Quantitative Genomics of Complex Traits and Common Disease



Xiaotong Wang
Somerville College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy in Psychiatry
Trinity 2025

Acknowledgement

As I write these words on a quiet Oxford night, on the eve of submitting this thesis, memories of the long journey that brought me here flood my mind—of challenges faced, lessons learned, and the kindness of those who have supported me along the way. Reaching this point has been both a privilege and a journey of growth. I owe my deepest gratitude to all the remarkable people who have guided, encouraged, and believed in me.

I would first like to express my profound gratitude to my primary supervisor, **Professor Naomi Wray**, for her unwavering guidance, support, and kindness throughout this journey. My academic path with Naomi began in the second year of my medical studies, when I came across a post from her group about a research elective opportunity. I still remember the anxiety of waiting for her reply—and the joy when, just a few hours later, she responded warmly and efficiently. That moment perfectly reflected her generosity and attentiveness toward her students. From Brisbane to Oxford, I have often thought that I could not imagine a better supervisor. Naomi combines world-class expertise with an extraordinary sense of care and encouragement. She is remarkably responsive, providing thoughtful feedback with astonishing promptness, and has consistently inspired me to explore my own ideas. Words often feel inadequate to express how much I admire and appreciate her—not only as a supervisor, but also as a role model in both science and life.

I am equally grateful to my secondary supervisor, **Professor Peter Visscher**, for his insightful guidance and inspiration. I continue to be amazed by his depth of knowledge in statistical genetics and his ability to identify key ideas and challenges in any project with remarkable clarity. His rigorous approach to science and genuine passion for discovery have set an example I deeply admire and aspire to follow.

I would also like to thank everyone who has supervised or collaborated with me since I first entered the field of statistical genetics. I am especially grateful to **Dr. Tian Lin**, who co-supervised my research elective project with Naomi during medical school. From setting up my first high-performance computing account to guiding me through the fundamentals of GWAS, Tian patiently taught me the foundations of statistical genetics through many insightful discussions.

I am also thankful to **Dr. Joana Revez** and **Dr. Valentin Hivert**, who supervised my vitamin D project during medical school. Joana taught me the art of communicating

science clearly and beautifully, while Valentin showed me how to approach research with methodological depth and conceptual clarity. Working with them helped me become a more thoughtful and mature researcher.

My sincere thanks also go to **Professor Loïc Yengo** and **Dr. Jian Zeng**. Loïc first proposed the idea of extending the COJO framework into multi-ancestry settings, which became a central chapter of this thesis. Jian collaborated with me on projects exploring interactions between genetic variants, which—though not included here due to word limits—greatly expanded my understanding of the field. Both Loïc and Jian are outstanding statistical geneticists whose dedication to methodological innovation continues to inspire me.

I would also like to thank all my colleagues and friends in the **Visscher–Wray groups**, both in Brisbane and Oxford. I am deeply grateful for the warmth, collaboration, and sense of community within our teams—how everyone works together to create not just a research group, but a family that supports one another academically and personally. I have made many lifelong friends during my time in the group, and though I will not list names for fear of missing anyone, please know that each of you has shaped who I am today.

I also wish to thank all those who work tirelessly behind the scenes, as well as the participants of large biobanks such as the UK Biobank**,** the China Kadoorie Biobank**,** and the All of Us study. None of the discoveries we celebrate today would be possible without the collective effort of these international collaborations and the generosity of participants who share their health data for the advancement of science. In a world often divided, such cooperation reminds us of what humanity can achieve when united by curiosity and compassion.

Outside the academic world, I would like to thank all my friends for their emotional support over the years. I have lived in Beijing, Toronto, Brisbane, and Oxford, and at each stage of my life I have been fortunate to meet wonderful people who have accompanied me along the way. I deeply cherish the memories we have shared—the laughter, warmth, and kindness that have filled my life. Thank you for creating an environment of love and encouragement that has allowed me to grow, both as a person and as a friend.

Finally, I would like to express my deepest gratitude to my **family** for their unconditional love and unwavering support. My mother, **Caixia Zhu**, has always been the *superwoman* in my heart since childhood. In my eyes, she could do

everything—building a successful career, taking care of our family with endless love, and facing every challenge with grace and strength. It was only after I grew up that I began to understand how much effort and sacrifice it took to achieve all that she has done. She has taught me the values of patience, perseverance, and courage in the face of difficulty. Her strength and determination continue to inspire me every single day. My father, **Shaowei Wang**, has always guided me with his calm wisdom and gentle perspective on life. Whenever I felt anxious about exams or uncertain about the future, he would remind me that life is a marathon—not defined by one single moment, but by endurance, balance, and persistence along the way. His words have quietly stayed with me through every challenge I've faced.

I am also deeply grateful to my grandparents, **Xiqing Dong** and **Qingyang Zhu**. When my parents were busy with work, I often lived with them, and they filled my childhood with warmth, imagination, and countless happy memories. I can still feel the joy of returning to their home each summer and winter break. Time passes so quickly, and sadly, my grandfather passed away last year. It is one of my deepest regrets that he cannot witness my graduation or hear my words of thanks in person. While he is no longer here to witness this moment, his love continues to live quietly within these pages and within me. I would also like to thank my partner and all my family members for creating such a warm and loving home. You have surrounded me with kindness and encouragement at every stage of my life.

To all my family, teachers, and friends who have stood by me and supported me through every step of this journey—thank you from the bottom of my heart**.**

Publications and Contributions

The following publication has arisen from this thesis.

Chapter 2

- **Wang, X.**, Walker, A., Revez, J. A., Ni, G., Adams, M. J., McIntosh, A. M., ... & Wray, N. R. (2023). Polygenic risk prediction: why and when out-of-sample prediction R2 can exceed SNP-based heritability. *The American Journal of Human Genetics*, *110*(7), 1207-1215.
- This study was initially motivated by N.R.W. and A.M. McI. I developed the theoretical framework, conducted all simulations and analyses, and drafted the manuscript, under the supervision and guidance of N.R.W. and P.M.V. The remaining authors contributed by preparing the data for analysis and provided final approval of the manuscript.

