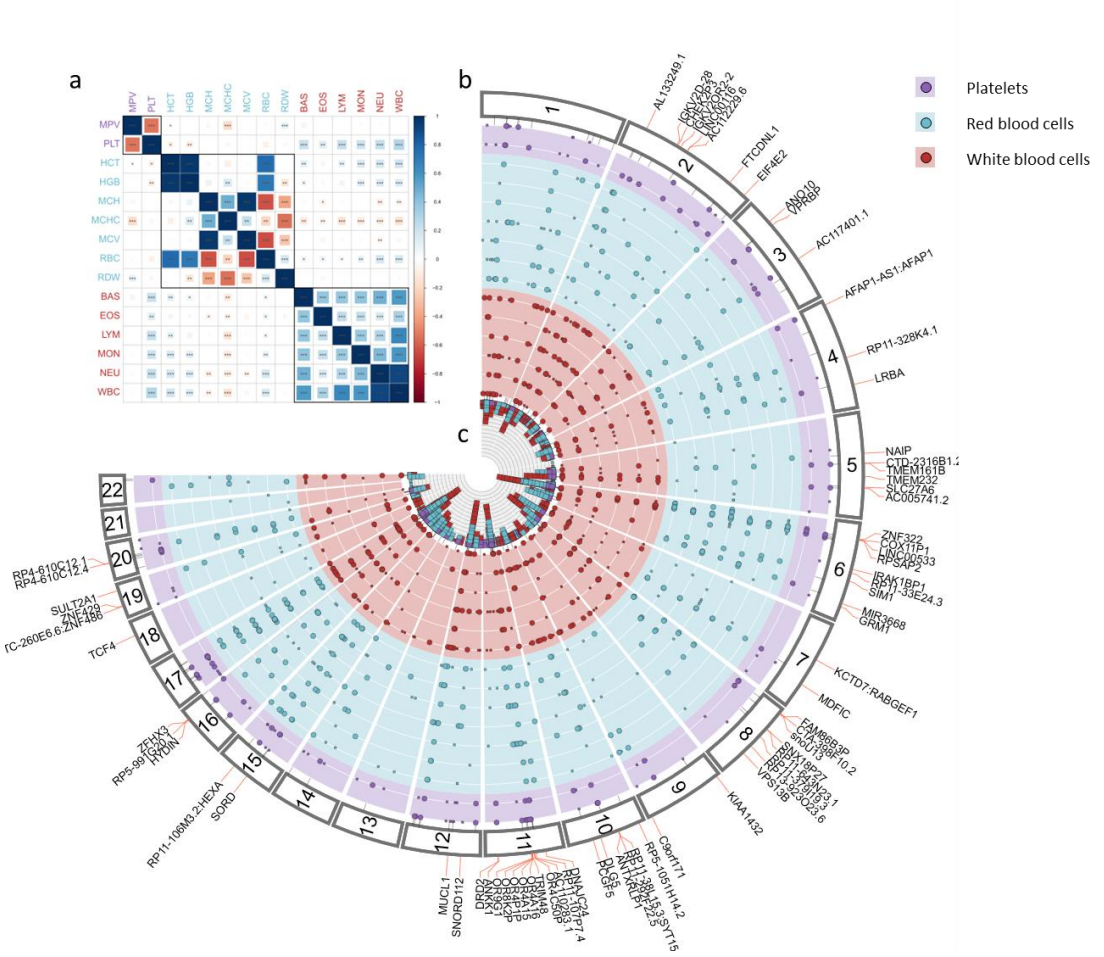


Figure 7.1 Results from MTAG on fifteen blood cell traits

a) Heatmap presenting the genetic correlation between traits. b) Circular plot showing the identified novel loci for each trait. The size of the dots indicates if the locus is shared across multiple traits (large) or is unique for each trait (small). The annotations show the identified genes associated with at least two different blood cell types.

MPV: Mean platelet volume, PLT: Platelet count, HCT: Haematocrit, HGB: Haemoglobin, MCH: Mean corpuscular haemoglobin, MCHC: Mean corpuscular haemoglobin concentration, MCV: Mean corpuscular volume, RBC: Red blood cell count, RDW: Red distribution width, BAS: Basophil count, EOS: Eosinophil count, NEU: Neutrophil count, LYM: Lymphocyte count, MON: Monocyte count, WBC: White blood cell count

Table 7.1 Novel loci that were not previously associated with any blood cell trait

Locus	Gene	SNP	Chr	Pos	OA	EA	MAF	Trait	*P*	Beta
1	*AL133249.1*	rs112881196	2	31982811	C	G	0.03	RDW	3.9×10^{-11}	0.04
2	*IGKV2OR2-2*	rs74175482	2	97720549	C	T	0.14	EOS	4.8×10^{-25}	-0.04
3	*LINC00116*	rs62160551	2	110971661	T	C	0.02	MON	5.5×10^{-14}	-0.08
	AC112229.6	rs62162561	2	111091622	G	A	0.06	RBC	2.5×10^{-11}	-0.04
								MCH	2.0×10^{-09}	0.03
								MCV	8.9×10^{-10}	0.03
4	*ANO10*	rs142641363	3	43637642	A	C	0.04	MON	1.4×10^{-15}	0.05
5	*VPRBP*	rs147715078	3	51490257	G	A	0.08	EOS	7.7×10^{-22}	-0.04
6	*AC117401.1*	rs149366143	3	123773622	G	A	0.01	MPV	2.5×10^{-68}	0.18

								PLT	1.3×10^{-15}	-0.08
7	*RP11-328K4.1*	rs6844726	4	104244730	C	G	0.05	LYM	1.7×10^{-09}	-0.03
8	*CTD-2316B1.2*	rs2037852	5	87237448	T	C	0.23	RBC	1.5×10^{-09}	0.02
	TMEM161B	rs4916889	5	87439691	C	T	0.32	MCH	8.7×10^{-12}	-0.02
								MCV	2.4×10^{-13}	-0.02
9	*TMEM232*	rs72788433	5	109814915	C	A	0.10	EOS	2.5×10^{-10}	0.02
10	*SLC27A6*	rs17790915	5	128118887	G	A	0.15	RDW	2.3×10^{-33}	0.04
11	*ZNF322*	rs77093797	6	26668278	A	G	0.06	MCHC	5.3×10^{-302}	0.17
	COX11P1	rs6902687	6	28413491	T	C	0.31	RBC	9.3×10^{-24}	-0.02
	LINC00533	rs114179634	6	28626101	C	G	0.08	MCH	7.0×10^{-305}	0.16
	RPSAP2	rs6456834	6	28700352	G	T	0.16	PLT	2.1×10^{-16}	-0.02
12	*MIR3668*	rs4896507	6	140563842	A	T	0.10	MCH	2.5×10^{-10}	0.02
								MCV	4.0×10^{-12}	0.02
13	*GRM1*	rs76173020	6	146649257	A	T	0.02	MPV	1.9×10^{-11}	0.04
14	*KCTD7:RABGEF1*	rs12698545	7	66203017	T	G	0.24	NEU	2.2×10^{-12}	0.02
		rs80289808	7	66219179	C	T	0.05	WBC	6.6×10^{-13}	-0.04
15	*RP11-379I19.3*	rs78521489	8	60581490	T	C	0.21	NEU	1.6×10^{-11}	0.02
								WBC	2.7×10^{-11}	0.02
16	*VPS13B*	rs76855829	8	100051894	G	A	0.04	MCH	3.9×10^{-09}	-0.03
		rs78889654	8	100163721	C	G	0.04	MCV	1.7×10^{-10}	-0.04
17	*C9orf171*	rs117516010	9	135314229	C	A	0.03	MPV	3.0×10^{-10}	-0.04
18	*RP5-1051H14.2*	rs78198807	10	9800835	C	T	0.03	EOS	3.1×10^{-09}	0.03
19	*RP11-38L15.3:SYT15*	rs7914056	10	46962991	G	A	0.48	MCH	2.2×10^{-11}	-0.02
								MCV	4.2×10^{-12}	-0.02
20	*RP11-292F22.5*	rs74622866	10	47541708	G	A	0.04	MCH	4.5×10^{-16}	0.05
								MCV	7.4×10^{-18}	0.05
	ANTXRLP1	rs139283651	10	47603424	G	A	0.01	RDW	3.5×10^{-12}	0.08
21	*DNAJC24*	rs6484504	11	31424823	T	C	0.26	HCT	9.0×10^{-20}	0.02
								HGB	5.8×10^{-19}	0.02
								RBC	5.2×10^{-20}	0.02
22	*OR8K2P*	rs4939033	11	56104645	C	T	0.35	EOS	1.4×10^{-12}	-0.02
	OR9G1	rs111517495	11	56478951	C	T	0.11	MCV	4.7×10^{-11}	-0.02
23	*ANKK1*	rs45612940	11	113268660	A	C	0.02	PLT	1.2×10^{-10}	0.05
	DRD2	rs4319541	11	113451055	T	C	0.07	MPV	1.7×10^{-09}	-0.03
24	*SNORD112*	rs11052592	12	33433752	C	T	0.01	MPV	1.9×10^{-09}	0.05
25	*MUCL1*	rs73124517	12	55268789	T	A	0.01	RDW	8.2×10^{-11}	0.06
26	*SORD*	rs3874239	15	45331689	C	G	0.11	MCHC	1.1×10^{-09}	0.02
		rs4774514	15	45374055	C	T	0.05	RDW	2.3×10^{-10}	0.03
27	*RP5-991G20.1*	rs34889159	16	72747685	T	A	0.12	RBC	1.4×10^{-09}	0.02
	ZFHX3	rs62053161	16	72877116	G	A	0.10	HGB	2.7×10^{-10}	0.02
28	*ZNF486*	rs118091526	19	20290149	T	C	0.01	PLT	7.6×10^{-20}	0.09
29	*RP4-610C12.4*	rs192359675	20	29512487	C	T	0.01	EOS	9.3×10^{-20}	-0.07
								PLT	8.5×10^{-40}	0.11
	RP4-610C12.1	rs2379715	20	29597785	C	T	0.43	MCH	4.8×10^{-14}	-0.02

OA: Other allele; EA: Effect allele; MAF: Minor allele frequency

The effects of the 2,317 blood-related SNPs indicated by MTAG were investigated across blood traits. Among these, 1,794 variants (550 loci) were associated ($P < 5\times10^{-9}$) with at least two traits, while 5 of them (2 loci) with up to fourteen traits. The variant

rs657197 (mapped to *ATXN2*) was associated with all traits at $P < 5\times10^{-7}$. Additionally, 48 variants (5 loci) were associated with all RBC-related traits, 210 variants (22 loci) with all WBC-related traits and 268 variants (80 loci) with all PLT-related traits. Furthermore, 13 variants (2 loci) were associated with all RBC and all WBC-related traits, 14 variants (2 loci) with all RBC and all PLT-related traits and 35 variants (5 loci) with all WBC and all PLT-related traits.

7.3.2 Genetic Pathways of the Novel Loci

Gene analysis identified 4,988 trait-gene associations at the Bonferroni significance threshold ($P < 4\times10^{-6}$). Across these associations, there were 2,557 unique genes linked to one or more blood cell traits. The number of the associated genes per trait ranged from 221 to 440. Among these genes, 1,062 were associated with two or more traits and 68 of them, showed associations with phenotypes of all blood cell types. Three genes within the major histocompatibility complex region (*GPX6*, *SCAND3*, and *TRIM27*), were associated with all fifteen examined blood cell traits.

Pathway analysis revealed 302 trait-specific associations involving 221 different gene-sets at the Bonferroni significance threshold ($P < 3.3\times10^{-6}$). The number of gene-sets per trait ranged from 2 to 70. Nine blood cell traits were positively associated with the downregulation of a gene-set in fibroblasts related to expressing mutant forms of *ERCC3* after ultraviolet irradiation while 5 other blood cell traits were positively associated with the upregulation of a gene-set in brain involved in the pathway of Alzheimer's disease.

Tissue expression analysis across 30 tissue types identified 43 Bonferroni significant trait-specific associations ($P < 1.7\times10^{-3}$) with 8 tissues (**Supplementary Table 1**). The most frequently associated tissues were blood and spleen (each associated with 12 blood cell traits), followed by lung (with 8 blood cell traits). The strongest associations were observed between white blood cell-related traits and spleen, blood, and lung.

7.3.3 Colocalisation

A multi-trait colocalisation analysis was performed on the 711 identified loci to prioritise variants with broad causal effects across blood traits. This analysis detected 9 variants with broad effect on 10 or more blood traits (**Figure 7.2**, **Supplementary Table 2**). The exonic variant rs3811444 (*TRIM58:OR2W3*) was the only one that colocalised with as many as 14 blood traits (all but MPV, PP = 0.88) explaining 100%

of PP of the shared association. Also, the intronic variant rs10849925 (*CUX2*) colocalised with 13 blood traits (all but MPV and MCV, PP = 0.76, 100% of PP explained by SNP). Furthermore, 11 variants showed global blood cell type-specific effects on all WBC-related (**Supplementary Table 3**), while 4 other variants on all RBC-related traits (**Supplementary Table 4**).

Figure 7.2 Results from colocalisation analysis

The plot presents the candidate variants with broad effects across blood cell traits. The dots show a variant-trait association. The rectangles on top show the Posterior Probability (PP), while the circles on bottom the percentage of the PP which is explained by the variant.

MPV: Mean platelet volume, PLT: Platelets count, HCT: Hematocrit, HGB: Hemoglobin, MCH: Mean corpuscular hemoglobin, MCHC: Mean corpuscular hemoglobin concentration, MCV: Mean corpuscular volume, RBC: Red blood cells count, RDW: Red distribution width (RDW), BAS: Basophils count, EOS: Eosinophils count, NEU: Neutrophils count, LYM: Lymphocytes count, MON: Monocytes count, WBC: White blood cells count

To identify shared genes across blood traits that are expressed in the three most strongly associated tissues from tissue expression analysis (whole blood, spleen and lung), a subsequent eQTL colocalisation analysis was conducted. This analysis on the 9 loci with broad effects detected 16 variant-gene-tissue associations in 4 loci highlighting 11 shared genes and 5 candidate causal variants (**Supplementary Table 5**). The variant rs7953257 exhibited the broadest effect, colocalising with 11 traits, with lower mean

levels of the blood traits and higher expression of *PHETA1* (PP = 0.85, 100% of PP explained by SNP) and higher expression of *ALDH2* (PP = 0.89, 100% of PP explained by SNP) in whole blood.

Another variant with a broad effect, rs615632, colocalized with a different set of 11 traits and was associated with higher mean levels of nine traits and higher expression of *ERI1* in lung (PP = 0.78, 100% of PP explained by SNP). Also, rs9977672 was the causal variant of a shared association for a set of 10 traits leading to lower mean levels of 9 blood traits and higher expression of *ETS2* in whole blood (PP = 0.91, 95% of PP explained by SNP) and spleen (PP = 0.91, 99% of PP explained by SNP).

The eQTL colocalisation on the 11 loci with WBC-specific global effects identified 5 variant-gene-tissue associations involving 4 genes in 3 loci and highlighted 4 candidate causal variants (**Supplementary Table 6**). The intronic variants rs11738827 showed a global causal effect on all white blood cell traits and was an eQTL for *SLC22A5* expressed in whole blood (PP = 1, 100% of PP explained by SNP) and for *NEDD8* expressed in spleen (PP = 0.81, 83% of PP explained by SNP). Another eQTL colocalisation on the 4 loci with RBC-specific global effects showed 5 variant-gene-tissue associations of 4 genes in 2 loci and prioritised 2 candidate causal variants (**Supplementary Table 7**). The intronic variant rs143875230 was an eQTL for *LCMT2* in spleen (PP = 0.76, %PP = 100%) and lung (PP = 0.8, 100% of PP explained by SNP) as well as for *ADAL* in lung (PP = 0.94, 100% of PP explained by SNP).

7.4 Discussion

Three multi-trait GWAS analyses were conducted using MTAG on red blood cell-related traits, white blood cell-related traits and platelet-related traits discovering 711 novel blood cell loci and highlighting 221 gene-sets associated with blood traits. Through subsequent colocalisation analyses, we detected 9 loci with broad causal effects across at least 10 blood traits, 11 loci with global effects on white blood cells and 4 loci with global effects on red-blood cells through shared genes expressed in blood-related tissues.

We leveraged the close genetic link and high genetic correlation between blood traits to boost statistical power and increase the number of identified loci. Particularly, this study discovered 29 loci that have not been previously associated with any blood trait

in European populations. Several genes within these 29 loci are involved in key cellular pathways, such as lipid metabolism and membrane dynamics, signal transduction or cellular communication, immune response, and protein homeostasis. These results highlight the diverse genetic mechanisms that affect blood traits through pathways critical to cellular functions. For instance, the variant rs657197 (*ATXN2*) demonstrated associations with all traits at $P < 5\times10^{-7}$. *ATXN2*, with an important role in RNA metabolism and regulation (Costa et al., 2024), has previously been associated with various blood cell traits. This study extended these associations and included associations with six additional traits. These findings are complementary to previous GWAS expanding the current knowledge of the genetic architecture of blood traits.

To gain a deeper understanding of the underlying mechanisms behind the novel findings, this study implemented a versatile analytical framework including colocalisation and pathway analyses. Several of the identified variants were associated with multiple blood traits possibly indicating a fundamental role of those loci with generic effects on eukaryotic cells or a blood-specific impact on premature blood cells at earlier stages of haematopoiesis. For example, the variant rs7953257 colocalised and was positively associated with eleven blood traits tissue and expression of PH domain containing endocytic trafficking adaptor 1 (*PHETA1*) in whole blood. This gene encodes a protein of the endocytic trafficking which is required for receptor recycling from endosomes (Noakes et al., 2011). Dysfunctionalities of this protein could affect several cellular activities including those of blood cells. This gene has previously been associated with eosinophil (Kichaev et al., 2019) and platelet count (Astle et al., 2016). This study expanded the associations for this gene with nine additional blood traits including all white blood cell traits. PHETA1, as an important part of the endosome function, plays a significant role in endocytic trafficking and hence, in several cellular activities. The same variant was additionally colocalised with higher mean levels for a set of eleven blood traits and higher expression of *ALDH2*. Aldehyde dehydrogenase 2 (*ALDH2*) encodes an enzyme essential for the oxidative pathway of alcohol metabolism (LeFort et al., 2024). The findings of this study support that this gene plays a significant role in haematopoiesis and the maintenance of blood cell integrity. Increased *ALDH2* expression may enhance the detoxification of harmful aldehydes, thereby promoting healthy blood cell function. This aligns with previous studies highlighting *ALDH2*'s

critical role in cellular protection and longevity (Wu and Ren, 2019, Ajoolabady et al., 2024).

Another variant exhibiting broad effects, rs615632, colocalised with 11 traits and was identified as an eQTL for *ERI1* in the lung. *ERI1* (Exoribonuclease 1) plays a vital role in RNA processing and degradation, particularly in regulating histone mRNA and rRNA. Increased *ERI1* expression may enhance the stability and functionality of essential RNA molecules, thereby supporting efficient cell cycle progression and optimal function of hematopoietic cells.

The exonic variant rs3811444, mapped within the genes tripartite motif containing 58 (*TRIM58*) and olfactory receptor family 2 subfamily W member 3 (*OR2W3*), colocalised with all blood traits except for the mean platelet volume. However, no shared gene was detected when using eQTL in whole blood or spleen tissues, possibly indicating either a horizontal pleiotropy of that locus on blood traits via different genes or an indirect association via shared genes expressed in other tissue types.

This study had several notable strengths. Firstly, its design ensured enhanced statistical power by employing appropriate multivariate methods on large-scale GWAS data encompassing over 500,000 participants. Secondly, it integrated advanced genetic epidemiology approaches, including multivariate analysis and colocalisation analysis. However, the study also had limitations. The analysis was confined to individuals of European ancestry, which may limit the generalisability of the findings to other populations. Additionally, the investigation excluded rare variants (MAF < 1%), leaving a significant portion of genetic predisposition unexplored. Furthermore, some associations might reflect false-positive findings or biased results. Finally, as MTAG utilised summary statistics, it was not possible to adjust for potential confounders.

In summary, using the largest GWAS on 15 blood traits to date, this study provides several novel genetic loci implicated in blood physiology. It also offers new insights into mechanisms underlying haematopoiesis, emphasising loci with potential effects at various stages of hematopoietic process.

CHAPTER 8

SHARED GENETIC ARCHITECTURE OF INFLAMMATION AND CARDIOMETABOLISM: A MULTI-TRAIT GWAS

8.1 Background

C-reactive protein (CRP) is an acute phase reactant secreted mainly by the liver and released in high concentrations in blood as a protective response to harmful irritants such as pathogens or damaged tissue. Genetic and environmental determinants of CRP levels have been widely studied as its levels have been associated with plethora of phenotypes and diseases including disorders of the immune system, cardiovascular diseases (CVD) and endocrine/metabolic-related disorders (Markozannes et al., 2021).

Genome-wide association studies (GWAS) on CRP levels have identified numerous, robustly associated, genetic loci associated with CRP levels (Dehghan et al., 2011) supporting a polygenic model for this trait. Several of the identified CRP loci are annotated not only to inflammation-related genes but also to lipid and glucose metabolism-related genes (Ligthart et al., 2018), providing evidence towards common biological pathways between inflammation and metabolic traits.

Here, we investigate further the pleiotropic nature of the genetic architecture of CRP. We performed MTAG on CRP levels with established cardiometabolic risk factors followed by a series of in silico analyses to identify novel pleiotropic genes, their tissue site of action and evidence for causal associations with a range of disease outcomes. Ultimately, the study provides additional biological insights into low-grade inflammation and highlights biological pathways that are likely to link inflammation to different chronic diseases.

8.2 Methods

8.2.1 Study Design & Population Samples

The summary statistics from six GWAS were used in this study: CRP (Said et al., 2022), high-density lipoprotein (HDL) levels (Neale lab team, 2018), low-density lipoprotein (LDL) levels (Neale lab team, 2018), triglyceride (TG) levels (Neale lab team, 2018), body mass index (BMI) (Pulit et al., 2019) and cigarettes per day (CPD) (Liu et al., 2019). The largest GWAS in European population for each trait was selected. All included GWAS have were conducted exclusively or partially in UK Biobank (UKB).

8.2.2 Genotypic Quality Control

The initial datasets contained 11,140,987 SNPs for CRP; 13,791,468 SNPs for HDL; 13,791,468 SNPs for LDL; 13,791,468 SNPs for TG; 27,381,303 SNPs for BMI and 12,003,614 SNPs for CPD. The datasets included both genotyped and imputed SNPs. SNPs that were any of insertions, deletions, rare (MAF < 0.01), with a sample size less than 2/3 of the 90^{th} percentile or palindromic were excluded from the analysis. Moreover, non-overlapping SNPs that were not present in at least one dataset were removed. After these exclusions, 6,206,408 SNPs remained for analysis.

8.2.3 Multi-Trait GWAS

A multi-trait GWAS analysis was performed using MTAG (Turley et al., 2018) to jointly analyse the summary statistics of the traits of interest. Genetic correlations between the traits were calculated and adjusted for sample overlap using bivariate LD score regression. The MTAG-generated summary statistics are trait-specific and can be interpreted similar to those from univariate analysis. Moreover, three additional bivariate MTAG were performed between CRP and lipids, CRP and BMI, and CRP and CPD, following a similar approach as described above.

8.2.4 Functional Mapping and Annotation

The MTAG results were subjected to functional analysis using Functional Mapping and Annotation of GWAS (FUMA) (Watanabe et al., 2017). Fourteen datasets were analysed: 6 from the multi-trait MTAG, 4 from the CRP-lipids MTAG, 2 from the CRP-BMI MTAG, and 2 from the CRP-CPD MTAG. The GWS SNPs ($P < 5\times10^{-8}$) were identified and clumped two times at different r^2 thresholds. The first clumping at $r^2 < 0.6$ was used to define the genomic risk loci coordinates. The second clumping at $r^2 <$

0.1 was performed to identify independent signals. SNPs in LD at $0.1 \leq r^2 < 0.6$ were grouped into the same LD block. LD blocks closer than 500 kb were merged. The top lead SNPs were those with the smallest *P*-value within genomic risk regions. Pairwise LD between SNPs was calculated with PLINK v1.923 using the European sample of 1000 Genome Project Phase 3 as a reference panel. SNPs were mapped to their nearest protein-coding genes from Ensembl build v92 within a maximum distance of 10kb using ANNOtate VARiation (ANNOVAR) (Wang et al., 2010).

8.2.5 Identification of Novel CRP Loci

CRP was included in four MTAG analyses (one multivariate and three bivariate). Independent CRP signals from all four analyses were combined and duplicate signals or proxies (either in distance ± 500 kb or in LD $r^2 > 0.1$) were excluded. The resulting top CRP signals were compared to previously identified (Said et al., 2022) and signals from this study either further than 500 kb or $r^2 < 0.1$ from previously reported signals were considered novel.

8.2.6 Gene-Based, Gene-Set and Tissue Expression Analysis

Gene, gene-set, and gene-property analyses were conducted using the Multi-marker Analysis of GenoMic Annotation software (MAGMA v1.08) (Mathur et al., 2018). In the gene analysis, the included SNPs were mapped to genes using the 1000 Genomes phase 3 reference panel. The gene-set analysis tested 15,477 gene-sets obtained from MsigDB v7.024. A Bonferroni approach was implemented to correct both gene (*P*-value threshold: 2.7×10^{-6}) and gene-set analysis (*P*-value threshold: 3.2×10^{-6}) for multiple tests.

The tissue expression analysis between tissue-specific gene expression profiles and the genes associated with the examined traits included 30 general and 53 more specific tissue types using data from Genotype-Tissue Expression version 8 (GTEx v8) and implementing a Bonferroni correction.

8.2.7 Colocalisation

A multi-trait colocalisation analysis was conducted using HyPrColoc R package (Foley et al., 2021). This approach was employed to detect colocalised loci between CRP and any combination of the other examined traits and to prioritise candidate causal variants that can explain the shared association in genomic regions associated with CRP. This

analysis was performed on each CRP locus identified by MTAG in a region ±200 kb around the top SNP. Variant-specific priors were incorporated, reflecting the assumption that the probability of a variant being colocalised with a set of traits decreases as the number of traits in the set increases. Two priors were specified for this model: the prior probability of a variant being associated with a single trait ($P = 1\times10^{-4}$) and the conditional prior probability of a variant being associated with an additional trait given its association with another trait ($Pc = 0.02$). A PP > 0.8 was considered evidence for colocalisation, while variants explaining the highest percentage of the shared association within a region were considered candidate causal variants as long as the percentage was > 80%.

8.2.8 Investigation of Direction of Genetic Effect and PheWAS

Variants colocalised with CRP and any combination of the other examined traits were further analysed to examine the direction of their genetic effects across all examined traits. Opposite directions of effect between CRP and any of LDL, TG, BMI, and CPD or the same directions between CRP and HDL were considered discordant if the associations were statistically significant for both traits ($P < 0.05$).

A PheWAS analysis was performed in the UKB for each SNP with discordant direction of effect. The analysis was restricted to participants of European ancestry and one participant from each pair of relatives (kinship coefficient > 0.0884) was randomly excluded. After quality control, 424,439 individuals were included in the analysis.

Data from the inpatient Hospital Episode Statistics (HES) records, cancer registries, and death registries were used, with diagnoses coded according to the World Health Organisation's (WHO) International Classification of Diseases, 9th Revision (ICD-9) and 10th Revision (ICD-10). These codes were translated into the phecode grouping system as implemented in the PheWAS R package (Carroll et al., 2014). For each phecode, case-control groups were defined, with cases identified as individuals meeting the specific phecode criteria and controls defined as those with no record of the corresponding outcome or related phecodes.

To ensure sufficient statistical power, only phecodes with at least 200 cases were analysed, as recommended by previous simulation studies (Verma et al., 2018). Logistic regression models were applied, adjusting for age, sex, and the first 15 genetic principal

components. To account for multiple testing and minimise false positives, the false discovery rate (FDR) method was applied (Benjamini and Hochberg, 1995).

8.2.9 Mendelian Randomisation

MR analysis was conducted to explore potential causal relationships between CRP and selected outcomes with plausible biological connections. Multiple MR analyses were performed, included the six examined traits as exposures (CRP, HDL, LDL, TG, BMI, and CPD) with 14 selected outcomes, including circulatory system diseases, neoplasms, mental disorders, and metabolomic traits (**Supplementary Table 8**). For each exposure, the genetic instrumental variables were the independent SNPs (LD $r^2 < 0.1$, $P < 5\times10^{-8}$) identified for the respective trait through multivariate MTAG.

The primary analysis employed the IVW method under a random-effects model (Burgess et al., 2013). Two additional MR methods were performed as sensitivity analyses: the weighted median MR (Bowden et al., 2016) and the MR-Egger (Bowden et al., 2015). To investigate a potential pleiotropic influence, a subgroup MR analysis was also conducted on two groups: one comprising colocalised CRP SNPs only, and another including non-colocalised SNPs that were non-significant for all other examined traits (MTAG *P*-value > 0.01).

8.3 Results

8.3.1 Multi-Trait GWAS and Novel CRP Loci

A schematic for the study design is provided in **Figure 8.1**. Multi-trait MTAG was performed to analyse CRP and five cardiometabolic risk factors including HDL levels, LDL levels, TG levels, BMI and CPD. Additionally, bivariate MTAG were conducted between CRP and each cardiometabolic trait. The multi-trait MTAG identified 797 independent signals across 283 genetic loci associated with CRP at GWS level ($P < 5\times10^{-8}$) (**Figure 8.2**). For the other traits, 549 independent signals (185 loci) were found for HDL, 527 (144) for LDL, 534 (173) for TG, 1,552 (740) for BMI and 108 (62) for CPD (**Figure 8.2**). Among the 797 CRP-associated variants, 295 (151 loci) were also associated with at least one other examined trait ($P < 5\times10^{-8}$) and 8 variants (8 loci) with at least 4 of the 5 traits (**Supplementary Figure 1**).

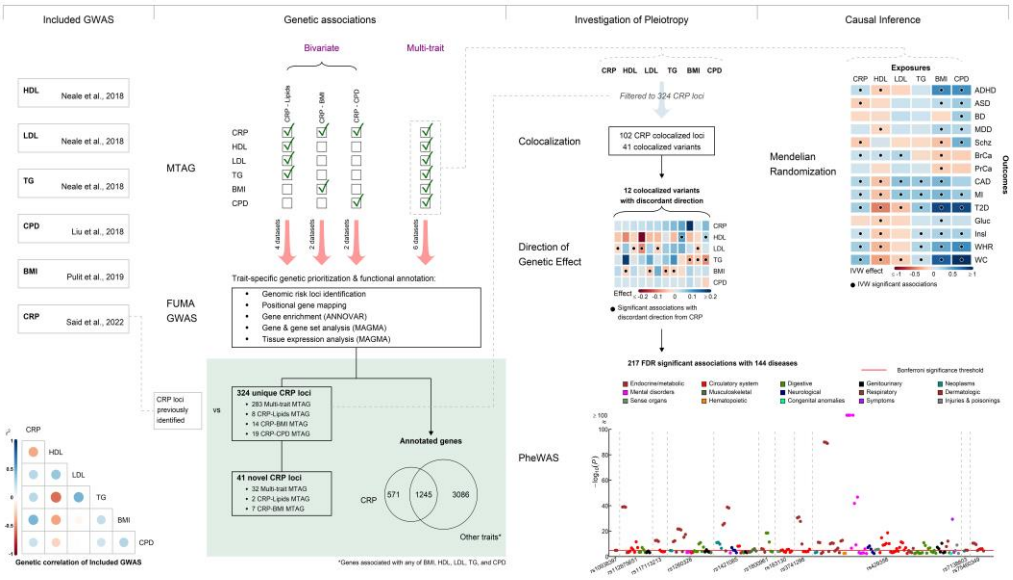


Figure 8.1 Schematic overview of the study

BMI: Body mass index, CPD: Cigarettes per day, CRP: C-reactive protein, HDL: High-density lipoprotein, LDL: Low-density lipoprotein, TG: Triglycerides, ADHD: Attention deficit hyperactivity disorder, ASD: Autism spectrum disorder, BD: Bipolar disease, MDD: Major depressive disorder, Schz: Schizophrenia, BrCa: Breast cancer, PrCa: Prostate Cancer, CAD: Coronary artery disease, MI: Myocardial infarction, T2D: Type 2 diabetes, Gluc: Fasting glucose, Insl: Fasting insulin, WHR: Waist-to-hip ratio, WC: Waist circumference

The bivariate MTAG identified 41 additional loci associated with CRP. Altogether, the 324 CRP loci (283 from multi-trait and 41 from bivariate MTAG) included 41 novel genomic loci (**Table 8.1, Supplementary Figure 2**). The most significant novel CRP locus was located in *LPL* (rs35237252, $P = 5\times10^{-20}$), which encodes lipoprotein lipase, an enzyme expressed in heart, muscle, and adipose tissue (Shang and Rodrigues, 2021).

8.3.2 Functional Annotation and Pathway Enrichment

Functional annotation identified 1,816 genes associated with CRP levels. Of these, 1,245 genes were additionally associated with one or more of the other traits, while 23 of them with all six traits, suggesting broad pleiotropic effects (**Supplementary Figure 3**). Pathway analysis revealed 19 CRP gene-sets at Bonferroni significance level ($P < 3.2\times10^{-6}$) (**Supplementary Table 9**). All examined traits were significantly associated with nucleic acid binding pathway. Tissue expression analysis showed that CRP genes were differentially expressed in the liver and pituitary, brain cerebellum and brain cerebellar hemisphere (**Supplementary Figure 4**).