The following manuscripts have arisen from this thesis, and are presented in chapter 3 and 4.

Chapter 3

- **Wang, X.**, Wang, Y., Visshcer, P.M., Wray, R.W., Yengo, L. Multi-ancestry conditional and joint analysis (Manc-COJO) applied to GWAS summary statistics. In preparation.
- This study was initially motivated by L.Y. I developed the theoretical framework, performed all simulations and analyses, and drafted the manuscript under the supervision and guidance of N.R.W., L.Y., and P.M.V. I created the initial version of the software, with Y.W. contributing to the C++ implementation and enhancing its computational performance. All authors reviewed and approved the final manuscript.

Chapter 4

- **Wang, X.**, Li, A., McGrath, I., Wray, R.W. Refined mitochondrial DNA copy number (mtCN) calling and analysis pipeline reveals nuclear genetic

regulation of mtCN in blood and its association with psychiatric disorders. In preparation.

- This study was initially motivated by N.R.W. (regarding the mtCN and disease association) and by me (regarding the whole-genome sequencing analysis pipeline). I conducted all analyses and drafted the manuscript under the supervision and guidance of N.R.W. All authors reviewed and approved the final manuscript.

Abstract

Quantitative genomics investigates the genetic basis of complex traits and diseases, which are shaped by numerous variants of small effect rather than single mutations, as in monogenic conditions such as cystic fibrosis or sickle-cell anaemia. Traits such as height and disorders including depression exemplify this polygenic architecture. Despite rapid progress, challenges remain in developing statistical methods for the analysis of genetic data and in translating findings into clinical applications.

One of the key challenges is cohort heterogeneity, which can arise from differences in ancestry, phenotype definition, or study design. Such variability can strongly influence analytical outcomes. For instance, in polygenic score (PGS) prediction, the coefficient of determination (R^2) measures predictive accuracy and is expected not to exceed SNP-based heritability (h^2_{SNP}). Yet empirical studies have reported R^2 values above this theoretical bound, accompanied by declining h^2_{SNP} estimates as increasing numbers of cohorts are combined in meta-analyses. The second data chapter of this thesis examines the conditions under which heterogeneity generates these patterns.

Although heterogeneities in allele frequencies and linkage disequilibrium (LD) structures across populations present a challenge for genetic analyses, they also offer a unique opportunity. By leveraging the diversity of LD patterns across ancestries, it becomes possible to increase our ability to identify independent trait- or disease-associated genetic variants. To this end, the third chapter introduces Manc-COJO, a method for trans-ancestry conditional and joint analysis. Manc-COJO increases power, reduces false positives relative to existing approaches, and accommodates

both genotype-level data and LD matrices, enabling global collaboration when restrictions on individual-level data apply.

Beyond sequence variation, genomic data yield valuable secondary measures. For example, mitochondrial copy number can be inferred from the number of sequencing reads annotated to the mitochondrial genome. In Chapter Four, mitochondrial copy number was estimated for half a million individuals, and its potential role in psychiatric disorders was investigated.

Overall, this thesis develops and applies new methodologies to address central challenges in quantitative genomics and advances understanding of the genetic architecture of human disease.

Table of Contents

Introduction ***- 1 -***

I.1 Background **- 2 -**

I.2 GWAS analysis **- 3 -**

I.2.1 pre-GWAS stage - 3 -

I.2.1.1 Study design - 3 -

I.2.1.2 Quality control - 7 -

I.2.1.3 Available databases - 7 -

I.2.2 GWAS stage - 7 -

I.2.2.1 Statistical model of GWAS - 8 -

I.2.2.2 Important considerations in GWAS - 9 -

I.2.2.3 Combining results from multiple GWAS studies - 11 -

I.2.3 Post-GWAS stage - 13 -

I.2.3.1 Gaining biological insight from the GWAS - 13 -

- Fine mapping of genetic variants - 14 -
- Trans-ancestry meta-analysis - 16 -
- Functional analysis of identified variants - 17 -

I.2.3.2 Understanding genetic architectures - 19 -

- Estimation of SNP-based heritability - 19 -
- Estimation of genetic correlation - 21 -

I.2.3.3 Translation into clinical care - 22 -

- Polygenic scores and out-of-sample predictions - 23 -
- Mendelian randomisation - 25 -

I.3 Expanding genomic association analyses **- 30 -**

I.3.1 Sequencing technique and genome assembly - 30 -

I.3.2 Variant calling - 31 -

I.3.3 Limitation of current approaches - 32 -

I.4 Thesis overview **- 34 -**

Polygenic Risk Prediction: why and when out-of-sample prediction R^2 can exceed SNP-based heritability ***- 35 -***

II.1 Introduction **- 37 -**

II.2 Method **- 39 -**

II.2.1 Derivation of SNP-based heritability (h^2_{SNP}) in the meta-analysis cohort and conditions when out-of-sample prediction R^2 will exceed h^2_{SNP} - 39 -

II.2.1.1 Assumptions and notation - 39 -

II.2.1.2 Meta-analysis and the SNP-based heritability from meta-analysis GWAS..... - 40 -
II.2.1.3 R^2 from out-of-sample prediction..... - 42 -
II.2.1.4 Condition when out-of-sample prediction R^2 will exceed the h^2_{SNP}..... - 42 -
II.2.1.5 The fixed-effect inverse variance meta-analysis model..... - 42 -
II.2.1.6 Special considerations in binary traits..... - 43 -
II.2.2 Simulating impacts of heterogeneities in cohort-specific heritability and between-cohorts genetic correlations on the h^2_{ma}..... - 44 -
II.2.2.1 Same underlying h^2_{SNP} for all cohorts; same underlying r_g for all pairs of cohorts... - 44 -
II.2.2.2 Same underlying h^2_{SNP} for all cohorts; different underlying r_g between pairs of cohorts - 45 -
II.2.2.3 Different underlying h^2_{SNP} for all cohorts; same underlying r_g between pairs of cohorts - 45 -
II.2.2.4 Different underlying h^2_{SNP} for all cohorts; different underlying r_g between pairs of cohorts..... - 45 -
II.2.3 Validate equation derivations with empirical data..... - 46 -
II.2.3.1 Estimation of cohort-specific SNP-based heritability and between cohort genetic correlations..... - 46 -
II.2.3.2 Calculating the h^2_{ma} and the $r_{g(ma,t)}$ directly from the meta-analysis summary statistics - 47 -
II.2.3.3 Calculating the h^2_{ma} and the $r_{g(ma,t)}$ with the derived equations..... - 48 -
II.2.4 Comparing out-of-sample predictions R^2 with h^2_{ma}..... - 48 -
II.2.4.1 Major depression..... - 48 -
II.2.4.2 Educational attainment..... - 49 -
II.3 Results..... - 49 -
II.3.1 Trends in SNP-based heritability in meta-analysis (h^2_{SNP}) with increasing meta-analysed cohorts under different genetic correlation and h^2_{SNP} settings..... - 49 -
II.3.2 Comparison of h^2_{ma} and genetic correlations derived from equations vs. estimates from meta-analysis summary statistics..... - 53 -
II.3.3 Empirical investigation using data on major depression and educational attainment.... - 53 -
II.4 Discussion..... - 55 -
Multi-ancestry conditional and joint analysis (Manc-COJO) applied to GWAS summary statistics..... - 56 -
III.1 Introduction..... - 57 -
III.2 Methods..... - 58 -
III.2.1 Concepts and assumptions..... - 58 -
III.2.1.1 Modelling true effect..... - 58 -
III.2.1.2 Modelling marginal effects in GWAS analysis..... - 59 -

III.2.1.3 Modelling joint effects ..... - 60 -
III.2.1.4 Modelling conditional effects ..... - 61 -
III.2.1.5 Assumption of a homogenous LD structure ..... - 62 -
III.2.2 Addressing the Challenges of Real-World Data ..... - 63 -
III.2.2.1 Considerations in dealing with real-world data ..... - 63 -
- Phenotypic variance ..... - 63 -
- Effective sample size ..... - 64 -
- Choice of reference sample to estimate LD correlation matrix ..... - 64 -
III.2.2.2 Overview of single ancestry COJO algorhithm ..... - 65 -
III.2.2.2 Extending COJO analysis into trans-ancestry settings ..... - 67 -
- Overview of the Manc-COJO algorithm ..... - 67 -
- Overview of the Manc-COJO --MDISA algorithm ..... - 68 -
- Optional R^2 parameter ..... - 68 -
- The benefit of trans-ancestry analyses ..... - 69 -
III.2.3 Simulations ..... - 71 -
III.2.3.1 Simulation settings ..... - 71 -
III.2.3.2 Simulation sensitivity analysis ..... - 74 -
III.2.3.3 Simulation for Manc-COJO --MDISA ..... - 75 -
III.2.4 Real data analysis ..... - 76 -
III.2.4.1 Data quality control ..... - 76 -
III.2.4.2 Data input ..... - 76 -
III.2.4.3 Out-of-sample prediction ..... - 78 -

III.3 Results ..... - 79 -
III.3.1 Overview of Manc-COJO ..... - 79 -
III.3.2 Simulation study results ..... - 81 -
III.3.3 Simulation sensitivity analysis ..... - 84 -
III.3.4 Manc-COJO --MDISA ..... - 87 -
III.3.5 Real data analysis ..... - 88 -
III.3.6 Ancestry-specific variants in real data analysis ..... - 91 -
III.3.7 SNP density ..... - 93 -
III.3.8 Choice of LD window size ..... - 93 -
III.3.9 Choice of R^2 parameter ..... - 94 -
III.3.10 Insights into Manc-COJO and Sanc-COJO outcomes ..... - 95 -

III.4 Discussion ..... - 99 -

Refined mitochondrial DNA copy number (mtCN) calling pipeline reveals nuclear genetic regulation of mtCN in blood and its association with mood disorders . - 103 -

IV.1 Introduction ..... - 104 -

IV.2 Methods ..... - 105 -
IV.2.1 Cohorts ..... - 105 -
IV.2.2 Calculation of mtCN ..... - 106 -
IV.2.3 Assessing coverage bias ..... - 107 -
IV.2.4 Comparing different copy number calling pipelines ..... - 109 -
IV.2.5 Calling mtCN ..... - 110 -
IV.2.6 Definition of MDD and BIP cases and control ..... - 111 -
IV.2.7 Sensitivity analysis ..... - 112 -
IV.2.8 GWAS of mtCN and downstream analysis ..... - 112 -
IV.2.9 Single cell RNAseq analysis ..... - 113 -
IV.3 Results ..... - 113 -
IV.3.1 Regional and chromosomal coverage bias ..... - 113 -
IV.3.2 Impact of coverage bias and QC pipelines ..... - 116 -
IV.3.3 Strategies in calling mtCN ..... - 118 -
IV.3.4 Association between mtCN and mood disorders ..... - 121 -
IV.3.5 Sensitivity analysis ..... - 123 -
IV.3.6 Genetic control over mtCN ..... - 125 -
IV.3.7 Cell types in blood and brain that are associated with mtCN ..... - 126 -
IV.4 Discussion ..... - 127 -
Discussion ..... - 129 -
V.1 Summary of Findings ..... - 130 -
V.2 Limitations ..... - 131 -
V.2.2 Limitations in the Manc-COJO algorhithm ..... - 132 -
V.2.3 Limitation in mitochondria copy number analyses and WGS analyses ..... - 133 -
V.3 Future directions ..... - 135 -
V.4 Conclusion ..... - 139 -
Appendix ..... - 141 -
A1. Supplementary files for Chapter II ..... - 142 -
..... - 144 -
A2. Supplementary files for Chapter III ..... - 148 -
A3. Supplementary files for Chapter IV ..... - 164 -
A4. Supplementary files for Chapter V ..... - 175 -
References ..... - 177 -