Table 8.1 Novel genomic loci associated with C-reactive protein

SNP	Chr:Pos	EA	OA	EAF	Beta	P	Gene
rs545152	1:96886504	T	C	0.366	0.01	4.7×10^{-8}	*UBE2WP1*
rs6587552	1:151018861	A	G	0.229	0.01	2.5×10^{-8}	*BNIPL*
rs8024	1:201845575	A	C	0.301	0.01	4.0×10^{-8}	*IPO9*
rs11122456	1:230305966	A	G	0.391	0.01	8.8×10^{-10}	*GALNT2*
rs754524	2:21311541	G	T	0.257	0.01	8.4×10^{-9}	*TDRD15*
rs76866386	2:44075483	C	T	0.079	−0.03	1.7×10^{-10}	*ABCG8*
rs4519576	2:48966146	C	T	0.450	0.01	4.2×10^{-8}	*STON1-GTF2A1L*
rs12998606	2:188725859	G	A	0.461	−0.01	2.3×10^{-8}	*LINC01090*
rs566279474	3:44135752	C	T	0.003	0.06	1.5×10^{-8}	*MIR138-1*
rs171390	3:154038412	C	T	0.424	0.01	4.7×10^{-8}	*DHX36*
rs2606227	3:183536836	T	C	0.361	0.01	4.5×10^{-8}	*MAP6D1*
rs34811474	4:25408838	A	G	0.217	−0.02	6.3×10^{-10}	*ANAPC4*
rs1229978	4:100256199	C	T	0.406	0.01	1.7×10^{-9}	*ADH1C*
rs1450786	4:112653076	G	A	0.367	−0.01	3.7×10^{-8}	*RP11-269F21.1*
rs10461497	5:63942398	C	T	0.494	−0.01	3.3×10^{-8}	*MRPL49P1*
rs6870983	5:87697533	T	C	0.236	−0.01	7.3×10^{-9}	*TMEM161B-AS1*
rs11135450	5:95554016	A	G	0.345	−0.01	2.4×10^{-8}	*CTD-2337A12.1*
rs2228213	6:12124855	A	G	0.333	−0.01	1.8×10^{-8}	*HIVEP1*
rs2635727	6:50820940	T	C	0.250	−0.01	2.5×10^{-9}	*RPS17P5*
rs57648913	7:21602065	A	G	0.145	0.02	2.4×10^{-9}	*DNAH11*
rs35237252	8:19870271	A	C	0.273	−0.02	5.1×10^{-20}	*LPL*
rs10464844	8:106419754	G	A	0.230	0.01	2.0×10^{-8}	*ZFPM2*
rs1411432	9:16728532	C	A	0.183	0.02	1.1×10^{-8}	*BNC2*
rs10968576	9:28414339	G	A	0.302	0.01	1.1×10^{-9}	*LINGO2*
rs722564	10:118550831	T	C	0.386	−0.01	4.3×10^{-8}	*RPL5P27*
rs36089024	11:67244644	T	C	0.401	−0.01	1.6×10^{-9}	*AIP*
rs10750096	11:116656788	C	A	0.093	0.02	4.1×10^{-8}	*ZNF259*
rs7138803	12:50247468	A	G	0.338	0.01	1.5×10^{-8}	*RP11-70F11.7*
rs56205943	12:57679414	A	G	0.193	−0.01	4.9×10^{-8}	*R3HDM2*
rs825457	12:124538302	C	A	0.161	−0.02	2.7×10^{-8}	*FAM101A*
rs17522122	14:33302882	T	G	0.489	0.01	2.6×10^{-9}	*AKAP6*
rs11856579	15:78012688	A	G	0.220	−0.01	2.1×10^{-8}	*LINGO1*
rs879620	16:4015729	C	T	0.389	−0.01	1.7×10^{-9}	*ADCY9*
rs12446515	16:56987015	T	C	0.292	−0.02	1.3×10^{-19}	*AC012181.1*
rs2000999	16:72108093	A	G	0.192	0.02	9.8×10^{-9}	*TXNL4B*
rs56823429	16:81533789	C	A	0.283	0.01	3.6×10^{-10}	*CMIP*
rs77542162	17:67081278	G	A	0.011	0.05	5.4×10^{-12}	*ABCA6*
rs9951447	18:20009691	C	T	0.439	−0.01	7.3×10^{-9}	*RP11-863N1.4*
rs2236707	18:21114997	T	C	0.430	−0.01	1.0×10^{-11}	*NPC1*
rs2147338	20:50320079	C	T	0.406	0.01	2.8×10^{-8}	*ATP9A*
rs3746778	20:61341472	A	G	0.392	−0.01	3.0×10^{-8}	*NTSR1*

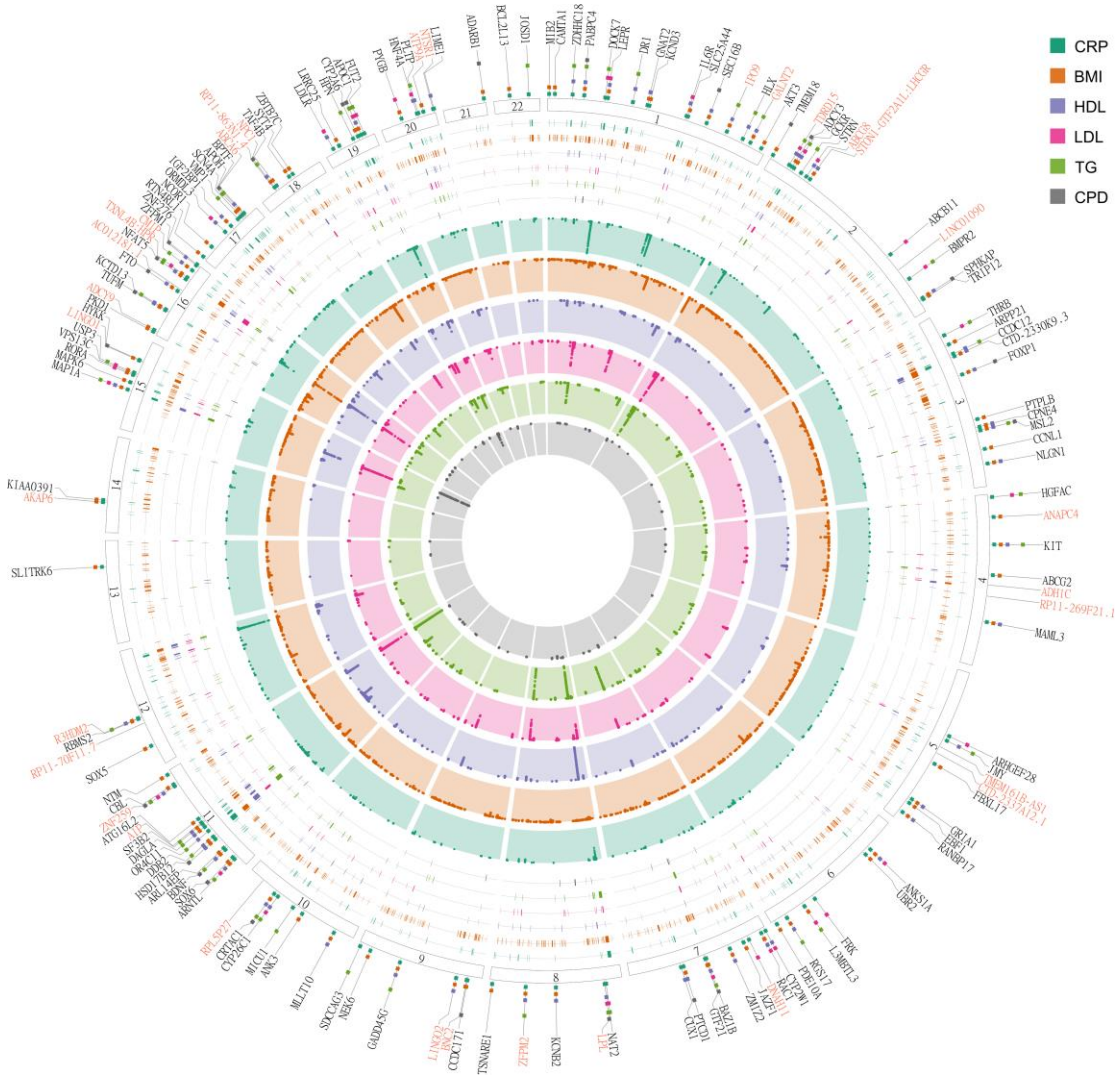


Figure 8.2 Circular plot with MTAG results

The inner part displays the Manhattan plots for each trait. The middle part highlights the genomic regions. The outer part shows the CRP genes with colourful squares on the gene lines indicating the associated traits. Red texts correspond to novel genes.

BMI: body mass index, CPD: cigarettes per day, CRP: C-reactive protein; HDL: high-density lipoprotein, LDL: low-density lipoprotein, TG: triglycerides

8.3.3 Colocalisation of CRP Loci

A colocalisation analysis on the 324 CRP loci found 102 colocalised loci between CRP and at least one other trait and prioritised 41 single candidate causal variants. Among these, 33 were CRP top signals including 9 novel (**Table 8.2**). The exonic variant rs1260326 within *GCKR* colocalised with all traits and was a candidate causal variant for the shared association (PP = 0.94, 100% of the PP explained by SNP). Also, the intergenic variant rs2211320, located 9kb upstream from the CRP gene, colocalised

with CRP, HDL, LDL, TG and CPD (PP = 0.97, 99.9% of PP explained by SNP). Novel CRP associated variants rs34811474 (*ANAPC4*), rs10968576 (*LINGO2*), and rs879620 (*ADCY9*), colocalised with HDL, BMI, and CPD, respectively.

Table 8.2 C-reactive protein candidate shared causal variants with other traits

Causal SNP	Chr:Pos	Traits	PP	PP%
rs75460349	1:27180088	CRP, TG	1.00	80.1%
rs61812598	1:154420087	CRP, HDL, LDL	0.96	88.3%
rs2211320	1:159693605	CRP, HDL, LDL, TG, CPD	0.97	100.0%
rs4658403	1:243832560	CRP, HDL, BMI	0.95	99.7%
rs1260326	2:27730940	CRP, HDL, LDL, TG, BMI, CPD	0.94	100.0%
rs17326656	2:48962291	CRP, HDL, TG	0.93	98.6%
rs2161037	2:169893419	CRP, LDL	1.00	99.9%
rs6792725	3:24520283	CRP, LDL, TG	1.00	100.0%
rs171390	3:154038412	CRP, BMI	0.98	87.0%
rs247975	3:173107443	CRP, HDL, TG, BMI, CPD	0.80	88.9%
rs34811474	4:25408838	CRP, HDL, BMI, CPD	0.95	100.0%
rs10938397	4:45182527	CRP, HDL, BMI	0.86	82.7%
rs6870983	5:87697533	CRP, BMI	1.00	100.0%
rs2228213	6:12124855	CRP, BMI	0.96	90.6%
rs5017416	6:18492350	CRP, BMI	1.00	94.0%
rs1490384	6:126851160	CRP, LDL	0.81	92.5%
rs35237252	8:19870271	CRP, CPD	0.98	92.8%
rs112875651	8:126506694	CRP, LDL, TG	1.00	100.0%
rs7031064	9:14455076	CRP, BMI	1.00	99.4%
rs10968576	9:28414339	CRP, HDL, BMI, CPD	0.87	85.6%
rs11012732	10:21830104	CRP, BMI	0.99	85.6%
rs6486122	11:13361524	CRP, HDL, LDL, TG	0.99	96.1%
rs6265	11:27679916	CRP, BMI, CPD	0.93	91.6%
rs4755720	11:43628749	CRP, HDL, TG	0.96	80.2%
rs3741298	11:116657561	CRP, LDL, TG	0.95	100.0%
rs7138803	12:50247468	CRP, HDL, BMI	0.99	100.0%
rs9604045	13:113927208	CRP, HDL	0.99	100.0%
rs2239222	14:73011885	CRP, LDL	1.00	100.0%
rs11635675	15:63793238	CRP, LDL, TG, BMI	0.94	81.3%
rs11852372	15:78801394	CRP, TG, CPD	0.83	99.7%
rs879620	16:4015729	CRP, HDL, BMI, CPD	0.98	100.0%
rs3814883	16:29994922	CRP, HDL, TG, BMI	0.93	100.0%
rs1421085	16:53800954	CRP, HDL, LDL, BMI, CPD	0.98	100.0%
rs183130	16:56991363	CRP, HDL, LDL, TG, CPD	0.89	100.0%
rs2000999	16:72108093	CRP, LDL, TG	0.99	88.1%

rs2925979	16:81534790	CRP, HDL, TG	0.93	100.0%
rs56113850	19:41353107	CRP, BMI, CPD	0.95	100.0%
rs429358	19:45411941	CRP, BMI, CPD	1.00	100.0%
rs117113213	20:39165692	CRP, LDL, TG	0.99	99.4%
rs1800961	20:43042364	CRP, HDL, LDL	1.00	100.0%
rs397092	21:46582564	CRP, HDL, TG, BMI, CPD	0.91	96.0%

Causal SNP: Candidate shared causal variant; Chr:Pos: Chromosome:Position; Traits: Traits that colocalise; PP: Posterior Probability; PP%: Percentage of PP explained by SNP; BMI: body mass index; CPD: cigarettes per day; CRP: C-reactive protein; LDL: high-density lipoprotein, LDL: low-density lipoprotein, TG: triglyceride

8.3.4 PheWAS on pleiotropic variants with discordant direction of effects

Of the 41 colocalised variants, 12 exhibited significant associations (MTAG $P < 0.05$) with discordant direction of effects between traits (**Figure 8.3**).

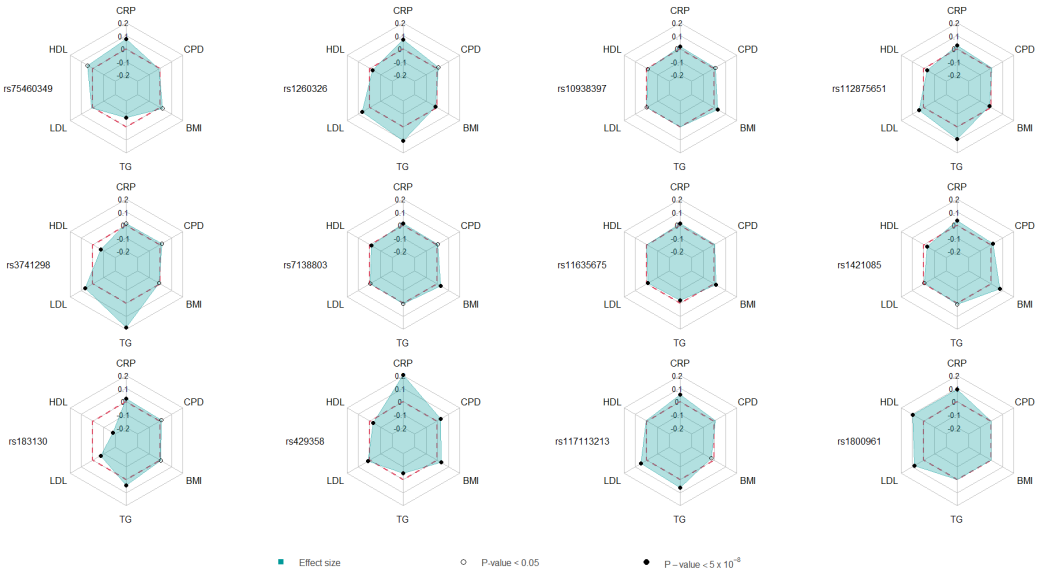


Figure 8.3 Effect sizes across traits for the twelve colocalised variants with discordant direction of effects

Inverse effect directions between CRP and any of LDL, TG, BMI, and CPD, or direct effect directions between CRP and HDL, were classified as discordant when the SNP's association with the discordant trait was statistically significant ($P < 0.05$) in the multi-trait MTAG analysis. The red dashed line in the figure represents the zero-effect value. Nominally significant associations ($P < 0.05$) are depicted with white-filled circles, while genome-wide significant associations are indicated by black-filled circles.

BMI: body mass index, CPD: cigarettes per day, CRP: C-reactive protein; HDL: high-density lipoprotein, LDL: low-density lipoprotein, TG: triglycerides

PheWAS on the 12 discordant variants identified discordant FDR associations for 7 variants translating the discordant genetic effects into clinical outcomes with various diseases (**Figure 8.4**). For instance, rs1260326 (*GCKR*) associated with increased levels of CRP and LDL and lower BMI in MTAG, was linked to 28 diseases in PheWAS including direct associations with lipometabolism disorders, gout, and angina pectoris and inverse associations with type 2 diabetes, cholelithiasis, alcoholism, and fasciitis. Also, rs1421085 (*FTO*) was associated with 27 diseases including direct associations with obesity, type 2 diabetes, and hypertension and inverse association with breast cancer, and fasciitis.

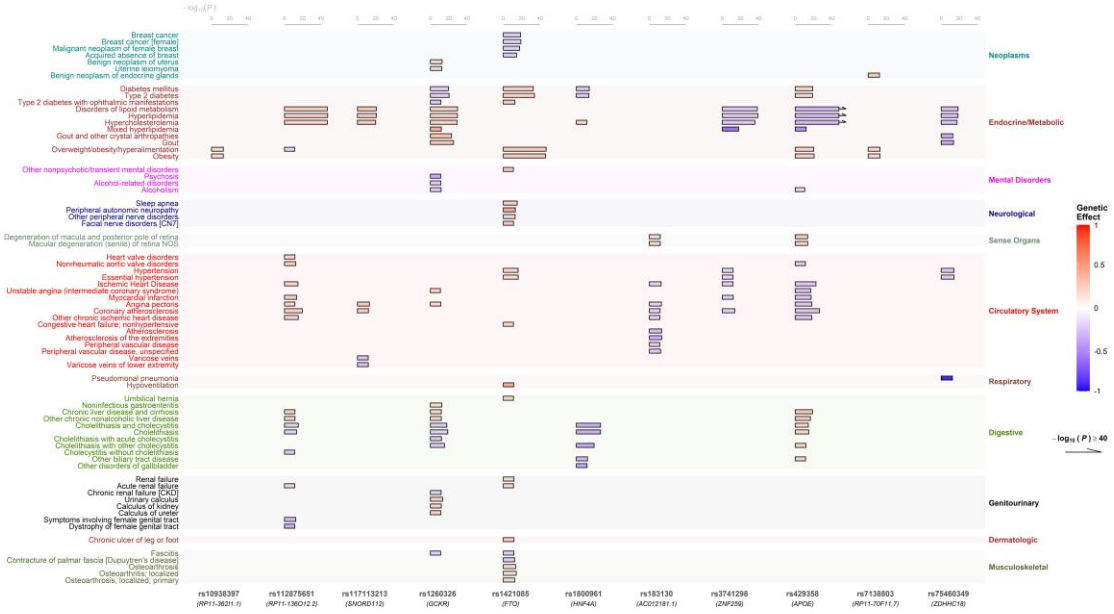


Figure 8.4 Results from phenome-wide association analysis for the twelve variants with discordant direction of effects

The plot shows FDR-significant associations for colocalised variants with discordant direction of effects between C-reactive protein levels and any of the other examined traits. The bar colour shows the genetic effect size while the bar length the statistical significance ($-\log_{10}P$). The variant rs429358 (*APOE*) showed 96 FDR-significant associations. To optimise the visualisation, diseases associated exclusively with this variant were excluded from the graph. Additionally, the variant rs11635675 (*USP3*) was excluded from the graph as it presented no FDR-significant associations.

8.3.5 Mendelian Randomisation

Two-sample MR showed that genetically predicted higher CRP levels were associated with a lower risk of schizophrenia with consistent results across sensitivity analyses (**Figure 8.5**). Weak evidence was observed for a direct association between genetically determined CRP levels and breast cancer. IVW analysis suggested potential pleiotropy for several other outcomes (e.g., ischemic heart disease and diabetes), but sensitivity analyses did not provide evidence of causality. In the subgroup analyses restricted to SNPs associated exclusively with CRP levels, no strong evidence of causal effects was observed for any outcome except for schizophrenia, which showed modest evidence of an inverse association with genetically determined higher CRP levels.

Figure 8.5 Results from Mendelian randomisation analysis

Each row represents a distinct exposure, while each column corresponds to a different outcome. Each Mendelian randomisation analysis is depicted within a four-sectioned box, summarising the estimates from the inverse-variance weighted (IVW) method (top left), weighted median (WM) method (top right), Egger method slope (bottom left), and Egger method intercept (bottom right). The colour of each section reflects the effect size, with asterisks denoting the level of statistical significance.

BMI: Body mass index, CPD: Cigarettes per day, CRP: C-reactive protein, HDL: High-density lipoprotein, LDL: Low-density lipoprotein, TG: Triglycerides, ADHD: Attention deficit hyperactivity disorder, ASD: Autism spectrum disorder

8.4 Discussion

This study used MTAG between CRP and five cardiometabolic risk factors identifying 41 novel CRP loci. Moreover, it provides a comprehensive exploration of the pleiotropic genetic basis of CRP, indicating 19 gene-sets as potential master regulators of chronic low-grade inflammation with extensive pleiotropic effects on lipids and other cardiometabolic pathways. A colocalisation analysis identified 41 potential shared causal variants between CRP and cardiometabolic risk factors and a PheWAS explored the associations across the phenome for 12 colocalised variants with discordant effect direction between CRP and other traits. An MR supported a causal association between genetically predicted low-grade chronic inflammation and lower risk of schizophrenia aligning with findings from previous studies (Said et al., 2022). However, evidence for causal associations with other diseases was limited when pleiotropic signals were excluded.

The multi-trait MTAG leveraged the correlation between CRP with its key determinants to increase the number of identified loci. The results reinforce the strong connection between CRP and lipometabolism. Several novel CRP loci overlapped with well-established lipid and BMI loci, such as variants mapped at *LPL*, *APOB* and *LINGO2*, which had not been previously linked to CRP in univariate GWAS. Additionally, numerous novel CRP loci have not been previously associated with any of the other examined traits, demonstrating the added statistical power of MTAG (Turley et al., 2018). The prioritised CRP loci were enriched for expression in the liver, pituitary, brain cerebellum and cerebellar hemisphere, highlighting the potential importance of these pleiotropic loci in brain.

An exonic variant (rs2228213) within *HIVEP1*, which encodes the human immunodeficiency virus type I enhancer-binding protein 1, colocalised between CRP and BMI and was identified as a novel CRP locus. *HIVEP1* is involved in regulation of the transcription of inflammatory target genes such as those belonging to the interleukin signalling pathway. Its deficiency has been connected to increased inflammation in septic conditions (Matsumoto et al., 2021). This colocalised exonic variant may provide insights into mechanisms connecting infection, inflammation and adiposity.

The non-synonymous exon variant rs34811474 in *ANAPC4*, encoding the anaphase-promoting complex subunit 4, was identified as the candidate causal variant for multiple

traits including CRP, HDL, BMI and CPD. While pleiotropy at the locus and gene level is common throughout the genome, variants with broad pleiotropic effects relatively rare (Watanabe et al., 2019). This variant has also been previously linked to cognitive performance and educational attainment (Lee et al., 2018), lung function (Kichaev et al., 2019), and osteoarthritis (Tachmazidou et al., 2019). Although the biological function of *ANAPC4* remains poorly understood, its extensive pleiotropic effects and low tissue specificity suggest a role in fundamental biological processes.

DHX36 (rs171390) colocalised with CRP and BMI and was identified as a novel locus for CRP. This gene is a highly conserved member of the DExD/H box helicase family, binds to and unfolds G-Quadruplex (G4) DNA structures, influencing DNA- and RNA-dependent processes (Antcliff et al., 2021). G4 structures and *DHX36* interactivity have been implicated in cancer and tumorigenesis, neurodegenerative diseases and aging mechanisms including cellular senescence (Antcliff et al., 2021). These findings suggest a potential role for this gene in CRP and BMI through pathways related to inflammation and adiposity.

The study demonstrated several strengths. Large-scale datasets were used with sample sizes exceeding 500,000 participants, ensuring high statistical power. This power was further enhanced through the application of suitable multivariate methods. Advanced genetic epidemiology techniques were combined allowing for a thorough exploration of the data. Additionally, associations between CRP and numerous risk factors for complex conditions, such as cardiovascular diseases, mental disorders, neoplasms, and other diseases, were investigated.

However, certain limitations were identified. The analysis was confined to individuals of European ancestry, potentially limiting the applicability of the findings to other populations. Rare variants (MAF < 1%) were excluded from the MTAG analysis and subsequent investigations, which may have led to the omission of rare variants with significant effects. The use of summary statistics in MTAG prevented adjustment for additional confounders. Sample overlap within each GWAS was also a limitation, though it was addressed using bivariate LD score regression (Bulik-Sullivan et al., 2015).

Although shared genetic associations suggest pleiotropy, they do not necessarily indicate common biological pathways. Pleiotropic genes might affect traits

independently via distinct pathways or be expressed in different tissues in response to various signals (Gratten and Visscher, 2016, Pickrell et al., 2016). Extensive pleiotropy at CRP loci could introduce bias into the MR assumption of no pleiotropy. Analyses restricting CRP-associated variants to those not linked with lipids or BMI might have been underpowered.

Furthermore, several novel genes associated with CRP were introduced by MTAG. However, the high polygenicity and pleiotropy of the trait present challenges in elucidating precise biological mechanisms, especially since many of the newly identified variants exhibit small individual effects on CRP.

In conclusion, this study conducted a comprehensive multi-trait analysis of CRP and cardiometabolic traits, identifying 41 novel CRP genetic loci. Colocalisation analysis further identified 41 shared causal variants linking inflammation and cardiometabolism, suggesting that disruptions in these loci could influence multiple traits. This detailed investigation of pleiotropic effects provides a foundation for identifying novel preventive and therapeutic targets and understanding potential side effects across traits. Functional studies to clarify the causal variants and mechanisms within these loci could yield valuable insights into the underlying etiological pathways connecting these traits.

CHAPTER 9

MULTI-TRAIT GWAS REVEALS SHARED GENETIC LOCI BETWEEN ALZHEIMER'S AND CARDIOVASCULAR DISEASES

9.1 Background

Alzheimer's disease (AD), the most frequent cause of dementia, is a brain disease that poses significant health challenges. In 2020, more than 55 million people worldwide were estimated to be living with dementia and 60%-70% of them are estimated to be AD patients (Rizzi et al., 2014, Gauthier S, 2022). Although AD's primary pathology is confined to the brain, yet epidemiological studies and recent genetic analyses supported mechanisms connecting AD and cardiovascular (CV) abnormalities (Newman et al., 2005, Breteler et al., 1994, Bleckwenn et al., 2017). One hypothesis for explaining this suggests that AD and CVD-related traits share common causal factors such as inflammation, obesity and diabetes, which may contribute to both diseases through independent (horizontal pleiotropy) or common biological pathways (Tublin et al., 2019). Other hypotheses include the indirect influence of impaired vascular functions that initiate or accelerate the progression of AD (Iadecola and Gottesman, 2018).

They also may share common genetic determinants (Broce et al., 2019). GWAS have previously identified common genetic risk loci for both AD and pathological CV phenotypes that may correspond to overlapping biological pathways. For example, apolipoprotein E (APOE), which encodes a lipid-transport protein crucial to cholesterol metabolism (Mahley, 2016), is an established genetic risk factor for AD (Jansen et al., 2019, Schwartzentruber et al., 2021), coronary artery disease (CAD) (van der Harst and Verweij, 2018) and myocardial infarction (MI) (Hartiala et al., 2021). A deeper exploration of the shared genetic structure of AD and CV traits can provide insights into potentially shared aetiologies and identify shared targets and mechanisms which can address both neurodegenerative and CVDs through suitable interventions.

This study aimed to investigate the genetic commonalities between AD and CV traits and identify loci with pleiotropic effects on both conditions. Multi-trait GWAS on AD and several CV traits were conducted, followed by genetic colocalisation to identify potential pleiotropic genes and the tissues that they are expressed.

9.2 Methods

9.2.1 Study Population

The study was restricted to a population of European ancestry. Summary statistics from seven GWAS were used: AD (Jansen et al., 2019), atrial fibrillation (AF) (Nielsen et al., 2018), CAD (Nikpay et al., 2015), carotid intima-media thickness (cIMT), stroke (Malik et al., 2018), systolic blood pressure (SBP) (Evangelou et al., 2018), and diastolic blood pressure (DBP) (Evangelou et al., 2018).

9.2.2 Genotypic Quality Control

The initial datasets contained genotyped and imputed variants ranging from 7 to 34 million SNPs. Only SNPs that were present in both datasets (AD and the examined CV trait) were included in the analysis. Furthermore, all insertions, deletions, rare variants (MAF < 0.01), variants with sample sizes less than 2/3 of the 90^{th} percentile and palindromic variants were excluded. Finally, approximately 5.75 million SNPs were included in the analysis.

9.2.3 Multi-Trait Association Analysis

Bivariate MTAG analyses were conducted pairwise between AD and CV traits (1. AF, 2. CAD, 3. cIMT, 4. Stroke, 5. SBP-DBP) (Turley et al., 2018). Genetic correlations between traits were calculated and the data were corrected for sample overlap using bivariate LD score regression (Bulik-Sullivan et al., 2015). MTAG generated trait-specific datasets (11 in total: 5 with AD plus 6 with CV traits). The MTAG summary statistics can be interpreted in a same way to those from single-trait GWAS (Turley et al., 2018).

9.2.4 Functional Mapping & Annotation

FUMA (Watanabe et al., 2017) was used to functionally analyse all the MTAG generated results. The GWS SNPs ($P < 5\times10^{-8}$) were initially clumped at $r^2 < 0.6$ to determine the genomic risk loci coordinates and then clumped once more at $r^2 < 0.1$ to

identify independent signals. SNPs in LD at $0.1 \leq r^2 < 0.6$ or SNPs closer than 500kb were assigned to the same LD block. The remaining SNPs after the second clumping were considered the independent and those with the most significant *P*-value in each LD block were the top signals. Annotation and gene prioritisation were performed on the SNPs that survived the first clumping. The European sample from 1000 Genome Project Phase 3 (Genomes Project et al., 2015) was used to calculate pairwise LD between SNPs. SNPs were positionally mapped to their nearest protein-coding genes (Ensembl build v92). To ensure robustness of the results, MTAG loci were considered only if the respective top variants were also associated with the examined trait at $P < 0.01$ in the original GWAS with a concordant direction of effect. To identify the unique AD top and secondary independent signals, AD independent signals from all pairwise MTAG were considered, excluding duplicate signals or proxies (either in distance ±500 kb or in LD $r^2 > 0.1$) and keeping the signal with the strongest *P*-value.

9.2.5 Novel loci definition and replication

A signal indicated by MTAG was considered novel if all the following criteria were met:

1. the top variant of the locus achieved genome-wide significance ($P < 5 \times 10^{-8}$) in the MTAG results
2. the top variant of the locus was significant ($P < 0.01$) in the included original GWAS with a concordant direction of effect
3. the top variant was not located within ±500 kb or in LD ($r^2 > 0.1$) with previously reported loci

For AD, in addition to the included GWAS study (Jansen et al., 2019), novel loci were also compared to two previously published key GWAS studies (Kunkle et al., 2019, Bellenguez et al., 2022).

To explore whether expressions of the novel AD genes in relevant brain tissues were causally associated with AD risk, summary-data-based MR (SMR) (Zhu et al., 2016) and Heterogeneity In Dependent Instruments (HEIDI) analysis were conducted. SMR was performed by integrating eQTL data from GTEx version 8 (Consortium, 2020) in hippocampus and cortex (SNPs within 1 Mb of the transcription start site with $P < 1 \times 10^{-5}$) and AD summary statistics from MTAG.

9.2.6 Trait-Trait and Trait-eQTL Colocalisation Analysis

Colocalisation was conducted using HyPrColoc (Foley et al., 2021), which is a Bayesian divisive clustering algorithm to detect shared genetic associations using GWAS summary statistics. This method was applied to identify colocalised loci between AD and CV traits and prioritise causal variants explaining the shared association. The colocalisation was conducted on regions ±200 kb from the top SNP from MTAG. The prior probability for a variant to be associated with a single trait only at $P = 1\times10^{-4}$ and a conditional prior probability for a variant to be associated with an additional trait given that it is already associated with another trait at $Pc = 0.02$.

Colocalisation evidence was categorised as considerable ($0.5 < PP < 0.75$) or strong ($PP \geq 0.75$). To limit spurious pleiotropy, analyses was restricted to regions with at least one SNP with $P < 5\times10^{-4}$ in the respective univariate GWAS. Variants explaining at least 80% of the shared association were considered candidate causal. To limit the probability of false positive findings, variants were considered causal only if they were associated ($P < 0.01$) with both AD and the respective CV trait in the univariate GWAS.

For colocalised loci, a expression quantitative trait loci (eQTL) colocalisation was performed using AD, CV trait and eQTL data from 48 tissues, obtained from GTEx v7, following the same approach as described before. The eQTL colocalisation analysis aimed to identify shared genes between the traits and tissues of expression.

9.3 Results

9.3.1 Multi-Trait Genetic Association Analysis

Five pairwise MTAG analyses were performed between AD and one each time of the following CV trait: CAD, AF, stroke, cIMT, and blood pressure (BP). The genetic correlations between AD and the examined CV traits were calculated (**Supplementary Table 10**) and the MTAG results were visually illustrated in comparison with the original GWAS (**Supplementary Figures 5-10**). The analysis identified 27 unique genetic loci associated with AD at GWS level ($P < 5\times10^{-8}$) that corresponded to114 unique SNPs. Among these loci, 5 were novel for AD with rs73069394 (*ULK4*) showing the strongest association (**Table 9.1**).

To validate the associations for the novel AD loci, a SMR analysis was conducted using eQTL data from relevant brain tissues. This analysis showed a potentially causal

association between *ULK4* expression in the hippocampus and AD risk (beta = 0.04, $P = 3.4\times10^{-10}$, $P_{HEIDI} = 0.22$). SMR was not applicable for the remaining loci due to unavailability of gene expression data.

Moreover, 740 genetic loci associated with CV traits at GWS level including 1222 independent signals. Among these, 13 novel loci were identified, including 4 for CAD, 8 for cIMT, and 1 for stroke (**Table 9.1**). Overall, 15 unique AD SNPs (in 9 loci) were also associated with at least one of the examined CV traits at GWS level.

Table 9.1 Novel genetic loci for Alzheimer's disease and cardiovascular traits

SNP	Chr	Pos	EA	OA	MAF	Beta	*P*	Gene
Alzheimer's disease								
rs7529220	1	22282619	T	C	0.139	-0.01	1.7×10^{-8}	*HSPG2*
rs11692604	2	19947507	C	T	0.483	0.01	4.1×10^{-8}	*AC019055.1*
rs73069394	3	41787233	A	G	0.188	0.03	1.5×10^{-29}	*ULK4*
rs77399788	5	123003001	G	A	0.058	0.03	1.8×10^{-10}	*KRT18P16*
rs56365761	19	39148103	G	A	0.436	-0.01	4.6×10^{-8}	*ACTN4*
Coronary artery disease								
rs2552527	2	218688596	G	T	0.409	-0.02	2.0×10^{-8}	*TNS1*
rs748431	3	14928077	G	T	0.395	0.02	3.5×10^{-8}	*FGD5*
rs11723436	4	120901336	G	A	0.327	0.02	2.7×10^{-8}	*RP11-700N1.1*
rs15052	19	41813375	C	T	0.158	0.02	3.6×10^{-8}	*HNRNPUL1:TGFB1*
Carotid intima-media thickness								
rs10064683	5	95567760	A	G	0.352	0.02	1.3×10^{-8}	*CTD-2337A12.1*
rs6904596	6	27491299	A	G	0.082	0.04	1.2×10^{-8}	*HNRNPA1P1*
rs56118607	9	127898024	A	G	0.124	0.03	3.8×10^{-8}	*SCAI*
rs1887182	10	97013497	G	T	0.467	0.02	1.0×10^{-8}	*PDLIM1*
rs11029956	11	27355804	A	G	0.340	0.02	7.4×10^{-9}	*CCDC34*
rs12370774	12	106510413	T	C	0.097	-0.04	3.4×10^{-8}	*NUAK1*
rs76064118	19	2235284	T	C	0.053	0.06	4.8×10^{-8}	*PLEKHJ1*
rs1034565	22	19984211	T	C	0.284	0.03	9.8×10^{-9}	*ARVCF*
Stroke								
rs2284665	10	124226630	T	G	0.197	-0.01	4.4×10^{-8}	*HTRA1*

SNP: Single-nucleotide polymorphism; Chr: Chromosome; Pos: Position; EA: Effect allele; OA: Other allele; MAF: Minor allele frequency; Beta: Effect size estimate; P: Two-sided P-value

9.3.2 Colocalisation Defines Genetic Loci Shared by AD and CV Traits

The colocalisation analysis on the 767 loci identified by MTAG (62 AD + 740 CV) identified 21 colocalised loci between AD and CV traits (**Figure 9. 1, Supplementary Table 11**).

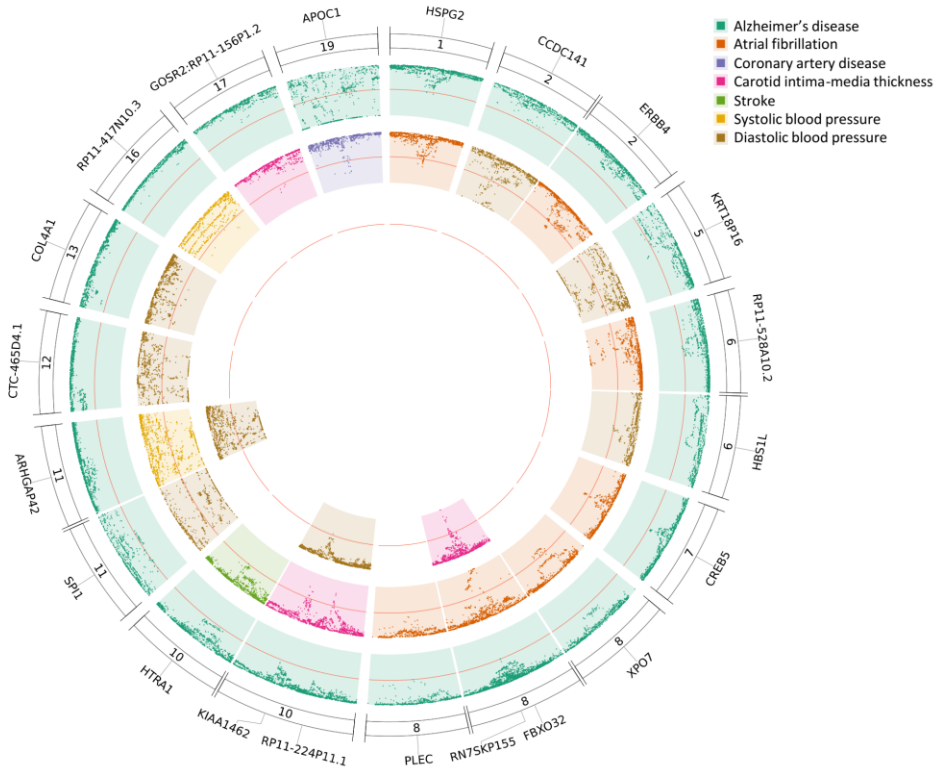


Figure 9.1 Circular scheme showing multiple regional plots for the colocalised loci between Alzheimer's and cardiovascular traits

The figure presents the distribution of *P*-values ($-\log_{10}P$) from MTAG with inner orientation. The annotations show the mapped genes of the lead variants

Most colocalised loci were found between AD and AF (7 loci) or between AD and DBP (7 loci). In three out of the twenty-one colocalised loci, a single candidate causal variant explained a sufficiently the shared association: rs11786896 (*PLEC*; colocalised with AD and AF; PP = 0.97; 86% of PP explained by SNP), rs7529220 (*HSPG2*; colocalised with AD and AF; PP = 1; 90% of PP explained by SNP) and rs429358 (*APOE*; colocalised with AD and CAD; PP = 0.57; 93% of PP explained by SNP). The variant rs11786896 in *PLEC* was not considered a novel AD locus due to its proximity to a previously reported variant (rs34173062 in *SHARPIN*). Nevertheless, the two loci may represent distinct and independent signals as supported by pairwise genetic correlation ($r^2 = 0.006$) and regional plots (**Supplementary Figure 11**).