I

Introduction

I.1 Background

Quantitative genomics is study of the genetic basis of the complex human disease and traits. Unlike the study of monogenic disease such as cystic fibrosis and sickle-cell anaemia, quantitative genomics focusses on traits or diseases that are controlled by multiple genes. Height is an example of a complex trait, with a recent study identifying ~7000 independent or near independent associated genetic loci[1].

In the past few decades, with emerging technologies such as the single nucleotide polymorphism (SNP) genotyping array, whole exome sequencing (WES), whole genome sequencing (WGS), and single cell RNA sequencing, huge advances have been made in the field of quantitative genomics in terms of both analytical methodology and their application. An example of these advances is the increasing popularity of genome-wide association studies (GWAS). Given the low cost of SNP arrays, the number of genome-wide association studies has increased exponentially over the last few years – by July 2024, more than 100,000 GWAS summary statistics covering over 15,000 traits had been published and the rate of growth is not slowing down[2]. GWAS have provided empirical evidence for the polygenic nature of complex traits/disease and enable us to generate polygenic scores (PGS) in independent samples which predict individual genetic risks by aggregating small estimated effects across the genome[3].

Advances in quantitative genomics have been translated into clinical care, in terms of disease prevention, diagnosis, and treatment. For example, when prescribing the antiepileptic medication carbamazepine, testing for HLA-B alleles has been recommended to reduce the risk of hypersensitivity skin reactions[4]. Another example is the repositioning of biological medications to target the IL23 pathways in the treatment of psoriasis. This repositioning was initially motivated by GWAS discoveries and has been proved to be effective in clinical practice[5,6].

Despite huge advances made in past years, significant challenges remain both in quantitative genomics methodology and in its translation to clinical care. For example, although PGS have proven highly effective in population genetics, PGS developed in one ancestry always exhibit reduced transferability to other ancestries[7]. Moreover, their discriminative accuracy is currently insufficient for routine clinical application[8].

The overarching aim of this thesis is to develop and apply new methodologies to address central challenges in quantitative genomics and to advance understanding of the genetic architecture of human disease. To achieve this aim, the remainder of this chapter will first provide a systematic review of the foundations and progress of quantitative genomics, encompassing both upstream and downstream aspects of GWAS analysis, which together form the backbone of quantitative genomic research. This is followed by a discussion of analyses focused on larger-scale genetic changes that extend beyond the scope of traditional GWAS. With this theoretical and methodological context established, the chapter concludes with an overview of the projects presented in this thesis.

I.2 GWAS analysis

GWAS is a powerful study design, where allele frequencies are compared between phenotypically different individuals within a group of similar ancestry[9]. It is the backbone of quantitative genomics, and enables a wide range of downstream clinical applications, such as genetic risk prediction. In this section, an overview of important steps and methods in pre-GWAS, GWAS and post- GWAS analyses is provided.

I.2.1 pre-GWAS stage

I.2.1.1 Study design

Before a GWAS is conducted, careful planning is needed to maximise statistical power and minimise confounding and bias.

Maximising statistical power The statistical power of a GWAS can be quantified with the non-centrality parameters (*NCP*)[6]:

$$NCP = N \times r^2 \times f^2/(1 - r^2 f^2)$$

Here N is the sample size of the experiment (note, in case-control studies, N is the effective sample size, which equals $\frac{4N_{case}N_{control}}{N_{case}+N_{control}}$, where N_{case} and $N_{control}$ are the number of cases and controls in a case-control study, respectively), f^2 is proportion of the phenotypic variance explained by a specific causal variant in the population, which under the Hardy-Weinberg equilibrium (HWE), can be calculated from minor allele frequency (p; q is the frequency for the alternative allele, where $q = 1 - p$) and the effect size (β, in standard deviation (SD) units):

$$f^2 = 2pq\beta^2$$

r^2 is a measure of linkage disequilibrium (LD), which quantifies the squared correlation between a pair of genetic markers[10]. In a SNP array study, the true causal variants are not always genotyped, but their effects can be captured by genotyped variants in LD with them. In such case, r^2 could be smaller than 1 but significantly greater than 0. In the context of WGS, the theoretical value of r^2 approaches 1 under ideal conditions—that is, in the absence of measurement error at the causal variant locus and assuming that the causal variants are both successfully genotyped and accurately called.

As shown in formulae above and in **Figure I.1**[6], the statistical power to detect associations between a causal variant and a trait depends on the sample size of the study, allele frequencies of causal variants, effect sizes of causal variants, and the density of genotyping probes (because r^2 is likely to be higher between genotyped variants that are physically close on the genome).

Allele frequencies and effect sizes of causal variants are intrinsic to the nature of a population (concepts of the idealised population are described in more detail in the Chapter II). Therefore, to increase the statistical power, researchers need either to increase the sample size of the study or increase the genotyping probes densities.

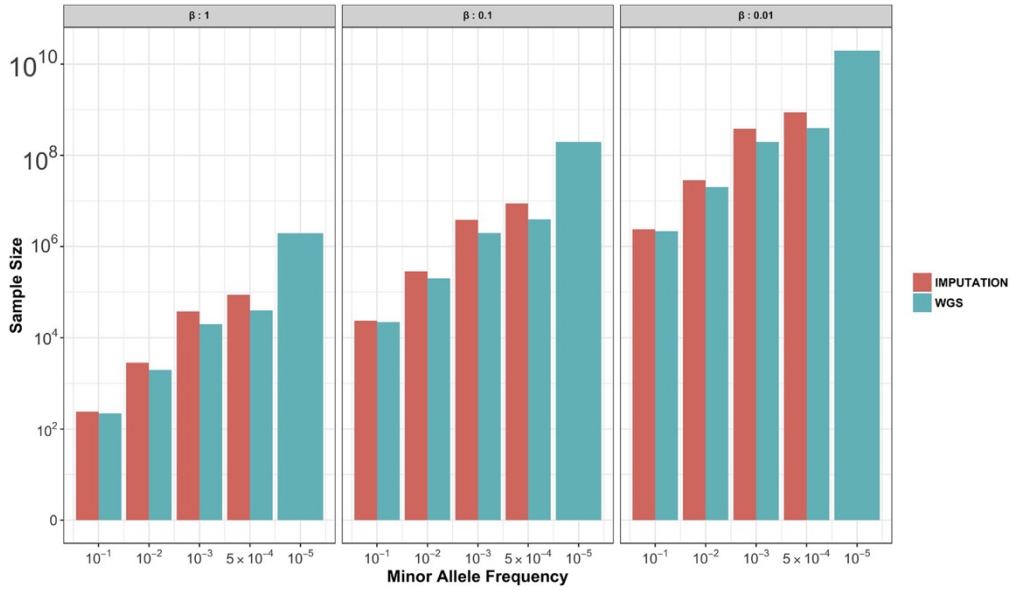


Figure I.1. Minimum Sample Sizes required for detecting Trait-SNP Associations from Imputed and WGS Data. This figure comes from the figure 1 of *Visscher, Wray et al. 2017*. It shows that under the assumption of HWE, type I error rate of 5 X 10^{-8} and power of 0.8, how many samples are needed to detect trait-SNP associations of certain effect size. For imputed data, r^2_{imp} are estimated from the average r^2_{imp} in the Haplotype Reference consortium. The figure shows that to detect causal variants with small effect sizes and low minor allele frequencies, more than a million samples must be included in the study even with the WGS data.

To increase the sample size, besides simply increasing the investments to include more participants in a single study, which is not always feasible, collaboration is often required in the field of quantitative genomics. Results from different studies can be meta-analysed to increase the statistical power. Several large biobanks exist where extensive phenotypic and genotypic data are available (these are briefly introduced in the section I.2.1.3). It is notable that biobanks may not have enough samples for a certain disease, especially rare diseases. Additionally, sometimes phenotypic data in large biobanks are *'proxy'* phenotypes (for example, there are lots of '*proxy'* phenotypes for depression in the UK Biobank, such as *anxiety/depression today* or *probable recurrent major depression*), and caution is needed in interpretation and meta-analysis of these data[11]. To address these issues, several consortia for specific diseases have been established. The Psychiatric Genomics Consortium[12] is a great example of international collaborations where results from 36 countries are combined to empower discoveries of hundreds of novel loci in psychiatric disorders.

A common strategy to improve probe densities in the SNP array data is imputation. Imputation is the process where ungenotyped variants are inferred from the genotyped SNPs integrated with a reference panel of haplotype or genotype panel such as the 1000 Genome project (1000G)[13], the Haplotype Reference consortium (HRC)[14] and TOPMed[15]. Standard quality-control (QC) pipelines such as RICOPILI[16] are commonly used before the imputation process because imputation algorithms rely on high-quality, well-calibrated, and accurately annotated genotype data as input. Nowadays, with the advances in WES and WGS technologies, whole exomes or genomes can be directly genotyped without the need for imputation. However to date, WES and WGS are much more costly than the SNP array data. Imputation from SNP array provide adequate resolution for common variants (here arbitrarily defined as variants with MAF > 0.01) - a simulation study showed that 97% of genetic variation can be captured by imputation (without applying any filtering based on imputation INFO score) when compared with WGS[17]. For rare variants, WGS indeed has much better resolution – imputation can only capture 68% of genetic variation comparing with WGS[17]. However, the required sample size to detect associations of rare variants in WGS is much larger than that of common variants, because power depends on variance explained ($f^2 = 2pq\beta^2$) such that the heterozygosity component (*2pq*) is maximised when *p*=*q*=0.5. Although some rare variants exhibit larger-than-average effect sizes, their effects are detectable only when the magnitude of the effect (β) is sufficiently large to offset the reduction in allele frequency—a condition that is rarely observed in empirical data[18]. Therefore, choice of the genotyping platform will depend on the aim of the study (i.e., whether rare

variants are of great importance for the study), the funding, and other available resources.

In summary, to maximise statistical power, careful planning is needed before conducting a GWAS in terms of considering the study purpose and available resources, maximising sample sizes (and the phenotypes that can be measured), and choosing appropriate genotyping platforms.

Minimising confounding and bias Biases occur when deviation from the truth happens because of systematic factors rather than random error, and confounding refers to conditions when the spurious associations between exposure and outcomes can be accounted by other (single or multiple, known or unknown) factors[19]. Bias and confounding can occur in every step of a quantitative genetics study. Therefore, before GWAS is conducted, careful planning is needed to minimise confounding and bias.

In the pre-GWAS stage, biases can be minimised by adequately controlling for batch effects and genotyping array effects, strictly following the QC guidelines, and adequate considerations of cryptic relatedness and population stratification[20]. Ideally, all individuals in the study should be genotyped as part of the same batch with the same platform to avoid confounding and bias. This is infeasible with biobank-level data and international consortia, moreover genotyping technologies are updated over time. Therefore, careful recording of genotyping batch and platform is needed in the pre-GWAS stage such that their effects can be tested for and, if needed, accounted for in the GWAS model, usually by fitting them as covariates. Strict sample QCs are also essential parts of reducing confounding and bias which will be discussed in section I.2.1.2.