9.3.3 Gene Expression Colocalisation Prioritises Shared Causal Genes

To identify pleiotropic causal genes for the 21 colocalised loci, a gene expression colocalisation was performed using eQTL data in 48 tissues. The analysis found 16 loci that colocalised with AD, at least one CV trait and expression of genes, for a total of 53 associations with 43 genes (**Supplementary Figure 12, Supplementary Table 12**). These associations included 20 for AF (in 6 loci with 17 genes), 22 for DBP (6 loci with 16 genes), 3 for stroke (1 locus with 3 genes) and 8 for cIMT (3 loci with 6 genes).

In two loci, a single candidate causal variant explained the colocalisation for AD, CV trait, and tissue-specific gene expression. The intronic variant rs11786896 (*PLEC*) colocalised with AD and AF and was an eQTL for *PLEC* in the cardiac left ventricle (PP = 0.99, 99% of PP explained by SNP) and skeletal muscle (PP = 0.92, 98%PP explained by SNP) (**Figure 9.2**). The variant was associated with increased risk of AD (Odds Ratio, OR = 1.02, $P = 5\times10^{-8}$), increased risk of AF (OR = 1.02, $P = 1.1\times10^{-6}$), and lower *PLEC* expression in both cardiac left ventricle (Beta = −0.71, $P = 5.9\times10^{-13}$) and skeletal muscle (Beta = −0.3, $P = 7.7\times10^{-7}$).

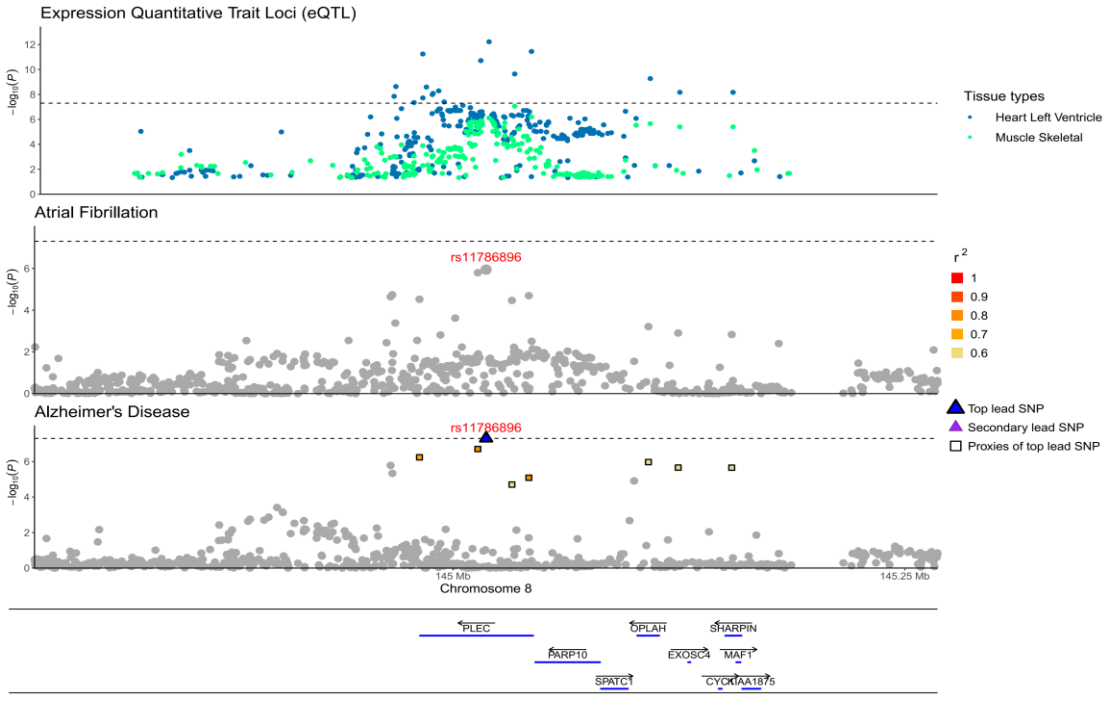


Figure 9.2 Regional plots for the candidate causal variant rs11786896 in *PLEC*

The intergenic variant rs7529220 (*HSPG2*) explained the colocalisation of AD and AF with the expression levels of *C1QA* (PP = 0.85, 82% of PP explained by SNP), *C1QB* (PP = 0.83, %PP = 97%), and *C1QC* (PP = 0.61, 99% of PP explained by SNP) in mammary tissue. This variant was associated with an increased risk of AD (OR = 1.01, $P = 1.7\times10^{-8}$), an increased risk of AF (OR = 1.01, $P = 2.7\times10^{-10}$), and elevated expression of *C1QA* (Beta = 0.19, $P = 2.4\times10^{-4}$), *C1QB* (Beta = 0.17, $P = 2.6\times10^{-4}$), and *C1QC* (Beta = 0.15, $P = 8.1\times10^{-4}$) in mammary tissue.

9.4 Discussion

A comprehensive approach was adopted to understand the co-occurrence between AD and various CV diseases and traits based on several multi-trait GWAS to characterise their shared genetic architecture. Convergent evidence from colocalisation between AD, AF and eQTLs prioritised two genetic regions that each included a single candidate causal variant (rs11786896 was eQTL for *PLEC* and rs7529220 for *C1QA*, *C1QB*, and *C1QC*). These findings provide new insights into genetic pleiotropic effects and potential shared mechanisms causally related to both AD and CVD.

This study identified five not previously reported AD loci. Among these, the intronic variant rs73069394 (*ULK4*) demonstrated the strongest association and was also GWS for DBP. However, the locus showed no evidence of colocalisation between AD and DBP, suggesting that it is not likely the causal for both traits. Beyond its association with DBP, previous GWAS have linked this locus to schizophrenia and bipolar disorder (Luo et al., 2022). The role of ULK4 protein in neurodegeneration is little studied, but its function suggests relevance to AD. It is involved in the regulation of autophagy and in multiple pathways related to AD pathology and brain functions such as neuronal growth, endocytosis, and myelination (Luo et al., 2022, Liu et al., 2018). Furthermore, this study provided additional evidence from SMR analysis supporting a potentially causal relationship between *ULK4* expression levels and AD.

Out of the several CV traits and diseases examined with AD, AF showed the largest number of pleiotropic signals with AD. Numerous observational studies, provide growing evidence that BP and AF are associated with cognitive impairment, risk of AD and other dementias (Rivard et al., 2022, Livingston et al., 2020). The suggested mechanisms linking these traits and AD involve a combination of cerebrovascular damage, neuroinflammation, amyloid-beta accumulation, oxidative stress, and

endothelial dysfunction (Iturria-Medina et al., 2016, Sweeney et al., 2019). However, it is unclear whether the diseases have a shared pathophysiology or whether the relationship arises as downstream consequences of BP and AF (e.g., stroke).

This study provides evidence suggesting shared genetic determinants for the two diseases. The colocalised intronic variant rs11786896 within the plectin gene (*PLEC*) was associated with lower expression of *PLEC* in the cardiac left ventricle and increased risk for both AD and AF. *PLEC* is a member of a protein family, named plakins, with a crucial structural role in the cytoskeleton including cell architecture and tissue integrity and a partially functional role in the assembly, positioning, and regulation of signalling complexes (Sonnenberg and Liem, 2007, Leung et al., 2002). Previous studies of human tissues or preclinical models provide independent evidence for an association of plectin with diseases including AD and AF (Thorolfsdottir et al., 2017, Lagisetty et al., 2022). Plectin deficiency in mice has been associated with diminished learning capabilities and reduced long-term memory compared to wild-type littermates (Valencia et al., 2021). A hypothesis supports that the risk of AD may be affected via functions of plectin in astrocytes (Potokar and Jorgacevski, 2021). Astrocytes play multiple roles, central to the pathology of AD, including metabolic support for neurons, modulation of brain microvascular function and, through activities associated with those of microglia, inflammatory responses (Potokar and Jorgacevski, 2021, Nedergaard et al., 2003). These functional roles may be mediated in part by interactions of plectin with intermediate filaments, microtubules and actin filaments (Potokar and Jorgacevski, 2021). Intermediate filaments are important structural components of the cytoskeleton with crucial roles in synaptic activity, neurogenesis and repair after brain injury (Potokar et al., 2020). Differences in expression of plectin modulate neuronal function and vesicular trafficking generally and interactions with tau suggest potential roles specific to AD (Valencia et al., 2021, Deak, 2014, Fuchs et al., 2009). The role of *PLEC* in AF has been largely hypothesised to act via structural effects on the heart and cause electrophysiological abnormalities (Thorolfsdottir et al., 2017). Therefore, in accordance with the hypothesised mechanisms linking *PLEC* to AD above, *PLEC* may play related roles in cardiomyocytes for assembling and mobilising the intermediate filaments and their networks. These effects further modulate contractile function in cardiomyocytes and inflammatory responses in macrophages which may further contribute to AF (Kamal et al., 2018).

Another colocalised variant between AD and AF, the intergenic rs7529220, which is located 19k upstream from Heparan Sulfate Proteoglycan 2 (*HSPG2*) and 21k downstream from Chymotrypsin Like Elastase 3B (*CELA3B*), was associated with increased risk of AD and AF and higher expression of three genes of the Complement Component 1, Q Subcomponent (*C1Q*) family (*C1QA*, *C1QB*, and *C1QC*) in breast mammary tissue (and, by inference, in brain vasculature) and is a previously unreported locus for AD. The variant is located 680kb downstream of *C1Q* genes. The complement system plays a central role in synaptic remodelling in the brain and in cellular damage response more generally in the body (Gomez-Arboledas et al., 2021, Carpanini et al., 2022). It is possible that greater expression of *C1Q* may lead to higher activity of the complement system which in turn may potentiate synapse loss in early AD (Dejanovic et al., 2022). Similarly, C1Q has roles in the genesis of atherosclerotic plaques (Haskard et al., 2008) and in the regulation of early stages of inflammatory responses to the cardiomyocyte injury associated with a range of cardiac traits (Mihlan et al., 2011).

This study had several strengths. First, it achieved high statistical power by incorporating GWAS with substantial sample sizes and by performing suitable multivariate methods. Second, it combined advanced methodologies from genetic epidemiology to provide supporting evidence.

However, a few limitations also must be acknowledged. We restricted our analyses to a population of European ancestry. The lack of genetic diversity may have hampered the possibility of detecting other relevant variants. Additionally, a considerable portion of the genetic predisposition coming from rare variants (MAF < 1%) was excluded from the analyses. However, including these variants might lead to false-positive findings and biased results. Moreover, statistical methods were used to detect pleiotropy, and therefore considered a genetic locus pleiotropic if it was statistically significantly associated with two or more phenotypes. However, this approach for identification of pleiotropic genes may not always highlight shared biological pathways, as the identified genes could affect the traits independently via different pathways (horizontal pleiotropy), or they could even be expressed in different tissues in response to different signals (Gratten and Visscher, 2016, Pickrell et al., 2016).

In conclusion, multi-trait GWAS analyses were performed on AD and CV traits and subsequent colocalisation analyses detected 16 shared genetic loci and prioritised two candidate shared causal variants. These findings define shared mechanisms for AD and

different CVDs. The complement system has been explored as a target for novel preventive or disease-modifying therapies in CVDs (Miyamoto et al., 2020) and AD (Spurrier et al., 2022). This work suggests that plectin could offer new and potentially promising targets for preventive and therapeutic medicines with benefits across these common comorbid disorders.

CONCLUSIONS

This thesis explores both the theoretical and practical aspects of multivariate methods in genetic epidemiology, including their statistical frameworks and conceptual classifications. Through applications to large genomic datasets, it expands knowledge on the genetic architecture of complex phenotypes discovering novel loci. It emphasises its focus on the investigation of pleiotropy providing new insights into shared genetic mechanisms across phenotypes.

A systematic review on multivariate methods that perform single variant – multiple phenotypes associations using GWAS summary statistics identified several methods of various statistical frameworks, each with its own advantages and limitations. MTAG was prioritised among others for a number of reasons including its special ability to generate trait-specific summary statistics that can be used to post-GWAS analyses.

Suitable multi-trait approaches including MTAG were applied on genomic data from 15 blood traits discovering 711 novel genetic loci involved in blood physiology and also providing insights into their effects on different stages of haematopoiesis.

Similarly, a multi-trait analytical framework on CRP and CV traits identified 41 novel CRP genetic loci, along with 41 shared causal variants linking inflammation and cardiometabolism. Further functional validation may elucidate aetiological pathways connecting inflammation with cardiometabolic disorders.

Another multi-trait analysis on AD and CV traits detected 5 novel AD loci and 19 shared genetic loci. Two potential causal variants prioritised through colocalisation analysis. These findings reveal shared biological mechanisms between AD and CV diseases. Targets within the complement system and plectin suggest promising opportunities for the development of therapies addressing the comorbidity of these disorders.

Overall, multivariate methods can be highly effective for applications in genetic epidemiology studies, as they are better equipped to capture the genetic correlation of the phenotypes under investigation. As such, they can offer significant advantages in GWAS, including, on the one hand, being more effective in identifying a larger number

of novel genetic loci due to increased statistical power, and on the other hand, providing a more comprehensive approach to the study of pleiotropy.

SUMMARY

Genetic association studies mainly focus on associations between single genetic variants and single phenotypes. However, multivariate methods allow the investigation of genetic effects on multiple phenotypes and, therefore, can better capture the shared genetic architecture between correlated traits. These multi-trait approaches offer many advantages, such as increased statistical power for detecting novel genetic loci, reduced multiple testing burden and a more holistic exploration of genetic pleiotropy.

The main objective of this thesis was to identify multivariate methods that utilise summary statistics from genome-wide association studies (GWAS), to compare and prioritise the most effective multivariate approaches for the discovery of novel genetic loci and exploration of pleiotropy. Specific aims included a systematic review of existing multivariate methods in order to identify the most efficient of them and practical applications of suitable multivariate methods on real GWAS data. This research applied advanced multivariate techniques to discover novel genetic loci for blood cell traits and prioritise genes with effects on different stages of haematopoiesis. It also examines the shared genetic architecture between inflammation and cardiometabolism and identified several pleiotropic loci and common mechanisms. Furthermore, the present thesis explored the genetic predisposition for comorbidity between Alzheimer's and cardiovascular diseases, identified novel loci and genes with pleiotropic effects and suggested new potential therapeutic targets.

The systematic review on multivariate methods of single variant - multiple traits genome-wide associations using GWAS summary statistics identified 24 distinct methods of various statistical frameworks. Multi-Trait Association of GWAS (MTAG) was indicated as the most influential method within the scientific community and was prioritised due to its special feature to generate trait-specific summary statistics that can be used to post-GWAS analyses.

Blood cells play a crucial role in the pathophysiology of several diseases, with their phenotypes exhibiting high heritability and polygenicity. However, previous GWAS have only partially explained the overall phenotypic variance of these traits. To further

explore the genetic structure of blood cells, a multi-trait GWAS was conducted on fifteen blood cell traits, which identified 3,100 novel associations across 711 genomic loci, including 29 loci not previously linked to any blood trait in the European population. A subsequent gene expression colocalisation analysis highlighted variants with broad or cell type-specific effects. These findings provide insights into underlying mechanisms involved into different stages of haematopoiesis and consist potential therapeutic targets for blood-related disorders.

C-reactive protein (CRP), a marker of chronic inflammation, is involved in a wide range of pathological conditions and likely shares biological pathways with metabolic traits. A multi-trait GWAS on CRP levels and cardiometabolic traits, including body mass inde, lipids and smoking, identified 41 novel CRP loci. A subsequent colocalisation analysis highlighted 41 shared variants between CRP and cardiometabolic risk factors, with 12 of them demonstrating unexpected discordant effects. Phenome-wide association studies linked these variants to clinical outcomes. These findings expand knowledge on the shared genetic architecture between inflammation and cardiometabolism, suggesting potential targets for preventive and therapeutic strategies.

The co-occurrence of cardiovascular and neurodegenerative diseases, such as Alzheimer's disease, indicates a shared genetic architecture between the two conditions. A multi-trait GWAS was performed on Alzheimer's disease and several cardiovascular diseases, identifying 5 novel genetic loci for Alzheimer's disease and 16 potentially pleiotropic loci affecting both conditions. Fine-mapping and colocalisation analyses suggested that lower expression of *PLEC* in the cardiac left ventricle or higher expression of *C1Q* genes can lead, through independent pathways, to increased risks of Alzheimer's disease and atrial fibrillation. These findings suggest *PLEC* and *C1Q* as potential therapeutic targets for addressing the comorbidity of cardiovascular and neurodegenerative diseases.

This thesis investigated theoretical foundations and practical applications of multivariate techniques in genetic epidemiology. It extends the current knowledge about the genetic architecture of several complex diseases and phenotypes, emphasising on the discovery of novel genetic loci and exploration of genetic pleiotropy.

ΠΕΡΙΛΗΨΗ

Οι μελέτες γενετικών συσχετίσεων επικεντρώνονται κυρίως στην αναζήτηση συσχετίσεων μεταξύ μεμονωμένων γενετικών παραλλαγών και μεμονωμένων φαινοτύπων. Ωστόσο, ορισμένες πολυμεταβλητές μέθοδοι επιτρέπουν τη διερεύνηση γενετικών επιδράσεων σε πολλαπλούς φαινοτύπους και, επομένως, μπορούν να συλλάβουν καλύτερα την κοινή γενετική αρχιτεκτονική μεταξύ συσχετιζόμενων χαρακτηριστικών. Αυτές οι πολυμεταβλητές προσεγγίσεις προσφέρουν σημαντικά πλεονεκτήματα, όπως αυξημένη στατιστική ισχύ για τον εντοπισμό νέων γενετικών τόπων, μειωμένο φορτίο πολλαπλών στατιστικών ελέγχων και μια πιο ολοκληρωμένη διερεύνηση της γενετικής πλειοτροπίας.

Ο κύριος στόχος της παρούσας διατριβής ήταν ο εντοπισμός μεθόδων πολυμεταβλητής ανάλυσης που χρησιμοποιούν συνοπτικά στατιστικά δεδομένα από μελέτες ευρείας σάρωσης του γονιδιώματος (GWAS), να συγκρίνει και να ιεραρχήσει τις πιο αποτελεσματικές πολυμεταβλητές προσεγγίσεις για την ανακάλυψη νέων γενετικών τόπων και την εξερεύνηση της πλειοτροπίας. Ειδικότεροι στόχοι συμπεριλάμβαναν, αφενός, τη συστηματική ανασκόπηση υφιστάμενων πολυμεταβλητών μεθόδων, ώστε να εντοπιστούν οι πιο αποδοτικές εξ αυτών και αφετέρου, την πρακτική εφαρμογή κατάλληλων πολυμεταβλητών μεθόδων σε πραγματικά δεδομένα GWAS. Στη συγκεκριμένη έρευνα, εφαρμόστηκαν προηγμένες πολυμεταβλητές τεχνικές για την ανακάλυψη νέων γενετικών τόπων των αιμοκυτταρικών χαρακτηριστικών και την ιεράρχηση γονιδίων με βάση την επίδρασή τους στα διάφορα στάδια της αιματοποίησης. Επιπλέον, διερευνήθηκε η κοινή γενετική αρχιτεκτονική μεταξύ της φλεγμονής και του καρδιομεταβολισμού και εντοπίστηκαν αρκετοί πλειοτροπικοί τόποι και κοινοί γενετικοί μηχανισμοί. Ακόμη, η παρούσα διατριβή μελέτησε τη γενετική προδιάθεση για συννοσηρότητα μεταξύ της νόσου Αλτσχάιμερ και καρδιαγγειακών παθήσεων, εντόπισε νέους τόπους και γονίδια με πλειοτροπικές επιδράσεις και πρότεινε νέους πιθανούς θεραπευτικούς στόχους.

Η συστηματική ανασκόπηση πάνω στις πολυμεταβλητές μεθόδους ευρυγονιδιωματικών συσχετίσεων μεταξύ ενός γενετικού τόπου και πολλαπλών φαινοτύπων με χρήση συνοπτικών στατιστικών από μελέτες GWAS εντόπισε 24

ξεχωριστές μεθόδους διαφόρων στατιστικών πλαισίων. Η μέθοδος Multi-Trait Association of GWAS (MTAG) αναδείχθηκε ως η πλέον επιδραστική στη επιστημονική κοινότητα και προτιμήθηκε λόγω της ειδικής της ιδιότητας να δημιουργεί συνοπτικά στατιστικά για κάθε φαινότυπο ξεχωριστά, τα οποία μπορούν να χρησιμοποιηθούν σε επακόλουθες της GWAS αναλύσεις.

Τα αιματικά κύτταρα παίζουν καθοριστικό ρόλο στην παθοφυσιολογία αρκετών ασθενειών, ενώ οι φαινότυποί τους εμφανίζουν υψηλή κληρονομησιμότητα και πολυγονικότητα. Ωστόσο, προηγούμενες GWAS μελέτες έχουν εξηγήσει μόνο εν μέρει τη συνολική φαινοτυπική διακύμανση αυτών των χαρακτηριστικών. Προκειμένου να διερευνηθεί περεταίρω η γενετική δομή των αιματικών κυττάρων, εφαρμόστηκε πολυμεταβλητή μελέτη GWAS σε δεκαπέντε χαρακτηριστικά αιματικών κυττάρων, η οποία εντόπισε 3.100 νέες συσχετίσεις σε 711 γονιδιωματικούς τόπους, μεταξύ των οποίων 29 τόποι που δεν είχαν συνδεθεί προηγουμένως με κανένα χαρακτηριστικό αίματος σε ευρωπαϊκό πληθυσμό. Επακόλουθη ανάλυση συνεντοπισμού γονιδιακής έκφρασης (gene expression colocalisation) υπέδειξε γενετικές παραλλαγές με γενικές ή κυτταροειδικές επιδράσεις. Τα συγκεκριμένα ευρήματα φωτίζουν πλευρές για την επίδραση υποκείμενων μηχανισμών που εμπλέκονται σε διάφορα στάδια της αιματοποίησης και συνιστούν πιθανούς θεραπευτικούς στόχους για διαταραχές που σχετίζονται με το αίμα.

Η C-αντιδρώσα πρωτεΐνη (CRP), ένας δείκτης χρόνιας φλεγμονής, εμπλέκεται σε ένα ευρύ φάσμα παθολογικών καταστάσεων και πιθανώς μοιράζεται βιολογικά μονοπάτια με χαρακτηριστικά του μεταβολισμού. Διεξήχθη πολυμεταβλητή μελέτη GWAS πάνω στα επίπεδα CRP και ορισμένους καρδιομεταβολικούς δείκτες, όπως ο δείκτης μάζας σώματος, τα λιπίδια και το κάπνισμα, εντοπίζοντας 41 νέους γενετικούς τόπους της CRP. Πραγματοποιήθηκε ανάλυση γενετικού συνεντοπισμού που ανέδειξε 41 κοινούς γενετικούς πολυμορφισμούς ανάμεσα στη CRP και καρδιομεταβολικούς παράγοντες κινδύνου, εκ των οποίων 12 εμφάνισαν μη αναμενόμενες αντιφατικές επιδράσεις. Μελέτες συσχέτισης του ευρύτερου φαινοτυπικού συνόλου (PheWAS) συνέδεσαν αυτές τις παραλλαγές με κλινικές εκβάσεις. Τα συγκεκριμένα αποτελέσματα επεκτείνουν τη γνώση σχετικά με την κοινή γενετική αρχιτεκτονική της φλεγμονής και του καρδιομεταβολισμού, προτείνοντας πιθανούς στόχους πρόληψης και θεραπείας.

Η συννοσηρότητα καρδιαγγειακών και νευροεκφυλιστικών παθήσεων, όπως η νόσος Αλτσχάιμερ, υποδηλώνει κοινή γενετική αρχιτεκτονική ανάμεσα στις δύο παθήσεις.

Εφαρμόστηκε πολυμεταβλητή ανάλυση GWAS πάνω στη νόσο Αλτσχάιμερ και διάφορες καρδιαγγειακές ασθένειες, εντοπίζοντας 5 νέους γενετικούς τόπους της νόσου Αλτσχάιμερ και 16 πιθανούς πλειοτροπικούς τόπους που επιδρούν και στις δύο παθήσεις. Επιπλέον αναλύσεις λεπτομερούς χαρτογράφησης (fine-mapping) και γενετικού συνεντοπισμού έδειξαν ότι μειωμένη έκφραση του γονιδίου *PLEC* στην αριστερή κοιλία της καρδιάς ή αυξημένη έκφραση των γονιδίων *C1Q* μπορεί να οδηγήσει, μέσω ανεξάρτητων μονοπατιών, σε αυξημένο κίνδυνο εμφάνισης της νόσου Αλτσχάιμερ και της κολπικής μαρμαρυγής. Τα ευρήματα αυτά υποδεικνύουν τα γονίδια *PLEC* και *C1Q* ως πιθανούς θεραπευτικούς στόχους για την αντιμετώπιση της συννοσηρότητας καρδιαγγειακών και νευροεκφυλιστικών ασθενειών.

Η παρούσα διατριβή διερεύνησε θεωρητικές βάσεις και πρακτικές εφαρμογές πολυμεταβλητών τεχνικών στη γενετική επιδημιολογία. Επεκτείνει την τρέχουσα γνώση σχετικά με τη γενετική αρχιτεκτονική διάφορων σύνθετων ασθενειών και φαινοτύπων, δίνοντας έμφαση στην ανακάλυψη νέων γενετικών τόπων και την εξερεύνηση της γενετικής πλειοτροπίας.

REFERENCES

AJOOLABADY, A., PRATICO, D. & REN, J. 2024. Angiotensin II: Role in oxidative stress, endothelial dysfunction, and diseases. *Mol Cell Endocrinol,* 592**,** 112309. DOI:10.1016/j.mce.2024.112309

AKAIKE, H. 1979. A Bayesian extension of the minimum AIC procedure of autoregressive model fitting. *Biometrika,* 66**,** 237-242. DOI:10.1093/biomet/66.2.237

ALLISON, D. B., THIEL, B., ST JEAN, P., ELSTON, R. C., INFANTE, M. C. & SCHORK, N. J. 1998. Multiple phenotype modeling in gene-mapping studies of quantitative traits: power advantages. *Am J Hum Genet,* 63**,** 1190-201. DOI:10.1086/302038

ANTCLIFF, A., MCCULLOUGH, L. D. & TSVETKOV, A. S. 2021. G-Quadruplexes and the DNA/RNA helicase DHX36 in health, disease, and aging. *Aging (Albany NY),* 13**,** 25578-25587. DOI:10.18632/aging.203738

ASTLE, W. J., ELDING, H., JIANG, T., ALLEN, D., RUKLISA, D., MANN, A. L., MEAD, D., BOUMAN, H., RIVEROS-MCKAY, F., KOSTADIMA, M. A., LAMBOURNE, J. J., SIVAPALARATNAM, S., DOWNES, K., KUNDU, K., BOMBA, L., BERENTSEN, K., BRADLEY, J. R., DAUGHERTY, L. C., DELANEAU, O., FRESON, K., GARNER, S. F., GRASSI, L., GUERRERO, J., HAIMEL, M., JANSSEN-MEGENS, E. M., KAAN, A., KAMAT, M., KIM, B., MANDOLI, A., MARCHINI, J., MARTENS, J. H. A., MEACHAM, S., MEGY, K., O'CONNELL, J., PETERSEN, R., SHARIFI, N., SHEARD, S. M., STALEY, J. R., TUNA, S., VAN DER ENT, M., WALTER, K., WANG, S. Y., WHEELER, E., WILDER, S. P., IOTCHKOVA, V., MOORE, C., SAMBROOK, J., STUNNENBERG, H. G., DI ANGELANTONIO, E., KAPTOGE, S., KUIJPERS, T. W., CARRILLO-DE-SANTA-PAU, E., JUAN, D., RICO, D., VALENCIA, A., CHEN, L., GE, B., VASQUEZ, L., KWAN, T., GARRIDO-MARTIN, D., WATT, S., YANG, Y., GUIGO, R., BECK, S., PAUL, D. S., PASTINEN, T., BUJOLD, D., BOURQUE, G., FRONTINI, M., DANESH, J., ROBERTS, D. J., OUWEHAND, W. H., BUTTERWORTH, A. S. & SORANZO, N. 2016. The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. *Cell,* 167**,** 1415-1429 e19. DOI:10.1016/j.cell.2016.10.042

BALDING, D. J., BISHOP, M. & CANNINGS, C. 2007. *Handbook of statistical genetics,* Chichester, John Wiley.

BAO, E. L., CHENG, A. N. & SANKARAN, V. G. 2019. The genetics of human hematopoiesis and its disruption in disease. *EMBO Mol Med,* 11**,** e10316. DOI:10.15252/emmm.201910316

BASELMANS, B. M. L., JANSEN, R., IP, H. F., VAN DONGEN, J., ABDELLAOUI, A., VAN DE WEIJER, M. P., BAO, Y., SMART, M., KUMARI, M., WILLEMSEN, G., HOTTENGA, J. J., CONSORTIUM, B., SOCIAL SCIENCE GENETIC ASSOCIATION, C., BOOMSMA, D. I., DE GEUS, E. J. C., NIVARD, M. G. & BARTELS, M. 2019. Multivariate genome-wide analyses of the well-being spectrum. *Nat Genet,* 51**,** 445-451. DOI:10.1038/s41588-018-0320-8

BELLENGUEZ, C., KUCUKALI, F., JANSEN, I. E., KLEINEIDAM, L., MORENO-GRAU, S., AMIN, N., NAJ, A. C., CAMPOS-MARTIN, R., GRENIER-BOLEY, B., ANDRADE, V., HOLMANS, P. A., BOLAND, A., DAMOTTE, V., VAN DER LEE, S. J., COSTA, M. R., KUULASMAA, T., YANG, Q., DE ROJAS, I., BIS, J. C., YAQUB, A., PROKIC, I., CHAPUIS, J., AHMAD, S., GIEDRAITIS, V., AARSLAND, D., GARCIA-GONZALEZ, P., ABDELNOUR, C., ALARCON-MARTIN, E., ALCOLEA, D., ALEGRET, M., ALVAREZ, I., ALVAREZ, V., ARMSTRONG, N. J., TSOLAKI, A., ANTUNEZ, C., APPOLLONIO, I., ARCARO, M., ARCHETTI, S., PASTOR, A. A., AROSIO, B., ATHANASIU, L., BAILLY, H., BANAJ, N., BAQUERO, M., BARRAL, S., BEISER, A., PASTOR, A. B., BELOW, J. E., BENCHEK, P., BENUSSI, L., BERR, C., BESSE, C., BESSI, V., BINETTI, G., BIZARRO, A., BLESA, R., BOADA, M., BOERWINKLE, E., BORRONI, B., BOSCHI, S., BOSSU, P., BRATHEN, G., BRESSLER, J., BRESNER, C., BRODATY, H., BROOKES, K. J., BRUSCO, L. I., BUIZA-RUEDA, D., BURGER, K., BURHOLT, V., BUSH, W. S., CALERO, M., CANTWELL, L. B., CHENE, G., CHUNG, J., CUCCARO, M. L., CARRACEDO, A., CECCHETTI, R., CERVERA-CARLES, L., CHARBONNIER, C., CHEN, H. H., CHILLOTTI, C., CICCONE, S., CLAASSEN, J., CLARK, C., CONTI, E., CORMA-GOMEZ, A., COSTANTINI, E., CUSTODERO, C., DAIAN, D., DALMASSO, M. C., DANIELE, A., DARDIOTIS, E., DARTIGUES, J. F., DE DEYN, P. P., DE PAIVA LOPES, K., DE WITTE, L. D., DEBETTE, S., DECKERT, J., DEL SER, T., et al. 2022. New insights into the genetic etiology of Alzheimer's disease and related dementias. *Nat Genet,* 54**,** 412-436. DOI:10.1038/s41588-022-01024-z

BENJAMINI, Y. & HOCHBERG, Y. 1995. Controlling the False Discovery Rate—a Practical and Powerful Approach to Multiple Testing. *J Roy Stat Soc B Met,* 57**,** 289-300. DOI:10.1111/j.2517-6161.1995.tb02031.x

BENNER, C., SPENCER, C. C., HAVULINNA, A. S., SALOMAA, V., RIPATTI, S. & PIRINEN, M. 2016. FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics,* 32**,** 1493-501. DOI:10.1093/bioinformatics/btw018

BLECKWENN, M., KLEINEIDAM, L., WAGNER, M., JESSEN, F., WEYERER, S., WERLE, J., WIESE, B., LUHMANN, D., POSSELT, T., KONIG, H. H., BRETTSCHNEIDER, C., MOSCH, E., WEEG, D., FUCHS, A., PENTZEK, M., LUCK, T., RIEDEL-HELLER, S. G., MAIER, W. & SCHERER, M. 2017. Impact of coronary heart disease on cognitive decline in Alzheimer's disease: a prospective longitudinal cohort study in primary care. *Br J Gen Pract,* 67**,** e111-e117. DOI:10.3399/bjgp16X688813

BOWDEN, J., DAVEY SMITH, G. & BURGESS, S. 2015. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *Int J Epidemiol,* 44**,** 512-25. DOI:10.1093/ije/dyv080

BOWDEN, J., DAVEY SMITH, G., HAYCOCK, P. C. & BURGESS, S. 2016. Consistent Estimation in Mendelian Randomization with Some Invalid Instruments Using a Weighted Median Estimator. *Genet Epidemiol,* 40**,** 304-14. DOI:10.1002/gepi.21965

BRETELER, M. M., CLAUS, J. J., GROBBEE, D. E. & HOFMAN, A. 1994. Cardiovascular disease and distribution of cognitive function in elderly people: the Rotterdam Study. *BMJ,* 308**,** 1604-8. DOI:10.1136/bmj.308.6944.1604

BROCE, I. J., TAN, C. H., FAN, C. C., JANSEN, I., SAVAGE, J. E., WITOELAR, A., WEN, N., HESS, C. P., DILLON, W. P., GLASTONBURY, C. M., GLYMOUR, M., YOKOYAMA, J. S., ELAHI, F. M., RABINOVICI, G. D., MILLER, B. L., MORMINO, E. C., SPERLING,

R. A., BENNETT, D. A., MCEVOY, L. K., BREWER, J. B., FELDMAN, H. H., HYMAN, B. T., PERICAK-VANCE, M., HAINES, J. L., FARRER, L. A., MAYEUX, R., SCHELLENBERG, G. D., YAFFE, K., SUGRUE, L. P., DALE, A. M., POSTHUMA, D., ANDREASSEN, O. A., KARCH, C. M. & DESIKAN, R. S. 2019. Dissecting the genetic relationship between cardiovascular risk factors and Alzheimer's disease. *Acta Neuropathol,* 137**,** 209-226. DOI:10.1007/s00401-018-1928-6

BULIK-SULLIVAN, B., FINUCANE, H. K., ANTTILA, V., GUSEV, A., DAY, F. R., LOH, P. R., REPROGEN, C., PSYCHIATRIC GENOMICS, C., GENETIC CONSORTIUM FOR ANOREXIA NERVOSA OF THE WELLCOME TRUST CASE CONTROL, C., DUNCAN, L., PERRY, J. R., PATTERSON, N., ROBINSON, E. B., DALY, M. J., PRICE, A. L. & NEALE, B. M. 2015. An atlas of genetic correlations across human diseases and traits. *Nat Genet,* 47**,** 1236-41. DOI:10.1038/ng.3406

BURGESS, S., BUTTERWORTH, A. & THOMPSON, S. G. 2013. Mendelian randomization analysis with multiple genetic variants using summarized data. *Genet Epidemiol,* 37**,** 658-65. DOI:10.1002/gepi.21758

BURGESS, S., THOMPSON, S. G. & BURGESS, S. 2015. *Mendelian randomization : methods for using genetic variants in causal estimation,* Boca Raton, CRC Press.