Cryptic relatedness occurs when the true genetic relationship between individuals is higher than documented from the phenotype data collection. It can happen when the pedigree information is missing/documented incorrectly, or when individuals are distantly related and such relatedness is unknown. Population stratification occurs when the sample comprises individuals drawn from populations whose allele frequencies are systematically different across many loci, and their mean phenotypic values are also different[9]. In the pre-GWAS stage, cryptic relatedness and population stratification can be addressed either by adopting a study design that restricts the sample to unrelated individuals of homogeneous ancestry, albeit at the expense of losing information on population heterogeneity. For example, in case–control studies

of binary traits, cases and controls can be closely matched with respect to ancestry to minimise confounding arising from population structure. Alternatively, researchers may explicitly document relatedness and ancestral background so that these factors can be appropriately modelled in subsequent association analyses.

I.2.1.2 Quality control

A standard QC pipeline[21] should be implemented at the pre-GWAS stage. This typically includes the removal of individuals with high genotyping failure rates and SNPs with substantial genotype missingness (a threshold of 10% is commonly applied, although more stringent cut-offs are often used). SNPs that deviate from Hardy–Weinberg equilibrium (HWE test $p < 1 \times 10^{-6}$), which frequently indicates technical artefacts, are excluded. Variants with low minor allele frequency (commonly MAF < 0.01 in standard GWAS) are also removed, as they are more susceptible to genotyping errors and batch effects. Although the study of rare variants is important, it requires specialised analytical approaches and is beyond the scope of this thesis. Additionally, samples are commonly excluded when the heterozygosity rate is high, and when the recorded sex is different from the genotype-inferred sex, which could reflect sample contaminations or mix-up.

I.2.1.3 Available databases

There are several publicly available databases that are commonly used in quantitative genomics research. The UK Biobank (UKB) is the most used biobank with a wide range of phenotypic and genetic data. The UKB has whole-genome SNP array data for 488,377 volunteer participants recruited in the United Kingdom. Other examples of biobanks where individual genotypically data are available include All of US research program[22], Our future health program[23], Dutch Lifeline cohort[24], Chinese Kadoorie Biobank[25], and Vanderbilt's BioVU biobank[26]. For certain traits, individual level data are available by submitting applications to certain consortia, such as depression data from the PGC. Additionally, GWAS summary statistics are available on several different sources. Examples include GWAS Catalog[27], GeneAtlas[28], FinnGen[29], dbGAP[30], and Biobank Japan[31].

I.2.2 GWAS stage

GWAS tests for associations between phenotypic values and genetic variants, evaluating one variant at a time. Linear and logistic models are the basis of most

GWAS tools for continuous traits and binary traits, respectively[32]. In this section, a review of statistical methods commonly used in GWAS is provided, followed by caveats that need to be considered in GWAS studies. Additionally, multiple GWAS can be combined to increase the statistical power, and common methods of meta-analysis are introduced.

I.2.2.1 Statistical model of GWAS

PLINK is a commonly used tool in the association testing using a standard linear regression model[33]. Assuming an idealised population in which the genotypes and phenotypes of every individual are fully observed. The model can be expressed as:

$$\mathbf{y} = \mathbf{X_m} b_m + \mathbf{e_m}$$

Here, let sample size of the study to be *N*,

- $\mathbf{y}$ is an *N*x1 vector of phenotype residuals (Confounding factors such as age and sex may be included as covariates, although their inclusion depends on the sampling design. For simplicity, assume that all relevant covariates have been regressed out of the phenotype; thus, the phenotype considered here is the residual from a linear regression on fixed-effect covariates. These residuals are typically standardised to have mean zero and variance one, so that effect size estimates are expressed in standard deviation units).
- $\mathbf{X_m}$ is an *N*x1 vector of genotypes at SNP *m*
- b_m is the **true** marginal effect size of the SNP *m*
- $\mathbf{e_m}$ is an *N*x1 residual error term vector, $\mathbf{e} \sim \boldsymbol{N}(\mathbf{0}, \mathbf{I}\sigma_e^2)$; $\mathbf{I}$ is the identity matrix, and σ_e^2 is the residual error variance

In practice, it is infeasible to collect data on every individual in the population. Therefore, the true marginal effects must be estimated from a sample. The ordinary least squares (OLS) estimator of b_m, denoted $\hat{b}_m$, is given by:

$$\hat{b}_m = \frac{\mathbf{X}_{\boldsymbol{m}}' \hat{\boldsymbol{y}}}{\mathbf{X}_{\boldsymbol{m}}' \mathbf{X}_{\boldsymbol{m}}}$$

Notably, $\mathbf{X}_{\boldsymbol{m}}$ now denotes the corresponding N_m x 1 vector of **observed** genotypes at SNP *m*, where N_m is the sample size for SNP m; $\hat{\boldsymbol{y}}$ is N_m x 1 vector of **observed** phenotypes. The standard error of the GWAS estimate at SNP m ($\hat{S}_m$) is given by:

$$Var(b_m - \hat{b}_m) = \hat{S}_m^2 \approx \frac{1}{2p_m q_m N_m}$$

Several mixed-linear-model (MLM) based approaches have been developed to account for the population stratification and cryptic relatedness. Examples of MLM-

based tools include GCTA-fastGWA [32] and BOLT-LMM [34]. In general, under the idealised population assumptions described above, the statistical models in the MLM-based approach can be expressed as:

$$\mathbf{y} = \mathbf{X_m} b_m + \mathbf{X_F}\mathbf{b_F} + \mathbf{g} + \boldsymbol{e}$$

Here,

- $\mathbf{X_F}$ is an *N*x*F* fixed-effect covariate matrix, where *F* is the number of fixed-effect covariates in the model
- $\mathbf{b_F}$ is a *F*x1 vector of the *F* fixed-effect covariates effect sizes
- $\mathbf{g}$ is a vector of total genetic effects accounted by the relatedness between samples, where $\mathbf{g} \sim \boldsymbol{N}(\mathbf{0}, \mathbf{A}\sigma_g^2)$. **A** is an *N*x*N* relationship matrix based on pedigree structure but can also be estimated from the genetic relationship matrix (GRM), or from a sparse GRM in which off-diagonal elements below a specified threshold (many studies arbitrarily use < 0.05) are set to zero. The parameter σ_g^2 denotes the corresponding genetic variance component.

In these methods, the estimate of the effect size b_m is given by:

$$\hat{b}_m = \frac{\mathbf{X_m'}\mathbf{V^{-1}}\hat{\boldsymbol{y}}}{\mathbf{X_m'}\mathbf{V^{-1}}\mathbf{X_m}}$$

And the standard error of the GWAS estimate is given by:

$$Var(b_m - \hat{b}_m) = \hat{S}_m^2 = (\mathbf{X_m'}\mathbf{V^{-1}}\mathbf{X_m})^{-1}$$

Here, $\mathbf{V}$ is the variance-covariance matrix of the phenotype, where $\mathbf{V} = \mathbf{A}\sigma_g^2 + \mathbf{I}\sigma_e^2$

Different MLM based methods use different strategies to estimate σ_g^2 and σ_e^2, resulting in different computational time. For example, fastGWA[32] uses an efficient restricted maximum likelihood (REML) algorithm termed fastGWA-REML, which is more time-efficient when compared with BOLT-LMM[34], at the expense of a slight reduction in power.

I.2.2.2 Important considerations in GWAS

Confounding and bias As discussed previously, confounding and bias need to be minimised throughout quantitative genetic analyses. Population stratification and cryptic relatedness are two major confounding factors that need to be considered in GWAS analysis.

Population stratification is commonly accounted for by fitting top PCs of the GRM as covariates or using MLM based approaches where all variants are fitted as random

effects, or combination of two strategies. In general, the controlling for the PCs is viewed as an approximation for MLM based approaches. Therefore, MLM based approaches should outperform PC based methods in terms of controlling population stratification in general, at the expense of computational efficiency. Under certain conditions, such as when a spatially confined unknown environmental confounder exists, PC controlling could outperform MLM based method because of different treatments of fixed and random effects[35]. In addition, in large samples, combining both MLM based approaches and PC controlling approaches will increase robustness[32].

Aside from aforementioned strategies in the planning stage, cryptic relatedness can also be accounted for by either removing individuals that are correlated or incorporating such correlations in a statistical model. Software such GCTA[36] enable users to remove one individual from a pair of individual that have genetic relatedness above certain threshold, and by repeating this process iteratively all remaining individuals will have genetic relatedness less than that threshold. Alternatively, relatedness can be accounted for within the MLM frameworks described above. For example, by using a sparse GRM to model relatedness, GCTA-fastGWA captures phenotypic covariance in a manner comparable to analyses based on pedigree-derived relationship matrices, while offering substantially greater computational efficiency[32].

Additional confounding factors must also be adjusted for in GWAS to avoid spurious associations. Common sources of confounding include[37]:

1. Technical artefacts, such as batch effects and genotyping array effects. These are often addressed using principal component analysis (PCA). Ancestry outliers are first identified and removed via PCA projection; subsequent PCA conducted within the ancestry-homogeneous sample typically captures residual technical variation and more localised cryptic relatedness.
2. Phenotypic and environmental heterogeneity, including variables such as age and sex.
3. Known biological or physiological correlates, as well as established disease status.

When the significance level of certain SNP is very high, the genotype of that SNP could be fitted as a covariate as well to unmask the true effect of surrounding SNPs. For example, in ankylosing spondylitis, conditioning on the primary SNP revealed significant associated secondary SNPs in IL23R and ERAP1 loci[38].

Notably, decisions regarding the inclusion of specific covariates should be guided by the overall objectives of the study, as their inclusion may have unintended consequences. Although adjusting for covariates can reduce confounding, it may also reduce statistical power, particularly for binary traits with a limited number of cases[39]. Another important consideration is that covariates included in the model should not be genetically correlated with the trait under investigation, as this may introduce overadjustment bias when the variable represents an intermediate step in the causal pathway linking the exposure to the outcome[40], or collider bias, which occurs when two (or more) variables influence a third variable and conditioning on that third variable induces a spurious statistical association between them. For example, leg length should not be included as a covariate in a study of height.

False discoveries A stringent significance threshold is used in GWAS studies to account for multiple testing – the HapMap Project has shown that there are approximately 10^6 independent common genetic variants (minor allele frequencies > 0.05) across the human genome[41,42]. To ensure family wise error rate of 0.05, Bonferroni correction is applied resulting in a significance threshold of $0.05/10^6 = 5x10^{-8}$, which is the traditional genome wide significant threshold used in many studies. However, an even more stringent threshold should be applied if the MAF threshold is lowered, as the multiple testing burden is increased in these cases.[9,43]

Another important consideration in GWAS studies is the replication of findings in an independent cohort. Replication is the traditional gold standard to avoid winners curse, where the effect size of newly identified locus is always overestimated. Notably, winner's curse will become less significant when the power of study is high[44]. As these days largest discovery size usually takes precedence, and replication samples are hard to find, study design should therefore be context specific.

I.2.2.3 Combining results from multiple GWAS studies

As outlined in the previous section, GWAS results from multiple studies can be combined to increase statistical power. Various tools are available for meta-analysis of GWAS summary statistics, such as METAL[45].