BUSH, W. S. & MOORE, J. H. 2012. Chapter 11: Genome-wide association studies. *PLoS Comput Biol,* 8**,** e1002822. DOI:10.1371/journal.pcbi.1002822

CARPANINI, S. M., TORVELL, M., BEVAN, R. J., BYRNE, R. A. J., DASKOULIDOU, N., SAITO, T., SAIDO, T. C., TAYLOR, P. R., HUGHES, T. R., ZELEK, W. M. & MORGAN, B. P. 2022. Terminal complement pathway activation drives synaptic loss in Alzheimer's disease models. *Acta Neuropathol Commun,* 10**,** 99. DOI:10.1186/s40478-022-01404-w

CARROLL, R. J., BASTARACHE, L. & DENNY, J. C. 2014. R PheWAS: data analysis and plotting tools for phenome-wide association studies in the R environment. *Bioinformatics,* 30**,** 2375-6. DOI:10.1093/bioinformatics/btu197

CHATTERJEE, N., SHI, J. & GARCIA-CLOSAS, M. 2016. Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nat Rev Genet,* 17**,** 392-406. DOI:10.1038/nrg.2016.27

CHAVALI, S., BARRENAS, F., KANDURI, K. & BENSON, M. 2010. Network properties of human disease genes with pleiotropic effects. *BMC Syst Biol,* 4**,** 78. DOI:10.1186/1752-0509-4-78

CHEN, M. H., RAFFIELD, L. M., MOUSAS, A., SAKAUE, S., HUFFMAN, J. E., MOSCATI, A., TRIVEDI, B., JIANG, T., AKBARI, P., VUCKOVIC, D., BAO, E. L., ZHONG, X., MANANSALA, R., LAPLANTE, V., CHEN, M., LO, K. S., QIAN, H., LAREAU, C. A., BEAUDOIN, M., HUNT, K. A., AKIYAMA, M., BARTZ, T. M., BEN-SHLOMO, Y., BESWICK, A., BORK-JENSEN, J., BOTTINGER, E. P., BRODY, J. A., VAN ROOIJ, F. J. A., CHITRALA, K., CHO, K., CHOQUET, H., CORREA, A., DANESH, J., DI ANGELANTONIO, E., DIMOU, N., DING, J., ELLIOTT, P., ESKO, T., EVANS, M. K., FLOYD, J. S., BROER, L., GRARUP, N., GUO, M. H., GREINACHER, A., HAESSLER, J., HANSEN, T., HOWSON, J. M. M., HUANG, Q. Q., HUANG, W., JORGENSON, E., KACPROWSKI, T., KAHONEN, M., KAMATANI, Y., KANAI, M., KARTHIKEYAN, S., KOSKERIDIS, F., LANGE, L. A., LEHTIMAKI, T., LERCH, M. M., LINNEBERG, A., LIU, Y., LYYTIKAINEN, L. P., MANICHAIKUL, A., MARTIN, H. C., MATSUDA, K., MOHLKE, K. L., MONONEN, N., MURAKAMI, Y., NADKARNI, G. N., NAUCK, M., NIKUS, K., OUWEHAND, W. H., PANKRATZ, N., PEDERSEN, O., PREUSS, M.,

PSATY, B. M., RAITAKARI, O. T., ROBERTS, D. J., RICH, S. S., RODRIGUEZ, B. A. T., ROSEN, J. D., ROTTER, J. I., SCHUBERT, P., SPRACKLEN, C. N., SURENDRAN, P., TANG, H., TARDIF, J. C., TREMBATH, R. C., GHANBARI, M., VOLKER, U., VOLZKE, H., WATKINS, N. A., ZONDERMAN, A. B., PROGRAM, V. A. M. V., WILSON, P. W. F., LI, Y., BUTTERWORTH, A. S., GAUCHAT, J. F., CHIANG, C. W. K., LI, B., et al. 2020. Trans-ethnic and Ancestry-Specific Blood-Cell Genetics in 746,667 Individuals from 5 Global Populations. *Cell,* 182**,** 1198-1213 e14. DOI:10.1016/j.cell.2020.06.045

CICHONSKA, A., ROUSU, J., MARTTINEN, P., KANGAS, A. J., SOININEN, P., LEHTIMAKI, T., RAITAKARI, O. T., JARVELIN, M. R., SALOMAA, V., ALA-KORPELA, M., RIPATTI, S. & PIRINEN, M. 2016. metaCCA: summary statistics-based multivariate meta-analysis of genome-wide association studies using canonical correlation analysis. *Bioinformatics,* 32**,** 1981-9. DOI:10.1093/bioinformatics/btw052

COLLINS, F. S., MORGAN, M. & PATRINOS, A. 2003. The Human Genome Project: lessons from large-scale biology. *Science,* 300**,** 286-90. DOI:10.1126/science.1084564

CONSORTIUM, G. T. 2020. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science,* 369**,** 1318-1330. DOI:10.1126/science.aaz1776

COSTA, R. G., CONCEICAO, A., MATOS, C. A. & NOBREGA, C. 2024. The polyglutamine protein ATXN2: from its molecular functions to its involvement in disease. *Cell Death Dis,* 15**,** 415. DOI:10.1038/s41419-024-06812-5

DAVIES, N. M., HOLMES, M. V. & DAVEY SMITH, G. 2018. Reading Mendelian randomisation studies: a guide, glossary, and checklist for clinicians. *BMJ,* 362**,** k601. DOI:10.1136/bmj.k601

DAWN TEARE, M. 2011. *Genetic Epidemiology,* New York, NY, Springer New York : Imprint: Humana.

DEAK, F. 2014. Neuronal vesicular trafficking and release in age-related cognitive impairment. *J Gerontol A Biol Sci Med Sci,* 69**,** 1325-30. DOI:10.1093/gerona/glu061

DEHGHAN, A., DUPUIS, J., BARBALIC, M., BIS, J. C., EIRIKSDOTTIR, G., LU, C., PELLIKKA, N., WALLASCHOFSKI, H., KETTUNEN, J., HENNEMAN, P., BAUMERT, J., STRACHAN, D. P., FUCHSBERGER, C., VITART, V., WILSON, J. F., PARE, G., NAITZA, S., RUDOCK, M. E., SURAKKA, I., DE GEUS, E. J., ALIZADEH, B. Z., GURALNIK, J., SHULDINER, A., TANAKA, T., ZEE, R. Y., SCHNABEL, R. B., NAMBI, V., KAVOUSI, M., RIPATTI, S., NAUCK, M., SMITH, N. L., SMITH, A. V., SUNDVALL, J., SCHEET, P., LIU, Y., RUOKONEN, A., ROSE, L. M., LARSON, M. G., HOOGEVEEN, R. C., FREIMER, N. B., TEUMER, A., TRACY, R. P., LAUNER, L. J., BURING, J. E., YAMAMOTO, J. F., FOLSOM, A. R., SIJBRANDS, E. J., PANKOW, J., ELLIOTT, P., KEANEY, J. F., SUN, W., SARIN, A. P., FONTES, J. D., BADOLA, S., ASTOR, B. C., HOFMAN, A., POUTA, A., WERDAN, K., GREISER, K. H., KUSS, O., MEYER ZU SCHWABEDISSEN, H. E., THIERY, J., JAMSHIDI, Y., NOLTE, I. M., SORANZO, N., SPECTOR, T. D., VOLZKE, H., PARKER, A. N., ASPELUND, T., BATES, D., YOUNG, L., TSUI, K., SISCOVICK, D. S., GUO, X., ROTTER, J. I., UDA, M., SCHLESSINGER, D., RUDAN, I., HICKS, A. A., PENNINX, B. W., THORAND, B., GIEGER, C., CORESH, J., WILLEMSEN, G., HARRIS, T. B., UITTERLINDEN, A. G., JARVELIN, M. R., RICE, K., RADKE, D., SALOMAA, V., WILLEMS VAN DIJK, K., BOERWINKLE, E., VASAN, R. S., FERRUCCI, L., GIBSON, Q. D., BANDINELLI, S., SNIEDER, H., BOOMSMA, D. I.,

XIAO, X., CAMPBELL, H., et al. 2011. Meta-analysis of genome-wide association studies in >80 000 subjects identifies multiple loci for C-reactive protein levels. *Circulation,* 123**,** 731-8. DOI:10.1161/CIRCULATIONAHA.110.948570

DEJANOVIC, B., WU, T., TSAI, M. C., GRAYKOWSKI, D., GANDHAM, V. D., ROSE, C. M., BAKALARSKI, C. E., NGU, H., WANG, Y., PANDEY, S., REZZONICO, M. G., FRIEDMAN, B. A., EDMONDS, R., DE MAZIERE, A., RAKOSI-SCHMIDT, R., SINGH, T., KLUMPERMAN, J., FOREMAN, O., CHANG, M. C., XIE, L., SHENG, M. & HANSON, J. E. 2022. Complement C1q-dependent excitatory and inhibitory synapse elimination by astrocytes and microglia in Alzheimer's disease mouse models. *Nat Aging,* 2**,** 837-850. DOI:10.1038/s43587-022-00281-1

DENG, Q., SONG, C. & LIN, S. 2024. An adaptive and robust method for multi-trait analysis of genome-wide association studies using summary statistics. *Eur J Hum Genet,* 32**,** 681-690. DOI:10.1038/s41431-023-01389-7

DENNY, J. C., RITCHIE, M. D., BASFORD, M. A., PULLEY, J. M., BASTARACHE, L., BROWN-GENTRY, K., WANG, D., MASYS, D. R., RODEN, D. M. & CRAWFORD, D. C. 2010. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics,* 26**,** 1205-10. DOI:10.1093/bioinformatics/btq126

DUNCAN, C. T. 2004. *Statistical methods in genetic epidemiology,* Oxford ; New York, Oxford University Press.

EICHLER, E. E., FLINT, J., GIBSON, G., KONG, A., LEAL, S. M., MOORE, J. H. & NADEAU, J. H. 2010. Missing heritability and strategies for finding the underlying causes of complex disease. *Nat Rev Genet,* 11**,** 446-50. DOI:10.1038/nrg2809

EVANGELOU, E. 2018. *Genetic Epidemiology : Methods and Protocols,* New York, NY, Springer New York : Imprint: Humana.

EVANGELOU, E., WARREN, H. R., MOSEN-ANSORENA, D., MIFSUD, B., PAZOKI, R., GAO, H., NTRITSOS, G., DIMOU, N., CABRERA, C. P., KARAMAN, I., NG, F. L., EVANGELOU, M., WITKOWSKA, K., TZANIS, E., HELLWEGE, J. N., GIRI, A., VELEZ EDWARDS, D. R., SUN, Y. V., CHO, K., GAZIANO, J. M., WILSON, P. W. F., TSAO, P. S., KOVESDY, C. P., ESKO, T., MAGI, R., MILANI, L., ALMGREN, P., BOUTIN, T., DEBETTE, S., DING, J., GIULIANINI, F., HOLLIDAY, E. G., JACKSON, A. U., LI-GAO, R., LIN, W. Y., LUAN, J., MANGINO, M., OLDMEADOW, C., PRINS, B. P., QIAN, Y., SARGURUPREMRAJ, M., SHAH, N., SURENDRAN, P., THERIAULT, S., VERWEIJ, N., WILLEMS, S. M., ZHAO, J. H., AMOUYEL, P., CONNELL, J., DE MUTSERT, R., DONEY, A. S. F., FARRALL, M., MENNI, C., MORRIS, A. D., NOORDAM, R., PARE, G., POULTER, N. R., SHIELDS, D. C., STANTON, A., THOM, S., ABECASIS, G., AMIN, N., ARKING, D. E., AYERS, K. L., BARBIERI, C. M., BATINI, C., BIS, J. C., BLAKE, T., BOCHUD, M., BOEHNKE, M., BOERWINKLE, E., BOOMSMA, D. I., BOTTINGER, E. P., BRAUND, P. S., BRUMAT, M., CAMPBELL, A., CAMPBELL, H., CHAKRAVARTI, A., CHAMBERS, J. C., CHAUHAN, G., CIULLO, M., COCCA, M., COLLINS, F., CORDELL, H. J., DAVIES, G., DE BORST, M. H., DE GEUS, E. J., DEARY, I. J., DEELEN, J., DEL GRECO, M. F., DEMIRKALE, C. Y., DORR, M., EHRET, G. B., ELOSUA, R., ENROTH, S., ERZURUMLUOGLU, A. M., FERREIRA, T., FRANBERG, M., FRANCO, O. H., GANDIN, I., et al. 2018. Genetic analysis of over 1 million people identifies 535 new loci associated with blood pressure traits. *Nat Genet,* 50**,** 1412-1425. DOI:10.1038/s41588-018-0205-x

FERREIRA, M. A. & PURCELL, S. M. 2009. A multivariate test of association. *Bioinformatics,* 25**,** 132-3. DOI:10.1093/bioinformatics/btn563

FOLEY, C. N., STALEY, J. R., BREEN, P. G., SUN, B. B., KIRK, P. D. W., BURGESS, S. & HOWSON, J. M. M. 2021. A fast and efficient colocalization algorithm for identifying shared genetic risk factors across multiple traits. *Nat Commun,* 12**,** 764. DOI:10.1038/s41467-020-20885-8

FUCHS, P., ZORER, M., REIPERT, S., REZNICZEK, G. A., PROPST, F., WALKO, G., FISCHER, I., BAUER, J., LESCHNIK, M. W., LUSCHER, B., THALHAMMER, J. G., LASSMANN, H. & WICHE, G. 2009. Targeted inactivation of a developmentally regulated neural plectin isoform (plectin 1c) in mice leads to reduced motor nerve conduction velocity. *J Biol Chem,* 284**,** 26502-9. DOI:10.1074/jbc.M109.018150

GAI, L. & ESKIN, E. 2018. Finding associated variants in genome-wide association studies on multiple traits. *Bioinformatics,* 34**,** i467-i474. DOI:10.1093/bioinformatics/bty249

GALESLOOT, T. E., VAN STEEN, K., KIEMENEY, L. A., JANSS, L. L. & VERMEULEN, S. H. 2014. A comparison of multivariate genome-wide association methods. *PLoS One,* 9**,** e95923. DOI:10.1371/journal.pone.0095923

GAUTHIER S, W. C., SERVAES S, MORAIS JA, ROSA-NETO P 2022. World Alzheimer Report 2022: Life after diagnosis: Navigating treatment, care and support. *London, England: Alzheimer's Disease International*.

GENOMES PROJECT, C., AUTON, A., BROOKS, L. D., DURBIN, R. M., GARRISON, E. P., KANG, H. M., KORBEL, J. O., MARCHINI, J. L., MCCARTHY, S., MCVEAN, G. A. & ABECASIS, G. R. 2015. A global reference for human genetic variation. *Nature,* 526**,** 68-74. DOI:10.1038/nature15393

GIAMBARTOLOMEI, C., VUKCEVIC, D., SCHADT, E. E., FRANKE, L., HINGORANI, A. D., WALLACE, C. & PLAGNOL, V. 2014. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet,* 10**,** e1004383. DOI:10.1371/journal.pgen.1004383

GOMEZ-ARBOLEDAS, A., ACHARYA, M. M. & TENNER, A. J. 2021. The Role of Complement in Synaptic Pruning and Neurodegeneration. *Immunotargets Ther,* 10**,** 373-386. DOI:10.2147/ITT.S305420

GRATTEN, J. & VISSCHER, P. M. 2016. Genetic pleiotropy in complex traits and diseases: implications for genomic medicine. *Genome Med,* 8**,** 78. DOI:10.1186/s13073-016-0332-x

GRIFFITHS, A. J. F. 2000. *An introduction to genetic analysis,* New York, W.H. Freeman.

GROTZINGER, A. D., RHEMTULLA, M., DE VLAMING, R., RITCHIE, S. J., MALLARD, T. T., HILL, W. D., IP, H. F., MARIONI, R. E., MCINTOSH, A. M., DEARY, I. J., KOELLINGER, P. D., HARDEN, K. P., NIVARD, M. G. & TUCKER-DROB, E. M. 2019. Genomic structural equation modelling provides insights into the multivariate genetic architecture of complex traits. *Nat Hum Behav,* 3**,** 513-525. DOI:10.1038/s41562-019-0566-x

GUO, B. & WU, B. 2019a. Integrate multiple traits to detect novel trait-gene association using GWAS summary data with an adaptive test approach. *Bioinformatics,* 35**,** 2251-2257. DOI:10.1093/bioinformatics/bty961

GUO, B. & WU, B. 2019b. Powerful and efficient SNP-set association tests across multiple phenotypes using GWAS summary data. *Bioinformatics,* 35**,** 1366-1372. DOI:10.1093/bioinformatics/bty811

HARTIALA, J. A., HAN, Y., JIA, Q., HILSER, J. R., HUANG, P., GUKASYAN, J., SCHWARTZMAN, W. S., CAI, Z., BISWAS, S., TREGOUET, D. A., SMITH, N. L., CONSORTIUM, I., GROUP, C. C. H. W., CONSORTIUM, G.-C., SELDIN, M., PAN, C., MEHRABIAN, M., LUSIS, A. J., BAZELEY, P., SUN, Y. V., LIU, C., QUYYUMI, A. A., SCHOLZ, M., THIERY, J., DELGADO, G. E., KLEBER, M. E., MARZ, W., HOWE, L. J., ASSELBERGS, F. W., VAN VUGT, M., VLACHOJANNIS, G. J., PATEL, R. S., LYYTIKAINEN, L. P., KAHONEN, M., LEHTIMAKI, T., NIEMINEN, T. V. M., KUUKASJARVI, P., LAURIKKA, J. O., CHANG, X., HENG, C. K., JIANG, R., KRAUS, W. E., HAUSER, E. R., FERGUSON, J. F., REILLY, M. P., ITO, K., KOYAMA, S., KAMATANI, Y., KOMURO, I., BIOBANK, J., STOLZE, L. K., ROMANOSKI, C. E., KHAN, M. D., TURNER, A. W., MILLER, C. L., AHERRAHROU, R., CIVELEK, M., MA, L., BJORKEGREN, J. L. M., KUMAR, S. R., TANG, W. H. W., HAZEN, S. L. & ALLAYEE, H. 2021. Genome-wide analysis identifies novel susceptibility loci for myocardial infarction. *Eur Heart J,* 42**,** 919-933. DOI:10.1093/eurheartj/ehaa1040

HASKARD, D. O., BOYLE, J. J. & MASON, J. C. 2008. The role of complement in atherosclerosis. *Curr Opin Lipidol,* 19**,** 478-82. DOI:10.1097/MOL.0b013e32830f4a06

HATTLE, M., BURKE, D. L., TRIKALINOS, T., SCHMID, C. H., CHEN, Y., JACKSON, D. & RILEY, R. D. 2022. Multivariate meta-analysis of multiple outcomes: characteristics and predictors of borrowing of strength from Cochrane reviews. *Syst Rev,* 11**,** 149. DOI:10.1186/s13643-022-01999-0

HEMANI, G., BOWDEN, J. & DAVEY SMITH, G. 2018. Evaluating the potential role of pleiotropy in Mendelian randomization studies. *Hum Mol Genet,* 27**,** R195-R208. DOI:10.1093/hmg/ddy163

HIRSCHHORN, J. N. & DALY, M. J. 2005. Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet,* 6**,** 95-108. DOI:10.1038/nrg1521

HUNTER, D. J. 2005. Gene-environment interactions in human diseases. *Nat Rev Genet,* 6**,** 287-98. DOI:10.1038/nrg1578

IADECOLA, C. & GOTTESMAN, R. F. 2018. Cerebrovascular Alterations in Alzheimer Disease. *Circ Res,* 123**,** 406-408. DOI:10.1161/CIRCRESAHA.118.313400

ITURRIA-MEDINA, Y., SOTERO, R. C., TOUSSAINT, P. J., MATEOS-PEREZ, J. M., EVANS, A. C. & ALZHEIMER'S DISEASE NEUROIMAGING, I. 2016. Early role of vascular dysregulation on late-onset Alzheimer's disease based on multifactorial data-driven analysis. *Nat Commun,* 7**,** 11934. DOI:10.1038/ncomms11934

JANSEN, I. E., SAVAGE, J. E., WATANABE, K., BRYOIS, J., WILLIAMS, D. M., STEINBERG, S., SEALOCK, J., KARLSSON, I. K., HAGG, S., ATHANASIU, L., VOYLE, N., PROITSI, P., WITOELAR, A., STRINGER, S., AARSLAND, D., ALMDAHL, I. S., ANDERSEN, F., BERGH, S., BETTELLA, F., BJORNSSON, S., BRAEKHUS, A., BRATHEN, G., DE LEEUW, C., DESIKAN, R. S., DJUROVIC, S., DUMITRESCU, L., FLADBY, T., HOHMAN, T. J., JONSSON, P. V., KIDDLE, S. J., RONGVE, A., SALTVEDT, I., SANDO, S. B., SELBAEK, G., SHOAI, M., SKENE, N. G., SNAEDAL, J., STORDAL, E., ULSTEIN, I. D., WANG, Y., WHITE, L. R., HARDY, J., HJERLING-LEFFLER, J., SULLIVAN, P. F., VAN DER FLIER, W. M., DOBSON, R., DAVIS, L. K., STEFANSSON, H., STEFANSSON, K., PEDERSEN, N. L., RIPKE, S., ANDREASSEN, O. A. & POSTHUMA, D. 2019. Genome-wide meta-analysis identifies new loci and functional

pathways influencing Alzheimer's disease risk. *Nat Genet,* 51**,** 404-413. DOI:10.1038/s41588-018-0311-9

JULIENNE, H., LECHAT, P., GUILLEMOT, V., LASRY, C., YAO, C., ARAUD, R., LAVILLE, V., VILHJALMSSON, B., MENAGER, H. & ASCHARD, H. 2020. JASS: command line and web interface for the joint analysis of GWAS results. *NAR Genom Bioinform,* 2**,** lqaa003. DOI:10.1093/nargab/lqaa003

KAMAL, A. H. M., CHAKRABARTY, J. K., UDDEN, S. M. N., ZAKI, M. H. & CHOWDHURY, S. M. 2018. Inflammatory Proteomic Network Analysis of Statin-treated and Lipopolysaccharide-activated Macrophages. *Sci Rep,* 8**,** 164. DOI:10.1038/s41598-017-18533-1

KAR, D., SARKAR, S. & SPRINGERLINK 2022. *Genetics Fundamentals Notes,* Singapore, Springer Nature Singapore : Imprint: Springer.

KHERA, A. V. & KATHIRESAN, S. 2017. Genetics of coronary artery disease: discovery, biology and clinical translation. *Nat Rev Genet,* 18**,** 331-344. DOI:10.1038/nrg.2016.160

KHOURY, M. J., IADEMARCO, M. F. & RILEY, W. T. 2016. Precision Public Health for the Era of Precision Medicine. *Am J Prev Med,* 50**,** 398-401. DOI:10.1016/j.amepre.2015.08.031

KICHAEV, G., BHATIA, G., LOH, P. R., GAZAL, S., BURCH, K., FREUND, M. K., SCHOECH, A., PASANIUC, B. & PRICE, A. L. 2019. Leveraging Polygenic Functional Enrichment to Improve GWAS Power. *Am J Hum Genet,* 104**,** 65-75. DOI:10.1016/j.ajhg.2018.11.008

KIM, J., BAI, Y. & PAN, W. 2015. An Adaptive Association Test for Multiple Phenotypes with GWAS Summary Statistics. *Genet Epidemiol,* 39**,** 651-63. DOI:10.1002/gepi.21931

KUNKLE, B. W., GRENIER-BOLEY, B., SIMS, R., BIS, J. C., DAMOTTE, V., NAJ, A. C., BOLAND, A., VRONSKAYA, M., VAN DER LEE, S. J., AMLIE-WOLF, A., BELLENGUEZ, C., FRIZATTI, A., CHOURAKI, V., MARTIN, E. R., SLEEGERS, K., BADARINARAYAN, N., JAKOBSDOTTIR, J., HAMILTON-NELSON, K. L., MORENO-GRAU, S., OLASO, R., RAYBOULD, R., CHEN, Y., KUZMA, A. B., HILTUNEN, M., MORGAN, T., AHMAD, S., VARDARAJAN, B. N., EPELBAUM, J., HOFFMANN, P., BOADA, M., BEECHAM, G. W., GARNIER, J. G., HAROLD, D., FITZPATRICK, A. L., VALLADARES, O., MOUTET, M. L., GERRISH, A., SMITH, A. V., QU, L., BACQ, D., DENNING, N., JIAN, X., ZHAO, Y., DEL ZOMPO, M., FOX, N. C., CHOI, S. H., MATEO, I., HUGHES, J. T., ADAMS, H. H., MALAMON, J., SANCHEZ-GARCIA, F., PATEL, Y., BRODY, J. A., DOMBROSKI, B. A., NARANJO, M. C. D., DANIILIDOU, M., EIRIKSDOTTIR, G., MUKHERJEE, S., WALLON, D., UPHILL, J., ASPELUND, T., CANTWELL, L. B., GARZIA, F., GALIMBERTI, D., HOFER, E., BUTKIEWICZ, M., FIN, B., SCARPINI, E., SARNOWSKI, C., BUSH, W. S., MESLAGE, S., KORNHUBER, J., WHITE, C. C., SONG, Y., BARBER, R. C., ENGELBORGHS, S., SORDON, S., VOIJNOVIC, D., ADAMS, P. M., VANDENBERGHE, R., MAYHAUS, M., CUPPLES, L. A., ALBERT, M. S., DE DEYN, P. P., GU, W., HIMALI, J. J., BEEKLY, D., SQUASSINA, A., HARTMANN, A. M., ORELLANA, A., BLACKER, D., RODRIGUEZ-RODRIGUEZ, E., LOVESTONE, S., GARCIA, M. E., DOODY, R. S., MUNOZ-FERNADEZ, C., SUSSAMS, R., LIN, H., FAIRCHILD, T. J., BENITO, Y. A., et al. 2019. Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates

Abeta, tau, immunity and lipid processing. *Nat Genet,* 51**,** 414-430. DOI:10.1038/s41588-019-0358-2

LAGISETTY, Y., BOURQUARD, T., AL-RAMAHI, I., MANGLEBURG, C. G., MOTA, S., SOLEIMANI, S., SHULMAN, J. M., BOTAS, J., LEE, K. & LICHTARGE, O. 2022. Identification of risk genes for Alzheimer's disease by gene embedding. *Cell Genom,* 2. DOI:10.1016/j.xgen.2022.100162

LANDER, E. S. 2011. Initial impact of the sequencing of the human genome. *Nature,* 470**,** 187-97. DOI:10.1038/nature09792

LANDER, E. S., LINTON, L. M., BIRREN, B., NUSBAUM, C., ZODY, M. C., BALDWIN, J., DEVON, K., DEWAR, K., DOYLE, M., FITZHUGH, W., FUNKE, R., GAGE, D., HARRIS, K., HEAFORD, A., HOWLAND, J., KANN, L., LEHOCZKY, J., LEVINE, R., MCEWAN, P., MCKERNAN, K., MELDRIM, J., MESIROV, J. P., MIRANDA, C., MORRIS, W., NAYLOR, J., RAYMOND, C., ROSETTI, M., SANTOS, R., SHERIDAN, A., SOUGNEZ, C., STANGE-THOMANN, Y., STOJANOVIC, N., SUBRAMANIAN, A., WYMAN, D., ROGERS, J., SULSTON, J., AINSCOUGH, R., BECK, S., BENTLEY, D., BURTON, J., CLEE, C., CARTER, N., COULSON, A., DEADMAN, R., DELOUKAS, P., DUNHAM, A., DUNHAM, I., DURBIN, R., FRENCH, L., GRAFHAM, D., GREGORY, S., HUBBARD, T., HUMPHRAY, S., HUNT, A., JONES, M., LLOYD, C., MCMURRAY, A., MATTHEWS, L., MERCER, S., MILNE, S., MULLIKIN, J. C., MUNGALL, A., PLUMB, R., ROSS, M., SHOWNKEEN, R., SIMS, S., WATERSTON, R. H., WILSON, R. K., HILLIER, L. W., MCPHERSON, J. D., MARRA, M. A., MARDIS, E. R., FULTON, L. A., CHINWALLA, A. T., PEPIN, K. H., GISH, W. R., CHISSOE, S. L., WENDL, M. C., DELEHAUNTY, K. D., MINER, T. L., DELEHAUNTY, A., KRAMER, J. B., COOK, L. L., FULTON, R. S., JOHNSON, D. L., MINX, P. J., CLIFTON, S. W., HAWKINS, T., BRANSCOMB, E., PREDKI, P., RICHARDSON, P., WENNING, S., SLEZAK, T., DOGGETT, N., CHENG, J. F., OLSEN, A., LUCAS, S., ELKIN, C., UBERBACHER, E., FRAZIER, M., et al. 2001. Initial sequencing and analysis of the human genome. *Nature,* 409**,** 860-921. DOI:10.1038/35057062

LAWLEY, D. N. & MAXWELL, A. E. 1962. Factor Analysis as a Statistical Method. *Journal of the Royal Statistical Society,* Series D (The Statistician), 12(3), 209–229. DOI:10.2307/2986915

LEE, J. J., WEDOW, R., OKBAY, A., KONG, E., MAGHZIAN, O., ZACHER, M., NGUYEN-VIET, T. A., BOWERS, P., SIDORENKO, J., KARLSSON LINNER, R., FONTANA, M. A., KUNDU, T., LEE, C., LI, H., LI, R., ROYER, R., TIMSHEL, P. N., WALTERS, R. K., WILLOUGHBY, E. A., YENGO, L., ANDME RESEARCH, T., COGENT, SOCIAL SCIENCE GENETIC ASSOCIATION, C., ALVER, M., BAO, Y., CLARK, D. W., DAY, F. R., FURLOTTE, N. A., JOSHI, P. K., KEMPER, K. E., KLEINMAN, A., LANGENBERG, C., MAGI, R., TRAMPUSH, J. W., VERMA, S. S., WU, Y., LAM, M., ZHAO, J. H., ZHENG, Z., BOARDMAN, J. D., CAMPBELL, H., FREESE, J., HARRIS, K. M., HAYWARD, C., HERD, P., KUMARI, M., LENCZ, T., LUAN, J., MALHOTRA, A. K., METSPALU, A., MILANI, L., ONG, K. K., PERRY, J. R. B., PORTEOUS, D. J., RITCHIE, M. D., SMART, M. C., SMITH, B. H., TUNG, J. Y., WAREHAM, N. J., WILSON, J. F., BEAUCHAMP, J. P., CONLEY, D. C., ESKO, T., LEHRER, S. F., MAGNUSSON, P. K. E., OSKARSSON, S., PERS, T. H., ROBINSON, M. R., THOM, K., WATSON, C., CHABRIS, C. F., MEYER, M. N., LAIBSON, D. I., YANG, J., JOHANNESSON, M., KOELLINGER, P. D., TURLEY, P., VISSCHER, P. M., BENJAMIN, D. J. & CESARINI, D. 2018. Gene discovery and polygenic prediction from a genome-wide association study of educational

attainment in 1.1 million individuals. *Nat Genet,* 50**,** 1112-1121. DOI:10.1038/s41588-018-0147-3

LEFORT, K. R., RUNGRATANAWANICH, W. & SONG, B. J. 2024. Contributing roles of mitochondrial dysfunction and hepatocyte apoptosis in liver diseases through oxidative stress, post-translational modifications, inflammation, and intestinal barrier dysfunction. *Cell Mol Life Sci,* 81**,** 34. DOI:10.1007/s00018-023-05061-7

LEUNG, C. L., GREEN, K. J. & LIEM, R. K. 2002. Plakins: a family of versatile cytolinker proteins. *Trends Cell Biol,* 12**,** 37-45. DOI:10.1016/s0962-8924(01)02180-8

LEWONTIN, R. C. 1964. The Interaction of Selection and Linkage. I. General Considerations; Heterotic Models. *Genetics,* 49**,** 49-67. DOI:10.1093/genetics/49.1.49

LIGTHART, S., VAEZ, A., VOSA, U., STATHOPOULOU, M. G., DE VRIES, P. S., PRINS, B. P., VAN DER MOST, P. J., TANAKA, T., NADERI, E., ROSE, L. M., WU, Y., KARLSSON, R., BARBALIC, M., LIN, H., POOL, R., ZHU, G., MACE, A., SIDORE, C., TROMPET, S., MANGINO, M., SABATER-LLEAL, M., KEMP, J. P., ABBASI, A., KACPROWSKI, T., VERWEIJ, N., SMITH, A. V., HUANG, T., MARZI, C., FEITOSA, M. F., LOHMAN, K. K., KLEBER, M. E., MILANESCHI, Y., MUELLER, C., HUQ, M., VLACHOPOULOU, E., LYYTIKAINEN, L. P., OLDMEADOW, C., DEELEN, J., PEROLA, M., ZHAO, J. H., FEENSTRA, B., LIFELINES COHORT, S., AMINI, M., GROUP, C. I. W., LAHTI, J., SCHRAUT, K. E., FORNAGE, M., SUKTITIPAT, B., CHEN, W. M., LI, X., NUTILE, T., MALERBA, G., LUAN, J., BAK, T., SCHORK, N., DEL GRECO, M. F., THIERING, E., MAHAJAN, A., MARIONI, R. E., MIHAILOV, E., ERIKSSON, J., OZEL, A. B., ZHANG, W., NETHANDER, M., CHENG, Y. C., ASLIBEKYAN, S., ANG, W., GANDIN, I., YENGO, L., PORTAS, L., KOOPERBERG, C., HOFER, E., RAJAN, K. B., SCHURMANN, C., DEN HOLLANDER, W., AHLUWALIA, T. S., ZHAO, J., DRAISMA, H. H. M., FORD, I., TIMPSON, N., TEUMER, A., HUANG, H., WAHL, S., LIU, Y., HUANG, J., UH, H. W., GELLER, F., JOSHI, P. K., YANEK, L. R., TRABETTI, E., LEHNE, B., VOZZI, D., VERBANCK, M., BIINO, G., SABA, Y., MEULENBELT, I., O'CONNELL, J. R., LAAKSO, M., GIULIANINI, F., MAGNUSSON, P. K. E., et al. 2018. Genome Analyses of >200,000 Individuals Identify 58 Loci for Chronic Inflammation and Highlight Pathways that Link Inflammation and Complex Disorders. *Am J Hum Genet,* 103**,** 691-706. DOI:10.1016/j.ajhg.2018.09.009

LIU, M., JIANG, Y., WEDOW, R., LI, Y., BRAZEL, D. M., CHEN, F., DATTA, G., DAVILA-VELDERRAIN, J., MCGUIRE, D., TIAN, C., ZHAN, X., ANDME RESEARCH, T., PSYCHIATRY, H. A.-I., CHOQUET, H., DOCHERTY, A. R., FAUL, J. D., FOERSTER, J. R., FRITSCHE, L. G., GABRIELSEN, M. E., GORDON, S. D., HAESSLER, J., HOTTENGA, J. J., HUANG, H., JANG, S. K., JANSEN, P. R., LING, Y., MAGI, R., MATOBA, N., MCMAHON, G., MULAS, A., ORRU, V., PALVIAINEN, T., PANDIT, A., REGINSSON, G. W., SKOGHOLT, A. H., SMITH, J. A., TAYLOR, A. E., TURMAN, C., WILLEMSEN, G., YOUNG, H., YOUNG, K. A., ZAJAC, G. J. M., ZHAO, W., ZHOU, W., BJORNSDOTTIR, G., BOARDMAN, J. D., BOEHNKE, M., BOOMSMA, D. I., CHEN, C., CUCCA, F., DAVIES, G. E., EATON, C. B., EHRINGER, M. A., ESKO, T., FIORILLO, E., GILLESPIE, N. A., GUDBJARTSSON, D. F., HALLER, T., HARRIS, K. M., HEATH, A. C., HEWITT, J. K., HICKIE, I. B., HOKANSON, J. E., HOPFER, C. J., HUNTER, D. J., IACONO, W. G., JOHNSON, E. O., KAMATANI, Y., KARDIA, S. L. R., KELLER, M. C., KELLIS, M., KOOPERBERG, C., KRAFT, P., KRAUTER, K. S., LAAKSO,

M., LIND, P. A., LOUKOLA, A., LUTZ, S. M., MADDEN, P. A. F., MARTIN, N. G., MCGUE, M., MCQUEEN, M. B., MEDLAND, S. E., METSPALU, A., MOHLKE, K. L., NIELSEN, J. B., OKADA, Y., PETERS, U., POLDERMAN, T. J. C., POSTHUMA, D., REINER, A. P., RICE, J. P., RIMM, E., ROSE, R. J., RUNARSDOTTIR, V., STALLINGS, M. C., STANCAKOVA, A., STEFANSSON, H., THAI, K. K., TINDLE, H. A., et al. 2019. Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use. *Nat Genet,* 51**,** 237-244. DOI:10.1038/s41588-018-0307-5

LIU, M., XU, P., GUAN, Z., QIAN, X., DOCKERY, P., FITZGERALD, U., O'BRIEN, T. & SHEN, S. 2018. Ulk4 deficiency leads to hypomyelination in mice. *Glia,* 66**,** 175-190. DOI:10.1002/glia.23236

LIU, W., XU, Y., WANG, A., HUANG, T. & LIU, Z. 2022. The eigen higher criticism and eigen Berk-Jones tests for multiple trait association studies based on GWAS summary statistics. *Genet Epidemiol,* 46**,** 89-104. DOI:10.1002/gepi.22439

LIU, Z. & LIN, X. 2019. A Geometric Perspective on the Power of Principal Component Association Tests in Multiple Phenotype Studies. *J Am Stat Assoc,* 114**,** 975-990. DOI:10.1080/01621459.2018.1513363

LIVINGSTON, G., HUNTLEY, J., SOMMERLAD, A., AMES, D., BALLARD, C., BANERJEE, S., BRAYNE, C., BURNS, A., COHEN-MANSFIELD, J., COOPER, C., COSTAFREDA, S. G., DIAS, A., FOX, N., GITLIN, L. N., HOWARD, R., KALES, H. C., KIVIMAKI, M., LARSON, E. B., OGUNNIYI, A., ORGETA, V., RITCHIE, K., ROCKWOOD, K., SAMPSON, E. L., SAMUS, Q., SCHNEIDER, L. S., SELBAEK, G., TERI, L. & MUKADAM, N. 2020. Dementia prevention, intervention, and care: 2020 report of the Lancet Commission. *Lancet,* 396**,** 413-446. DOI:10.1016/S0140-6736(20)30367-6

LUO, L., SHEN, J., ZHANG, H., CHHIBBER, A., MEHROTRA, D. V. & TANG, Z. Z. 2020. Multi-trait analysis of rare-variant association summary statistics using MTAR. *Nat Commun,* 11**,** 2850. DOI:10.1038/s41467-020-16591-0

LUO, S., ZHENG, N. & LANG, B. 2022. ULK4 in Neurodevelopmental and Neuropsychiatric Disorders. *Front Cell Dev Biol,* 10**,** 873706. DOI:10.3389/fcell.2022.873706

MAHLEY, R. W. 2016. Central Nervous System Lipoproteins: ApoE and Regulation of Cholesterol Metabolism. *Arterioscler Thromb Vasc Biol,* 36**,** 1305-15. DOI:10.1161/ATVBAHA.116.307023

MALIK, R., CHAUHAN, G., TRAYLOR, M., SARGURUPREMRAJ, M., OKADA, Y., MISHRA, A., RUTTEN-JACOBS, L., GIESE, A. K., VAN DER LAAN, S. W., GRETARSDOTTIR, S., ANDERSON, C. D., CHONG, M., ADAMS, H. H. H., AGO, T., ALMGREN, P., AMOUYEL, P., AY, H., BARTZ, T. M., BENAVENTE, O. R., BEVAN, S., BONCORAGLIO, G. B., BROWN, R. D., JR., BUTTERWORTH, A. S., CARRERA, C., CARTY, C. L., CHASMAN, D. I., CHEN, W. M., COLE, J. W., CORREA, A., COTLARCIUC, I., CRUCHAGA, C., DANESH, J., DE BAKKER, P. I. W., DESTEFANO, A. L., DEN HOED, M., DUAN, Q., ENGELTER, S. T., FALCONE, G. J., GOTTESMAN, R. F., GREWAL, R. P., GUDNASON, V., GUSTAFSSON, S., HAESSLER, J., HARRIS, T. B., HASSAN, A., HAVULINNA, A. S., HECKBERT, S. R., HOLLIDAY, E. G., HOWARD, G., HSU, F. C., HYACINTH, H. I., IKRAM, M. A., INGELSSON, E., IRVIN, M. R., JIAN, X., JIMENEZ-CONDE, J., JOHNSON, J. A., JUKEMA, J. W., KANAI, M., KEENE, K. L., KISSELA, B. M., KLEINDORFER, D. O., KOOPERBERG, C., KUBO, M., LANGE, L. A.,

LANGEFELD, C. D., LANGENBERG, C., LAUNER, L. J., LEE, J. M., LEMMENS, R., LEYS, D., LEWIS, C. M., LIN, W. Y., LINDGREN, A. G., LORENTZEN, E., MAGNUSSON, P. K., MAGUIRE, J., MANICHAIKUL, A., MCARDLE, P. F., MESCHIA, J. F., MITCHELL, B. D., MOSLEY, T. H., NALLS, M. A., NINOMIYA, T., O'DONNELL, M. J., PSATY, B. M., PULIT, S. L., RANNIKMAE, K., REINER, A. P., REXRODE, K. M., RICE, K., RICH, S. S., RIDKER, P. M., ROST, N. S., ROTHWELL, P. M., ROTTER, J. I., RUNDEK, T., SACCO, R. L., SAKAUE, S., SALE, M. M., et al. 2018. Multiancestry genome-wide association study of 520,000 subjects identifies 32 loci associated with stroke and stroke subtypes. *Nat Genet,* 50**,** 524-537. DOI:10.1038/s41588-018-0058-3

MANOLIO, T. A. & COLLINS, F. S. 2009. The HapMap and genome-wide association studies in diagnosis and therapy. *Annu Rev Med,* 60**,** 443-56. DOI:10.1146/annurev.med.60.061907.093117

MANOLIO, T. A., COLLINS, F. S., COX, N. J., GOLDSTEIN, D. B., HINDORFF, L. A., HUNTER, D. J., MCCARTHY, M. I., RAMOS, E. M., CARDON, L. R., CHAKRAVARTI, A., CHO, J. H., GUTTMACHER, A. E., KONG, A., KRUGLYAK, L., MARDIS, E., ROTIMI, C. N., SLATKIN, M., VALLE, D., WHITTEMORE, A. S., BOEHNKE, M., CLARK, A. G., EICHLER, E. E., GIBSON, G., HAINES, J. L., MACKAY, T. F., MCCARROLL, S. A. & VISSCHER, P. M. 2009. Finding the missing heritability of complex diseases. *Nature,* 461**,** 747-53. DOI:10.1038/nature08494

MARCHINI, J., CARDON, L. R., PHILLIPS, M. S. & DONNELLY, P. 2004. The effects of human population structure on large genetic association studies. *Nat Genet,* 36**,** 512-7. DOI:10.1038/ng1337

MARCHINI, J., HOWIE, B., MYERS, S., MCVEAN, G. & DONNELLY, P. 2007. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet,* 39**,** 906-13. DOI:10.1038/ng2088

MARKOZANNES, G., KOUTSIOUMPA, C., CIVIDINI, S., MONORI, G., TSILIDIS, K. K., KRETSAVOS, N., THEODORATOU, E., GILL, D., IOANNIDIS, J. P. & TZOULAKI, I. 2021. Global assessment of C-reactive protein and health-related outcomes: an umbrella review of evidence from observational studies and Mendelian randomization studies. *Eur J Epidemiol,* 36**,** 11-36. DOI:10.1007/s10654-020-00681-w

MATHUR, R., ROTROFF, D., MA, J., SHOJAIE, A. & MOTSINGER-REIF, A. 2018. Gene set analysis methods: a systematic comparison. *BioData Min,* 11**,** 8. DOI:10.1186/s13040-018-0166-8

MATOBA, N., LOVE, M. I. & STEIN, J. L. 2022. Evaluating brain structure traits as endophenotypes using polygenicity and discoverability. *Hum Brain Mapp,* 43**,** 329-340. DOI:10.1002/hbm.25257

MATSUMOTO, H., SCICLUNA, B. P., JIM, K. K., FALAHI, F., QIN, W., GURKAN, B., MALMSTROM, E., MEIJER, M. T., BUTLER, J. M., KHAN, H. N., TAKAGI, T., ISHII, S., SCHULTZ, M. J., VAN DE BEEK, D., DE VOS, A. F., VAN 'T VEER, C. & VAN DER POLL, T. 2021. HIVEP1 Is a Negative Regulator of NF-kappaB That Inhibits Systemic Inflammation in Sepsis. *Front Immunol,* 12**,** 744358. DOI:10.3389/fimmu.2021.744358

MAURANO, M. T., HUMBERT, R., RYNES, E., THURMAN, R. E., HAUGEN, E., WANG, H., REYNOLDS, A. P., SANDSTROM, R., QU, H., BRODY, J., SHAFER, A., NERI, F., LEE, K., KUTYAVIN, T., STEHLING-SUN, S., JOHNSON, A. K., CANFIELD, T. K., GISTE, E.,

DIEGEL, M., BATES, D., HANSEN, R. S., NEPH, S., SABO, P. J., HEIMFELD, S., RAUBITSCHEK, A., ZIEGLER, S., COTSAPAS, C., SOTOODEHNIA, N., GLASS, I., SUNYAEV, S. R., KAUL, R. & STAMATOYANNOPOULOS, J. A. 2012. Systematic localization of common disease-associated variation in regulatory DNA. *Science,* 337**,** 1190-5. DOI:10.1126/science.1222794

MCCARTHY, M. I., ABECASIS, G. R., CARDON, L. R., GOLDSTEIN, D. B., LITTLE, J., IOANNIDIS, J. P. & HIRSCHHORN, J. N. 2008. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet,* 9**,** 356-69. DOI:10.1038/nrg2344

MIHLAN, M., BLOM, A. M., KUPREISHVILI, K., LAUER, N., STELZNER, K., BERGSTROM, F., NIESSEN, H. W. & ZIPFEL, P. F. 2011. Monomeric C-reactive protein modulates classic complement activation on necrotic cells. *FASEB J,* 25**,** 4198-210. DOI:10.1096/fj.11-186460

MIYAMOTO, T., OZAKI, S., INUI, A., TANAKA, Y., YAMADA, Y. & MATSUMOTO, N. 2020. C1 esterase inhibitor in pediatric cardiac surgery with cardiopulmonary bypass plays a vital role in activation of the complement system. *Heart Vessels,* 35**,** 46-51. DOI:10.1007/s00380-019-01466-2

MORTON, N. E. & CHUNG, C. S. 1978. *Genetic epidemiology,* New York ; London, Academic Press.

NEALE LAB TEAM. 2018. *GWAS RESULTS round 2* [Online]. Available: http://www.nealelab.is/uk-biobank/ [Accessed].

NEDERGAARD, M., RANSOM, B. & GOLDMAN, S. A. 2003. New roles for astrocytes: redefining the functional architecture of the brain. *Trends Neurosci,* 26**,** 523-30. DOI:10.1016/j.tins.2003.08.008

NEWMAN, A. B., FITZPATRICK, A. L., LOPEZ, O., JACKSON, S., LYKETSOS, C., JAGUST, W., IVES, D., DEKOSKY, S. T. & KULLER, L. H. 2005. Dementia and Alzheimer's disease incidence in relationship to cardiovascular disease in the Cardiovascular Health Study cohort. *J Am Geriatr Soc,* 53**,** 1101-7. DOI:10.1111/j.1532-5415.2005.53360.x

NIELSEN, J. B., THOROLFSDOTTIR, R. B., FRITSCHE, L. G., ZHOU, W., SKOV, M. W., GRAHAM, S. E., HERRON, T. J., MCCARTHY, S., SCHMIDT, E. M., SVEINBJORNSSON, G., SURAKKA, I., MATHIS, M. R., YAMAZAKI, M., CRAWFORD, R. D., GABRIELSEN, M. E., SKOGHOLT, A. H., HOLMEN, O. L., LIN, M., WOLFORD, B. N., DEY, R., DALEN, H., SULEM, P., CHUNG, J. H., BACKMAN, J. D., ARNAR, D. O., THORSTEINSDOTTIR, U., BARAS, A., O'DUSHLAINE, C., HOLST, A. G., WEN, X., HORNSBY, W., DEWEY, F. E., BOEHNKE, M., KHETERPAL, S., MUKHERJEE, B., LEE, S., KANG, H. M., HOLM, H., KITZMAN, J., SHAVIT, J. A., JALIFE, J., BRUMMETT, C. M., TESLOVICH, T. M., CAREY, D. J., GUDBJARTSSON, D. F., STEFANSSON, K., ABECASIS, G. R., HVEEM, K. & WILLER, C. J. 2018. Biobank-driven genomic discovery yields new insight into atrial fibrillation biology. *Nat Genet,* 50**,** 1234-1239. DOI:10.1038/s41588-018-0171-3

NIKPAY, M., GOEL, A., WON, H. H., HALL, L. M., WILLENBORG, C., KANONI, S., SALEHEEN, D., KYRIAKOU, T., NELSON, C. P., HOPEWELL, J. C., WEBB, T. R., ZENG, L., DEHGHAN, A., ALVER, M., ARMASU, S. M., AURO, K., BJONNES, A., CHASMAN, D. I., CHEN, S., FORD, I., FRANCESCHINI, N., GIEGER, C., GRACE, C., GUSTAFSSON, S., HUANG, J., HWANG, S. J., KIM, Y. K., KLEBER, M. E., LAU, K. W., LU, X., LU, Y., LYYTIKAINEN, L. P., MIHAILOV, E., MORRISON, A. C., PERVJAKOVA,

N., QU, L., ROSE, L. M., SALFATI, E., SAXENA, R., SCHOLZ, M., SMITH, A. V., TIKKANEN, E., UITTERLINDEN, A., YANG, X., ZHANG, W., ZHAO, W., DE ANDRADE, M., DE VRIES, P. S., VAN ZUYDAM, N. R., ANAND, S. S., BERTRAM, L., BEUTNER, F., DEDOUSSIS, G., FROSSARD, P., GAUGUIER, D., GOODALL, A. H., GOTTESMAN, O., HABER, M., HAN, B. G., HUANG, J., JALILZADEH, S., KESSLER, T., KONIG, I. R., LANNFELT, L., LIEB, W., LIND, L., LINDGREN, C. M., LOKKI, M. L., MAGNUSSON, P. K., MALLICK, N. H., MEHRA, N., MEITINGER, T., MEMON, F. U., MORRIS, A. P., NIEMINEN, M. S., PEDERSEN, N. L., PETERS, A., RALLIDIS, L. S., RASHEED, A., SAMUEL, M., SHAH, S. H., SINISALO, J., STIRRUPS, K. E., TROMPET, S., WANG, L., ZAMAN, K. S., ARDISSINO, D., BOERWINKLE, E., BORECKI, I. B., BOTTINGER, E. P., BURING, J. E., CHAMBERS, J. C., COLLINS, R., CUPPLES, L. A., DANESH, J., DEMUTH, I., ELOSUA, R., EPSTEIN, S. E., ESKO, T., FEITOSA, M. F., et al. 2015. A comprehensive 1,000 Genomes-based genome-wide association meta-analysis of coronary artery disease. *Nat Genet,* 47**,** 1121-1130. DOI:10.1038/ng.3396

NOAKES, C. J., LEE, G. & LOWE, M. 2011. The PH domain proteins IPIP27A and B link OCRL1 to receptor recycling in the endocytic pathway. *Mol Biol Cell,* 22**,** 606-23. DOI:10.1091/mbc.E10-08-0730

O'REILLY, P. F., HOGGART, C. J., POMYEN, Y., CALBOLI, F. C., ELLIOTT, P., JARVELIN, M. R. & COIN, L. J. 2012. MultiPhen: joint model of multiple phenotypes can increase discovery in GWAS. *PLoS One,* 7**,** e34861. DOI:10.1371/journal.pone.0034861

OTT, J., KAMATANI, Y. & LATHROP, M. 2011. Family-based designs for genome-wide association studies. *Nat Rev Genet,* 12**,** 465-74. DOI:10.1038/nrg2989

PE'ER, I., YELENSKY, R., ALTSHULER, D. & DALY, M. J. 2008. Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. *Genet Epidemiol,* 32**,** 381-5. DOI:10.1002/gepi.20303

PICKRELL, J. K., BERISA, T., LIU, J. Z., SEGUREL, L., TUNG, J. Y. & HINDS, D. A. 2016. Detection and interpretation of shared genetic influences on 42 human traits. *Nat Genet,* 48**,** 709-17. DOI:10.1038/ng.3570

PLOMIN, R., DEFRIES, J. C., KNOPIK, V. S. & NEIDERHISER, J. M. 2016. Top 10 Replicated Findings From Behavioral Genetics. *Perspect Psychol Sci,* 11**,** 3-23. DOI:10.1177/1745691615617439

POTOKAR, M. & JORGACEVSKI, J. 2021. Plectin in the Central Nervous System and a Putative Role in Brain Astrocytes. *Cells,* 10. DOI:10.3390/cells10092353

POTOKAR, M., MORITA, M., WICHE, G. & JORGACEVSKI, J. 2020. The Diversity of Intermediate Filaments in Astrocytes. *Cells,* 9. DOI:10.3390/cells9071604

PRICE, A. L., PATTERSON, N. J., PLENGE, R. M., WEINBLATT, M. E., SHADICK, N. A. & REICH, D. 2006. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet,* 38**,** 904-9. DOI:10.1038/ng1847

PULIT, S. L., STONEMAN, C., MORRIS, A. P., WOOD, A. R., GLASTONBURY, C. A., TYRRELL, J., YENGO, L., FERREIRA, T., MAROULI, E., JI, Y., YANG, J., JONES, S., BEAUMONT, R., CROTEAU-CHONKA, D. C., WINKLER, T. W., CONSORTIUM, G., HATTERSLEY, A. T., LOOS, R. J. F., HIRSCHHORN, J. N., VISSCHER, P. M., FRAYLING, T. M., YAGHOOTKAR, H. & LINDGREN, C. M. 2019. Meta-analysis of genome-wide association studies for body fat distribution in 694 649 individuals of European ancestry. *Hum Mol Genet,* 28**,** 166-174. DOI:10.1093/hmg/ddy327

QI, G. & CHATTERJEE, N. 2018. Heritability informed power optimization (HIPO) leads to enhanced detection of genetic associations across multiple traits. *PLoS Genet,* 14**,** e1007549. DOI:10.1371/journal.pgen.1007549

RAY, D. & BOEHNKE, M. 2018. Methods for meta-analysis of multiple traits using GWAS summary statistics. *Genet Epidemiol,* 42**,** 134-145. DOI:10.1002/gepi.22105

RAY, D. & CHATTERJEE, N. 2020. A powerful method for pleiotropic analysis under composite null hypothesis identifies novel shared loci between Type 2 Diabetes and Prostate Cancer. *PLoS Genet,* 16**,** e1009218. DOI:10.1371/journal.pgen.1009218

RICHMOND, R. C. & DAVEY SMITH, G. 2022. Mendelian Randomization: Concepts and Scope. *Cold Spring Harb Perspect Med,* 12. DOI:10.1101/cshperspect.a040501

RISCH, N. & MERIKANGAS, K. 1996. The future of genetic studies of complex human diseases. *Science,* 273**,** 1516-7. DOI:10.1126/science.273.5281.1516

RIVARD, L., FRIBERG, L., CONEN, D., HEALEY, J. S., BERGE, T., BORIANI, G., BRANDES, A., CALKINS, H., CAMM, A. J., YEE CHEN, L., LLUIS CLUA ESPUNY, J., COLLINS, R., CONNOLLY, S., DAGRES, N., ELKIND, M. S. V., ENGDAHL, J., FIELD, T. S., GERSH, B. J., GLOTZER, T. V., HANKEY, G. J., HARBISON, J. A., HAEUSLER, K. G., HILLS, M. T., JOHNSON, L. S. B., JOUNG, B., KHAIRY, P., KIRCHHOF, P., KRIEGER, D., LIP, G. Y. H., LOCHEN, M. L., MADHAVAN, M., MAIRESSE, G. H., MONTANER, J., NTAIOS, G., QUINN, T. J., RIENSTRA, M., ROSENQVIST, M., SANDHU, R. K., SMYTH, B., SCHNABEL, R. B., STAVRAKIS, S., THEMISTOCLAKIS, S., VAN GELDER, I. C., WANG, J. G. & FREEDMAN, B. 2022. Atrial Fibrillation and Dementia: A Report From the AF-SCREEN International Collaboration. *Circulation,* 145**,** 392-409. DOI:10.1161/CIRCULATIONAHA.121.055018

RIZZI, L., ROSSET, I. & RORIZ-CRUZ, M. 2014. Global epidemiology of dementia: Alzheimer's and vascular types. *Biomed Res Int,* 2014**,** 908915. DOI:10.1155/2014/908915

ROADMAP EPIGENOMICS CONSORTIUM, KUNDAJE, A., MEULEMAN, W., ERNST, J., BILENKY, M., YEN, A., HERAVI-MOUSSAVI, A., KHERADPOUR, P., ZHANG, Z., WANG, J., ZILLER, M. J., AMIN, V., WHITAKER, J. W., SCHULTZ, M. D., WARD, L. D., SARKAR, A., QUON, G., SANDSTROM, R. S., EATON, M. L., WU, Y. C., PFENNING, A. R., WANG, X., CLAUSSNITZER, M., LIU, Y., COARFA, C., HARRIS, R. A., SHORESH, N., EPSTEIN, C. B., GJONESKA, E., LEUNG, D., XIE, W., HAWKINS, R. D., LISTER, R., HONG, C., GASCARD, P., MUNGALL, A. J., MOORE, R., CHUAH, E., TAM, A., CANFIELD, T. K., HANSEN, R. S., KAUL, R., SABO, P. J., BANSAL, M. S., CARLES, A., DIXON, J. R., FARH, K. H., FEIZI, S., KARLIC, R., KIM, A. R., KULKARNI, A., LI, D., LOWDON, R., ELLIOTT, G., MERCER, T. R., NEPH, S. J., ONUCHIC, V., POLAK, P., RAJAGOPAL, N., RAY, P., SALLARI, R. C., SIEBENTHALL, K. T., SINNOTT-ARMSTRONG, N. A., STEVENS, M., THURMAN, R. E., WU, J., ZHANG, B., ZHOU, X., BEAUDET, A. E., BOYER, L. A., DE JAGER, P. L., FARNHAM, P. J., FISHER, S. J., HAUSSLER, D., JONES, S. J., LI, W., MARRA, M. A., MCMANUS, M. T., SUNYAEV, S., THOMSON, J. A., TLSTY, T. D., TSAI, L. H., WANG, W., WATERLAND, R. A., ZHANG, M. Q., CHADWICK, L. H., BERNSTEIN, B. E., COSTELLO, J. F., ECKER, J. R., HIRST, M., MEISSNER, A., MILOSAVLJEVIC, A., REN, B., STAMATOYANNOPOULOS, J. A., WANG, T. & KELLIS, M. 2015. Integrative analysis of 111 reference human epigenomes. *Nature,* 518**,** 317-30. DOI:10.1038/nature14248

ROTHMAN, K. J., GREENLAND, S. & LASH, T. L. 2008. *Modern epidemiology,* Philadelphia, Wolters Kluwer Health/Lippincott Williams & Wilkins.

ROWE, S. J. & TENESA, A. 2012. Human complex trait genetics: lifting the lid of the genomics toolbox - from pathways to prediction. *Curr Genomics,* 13**,** 213-24. DOI:10.2174/138920212800543101

SAID, S., PAZOKI, R., KARHUNEN, V., VOSA, U., LIGTHART, S., BODINIER, B., KOSKERIDIS, F., WELSH, P., ALIZADEH, B. Z., CHASMAN, D. I., SATTAR, N., CHADEAU-HYAM, M., EVANGELOU, E., JARVELIN, M. R., ELLIOTT, P., TZOULAKI, I. & DEHGHAN, A. 2022. Genetic analysis of over half a million people characterises C-reactive protein loci. *Nat Commun,* 13**,** 2198. DOI:10.1038/s41467-022-29650-5

SANCHEZ-VALLE, J. & VALENCIA, A. 2023. Molecular bases of comorbidities: present and future perspectives. *Trends Genet,* 39**,** 773-786. DOI:10.1016/j.tig.2023.06.003

SANDERSON, E. 2021. Multivariable Mendelian Randomization and Mediation. *Cold Spring Harb Perspect Med,* 11. DOI:10.1101/cshperspect.a038984

SCHAID, D. J., CHEN, W. & LARSON, N. B. 2018. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat Rev Genet,* 19**,** 491-504. DOI:10.1038/s41576-018-0016-z

SCHWARTZENTRUBER, J., COOPER, S., LIU, J. Z., BARRIO-HERNANDEZ, I., BELLO, E., KUMASAKA, N., YOUNG, A. M. H., FRANKLIN, R. J. M., JOHNSON, T., ESTRADA, K., GAFFNEY, D. J., BELTRAO, P. & BASSETT, A. 2021. Genome-wide meta-analysis, fine-mapping and integrative prioritization implicate new Alzheimer's disease risk genes. *Nat Genet,* 53**,** 392-402. DOI:10.1038/s41588-020-00776-w

SHANG, R. & RODRIGUES, B. 2021. Lipoprotein Lipase and Its Delivery of Fatty Acids to the Heart. *Biomolecules,* 11. DOI:10.3390/biom11071016

SIVAKUMARAN, S., AGAKOV, F., THEODORATOU, E., PRENDERGAST, J. G., ZGAGA, L., MANOLIO, T., RUDAN, I., MCKEIGUE, P., WILSON, J. F. & CAMPBELL, H. 2011. Abundant pleiotropy in human complex diseases and traits. *Am J Hum Genet,* 89**,** 607-18. DOI:10.1016/j.ajhg.2011.10.004

SLATKIN, M. 2008. Linkage disequilibrium--understanding the evolutionary past and mapping the medical future. *Nat Rev Genet,* 9**,** 477-85. DOI:10.1038/nrg2361

SMITH, G. D. & EBRAHIM, S. 2003. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *Int J Epidemiol,* 32**,** 1-22. DOI:10.1093/ije/dyg070

SOLOVIEFF, N., COTSAPAS, C., LEE, P. H., PURCELL, S. M. & SMOLLER, J. W. 2013. Pleiotropy in complex traits: challenges and strategies. *Nat Rev Genet,* 14**,** 483-95. DOI:10.1038/nrg3461

SONNENBERG, A. & LIEM, R. K. 2007. Plakins in development and disease. *Exp Cell Res,* 313**,** 2189-203. DOI:10.1016/j.yexcr.2007.03.039

SPIELMAN, R. S., MCGINNIS, R. E. & EWENS, W. J. 1993. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am J Hum Genet,* 52**,** 506-16.

SPURRIER, J., NICHOLSON, L., FANG, X. T., STONER, A. J., TOYONAGA, T., HOLDEN, D., SIEGERT, T. R., LAIRD, W., ALLNUTT, M. A., CHIASSEU, M., BRODY, A. H., TAKAHASHI, H., NIES, S. H., PEREZ-CANAMAS, A., SADASIVAM, P., LEE, S., LI, S., ZHANG, L., HUANG, Y. H., CARSON, R. E., CAI, Z. & STRITTMATTER, S. M. 2022. Reversal of synapse loss in Alzheimer mouse models by targeting mGluR5 to

prevent synaptic tagging by C1Q. *Sci Transl Med,* 14**,** eabi8593. DOI:10.1126/scitranslmed.abi8593

STEIN, C. M., MORRIS, N. J. & NOCK, N. L. 2012. Structural equation modeling. *Methods Mol Biol,* 850**,** 495-512. DOI:10.1007/978-1-61779-555-8_27

STEPHENS, M. 2013. A unified framework for association analysis with multiple related phenotypes. *PLoS One,* 8**,** e65245. DOI:10.1371/journal.pone.0065245

SWEENEY, M. D., MONTAGNE, A., SAGARE, A. P., NATION, D. A., SCHNEIDER, L. S., CHUI, H. C., HARRINGTON, M. G., PA, J., LAW, M., WANG, D. J. J., JACOBS, R. E., DOUBAL, F. N., RAMIREZ, J., BLACK, S. E., NEDERGAARD, M., BENVENISTE, H., DICHGANS, M., IADECOLA, C., LOVE, S., BATH, P. M., MARKUS, H. S., AL-SHAHI SALMAN, R., ALLAN, S. M., QUINN, T. J., KALARIA, R. N., WERRING, D. J., CARARE, R. O., TOUYZ, R. M., WILLIAMS, S. C. R., MOSKOWITZ, M. A., KATUSIC, Z. S., LUTZ, S. E., LAZAROV, O., MINSHALL, R. D., REHMAN, J., DAVIS, T. P., WELLINGTON, C. L., GONZALEZ, H. M., YUAN, C., LOCKHART, S. N., HUGHES, T. M., CHEN, C. L. H., SACHDEV, P., O'BRIEN, J. T., SKOOG, I., PANTONI, L., GUSTAFSON, D. R., BIESSELS, G. J., WALLIN, A., SMITH, E. E., MOK, V., WONG, A., PASSMORE, P., BARKOF, F., MULLER, M., BRETELER, M. M. B., ROMAN, G. C., HAMEL, E., SESHADRI, S., GOTTESMAN, R. F., VAN BUCHEM, M. A., ARVANITAKIS, Z., SCHNEIDER, J. A., DREWES, L. R., HACHINSKI, V., FINCH, C. E., TOGA, A. W., WARDLAW, J. M. & ZLOKOVIC, B. V. 2019. Vascular dysfunction-The disregarded partner of Alzheimer's disease. *Alzheimers Dement,* 15**,** 158-167. DOI:10.1016/j.jalz.2018.07.222

TABOR, H. K., RISCH, N. J. & MYERS, R. M. 2002. Candidate-gene approaches for studying complex genetic traits: practical considerations. *Nat Rev Genet,* 3**,** 391-7. DOI:10.1038/nrg796

TACHMAZIDOU, I., HATZIKOTOULAS, K., SOUTHAM, L., ESPARZA-GORDILLO, J., HABERLAND, V., ZHENG, J., JOHNSON, T., KOPRULU, M., ZENGINI, E., STEINBERG, J., WILKINSON, J. M., BHATNAGAR, S., HOFFMAN, J. D., BUCHAN, N., SUVEGES, D., ARC, O. C., YERGES-ARMSTRONG, L., SMITH, G. D., GAUNT, T. R., SCOTT, R. A., MCCARTHY, L. C. & ZEGGINI, E. 2019. Identification of new therapeutic targets for osteoarthritis through genome-wide analyses of UK Biobank data. *Nat Genet,* 51**,** 230-236. DOI:10.1038/s41588-018-0327-1

TARASZKA, K., ZAITLEN, N. & ESKIN, E. 2022. Leveraging pleiotropy for joint analysis of genome-wide association studies with per trait interpretations. *PLoS Genet,* 18**,** e1010447. DOI:10.1371/journal.pgen.1010447

THOROLFSDOTTIR, R. B., SVEINBJORNSSON, G., SULEM, P., HELGADOTTIR, A., GRETARSDOTTIR, S., BENONISDOTTIR, S., MAGNUSDOTTIR, A., DAVIDSSON, O. B., RAJAMANI, S., RODEN, D. M., DARBAR, D., PEDERSEN, T. R., SABATINE, M. S., JONSDOTTIR, I., ARNAR, D. O., THORSTEINSDOTTIR, U., GUDBJARTSSON, D. F., HOLM, H. & STEFANSSON, K. 2017. A Missense Variant in PLEC Increases Risk of Atrial Fibrillation. *J Am Coll Cardiol,* 70**,** 2157-2168. DOI:10.1016/j.jacc.2017.09.005

TIMPSON, N. J., GREENWOOD, C. M. T., SORANZO, N., LAWSON, D. J. & RICHARDS, J. B. 2018. Genetic architecture: the shape of the genetic contribution to human traits and disease. *Nat Rev Genet,* 19**,** 110-124. DOI:10.1038/nrg.2017.101

TORKAMANI, A., WINEINGER, N. E. & TOPOL, E. J. 2018. The personal and clinical utility of polygenic risk scores. *Nat Rev Genet,* 19**,** 581-590. DOI:10.1038/s41576-018-0018-x

TUBLIN, J. M., ADELSTEIN, J. M., DEL MONTE, F., COMBS, C. K. & WOLD, L. E. 2019. Getting to the Heart of Alzheimer Disease. *Circ Res,* 124**,** 142-149. DOI:10.1161/CIRCRESAHA.118.313563

TURLEY, P., WALTERS, R. K., MAGHZIAN, O., OKBAY, A., LEE, J. J., FONTANA, M. A., NGUYEN-VIET, T. A., WEDOW, R., ZACHER, M., FURLOTTE, N. A., ANDME RESEARCH, T., SOCIAL SCIENCE GENETIC ASSOCIATION, C., MAGNUSSON, P., OSKARSSON, S., JOHANNESSON, M., VISSCHER, P. M., LAIBSON, D., CESARINI, D., NEALE, B. M. & BENJAMIN, D. J. 2018. Multi-trait analysis of genome-wide association summary statistics using MTAG. *Nat Genet,* 50**,** 229-237. DOI:10.1038/s41588-017-0009-4

TYLER, A. L., CRAWFORD, D. C. & PENDERGRASS, S. A. 2016. The detection and characterization of pleiotropy: discovery, progress, and promise. *Brief Bioinform,* 17**,** 13-22. DOI:10.1093/bib/bbv050

UFFELMANN, E., HUANG, Q. Q., MUNUNG, N. S., DE VRIES, J., OKADA, Y., MARTIN, A. R., MARTIN, H. C., LAPPALAINEN, T. & POSTHUMA, D. 2021. Genome-wide association studies. *Nat Rev Methods Primers,* 59. DOI:10.1038/s43586-021-00056-9

VALDERAS, J. M., STARFIELD, B., SIBBALD, B., SALISBURY, C. & ROLAND, M. 2009. Defining comorbidity: implications for understanding health and health services. *Ann Fam Med,* 7**,** 357-63. DOI:10.1370/afm.983

VALENCIA, R. G., MIHAILOVSKA, E., WINTER, L., BAUER, K., FISCHER, I., WALKO, G., JORGACEVSKI, J., POTOKAR, M., ZOREC, R. & WICHE, G. 2021. Plectin dysfunction in neurons leads to tau accumulation on microtubules affecting neuritogenesis, organelle trafficking, pain sensitivity and memory. *Neuropathol Appl Neurobiol,* 47**,** 73-95. DOI:10.1111/nan.12635

VAN DER HARST, P. & VERWEIJ, N. 2018. Identification of 64 Novel Genetic Loci Provides an Expanded View on the Genetic Architecture of Coronary Artery Disease. *Circ Res,* 122**,** 433-443. DOI:10.1161/CIRCRESAHA.117.312086

VAN DER SLUIS, S., POSTHUMA, D. & DOLAN, C. V. 2013. TATES: efficient multivariate genotype-phenotype analysis for genome-wide association studies. *PLoS Genet,* 9**,** e1003235. DOI:10.1371/journal.pgen.1003235

VAN HOUWELINGEN, H. C., ARENDS, L. R. & STIJNEN, T. 2002. Advanced methods in meta-analysis: multivariate approach and meta-regression. *Stat Med,* 21**,** 589-624. DOI:10.1002/sim.1040

VENKATARAMAN, G. R., DEBOEVER, C., TANIGAWA, Y., AGUIRRE, M., IOANNIDIS, A. G., MOSTAFAVI, H., SPENCER, C. C. A., POTERBA, T., BUSTAMANTE, C. D., DALY, M. J., PIRINEN, M. & RIVAS, M. A. 2021. Bayesian model comparison for rare-variant association studies. *Am J Hum Genet,* 108**,** 2354-2367. DOI:10.1016/j.ajhg.2021.11.005

VERMA, A., BRADFORD, Y., DUDEK, S., LUCAS, A. M., VERMA, S. S., PENDERGRASS, S. A. & RITCHIE, M. D. 2018. A simulation study investigating power estimates in phenome-wide association studies. *BMC Bioinformatics,* 19**,** 120. DOI:10.1186/s12859-018-2135-0

VISSCHER, P. M., WRAY, N. R., ZHANG, Q., SKLAR, P., MCCARTHY, M. I., BROWN, M. A. & YANG, J. 2017. 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am J Hum Genet,* 101**,** 5-22. DOI:10.1016/j.ajhg.2017.06.005

VON BERG, J., TEN DAM, M., VAN DER LAAN, S. W. & DE RIDDER, J. 2022. PolarMorphism enables discovery of shared genetic variants across multiple traits from GWAS summary statistics. *Bioinformatics,* 38**,** i212-i219. DOI:10.1093/bioinformatics/btac228

VROOM, C., DE LEEUW, C., POSTHUMA, D., DOLAN, C. & VAN DER SLUIS, S. 2019. The more the merrier? Multivariate approaches to genome-wide association analysis. *bioRxiv*. DOI:10.1101/610287

VUCKOVIC, D., BAO, E. L., AKBARI, P., LAREAU, C. A., MOUSAS, A., JIANG, T., CHEN, M. H., RAFFIELD, L. M., TARDAGUILA, M., HUFFMAN, J. E., RITCHIE, S. C., MEGY, K., PONSTINGL, H., PENKETT, C. J., ALBERS, P. K., WIGDOR, E. M., SAKAUE, S., MOSCATI, A., MANANSALA, R., LO, K. S., QIAN, H., AKIYAMA, M., BARTZ, T. M., BEN-SHLOMO, Y., BESWICK, A., BORK-JENSEN, J., BOTTINGER, E. P., BRODY, J. A., VAN ROOIJ, F. J. A., CHITRALA, K. N., WILSON, P. W. F., CHOQUET, H., DANESH, J., DI ANGELANTONIO, E., DIMOU, N., DING, J., ELLIOTT, P., ESKO, T., EVANS, M. K., FELIX, S. B., FLOYD, J. S., BROER, L., GRARUP, N., GUO, M. H., GUO, Q., GREINACHER, A., HAESSLER, J., HANSEN, T., HOWSON, J. M. M., HUANG, W., JORGENSON, E., KACPROWSKI, T., KAHONEN, M., KAMATANI, Y., KANAI, M., KARTHIKEYAN, S., KOSKERIDIS, F., LANGE, L. A., LEHTIMAKI, T., LINNEBERG, A., LIU, Y., LYYTIKAINEN, L. P., MANICHAIKUL, A., MATSUDA, K., MOHLKE, K. L., MONONEN, N., MURAKAMI, Y., NADKARNI, G. N., NIKUS, K., PANKRATZ, N., PEDERSEN, O., PREUSS, M., PSATY, B. M., RAITAKARI, O. T., RICH, S. S., RODRIGUEZ, B. A. T., ROSEN, J. D., ROTTER, J. I., SCHUBERT, P., SPRACKLEN, C. N., SURENDRAN, P., TANG, H., TARDIF, J. C., GHANBARI, M., VOLKER, U., VOLZKE, H., WATKINS, N. A., WEISS, S., PROGRAM, V. A. M. V., CAI, N., KUNDU, K., WATT, S. B., WALTER, K., ZONDERMAN, A. B., CHO, K., LI, Y., LOOS, R. J. F., KNIGHT, J. C., GEORGES, M., STEGLE, O., et al. 2020. The Polygenic and Monogenic Basis of Blood Traits and Diseases. *Cell,* 182**,** 1214-1231 e11. DOI:10.1016/j.cell.2020.08.008

VUCKOVIC, D., GASPARINI, P., SORANZO, N. & IOTCHKOVA, V. 2015. MultiMeta: an R package for meta-analyzing multi-phenotype genome-wide association studies. *Bioinformatics,* 31**,** 2754-6. DOI:10.1093/bioinformatics/btv222

WALLACE, C. 2021. A more accurate method for colocalisation analysis allowing for multiple causal variants. *PLoS Genet,* 17**,** e1009440. DOI:10.1371/journal.pgen.1009440

WANG, C., WU, W., YANG, H., YE, Z., ZHAO, Y., LIU, J. & MU, L. 2022. Mendelian randomization analyses for PCOS: evidence, opportunities, and challenges. *Trends Genet,* 38**,** 468-482. DOI:10.1016/j.tig.2022.01.005

WANG, K., LI, M. & HAKONARSON, H. 2010. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res,* 38**,** e164. DOI:10.1093/nar/gkq603

WANG, L., ZHANG, X., MENG, X., KOSKERIDIS, F., GEORGIOU, A., YU, L., CAMPBELL, H., THEODORATOU, E. & LI, X. 2021. Methodology in phenome-wide association studies: a systematic review. *J Med Genet,* 58**,** 720-728. DOI:10.1136/jmedgenet-2021-107696

WANG, N., AKEY, J. M., ZHANG, K., CHAKRABORTY, R. & JIN, L. 2002. Distribution of recombination crossovers and the origin of haplotype blocks: the interplay of population history, recombination, and mutation. *Am J Hum Genet,* 71**,** 1227-34. DOI:10.1086/344398

WATANABE, K., STRINGER, S., FREI, O., UMICEVIC MIRKOV, M., DE LEEUW, C., POLDERMAN, T. J. C., VAN DER SLUIS, S., ANDREASSEN, O. A., NEALE, B. M. & POSTHUMA, D. 2019. A global overview of pleiotropy and genetic architecture in complex traits. *Nat Genet,* 51**,** 1339-1348. DOI:10.1038/s41588-019-0481-0

WATANABE, K., TASKESEN, E., VAN BOCHOVEN, A. & POSTHUMA, D. 2017. Functional mapping and annotation of genetic associations with FUMA. *Nat Commun,* 8**,** 1826. DOI:10.1038/s41467-017-01261-5

WU, C. 2020. Multi-trait Genome-Wide Analyses of the Brain Imaging Phenotypes in UK Biobank. *Genetics,* 215**,** 947-958. DOI:10.1534/genetics.120.303242

WU, N. N. & REN, J. 2019. Aldehyde Dehydrogenase 2 (ALDH2) and Aging: Is There a Sensible Link? *Adv Exp Med Biol,* 1193**,** 237-253. DOI:10.1007/978-981-13-6260-6_15

XIONG, Z., GAO, X., CHEN, Y., FENG, Z., PAN, S., LU, H., UITTERLINDEN, A. G., NIJSTEN, T., IKRAM, A., RIVADENEIRA, F., GHANBARI, M., WANG, Y., KAYSER, M. & LIU, F. 2022. Combining genome-wide association studies highlight novel loci involved in human facial variation. *Nat Commun,* 13**,** 7832. DOI:10.1038/s41467-022-35328-9

YANG, J., LEE, S. H., GODDARD, M. E. & VISSCHER, P. M. 2011. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet,* 88**,** 76-82. DOI:10.1016/j.ajhg.2010.11.011

YANG, J. J., LI, J., WILLIAMS, L. K. & BUU, A. 2016. An efficient genome-wide association test for multivariate phenotypes based on the Fisher combination function. *BMC Bioinformatics,* 17**,** 19. DOI:10.1186/s12859-015-0868-6

YANG, Q. & WANG, Y. 2012. Methods for Analyzing Multivariate Phenotypes in Genetic Association Studies. *J Probab Stat,* 2012**,** 652569. DOI:10.1155/2012/652569

YANG, Y., BASU, S. & ZHANG, L. 2022. A Bayesian hierarchically structured prior for gene-based association testing with multiple traits in genome-wide association studies. *Genet Epidemiol,* 46**,** 63-72. DOI:10.1002/gepi.22437

ZENG, P., HAO, X. & ZHOU, X. 2018. Pleiotropic mapping and annotation selection in genome-wide association studies with penalized Gaussian mixture models. *Bioinformatics,* 34**,** 2797-2807. DOI:10.1093/bioinformatics/bty204

ZHENG, G., YANG, Y., ZHU, X., ELSTON, R. C. & SPRINGERLINK 2012. *Analysis of Genetic Association Studies,* New York, NY, Springer New York : Imprint: Springer.

ZHOU, X. & STEPHENS, M. 2014. Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nat Methods,* 11**,** 407-9. DOI:10.1038/nmeth.2848

ZHU, H., ZHANG, S. & SHA, Q. 2015a. Power Comparisons of Methods for Joint Association Analysis of Multiple Phenotypes. *Hum Hered,* 80**,** 144-52. DOI:10.1159/000446239

ZHU, X., FENG, T., TAYO, B. O., LIANG, J., YOUNG, J. H., FRANCESCHINI, N., SMITH, J. A., YANEK, L. R., SUN, Y. V., EDWARDS, T. L., CHEN, W., NALLS, M., FOX, E., SALE, M., BOTTINGER, E., ROTIMI, C., CONSORTIUM, C. B., LIU, Y., MCKNIGHT, B., LIU, K., ARNETT, D. K., CHAKRAVATI, A., COOPER, R. S. & REDLINE, S. 2015b. Meta-

analysis of correlated traits via summary statistics from GWASs with an application in hypertension. *Am J Hum Genet,* 96**,** 21-36. DOI:10.1016/j.ajhg.2014.11.011

ZHU, Z., ZHANG, F., HU, H., BAKSHI, A., ROBINSON, M. R., POWELL, J. E., MONTGOMERY, G. W., GODDARD, M. E., WRAY, N. R., VISSCHER, P. M. & YANG, J. 2016. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat Genet,* 48**,** 481-7. DOI:10.1038/ng.3538

ZUBER, V., GRINBERG, N. F., GILL, D., MANIPUR, I., SLOB, E. A. W., PATEL, A., WALLACE, C. & BURGESS, S. 2022. Combining evidence from Mendelian randomization and colocalization: Review and comparison of approaches. *Am J Hum Genet,* 109**,** 767-782. DOI:10.1016/j.ajhg.2022.04.001

ZUBER, V., LEWIN, A., LEVIN, M. G., HAGLUND, A., BEN-AICHA, S., EMANUELI, C., DAMRAUER, S., BURGESS, S., GILL, D. & BOTTOLO, L. 2023. Multi-response Mendelian randomization: Identification of shared and distinct exposures for multimorbidity and multiple related disease outcomes. *Am J Hum Genet,* 110**,** 1177-1199. DOI:10.1016/j.ajhg.2023.06.005

APPENDIX

A. SUPPLEMENTARY TABLES

Supplementary Table 1 Tissue enrichment of the genes associated with blood cell traits after excluding previously known loci

Phenotype	Tissue	N Genes	Beta	SE	P	FDR
BAS	Spleen	15411	0.116	0.012	3.3E-22	1.0E-20
BAS	Blood	15411	0.081	0.009	3.5E-18	5.3E-17
LYM	Spleen	12044	0.108	0.014	1.2E-15	3.5E-14
EOS	Spleen	12536	0.095	0.013	8.7E-14	2.6E-12
WBC	Spleen	11995	0.102	0.014	1.1E-13	3.2E-12
BAS	Lung	15411	0.113	0.016	5.5E-13	5.5E-12
MON	Spleen	12144	0.091	0.014	1.4E-11	4.1E-10
EOS	Blood	12536	0.063	0.010	1.9E-10	2.9E-09
WBC	Blood	11995	0.064	0.011	1.7E-09	2.5E-08
MPV	Spleen	12205	0.075	0.013	2.7E-09	8.1E-08
LYM	Blood	12044	0.061	0.011	5.1E-09	7.6E-08
EOS	Lung	12536	0.097	0.017	5.8E-09	5.8E-08
NEU	Blood	12899	0.058	0.010	6.9E-09	2.1E-07
MPV	Blood	12205	0.058	0.010	1.1E-08	1.6E-07
WBC	Lung	11995	0.102	0.018	1.8E-08	1.8E-07
NEU	Spleen	12899	0.071	0.013	2.6E-08	3.9E-07
PLT	Spleen	11253	0.074	0.014	7.7E-08	2.3E-06
PLT	Lung	11253	0.095	0.019	1.7E-07	2.5E-06
LYM	Small Intestine	12044	0.094	0.019	1.8E-07	1.8E-06
NEU	Lung	12899	0.087	0.017	3.0E-07	3.0E-06
MCH	Blood	11932	0.051	0.010	4.2E-07	1.3E-05

MON	Lung	12144	0.088	0.018	4.4E-07	6.0E-06
MON	Blood	12144	0.052	0.011	6.0E-07	6.0E-06
MCHC	Blood	14569	0.043	0.009	7.1E-07	2.1E-05
BAS	Small Intestine	15411	0.077	0.016	1.0E-06	7.8E-06
LYM	Lung	12044	0.084	0.018	1.3E-06	9.9E-06
RDW	Blood	12500	0.045	0.010	4.5E-06	1.3E-04
BAS	Adipose Tissue	15411	0.082	0.020	2.5E-05	1.5E-04
MCV	Blood	11707	0.040	0.011	8.1E-05	2.4E-03
MCHC	Spleen	14569	0.043	0.011	9.2E-05	1.4E-03
WBC	Small Intestine	11995	0.067	0.018	1.4E-04	1.1E-03
RDW	Spleen	12500	0.046	0.013	1.7E-04	2.6E-03
MCHC	Lung	14569	0.054	0.015	1.9E-04	1.9E-03
MON	Adipose Tissue	12144	0.076	0.022	2.6E-04	2.0E-03
EOS	Small Intestine	12536	0.061	0.018	2.7E-04	2.0E-03
RBC	Spleen	11907	0.045	0.013	3.5E-04	8.0E-03
RBC	Ovary	11907	0.056	0.017	5.3E-04	8.0E-03
MON	Uterus	12144	0.064	0.020	6.3E-04	3.8E-03
MCH	Spleen	11932	0.041	0.013	7.6E-04	1.1E-02
PLT	Blood	11253	0.036	0.011	7.9E-04	7.9E-03
MCHC	Small Intestine	14569	0.046	0.016	1.4E-03	1.1E-02
MCV	Testis	11707	0.028	0.009	1.5E-03	1.7E-02
PLT	Small Intestine	11253	0.056	0.019	1.6E-03	1.2E-02

Phenotype: Blood cell phenotype associated with the genes enriched in the tissue, Tissue: General tissue type, N Genes: Number of genes enriched in the tissue, Beta: Regression coefficient of the genes associated with the phenotype on the tissue, SE: Standard error of beta, P: Nominal P-value of the association, FDR: P-value after correcting with False Discovery Rate method, MPV: Mean platelet volume, PLT: Platelets count, HCT: Haematocrit, HGB: Haemoglobin, MCH: Mean corpuscular haemoglobin, MCHC: Mean corpuscular haemoglobin concentration, MCV: Mean corpuscular volume, RBC: Red blood cells count, RDW: Red distribution width (RDW), BAS: Basophils count, EOS: Eosinophils count, NEU: Neutrophils count, LYM: Lymphocytes count, MON: Monocytes count, WBC: White blood cells count

Supplementary Table 2 Colocalized variants with broad effect across blood cell traits

Locus	SNP	Chr	Pos	Colocalised traits	N traits	PP	%PP explained
rs35014299	rs7554335	1	203655743	PLT, HCT, HGB, MCHC, MCV, RBC, RDW, BAS, LYM, MON, NEU, WBC	12	0.76	95%

rs3811444	rs3811444	1	248039451	PLT, HCT, HGB, MCH, MCHC, MCV, RBC, RDW, BAS, EOS, LYM, MON, NEU, WBC	14	0.88	100%
rs615632	rs615632	8	9796321	PLT, MCH, MCHC, MCV, RBC, RDW, BAS, LYM, MON, NEU, WBC	11	0.75	100%
rs10828725	rs11014291	10	25198687	PLT, MCH, MCHC, MCV, RDW, BAS, EOS, LYM, MON, NEU, WBC	11	0.79	87%
rs2283351	rs10849925	12	111495518	PLT, HCT, HGB, MCH, MCHC, RBC, RDW, BAS, EOS, LYM, MON, NEU, WBC	13	0.76	100%
rs111707684	rs11065987	12	112072424	PLT, HCT, HGB, MCH, RBC, BAS, EOS, LYM, MON, NEU, WBC	11	0.94	99%
rs11066309	rs11066309	12	112883476	PLT, HCT, HGB, MCH, RBC, BAS, EOS, LYM, MON, NEU, WBC	11	0.98	100%
rs17758695	rs17758695	18	60920854	MPV, PLT, HCT, MCH, MCHC, MCV, RBC, BAS, EOS, MON, NEU, WBC	12	0.98	100%
rs9977672	rs9977672	21	40463283	HCT, HGB, MCH, MCHC, MCV, RDW, BAS, LYM, NEU, WBC	10	0.91	97%

Locus: Genomic region represented by the top variant associated with blood cell traits identified by MTAG, Candidate causal SNP: Candidate variant explaining the shared association, Chr: Chromosome of the causal variant, Pos: Position of the causal variant, Colocalized traits: Traits that colocalize, N traits: Number of traits that colocalize, PP: Posterior Probability, %PP explained: Percentage of the PP explained by the SNP, MPV: Mean platelet volume, PLT: Platelets count, HCT: Haematocrit, HGB: Haemoglobin, MCH: Mean corpuscular haemoglobin, MCHC: Mean corpuscular haemoglobin concentration, MCV: Mean corpuscular volume, RBC: Red blood cells count, RDW: Red distribution width (RDW), BAS: Basophils count, EOS: Eosinophils count, NEU: Neutrophils count, LYM: Lymphocytes count, MON: Monocytes count, WBC: White blood cells count

Supplementary Table 3 Colocalized variants with global effects on white blood cells

Locus	Candidate causal SNP	Chr	Pos	Colocalised traits	N traits	PP	%PP explained
rs2476601	rs2476601	1	114377568	BAS, EOS, LYM, MON, NEU, WBC	6	0.93	100%
rs3811444	rs3811444	1	248039451	BAS, EOS, LYM, MON, NEU, WBC	6	0.84	100%
rs4706020	rs4706020	5	130674076	BAS, EOS, LYM, MON, NEU, WBC	6	0.99	96%
rs9257800	rs1235162	6	29537224	BAS, EOS, LYM, MON, NEU, WBC	6	1.00	100%
rs13207171	rs13207171	6	41165819	BAS, EOS, LYM, MON, NEU, WBC	6	0.93	96%
rs17710008	rs17710008	6	153043035	BAS, EOS, LYM, MON, NEU, WBC	6	0.91	100%
rs6952262	rs6952262	7	139719439	BAS, EOS, LYM, MON, NEU, WBC	6	0.90	93%
rs17041439	rs17041439	12	101873240	BAS, EOS, LYM, MON, NEU, WBC	6	0.97	100%
rs2283351	rs10849925	12	111495518	BAS, EOS, LYM, MON, NEU, WBC	6	0.99	100%
rs2891403	rs233701	12	113012149	BAS, EOS, LYM, MON, NEU, WBC	6	1.00	100%
rs7141943	rs2038700	14	25461989	BAS, EOS, LYM, MON, NEU, WBC	6	0.97	86%
rs247826	rs247826	16	84582965	BAS, EOS, LYM, MON, NEU, WBC	6	0.95	100%
rs571497	rs571497	19	7827830	BAS, EOS, LYM, MON, NEU, WBC	6	0.97	100%
rs150665764	rs150665764	19	14499614	BAS, EOS, LYM, MON, NEU, WBC	6	0.99	93%

Locus: Genomic region represented by the top variant associated with blood cell traits identified by MTAG, Candidate causal SNP: Candidate variant explaining the shared association, Chr: Chromosome of the causal variant, Pos: Position of the causal variant, Colocalized traits: Traits that colocalize, N traits: Number of traits that colocalize PP: Posterior Probability, %PP explained: Percentage of the PP explained by the SNP, BAS: Basophils count, EOS: Eosinophils count, NEU: Neutrophils count, LYM: Lymphocytes count, MON: Monocytes count, WBC: White blood cells count

Supplementary Table 4 Colocalized variants with global effects on red blood cells

Locus	Candidate causal SNP	Chr	Pos (GRCh37)	Colocalised traits	N traits	PP	%PP explained
rs3811444	rs3811444	1	248039451	HCT, HGB, MCH, MCHC, MCV, RBC, RDW	7	1.00	100%
rs218265	rs218264	4	55408875	HCT, HGB, MCH, MCHC, MCV, RBC, RDW	7	1.00	100%
rs11756844	rs78912080	6	25616453	HCT, HGB, MCH, MCHC, MCV, RBC, RDW	7	1.00	100%
rs35416932	rs71559014	6	27122444	HCT, HGB, MCH, MCHC, MCV, RBC, RDW	7	1.00	100%
rs143875230	rs143875230	15	43278726	HCT, HGB, MCH, MCHC, MCV, RBC, RDW	7	1.00	100%

Locus: Genomic region represented by the top variant associated with blood cell traits identified by MTAG, Candidate causal SNP: Candidate variant explaining the shared association, Chr: Chromosome of the causal variant, Pos: Position of the causal variant, Colocalized traits: Traits that colocalize, N traits: Number of traits that colocalize, PP: Posterior Probability, %PP explained: Percentage of the PP explained by the SNP, HCT: Haematocrit, HGB: Haemoglobin, MCH: Mean corpuscular haemoglobin, MCHC: Mean corpuscular haemoglobin concentration, MCV: Mean corpuscular volume, RBC: Red blood cells count, RDW: Red distribution width (RDW)

Supplementary Table 5 Causal variants with broad effect across blood cell traits with a shared gene identified by expression quantitative trait loci colocalization analysis

Locus	SNP	Chr	Pos (GRCh37)	Colocalised traits	Gene	Tissue	PP	%PP explained
rs615632	rs615171	8	9796189	PLT, MCH, MCHC, MCV, RBC, RDW, BAS, LYM, MON, NEU, WBC	RPL10P19	Spleen	0.79	100%
rs615632	rs615632	8	9796321	PLT, MCH, MCHC, MCV, RBC, RDW, BAS, LYM, MON, NEU, WBC	RPL10P19	Lung	0.79	100%
rs615632	rs615632	8	9796321	PLT, MCH, MCHC, MCV, RBC, RDW, BAS, LYM, MON, NEU, WBC	ERI1	Lung	0.78	100%
rs111707684	rs3184504	12	111884608	PLT, HCT, HGB, RBC, BAS, EOS, LYM, MON, NEU, WBC	SH2B3	Spleen	0.83	100%
rs111707684	rs3184504	12	111884608	PLT, HCT, HGB, RBC, BAS, EOS, LYM, MON, NEU, WBC	PHETA1	Spleen	0.78	100%
rs111707684	rs3184504	12	111884608	PLT, HCT, HGB, RBC, BAS, EOS, LYM, MON, NEU, WBC	LINC02356	Spleen	0.76	100%
rs11066309	rs7953257	12	112684221	PLT, HCT, HGB, MCH, RBC, BAS, EOS, LYM, MON, NEU, WBC	ALDH2	Whole Blood	0.89	100%
rs11066309	rs7953257	12	112684221	PLT, HCT, HGB, MCH, RBC, BAS, EOS, LYM, MON, NEU, WBC	PHETA1	Whole Blood	0.85	100%
rs9977672	rs9977672	21	40463283	HCT, HGB, MCH, MCHC, MCV, RDW, BAS, LYM, NEU, WBC	NA	Whole Blood	0.92	100%

rs9977672	rs9977672	21	40463283	HCT, HGB, MCH, MCHC, MCV, RDW, BAS, LYM, NEU, WBC	LINC02940	Whole Blood	0.91	99%
rs9977672	rs9977672	21	40463283	HCT, HGB, MCH, MCHC, MCV, RDW, BAS, LYM, NEU, WBC	ETS2	Spleen	0.91	99%
rs9977672	rs9977672	21	40463283	HCT, HGB, MCH, MCHC, MCV, RDW, BAS, LYM, NEU, WBC	ETS2	Whole Blood	0.91	95%
rs9977672	rs9977672	21	40463283	HCT, HGB, MCH, MCHC, MCV, RDW, BAS, LYM, NEU, WBC	LINC01700	Whole Blood	0.90	92%
rs9977672	rs9977672	21	40463283	HCT, HGB, MCH, MCHC, MCV, RDW, BAS, LYM, NEU, WBC	LINC02943	Whole Blood	0.85	94%
rs9977672	rs9977672	21	40463283	HCT, HGB, MCH, MCHC, MCV, RDW, BAS, LYM, NEU, WBC	NA	Spleen	0.79	98%
rs9977672	rs9977672	21	40463283	HCT, HGB, MCH, MCHC, MCV, RDW, BAS, LYM, NEU, WBC	LINC01700	Spleen	0.75	97%

Locus: Genomic region represented by the top variant associated with blood cell traits identified by MTAG, Candidate causal SNP: Candidate variant explaining the shared association, Chr: Chromosome of the causal variant, Pos: Position of the causal variant, Colocalized traits: Traits that colocalize, N traits: Number of traits that colocalize, Ensembl gene: Gene which is shared across the colocalized traits using the Ensembl ID, HUGO gene: Gene which is shared across the colocalized traits using the HUGO ID, Tissue: Tissue showing expression of the shared gene, PP: Posterior Probability, %PP explained: Percentage of the PP explained by the SNP, MPV: Mean platelet volume, PLT: Platelets count, HCT: Haematocrit, HGB: Haemoglobin, MCH: Mean corpuscular haemoglobin, MCHC: Mean corpuscular haemoglobin concentration, MCV: Mean corpuscular volume, RBC: Red blood cells count, RDW: Red distribution width (RDW), BAS: Basophils count, EOS: Eosinophils count, NEU: Neutrophils count, LYM: Lymphocytes count, MON: Monocytes count, WBC: White blood cells count

Supplementary Table 6 Causal variants with global effects on white blood cell traits with a shared gene identified by expression quantitative trait loci colocalization analysis

Locus	SNP	Chr	Pos (GRCh37)	Colocalised traits	HUGO gene	Tissue	PP	%PP explained
rs2476601	rs2476601	1	114377568	BAS, EOS, LYM, MON, NEU, WBC	AP4B1-AS1	Lung	0.85	100%
rs4706020	rs11738827	5	130717090	BAS, EOS, LYM, MON, NEU, WBC	SLC22A5	Lung	1.00	100%
rs4706020	rs11738827	5	130717090	BAS, EOS, LYM, MON, NEU, WBC	SLC22A5	Whole Blood	1.00	100%
rs4706020	rs31251	5	130833946	BAS, EOS, LYM, MON, NEU, WBC	AC116366.6	Whole Blood	0.98	100%
rs7141943	rs2038700	14	25461989	BAS, EOS, LYM, MON, NEU, WBC	NEDD8	Spleen	0.81	84%

Locus: Genomic region represented by the top variant associated with blood cell traits identified by MTAG, Candidate causal SNP: Candidate variant explaining the shared association, Chr: Chromosome of the causal variant, Pos: Position of the causal variant, Colocalized traits: Traits that colocalize, N traits: Number of traits that colocalize, Ensembl gene: Gene which is shared across the colocalized traits using the Ensembl ID, HUGO gene: Gene which is shared across the colocalized traits using the HUGO ID, Tissue: Tissue showing expression of the shared gene, PP: Posterior Probability, %PP explained: Percentage of the PP explained by the SNP, BAS: Basophils count, EOS: Eosinophils count, NEU: Neutrophils count, LYM: Lymphocytes count, MON: Monocytes count, WBC: White blood cells count

Supplementary Table 7 Causal variants with global effects on red blood cell traits with a shared gene identified by expression quantitative trait loci colocalization analysis

Locus	SNP	Chr	Pos (GRCh37)	Colocalised traits	HUGO gene	Tissue	PP	%PP explained
rs35416932	rs13191659	6	27001055	HCT, HGB, MCH, MCHC, MCV, RBC, RDW	HIST1H2BG	Whole Blood	0.77	100%
rs143875230	rs143875230	15	43278726	HCT, HGB, MCH, MCHC, MCV, RBC, RDW	ADAL	Lung	0.93	100%
rs143875230	rs143875230	15	43278726	HCT, HGB, MCH, MCHC, MCV, RBC, RDW	LCMT2	Lung	0.81	100%
rs143875230	rs143875230	15	43278726	HCT, HGB, MCH, MCHC, MCV, RBC, RDW	LCMT2	Spleen	0.76	100%
rs143875230	rs143875230	15	43278726	HCT, HGB, MCH, MCHC, MCV, RBC, RDW	CKMT1A	Lung	0.76	100%

Locus: Genomic region represented by the top variant associated with blood cell traits identified by MTAG, Candidate causal SNP: Candidate variant explaining the shared association, Chr: Chromosome of the causal variant, Pos: Position of the causal variant, Colocalized traits: Traits that colocalize, N traits: Number of traits that colocalize, Ensembl gene: Gene which is shared across the colocalized traits using the Ensembl ID, HUGO gene: Gene which is shared across the colocalized traits using the HUGO ID, Tissue: Tissue showing expression of the shared gene, PP: Posterior Probability, %PP explained: Percentage of the PP explained by the SNP, HCT: Haematocrit, HGB: Haemoglobin, MCH: Mean corpuscular haemoglobin, MCHC: Mean corpuscular haemoglobin concentration, MCV: Mean corpuscular volume, RBC: Red blood cells count, RDW: Red distribution width (RDW)

Supplementary Table 8 Information of the studies used in Mendelian Randomization analysis as outcomes

Trait	Authors	Journal	Year	Source
Attention Deficit Hyperactivity Disorder	Demontis et al.	Nature Genetics	2019	doi: 10.1038/s41588-018-0269-7
Autism Spectrum Disorder - Autism	Grove et al.	Nature Genetics	2019	doi: 10.1038/s41588-019-0344-8
Bipolar Disorder	Stahl et al.	Nature Genetics	2019	doi: 10.1038/s41588-019-0397-8
Breast Cancer	Michailidou et al.	Nature	2017	doi: 10.1038/nature24284
Coronary Artery Disease	Nikpay et al.	Nature Genetics	2015	doi: 10.1038/ng.3396
Fasting Glucose	Dupuis et al.	Nature Genetics	2010	doi: 10.1038/ng.520
Fasting Insulin	Dupuis et al.	Nature Genetics	2010	doi: 10.1038/ng.520
Major Depressive Disorder	Wray et al.	Nature Genetics	2018	doi: 10.1038/s41588-018-0090-3
Myocardial Infarction	Nikpay et al.	Nature Genetics	2015	doi: 10.1038/ng.3396
Prostate Cancer	Schumacher et al.	Nature Genetics	2018	doi: 10.1038/s41588-018-0142-8
Schizophrenia	Ripke et al.	Nature	2014	doi: 10.1038/nature13595
Type 2 Diabetes	Xue et al.	Nature Communications	2018	doi: 10.1038/s41467-018-04951-w
Waist Circumference	Shungin et al.	Nature	2015	doi: 10.1038/nature14132
Waist to Hip Ratio	Shungin et al.	Nature	2015	doi: 10.1038/nature14132

Supplementary Table 9 Gene-sets associated with C-reactive protein (CRP; N=19) in a Bonferroni significance level

Trait	Gene-Set	N Genes	Beta	SE	P
CRP	Curated_gene_sets:dacosta_uv_response_via_ercc3_common_dn	452	0.28117	0.060631	1.78E-06
CRP	Curated_gene_sets:biocarta_fxr_pathway	7	2.6213	0.4892	4.25E-08
CRP	GO_bp:go_acylglycerol_homeostasis	33	1.2367	0.20629	1.04E-09
CRP	GO_bp:go_positive_regulation_of_biosynthetic_process	1831	0.20198	0.029186	2.33E-12
CRP	GO_bp:go_positive_regulation_of_gene_expression	1811	0.19671	0.029515	1.37E-11
CRP	GO_bp:go_negative_regulation_of_transcription_by_rna_polymerase_ii	754	0.24564	0.044501	1.72E-08
CRP	GO_bp:go_negative_regulation_of_biosynthetic_process	1376	0.18191	0.033124	2.02E-08
CRP	GO_bp:go_positive_regulation_of_transcription_by_rna_polymerase_ii	1112	0.26528	0.037436	7.16E-13
CRP	GO_bp:go_negative_regulation_of_rna_biosynthetic_process	1096	0.20375	0.037092	2.00E-08
CRP	GO_bp:go_positive_regulation_of_rna_biosynthetic_process	1492	0.20615	0.03229	8.83E-11
CRP	GO_bp:go_regulation_of_lipid_localization	131	0.4999	0.11059	3.11E-06
CRP	GO_bp:go_regulation_of_interleukin_1_mediated_signaling_pathway	7	2.8943	0.57716	2.68E-07
CRP	GO_mf:go_regulatory_region_nucleic_acid_binding	878	0.23889	0.042336	8.50E-09
CRP	GO_mf:go_double_stranded_dna_binding	890	0.20078	0.041811	7.92E-07
CRP	GO_mf:go_transcription_factor_binding	608	0.22524	0.048751	1.93E-06
CRP	GO_mf:go_sequence_specific_dna_binding	1039	0.21288	0.039268	3.00E-08
CRP	GO_mf:go_dna_binding_transcription_factor_activity	1578	0.16736	0.032986	1.97E-07
CRP	GO_mf:go_interleukin_1_receptor_activity	6	2.9436	0.58831	2.84E-07
CRP	GO_mf:go_sequence_specific_double_stranded_dna_binding	811	0.21406	0.044342	6.97E-07

Supplementary Table 10 Genetic correlation between Alzheimer's disease and the examined cardiovascular traits

	r	P
Atrial fibrillation	0.096	0.328
Coronary artery disease	0.046	0.653
Carotid intima-media thickness	-0.148	0.367
Stroke	0.198	0.238

Systolic blood pressure	-0.011	0.879
Diastolic blood pressure	0.050	0.513

r: genetic correlation value; P: P-value

Supplementary Table 11 Colocalised genetic loci between Alzheimer's disease and cardiovascular traits

SNP	Chr	Pos	Gene	Traits	PP	Candidate Causal SNP	%PP explained
rs7529220	1	22282619	HSPG2	AD, AF	0.99	rs7529220	90%
rs17362588	2	179721046	CCDC141	AD, DBP	0.98	rs55844607	38%
rs6738011	2	213259857	ERBB4	AD, AF	0.62	rs6738011	22%
rs77399788	5	123003001	KRT18P16	AD, DBP	0.66	rs11959808	42%
rs10949499	6	18361663	RP11-528A10.2	AD, AF	0.96	rs10949499	40%
rs9399136	6	135402339	HBS1L	AD, DBP	0.88	rs9399136	26%
rs6462079	7	28415827	CREB5	AD, AF	0.96	rs6462079	20%
rs7834729	8	21821778	XPO7	AD, AF	0.81	rs56177941	16%
rs62521286	8	124551975	FBXO32	AD, AF	0.82	rs62521286	67%
rs7006122	8	124608614	RN7SKP155	AD, CIMT	1.00	rs7006122	53%
rs11786896	8	145018354	PLEC	AD, AF	0.97	rs11786896	86%
rs10740811	10	30167754	RP11-224P11.1	AD, CIMT	0.53	rs10740811	35%
rs2478835	10	30317949	KIAA1462	AD, DBP	0.80	rs2505084	45%
rs2284665	10	124226630	HTRA1	AD, STRK	0.87	rs2284665	46%
rs10437655	11	47391948	SPI1	AD, DBP	0.95	rs10838702	17%
rs604723	11	100610546	ARHGAP42	AD, SBP, DBP	0.56	rs604723	100%
rs73075659	12	20373541	CTC-465D4.1	AD, DBP	0.64	rs10743353	25%
rs4773140	13	110954237	COL4A1	AD, DBP	0.87	rs650724	34%
rs4788444	16	71856547	RP11-417N10.3	AD, SBP	0.61	rs62053796	69%
rs17608766	17	45013271	GOSR2:RP11-156P1.2	AD, CIMT	0.94	rs17608766	94%
rs438811	19	45416741	APOC1	AD, CAD	0.57	rs429358	93%

SNP: Single-nucleotide polymorphism; Chr: Chromosome; Pos: Position; Gene: Mapped gene; Traits: Traits which colocalize; PP: Posterior Probability; Causal SNP: Candidate variant causing the colocalization; %PP explained: Percentage of the PP explained by the SNP; AD: Alzheimer's disease; CAD: Coronary artery disease; CIMT: Carotid intima-media thickness; DBP: Diastolic blood pressure; SBP: Systolic blood pressure; STRK: Stroke

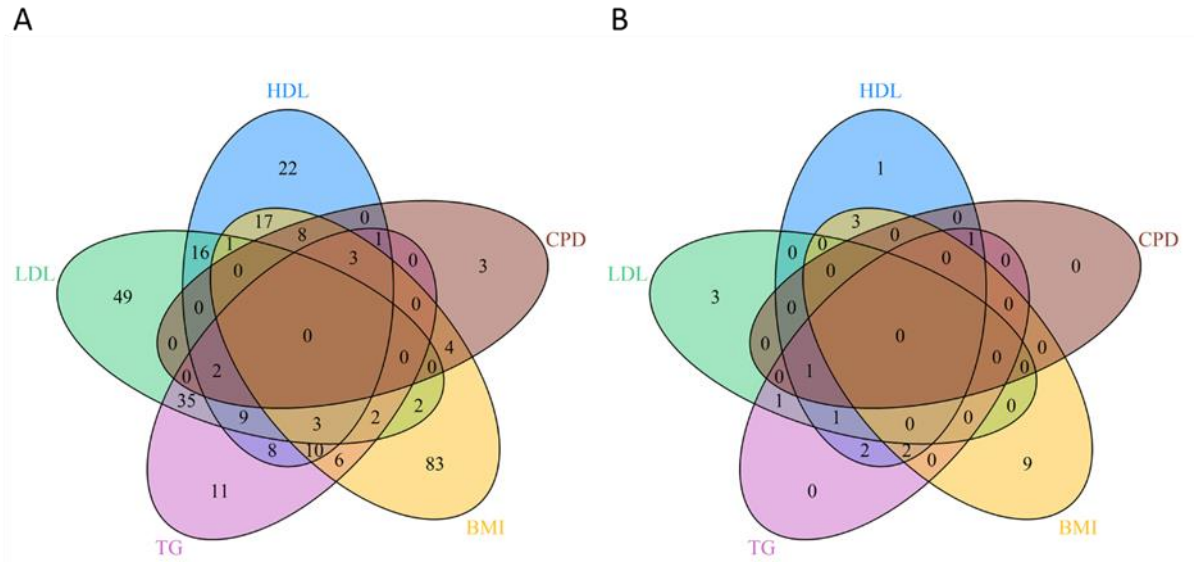
Supplementary Table 12 Genetic loci that colocalise between Alzheimer's disease and cardiovascular traits with a shared gene

SNP	Chr	Pos	Traits	Gene symbol	Tissue	PP	Causal SNP
rs7529220	1	22282619	AD, AF	C1QA	Breast Mammary Tissue	0.85	rs7529220
rs7529220	1	22282619	AD, AF	C1QB	Breast Mammary Tissue	0.83	rs7529220
rs7529220	1	22282619	AD, AF	C1QC	Breast Mammary Tissue	0.61	rs7529220
rs7529220	1	22282619	AD, AF	CDC42	Brain Putamen basal ganglia	0.64	rs4406609
rs17362588	2	179721046	AD, DBP	FKBP7	Esophagus Gastroesophageal Junction	0.58	rs55844607
rs17362588	2	179721046	AD, DBP	PJVK	Brain Cortex	0.57	rs55844607
rs17362588	2	179721046	AD, DBP	PJVK	Esophagus Muscularis	0.52	rs151041685
rs17362588	2	179721046	AD, DBP	PJVK	Testis	0.93	rs17362588
rs17362588	2	179721046	AD, DBP	PRKRA	Brain Cerebellum	0.90	rs55844607
rs6738011	2	213259857	AD, AF	ERBB4	Adipose Visceral Omentum	0.53	rs35880620
rs6738011	2	213259857	AD, AF	ERBB4	Heart Left Ventricle	0.53	rs6738011
rs77399788	5	123003001	AD, DBP	CSNK1G3	Lung	0.54	rs1579036
rs10949499	6	18361663	AD, AF	DEK	Heart Left Ventricle	0.57	rs9396839
rs10949499	6	18361663	AD, AF	IMPDH1P9	Thyroid	0.57	rs10949499
rs10949499	6	18361663	AD, AF	KDM1B	Heart Atrial Appendage	0.71	rs10949499
rs10949499	6	18361663	AD, AF	KDM1B	Heart Left Ventricle	0.64	rs9396839
rs10949499	6	18361663	AD, AF	NUP153-AS1	Artery Aorta	0.51	rs10949499
rs10949499	6	18361663	AD, AF	RNF144B	Heart Atrial Appendage	0.62	rs10949499
rs10949499	6	18361663	AD, AF	TPMT	Heart Atrial Appendage	0.66	rs10949499
rs9399136	6	135402339	AD, DBP	HBS1L	Cells EBV-transformed lymphocytes	0.61	rs6913541
rs7834729	8	21821778	AD, AF	BIN3	Skin Sun Exposed Lower leg	0.52	rs56177941
rs7834729	8	21821778	AD, AF	PDLIM2	Brain Amygdala	0.63	rs56177941
rs7834729	8	21821778	AD, AF	SLC39A14	Brain Nucleus accumbens basal ganglia	0.56	rs17060677
rs62521286	8	124551975	AD, AF	ZHX2	Brain Frontal Cortex BA9	0.57	rs2280919
rs7006122	8	124608614	AD, CIMT		Artery Aorta	0.94	rs7006122
rs11786896	8	145018354	AD, AF	DGAT1	Esophagus Mucosa	0.93	rs11786896
rs11786896	8	145018354	AD, AF	PLEC	Heart Left Ventricle	0.99	rs11786896
rs11786896	8	145018354	AD, AF	PLEC	Muscle Skeletal	0.92	rs11786896
rs10740811	10	30167754	AD, CIMT	JCAD	Artery Tibial	0.78	rs2505084

rs10740811	10	30167754	AD, CIMT	SVIL	Artery Tibial	0.56	rs10740811
rs2478835	10	30317949	AD, DBP	JCAD	Artery Tibial	0.90	rs2505084
rs2478835	10	30317949	AD, DBP	JCAD	Esophagus Mucosa	0.54	rs2505084
rs2284665	10	124226630	AD, STROKE	ARMS2	Testis	0.74	rs11200633
rs2284665	10	124226630	AD, STROKE	HTRA1	Testis	0.79	rs3750847
rs2284665	10	124226630	AD, STROKE	PLEKHA1	Brain Nucleus accumbens basal ganglia	0.62	rs3750847
rs10437655	11	47391948	AD, DBP	ARHGAP1	Brain Amygdala	0.90	rs10838702
rs10437655	11	47391948	AD, DBP	C11orf49	Muscle Skeletal	0.63	rs10838702
rs10437655	11	47391948	AD, DBP	CELF1	Whole Blood	0.80	rs10838702
rs10437655	11	47391948	AD, DBP	FNBP4	Whole Blood	0.80	rs10838702
rs10437655	11	47391948	AD, DBP	NDUFS3	Brain Cerebellar Hemisphere	0.87	rs10838702
rs10437655	11	47391948	AD, DBP	NUP160	Skin Not Sun Exposed Suprapubic	0.71	rs10838702
rs10437655	11	47391948	AD, DBP	PTPRJ	Adrenal Gland	0.73	rs1534576
rs10437655	11	47391948	AD, DBP	RP11-750H9.5	Brain Cerebellar Hemisphere	0.87	rs896817
rs10437655	11	47391948	AD, DBP	RP11-750H9.5	Brain Cerebellum	0.86	rs896817
rs10437655	11	47391948	AD, DBP	RP11-750H9.5	Heart Atrial Appendage	0.73	rs10838702
rs10437655	11	47391948	AD, DBP	SLC39A13	Cells Transformed fibroblasts	0.77	rs896817
rs10437655	11	47391948	AD, DBP	SLC39A13	Thyroid	0.75	rs896817
rs73075659	12	20373541	AD, DBP	LINC02398	Artery Aorta	0.61	rs10743353
rs17608766	17	45013271	AD, CIMT	FAM215B	Esophagus Mucosa	0.63	rs17608766
rs17608766	17	45013271	AD, CIMT	GOSR2	Adrenal Gland	0.80	rs17608766
rs17608766	17	45013271	AD, CIMT	KPNB1	Uterus	0.55	rs17608766
rs17608766	17	45013271	AD, CIMT	TBX21	Brain Anterior cingulate cortex BA24	0.62	rs55657964
rs17608766	17	45013271	AD, CIMT		Testis	0.51	rs17608766

SNP: Single-nucleotide polymorphism; Chr: Chromosome; Pos: Position; Traits: Traits which colocalize; Gene (Ensembl): Gene ID (Ensembl) that is shared between the traits; Gene symbol: HUGO (HGNC) gene symbol; Tissue: Tissue type that the shared gene between the traits is expressed; PP: Posterior Probability; Causal SNP: Candidate variant causing the colocalization; PP%: Percentage of the PP explained by the SNP; AD: Alzheimer's disease; CAD: Coronary artery disease; CIMT: Carotid intima-media thickness; DBP: Diastolic blood pressure; SBP: Systolic blood pressure; STRK: Stroke

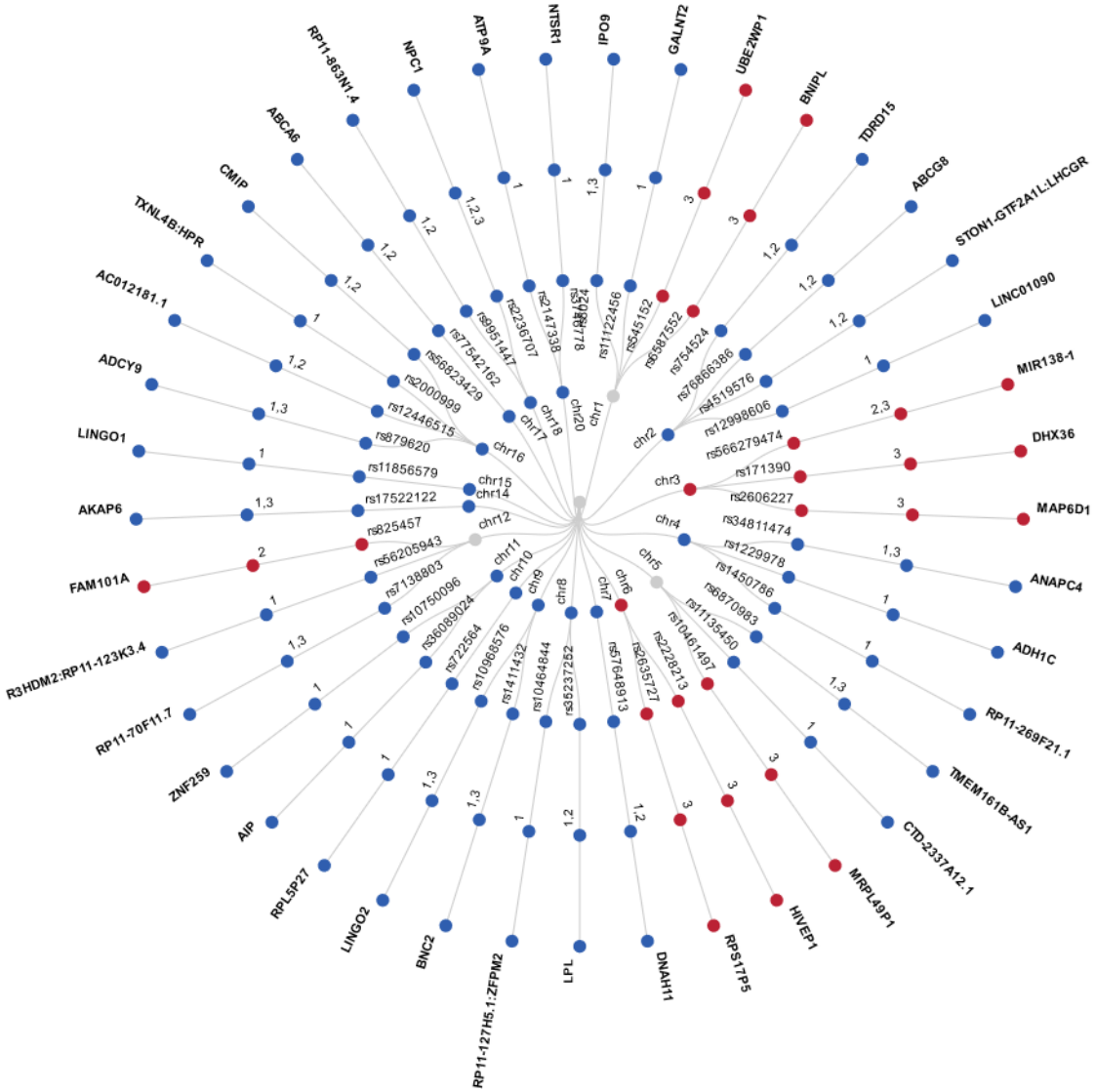
B. SUPPLEMENTARY FIGURES



Supplementary Figure 1 Venn diagrams

A) 295 C-reactive protein independent SNPs which are also genome-wide significantly associated with any of the other examined traits and B) the same focusing on the novel CRP SNPs only

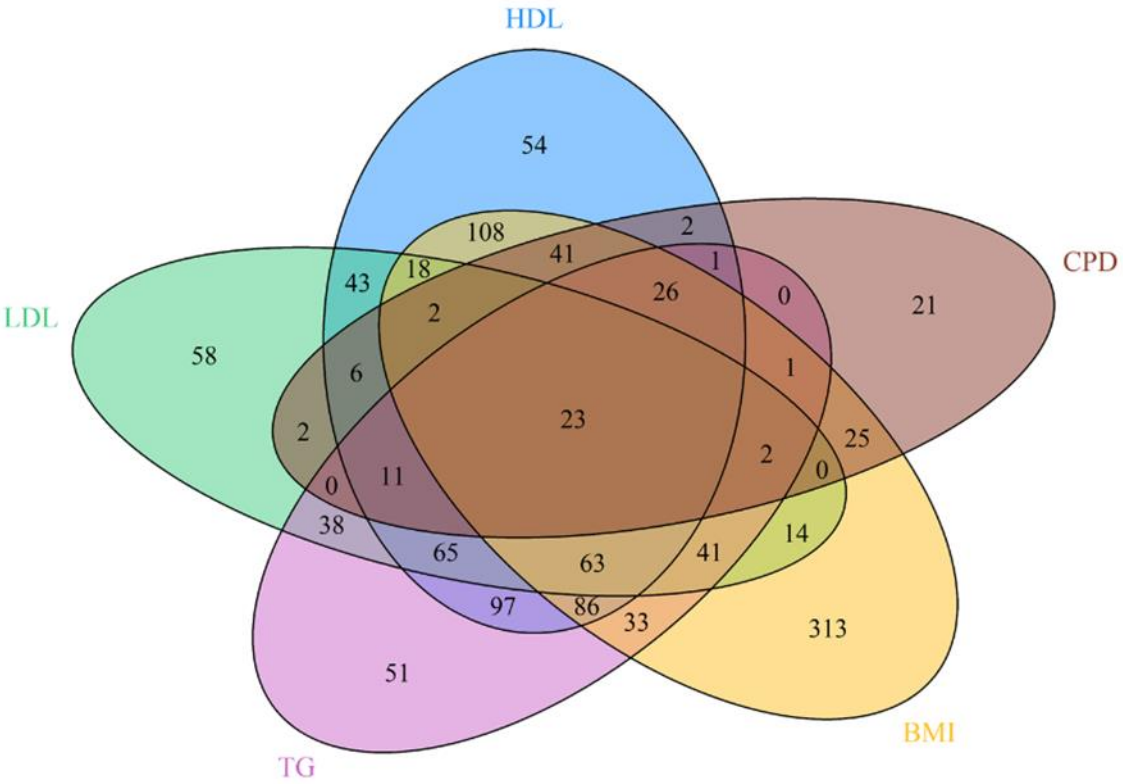
HDL: high density lipoprotein levels, LDL: low density lipoprotein levels, TG: triglyceride levels, BMI: body mass index, CPD: cigarettes per day



Supplementary Figure 2 Circular dendrogram presenting the novel C-reactive protein (CRP) loci

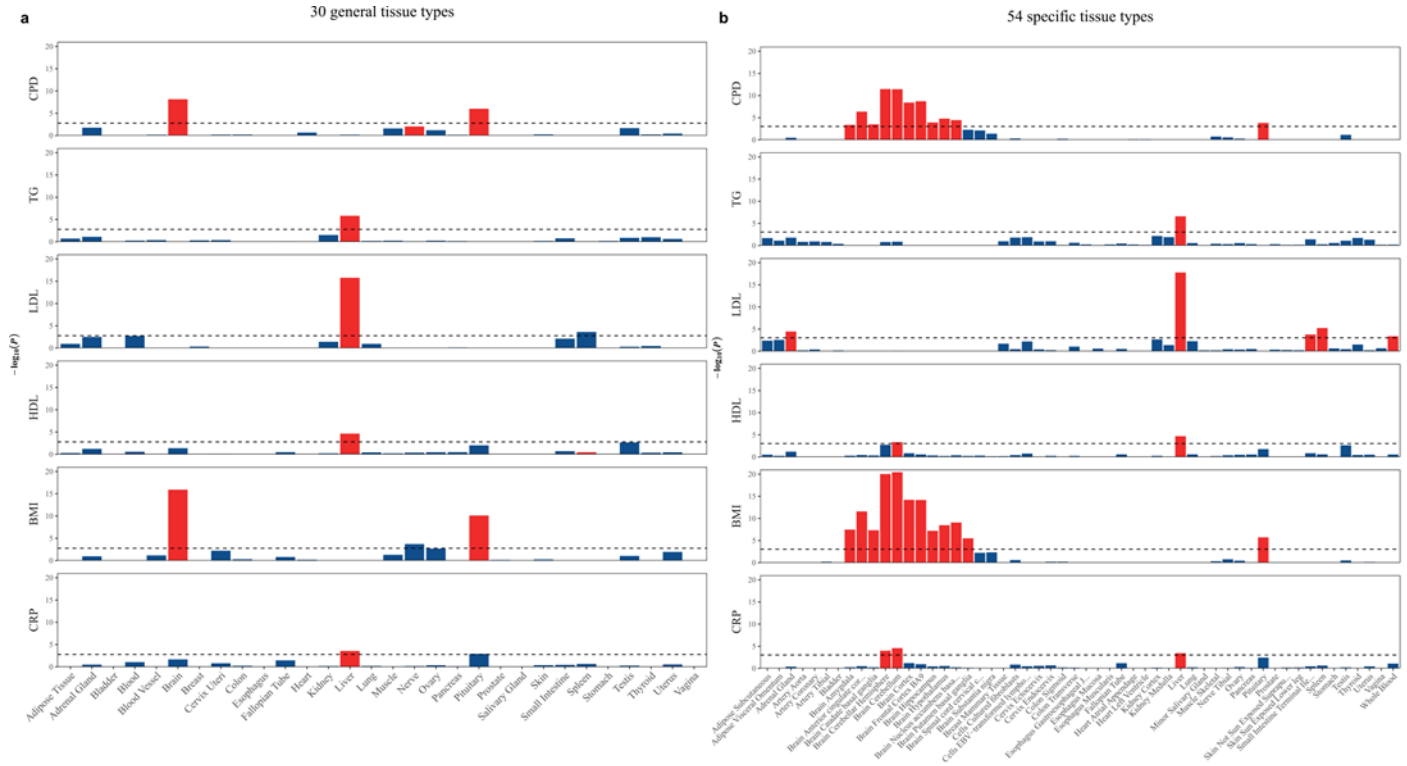
Starting from the centre, the layers represent chromosome, rsID, MTAG analysis (1: multi-trait, 2: bivariate on CRP-lipids, 3: bivariate on CRP-BMI), and the mapped genes, respectively. The loci from multi-trait MTAG are coloured with blue and from bivariate with red

BMI: Body mass index; CPD: Cigarettes per day; CRP: C-reactive protein; HDL: High-Density lipoprotein; LDL: Low-Density lipoprotein; TG: Triglycerides



Supplementary Figure 3 Venn diagram of C-Reactive Protein associated genes which are also associated with any of the other examined traits

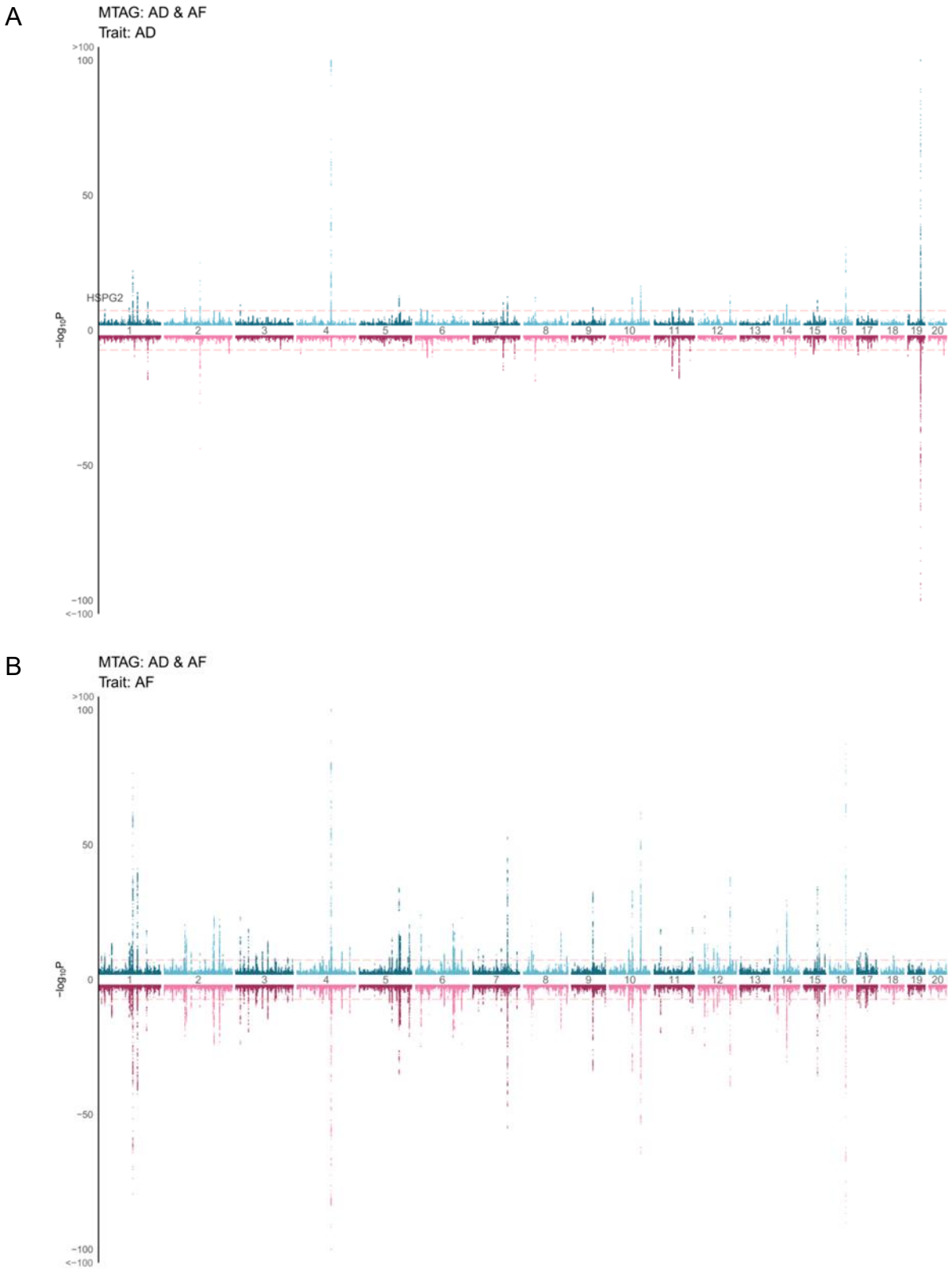
HDL: High-Density Lipoprotein, LDL: Low-Density Lipoprotein, TG: Triglycerides, BMI: Body Mass Index, CPD: Cigarettes per day



Supplementary Figure 4 Tissue expression analysis of a) 30 general and b) 54 more specific tissue types for each examined trait

Horizontal line represents the Bonferroni threshold (a: $P=1.7\times10^{-3}$; b: $P=9.3\times10^{-4}$). Bonferroni significant tissues are colored with red.

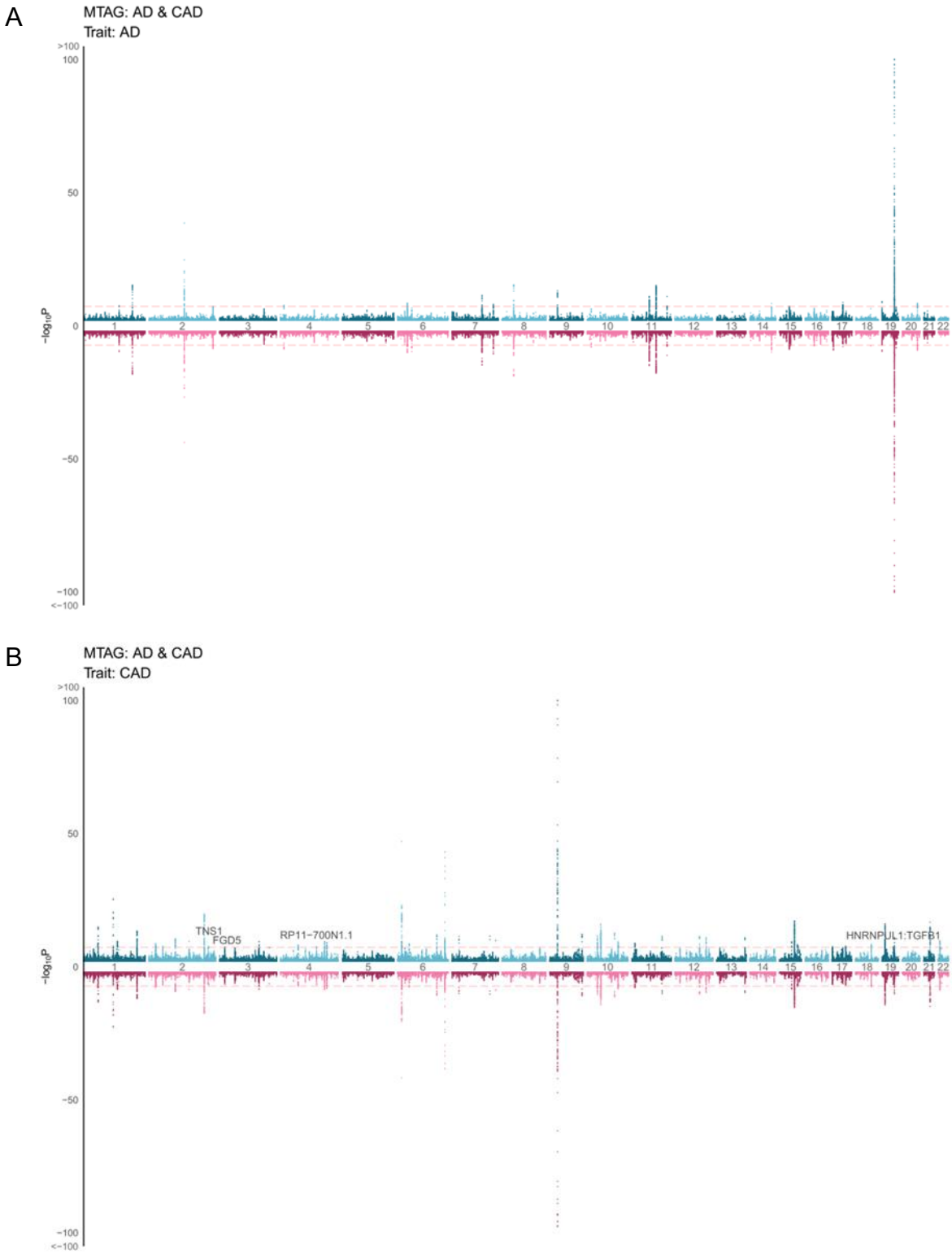
CRP: C-Reactive Protein, BMI: Body Mass Index, HDL: High-Density Lipoprotein, LDL: Low-Density Lipoprotein, TG: Triglycerides, CPD: Cigarettes per day



Supplementary Figure 5 Manhattan plots of MTAG results for A) Alzheimer's disease and B) atrial fibrillation.

Each mirror Manhattan plot illustrates the results from the MTAG analysis (upper part in turquoise) compared with the original GWAS results for the same trait and the same set of SNPs (lower part in red). Annotated genes denote novel discoveries identified by the MTAG analysis.

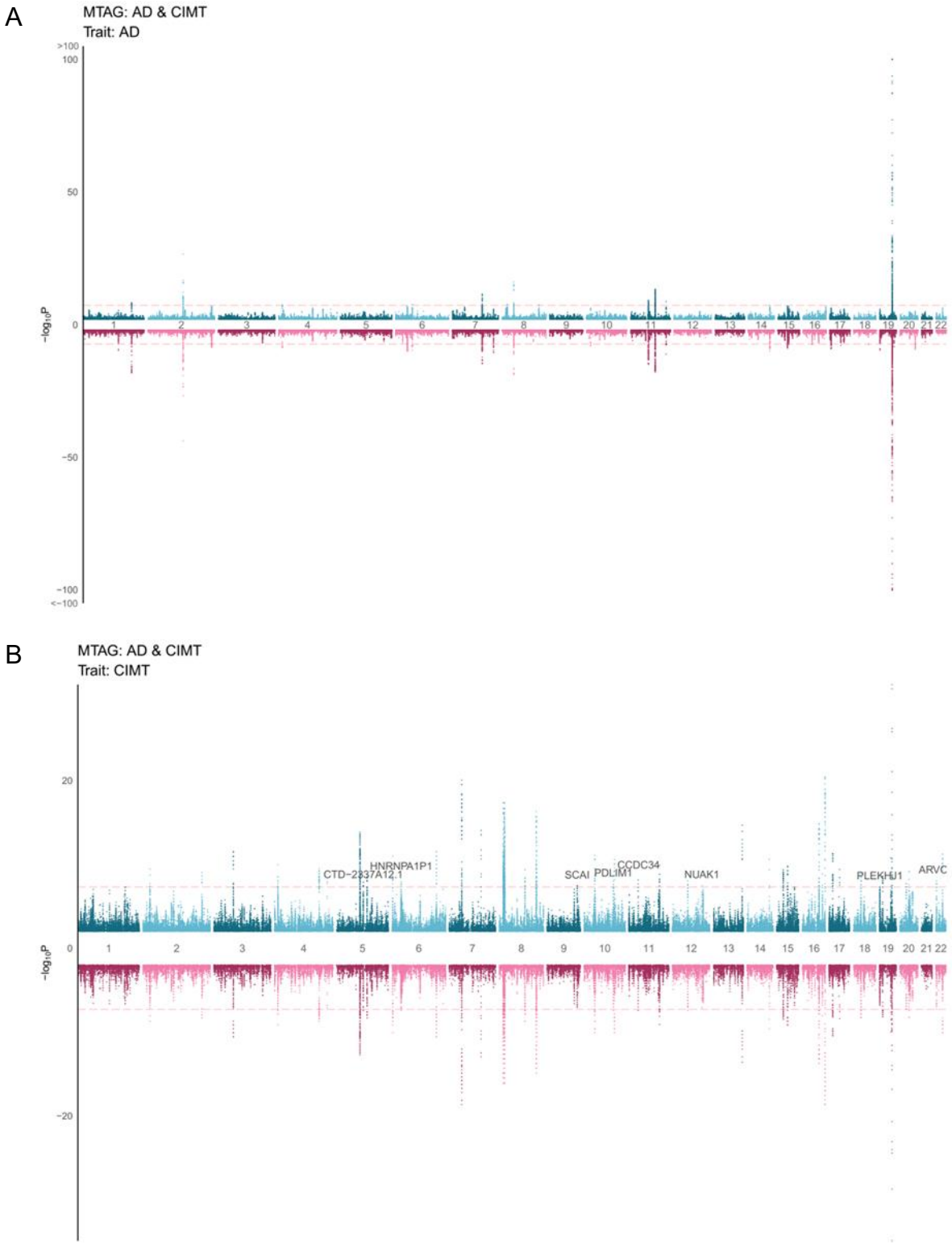
AD Alzheimer's disease, AF atrial fibrillation.



Supplementary Figure 6 Manhattan plots of MTAG results for A) Alzheimer's disease and B) coronary artery disease.

Each mirror Manhattan plot illustrates the results from the MTAG analysis (upper part in turquoise) compared with the original GWAS results for the same trait and the same set of SNPs (lower part in red). Annotated genes represent novel discoveries identified by the MTAG analysis.

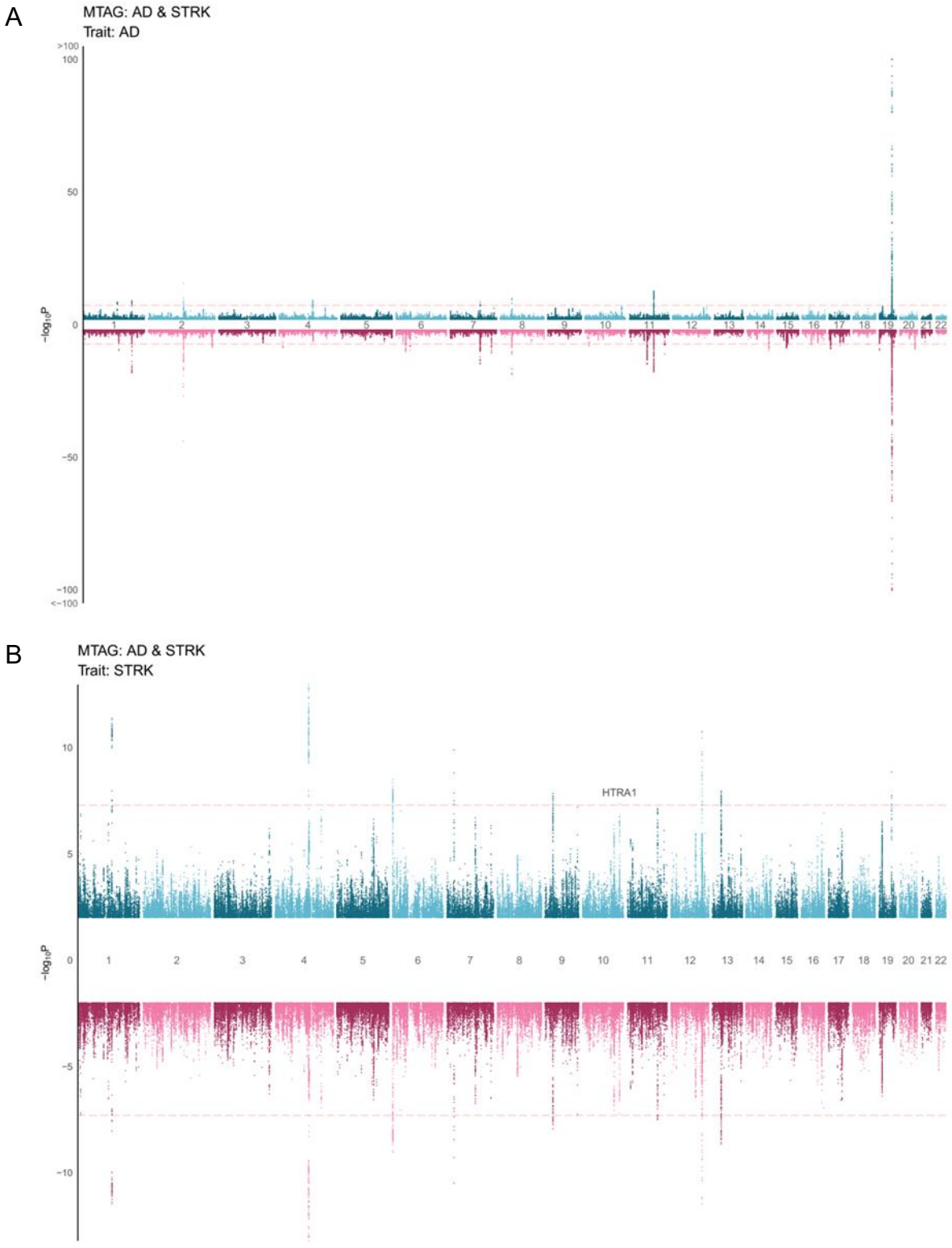
AD Alzheimer's disease, CAD coronary artery disease.



Supplementary Figure 7 Manhattan plots of MTAG results for A) Alzheimer's disease and B) carotid intima-media thickness.

Each mirror Manhattan plot illustrates the results from the MTAG analysis (upper part in turquoise) compared with the original GWAS results for the same trait and the same set of SNPs (lower part in red). Annotated genes represent novel discoveries identified by the MTAG analysis.

AD Alzheimer's disease, CIMT carotid intima-media thickness.

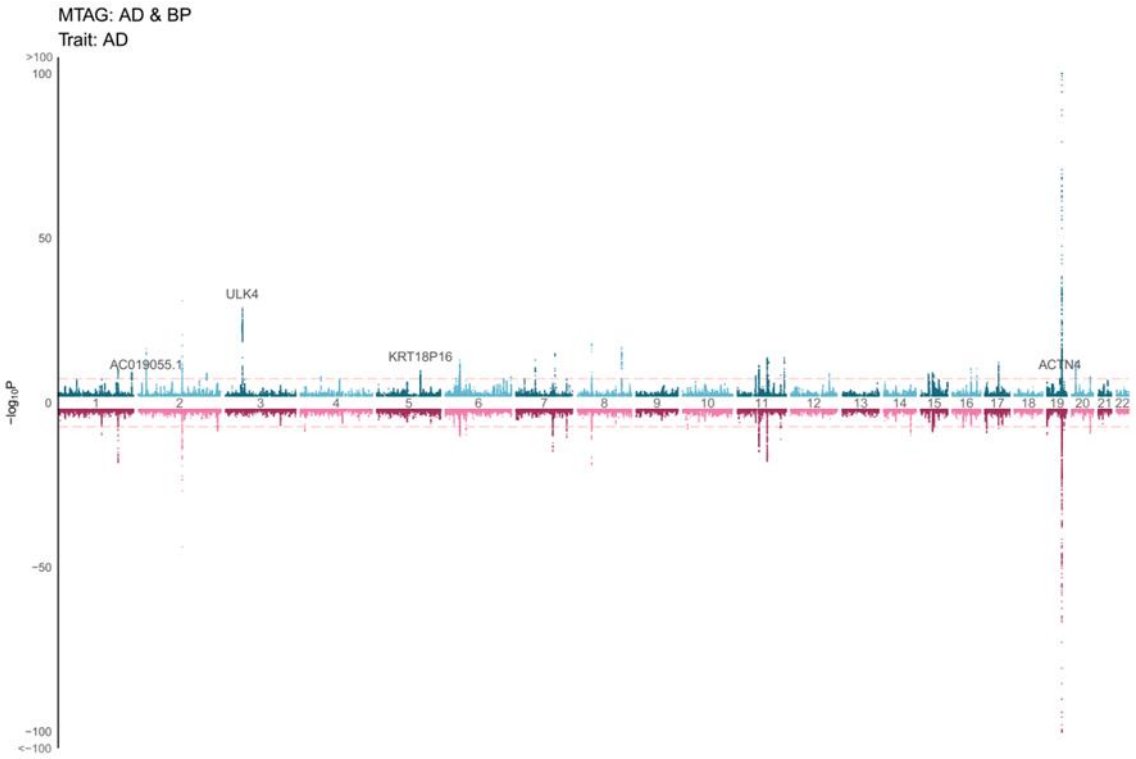


Supplementary Figure 8 Manhattan plots of MTAG results for A) Alzheimer's disease and B) stroke.

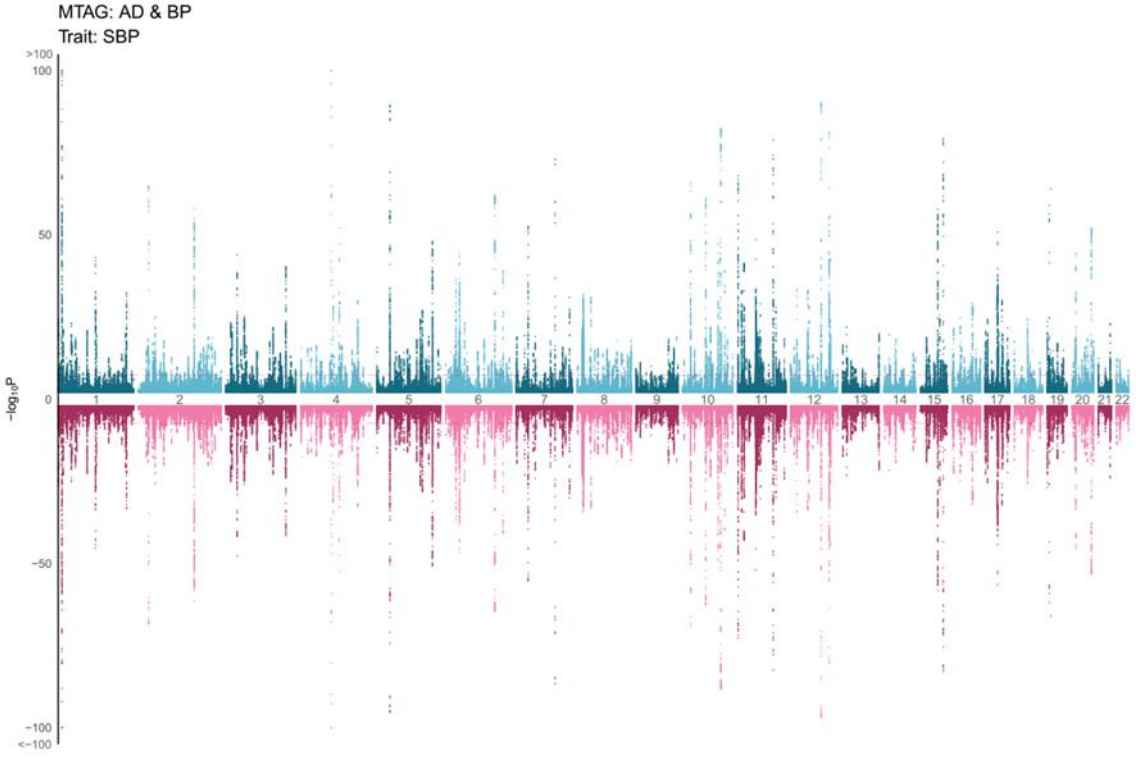
Each mirror Manhattan plot illustrates the results from the MTAG analysis (upper part in turquoise) compared with the original GWAS results for the same trait and the same set of SNPs (lower part in red). Annotated genes represent novel discoveries identified by the MTAG analysis.

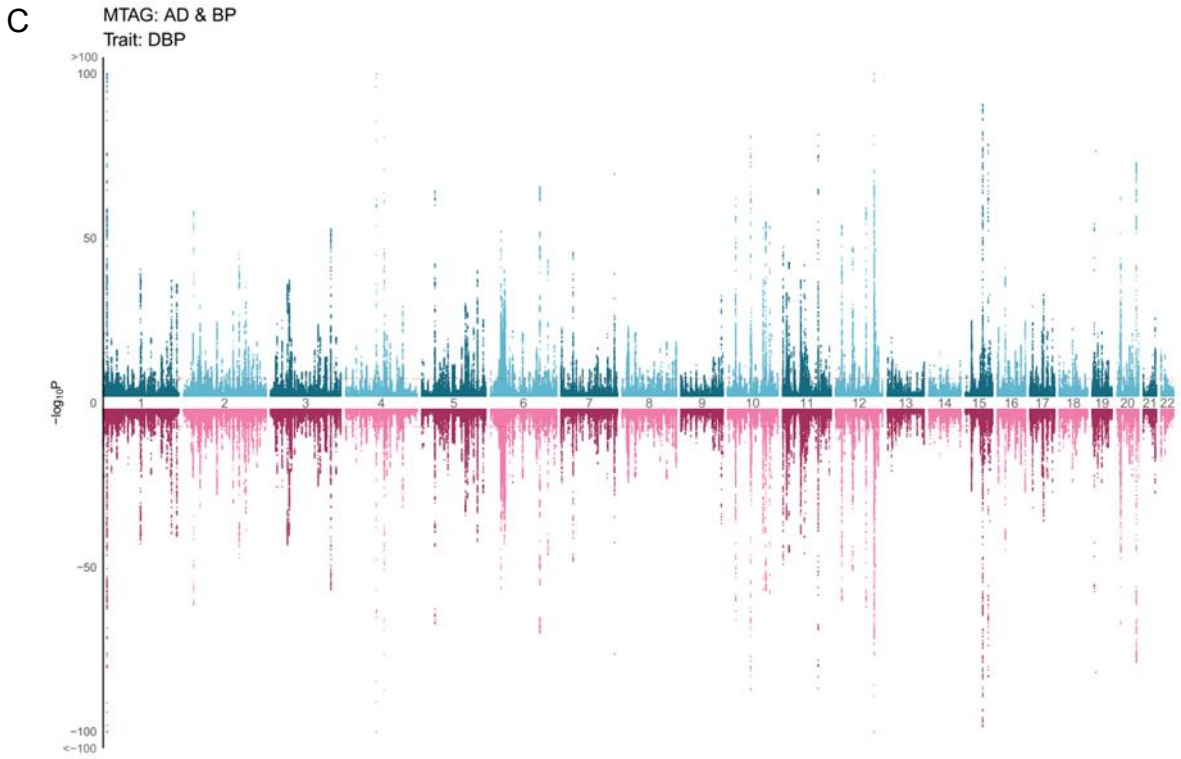
AD Alzheimer's disease, STRK stroke.

A


MTAG: AD & BP
Trait: AD
>100
100
50
0
−50
−100
<−100
−log10P
ULK4
AC019055.1
KRT18P16
ACTN4
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22

B


MTAG: AD & BP
Trait: SBP
>100
100
50
0
−50
−100
<−100
−log10P
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22



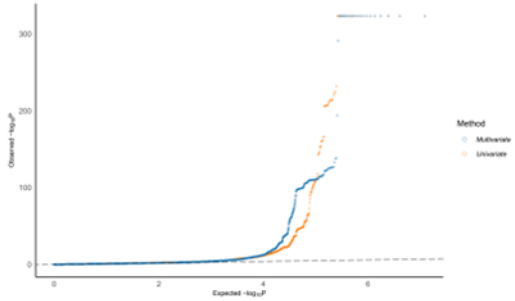
Supplementary Figure 9 Manhattan plots of MTAG results for A) Alzheimer's disease, B) systolic blood pressure and C) diastolic blood pressure.

Each mirror Manhattan plot illustrates the results from the MTAG analysis (upper part in turquoise) compared with the original GWAS results for the same trait and the same set of SNPs (lower part in red). Annotated genes represent novel discoveries identified by the MTAG analysis.

AD Alzheimer's disease, BP blood pressure, SBP systolic blood pressure, DBP diastolic blood pressure.

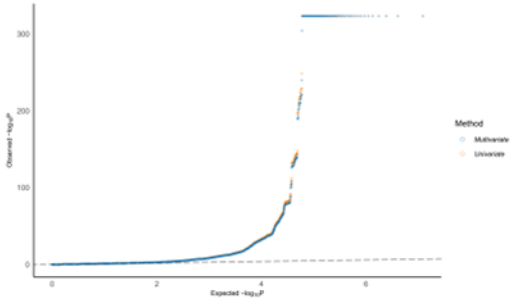
A MTAG: AD & AF

Trait: AD



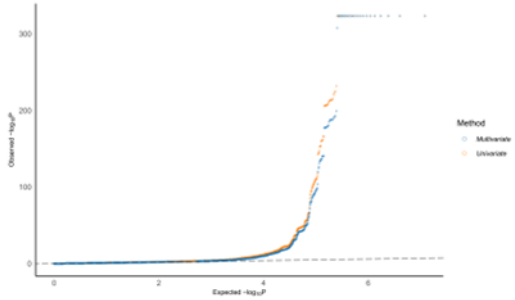
B MTAG: AD & AF

Trait: AF



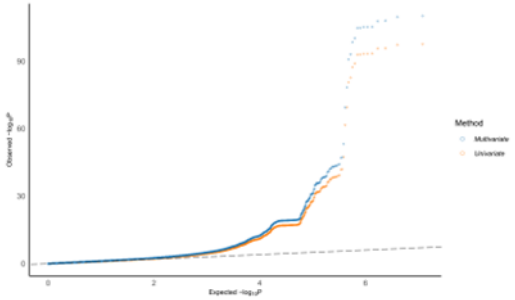
C MTAG: AD & CAD

Trait: AD



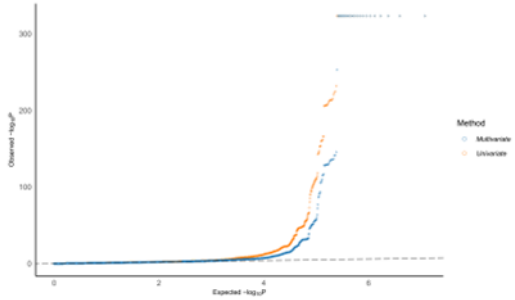
D MTAG: AD & CAD

Trait: CAD



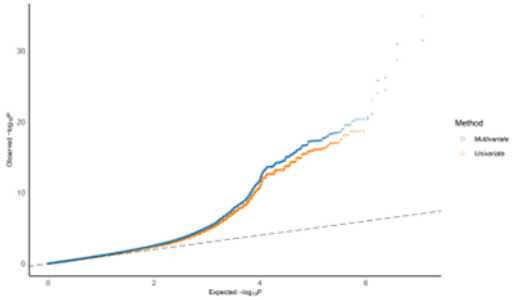
E MTAG: AD & CIMT

Trait: AD



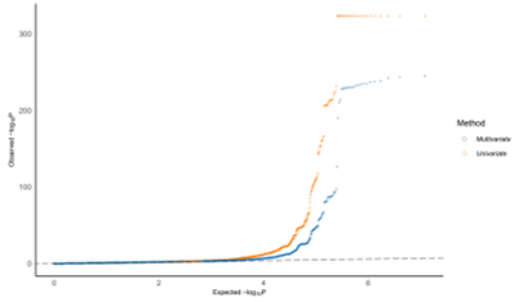
F MTAG: AD & CIMT

Trait: CIMT



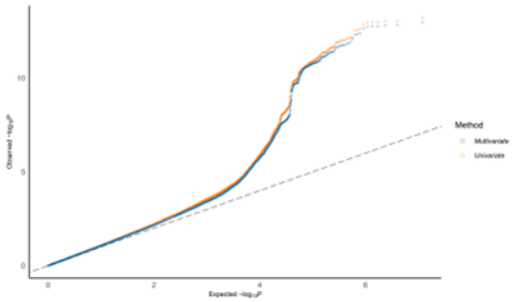
G MTAG: AD & STRK

Trait: AD



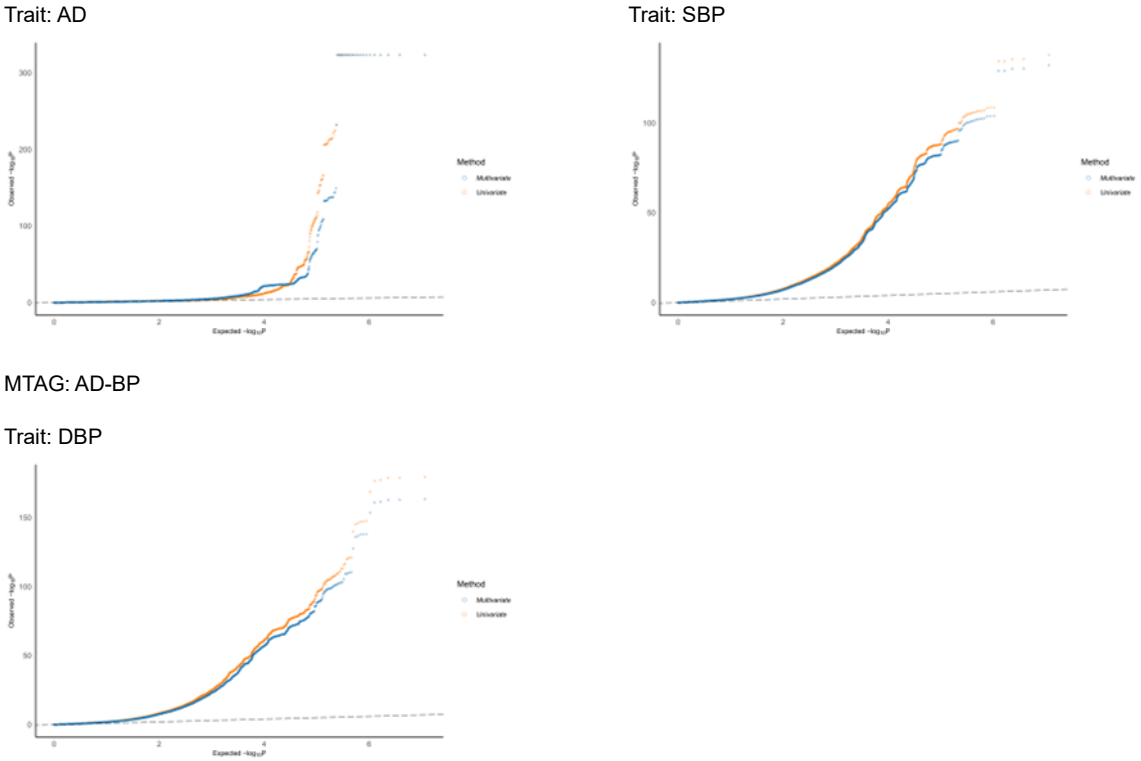
H MTAG: AD & STRK

Trait: STRK



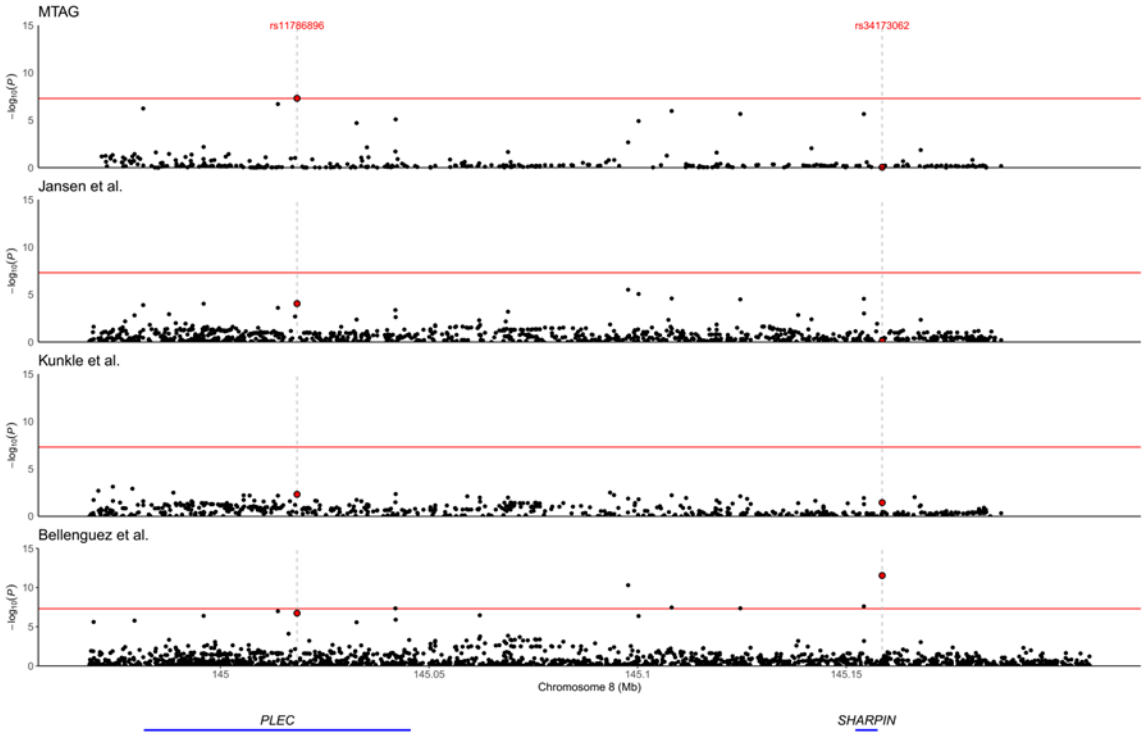
I MTAG: AD-BP

J MTAG: AD-BP



Supplementary Figure 10 QQ-plots of bivariate MTAG analyses for Alzheimer's disease and the examined cardiovascular traits

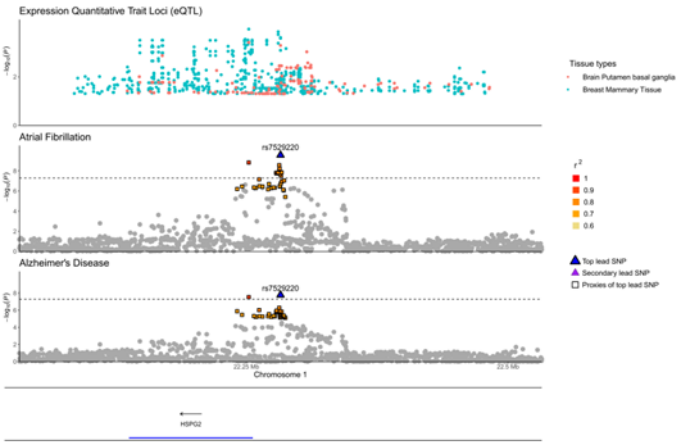
This figure presents the QQ-plots resulting from five bivariate MTAG analyses, each comparing Alzheimer's disease (AD) with a different cardiovascular trait. The examined traits alongside AD include atrial fibrillation (AF), coronary artery disease (CAD), carotid intima-media thickness (cIMT), stroke (STRK), and blood pressure (BP). Each QQ-plot displays the observed versus expected -log10(P-values) for genetic associations, comparing the MTAG results (blue) with the original GWAS results (orange) for the same trait and the same set of SNPs. The plots are organised as follows: A) AD results from the AD-AF MTAG analysis, B) AF results from the AD-AF MTAG analysis, C) AD results from the AD-CAD MTAG analysis, D) CAD results from the AD-CAD MTAG analysis, E) AD results from the AD-CIMT MTAG analysis, F) CIMT results from the AD-CIMT MTAG analysis, G) AD results from the AD-STRK MTAG analysis, H) STRK results from the AD-STRK MTAG analysis, I) AD results from the AD-BP MTAG analysis, J) SBP results from the AD-BP MTAG analysis, K) DBP results from the AD-BP MTAG analysis.



Supplementary Figure 11 Regional plots of *PLEC* region across different genome-wide association studies on Alzheimer's disease

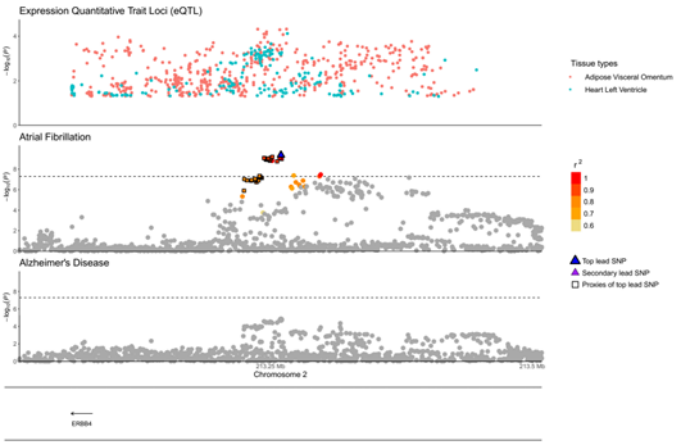
The variant rs11786896 (*PLEC*), which was indicated by MTAG as a top signal associated with Alzheimer's disease is located approximately 150kb upstream from another previously identified variant (rs34173062). The two variants likely represent two independent signals (linkage disequilibrium $r^2 = 0.006$)

rs7529220

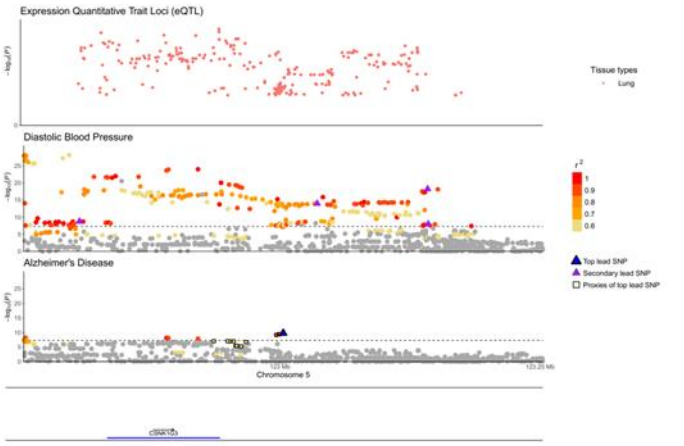


rs17362588

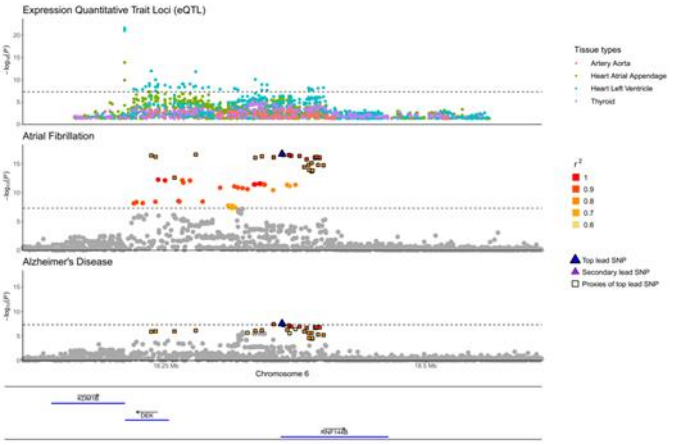
rs6738011



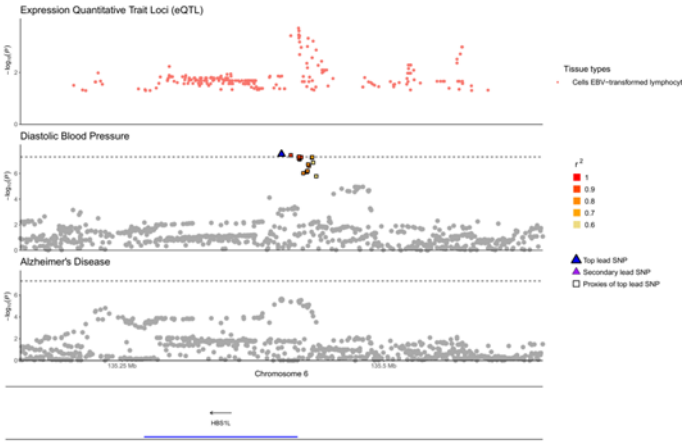
rs77399788



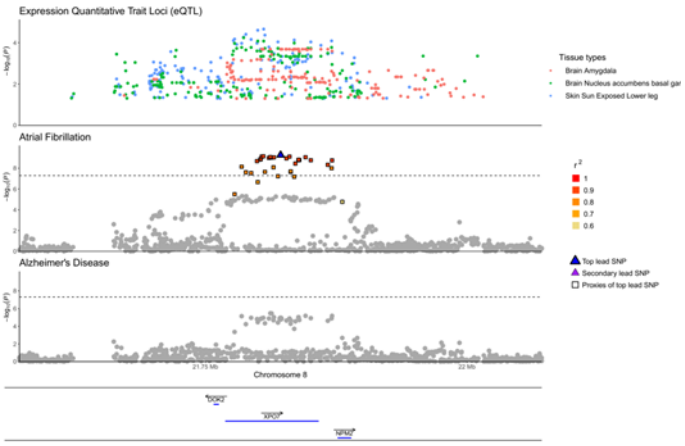
rs10949499



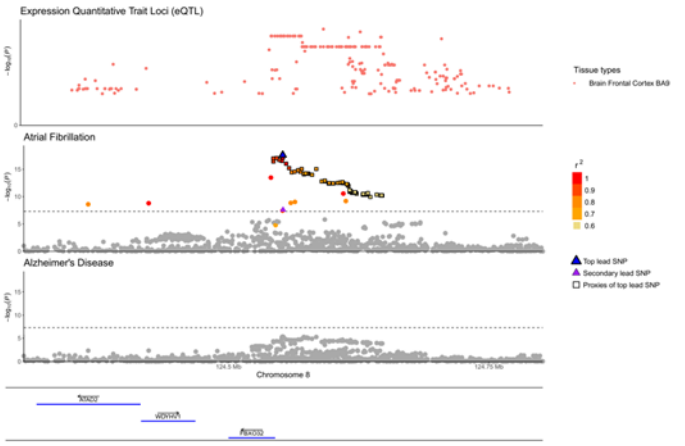
rs9399136



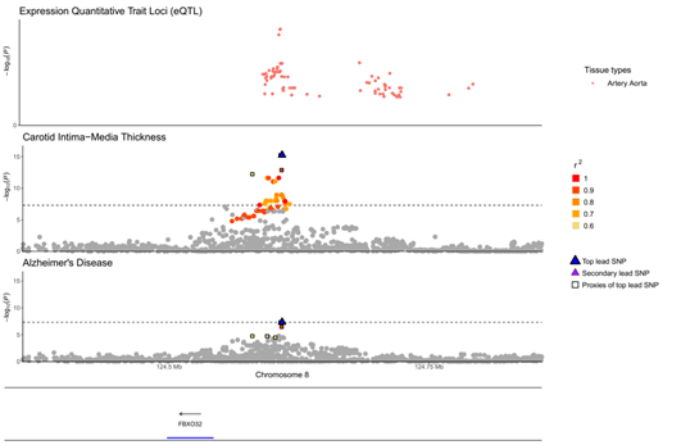
rs7834729



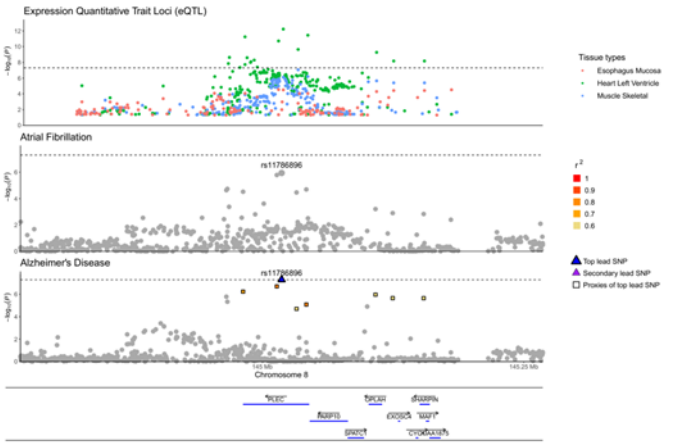
rs62521286



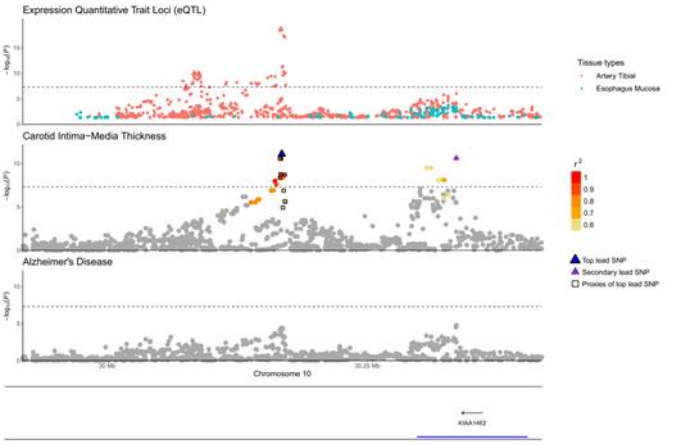
rs7006122



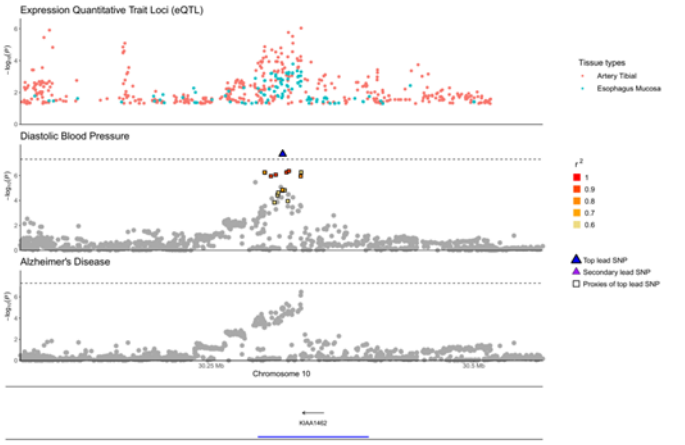
rs11786896

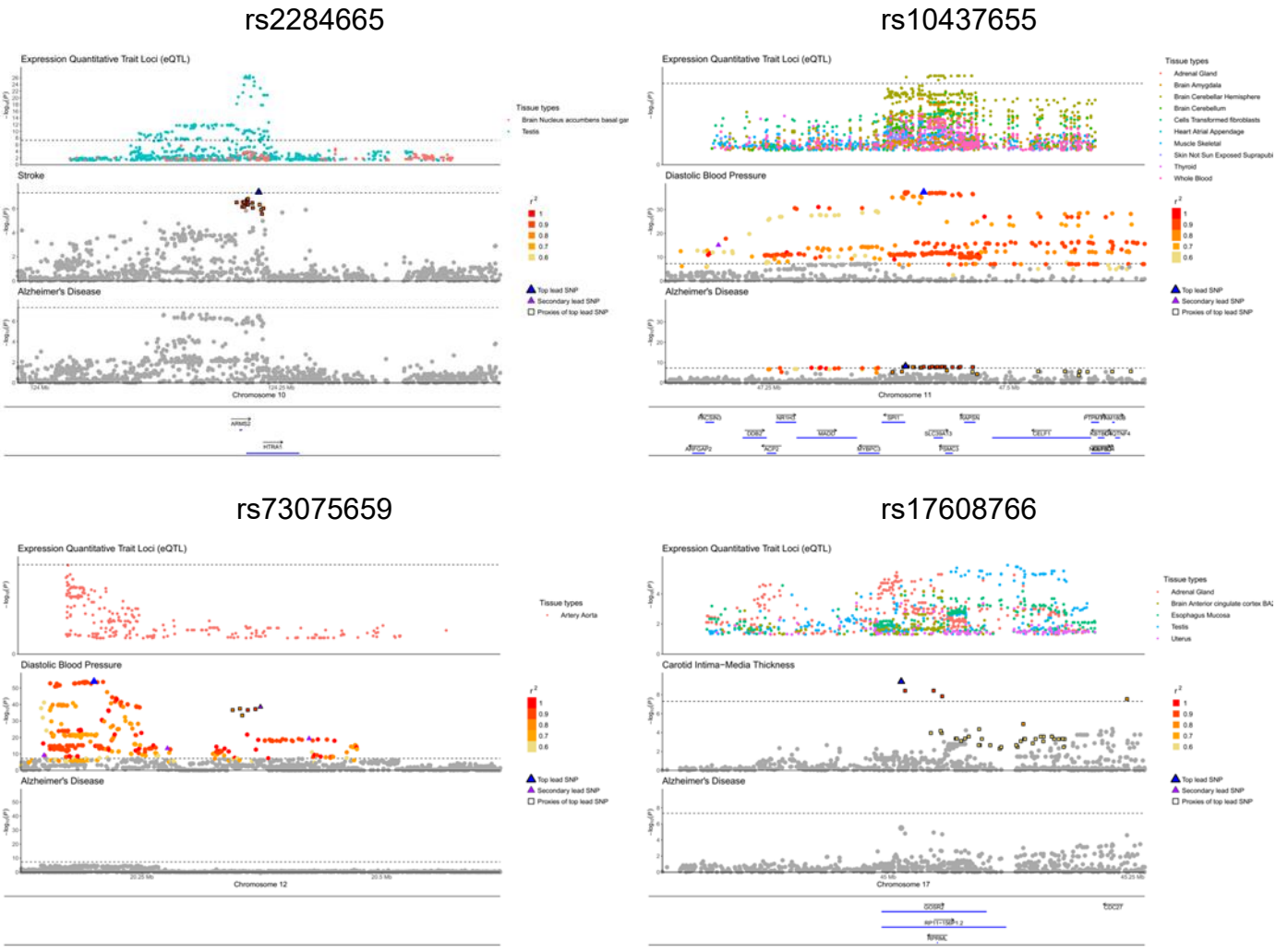


rs10740811



rs2478835





Supplementary Figure 12 Regional plots on the colocalized locı between Alzheimer's disease (bottom), cardiovascular trait (middle) and the expression quantitative trait loci for the respective tissues (top)