METAL uses sample size or inverse variance-based weighting of summary statistics[45]. In the *sample size based approach*, the overall Z score for the jth SNP:

$$Z_j = \frac{\sum_i Z_{i,j} w_{i,j}}{\sqrt{\sum_i w_{i,j}^2}}$$

Here,

- $w_{i,j}$ is the weight applied to the jth SNP in the ith study, $w_{i,j} = \sqrt{N_i}$, N_i is the sample size of *i*th study
- $Z_{i,j} = \Phi^{-1}(P_{i,j}/2)sign(\Delta_{i,j})$, $P_{i,j}$ is the p-value of the *j*th SNP in the *i*th study, and $sign(\Delta_{i,j})$ is the direction of the association of the *j*th SNP in the *i*th study, i.e. +1 or -1.

In an *inverse variance-based approach*,

$$\hat{b}_j = \sum_i \hat{b}_{i,j} w_{IVM(i,j)} / \sum_i w_{IVM(i,j)}$$

Here,

- $\hat{b}_j$ is the observed marginal effect size of SNP *j* in the meta-analysis
- $\hat{b}_{i,j}$ is the observed marginal effect size of the *j*th SNP in the *i*th study
- $w_{IVM(i,j)} = 1/\hat{S}_{i,j}^2$; $\hat{S}_{i,j}$ is the standard error of the *j*th SNP in the *i*th study

The standard error of the meta-analysis for SNP *j*, $\hat{S}_j$, is:

$$\hat{S}_j = \sqrt{1/\sum_i w_{IVM(i,j)}}$$

The overall Z score of the SNP *j* in the meta-analysis is thus:

$$Z_j = \hat{b}_j/\hat{S}_j = \sum_i \hat{b}_{i,j} w_{IVM(i,j)} \Big/ \sqrt{\sum_i w_{IVM(i,j)}}$$

In principle, the inverse-variance–weighted (IVM) approach is preferable to a sample size–based method when cohort-specific variance estimates are available, as it minimises the variance of the combined estimator[46]. However, in the presence of sample overlap across cohorts, further adjustments are required to prevent inflation of false positive discoveries[47,48].

Additionally, GWAS results from different but correlated traits can be combined to increase statistical power. MTAG[48] is a commonly used tool in this category.

MTAG is able to perform meta-analysis across correlated traits. In MTAG, a focal trait is defined (for example, depressive symptoms), and information from other genetically correlated traits (such as neuroticism and subjective well-being) can be borrowed to increase the statistical power. The core assumption is that traits

included in the analysis have same variance-covariance matrix of effect sizes across all SNPs. MTAG is an estimator of generalised method of moments (GMM), and its moment condition is given by[48]:

$$\mathrm{E}\left(\hat{\mathbf{b}}_{\mathbf{j}} - \frac{\boldsymbol{\omega}_{\mathbf{t}}}{\omega_{tt}} b_{j,t}\right) = \mathbf{0}$$

And MTAG estimates ($\hat{b}_{\mathrm{MTAG},j,t}$) for SNP *j* on the trait *t* is given by:

$$\hat{b}_{\mathrm{MTAG},j,t} = \frac{\frac{\boldsymbol{\omega}_{\mathbf{t}}^{\mathbf{T}}}{\omega_{tt}}\left(\boldsymbol{\Omega} - \frac{\boldsymbol{\omega}_{\mathbf{t}}\boldsymbol{\omega}_{\mathbf{t}}^{\mathbf{T}}}{\omega_{tt}} + \sum_{j}\right)^{-1}}{\frac{\boldsymbol{\omega}_{\mathbf{t}}^{\mathbf{T}}}{\omega_{tt}}\left(\boldsymbol{\Omega} - \frac{\boldsymbol{\omega}_{\mathbf{t}}\boldsymbol{\omega}_{\mathbf{t}}^{\mathbf{T}}}{\omega_{tt}} + \sum_{j}\right)^{-1}\frac{\boldsymbol{\omega}_{\mathbf{t}}}{\omega_{tt}}} \hat{\mathbf{b}}_{\mathbf{j}}$$

Here,

- $\mathbf{b_j}$ is a *T*x1 vector of true marginal effect size for SNP *j*, where *T* is the number of traits included in the MTAG analysis
- $\hat{\mathbf{b}}_{\mathbf{j}}$ is a *T*x1 vector of estimated marginal effect size for SNP *j*
- $\boldsymbol{\Omega}$ is a *T*x*T* variance-covariance matrix of $\mathbf{b_j}$
- $\boldsymbol{\omega}_t$ is a *T*x1 vector which is the *t*th column of $\boldsymbol{\Omega}$
- ω_{tt} is the *t*th diagonal column of $\boldsymbol{\Omega}$
- $b_{j,t}$ is the true marginal effect size of SNP *j* on the trait *t*
- $\boldsymbol{\Sigma}_j$ is a *T*x*T* variance-covariance matrix of the estimation error, which is estimated by LDSC

When applied to summary statistics of the same trait with no sample overlap, MTAG reduces to an IVM. It can also be used to account for sample overlap when combining GWAS of the same trait. In general, MTAG tends to improve polygenic prediction accuracy. However, the false-positive discovery rate may increase when many low-powered studies are combined or when the genetic correlations between traits are weak[48].

I.2.3 Post-GWAS stage

There are several powerful downstream analyses of the GWAS which enable us to gain better biological insight, better understand the genetic architecture of traits, and translate our findings into clinical care.

I.2.3.1 Gaining biological insight from the GWAS

An outcome of a GWAS analysis is a list of SNPs with allele-associated effect sizes and p-values, which are called the GWAS summary statistics. GWAS summary statistics

can be used for fine-mapping to prioritise genetic variants that are more likely to be causal within a genomic region. The functions of these prioritised variants can then be interpreted at the gene and biological pathway levels through integration with multi-omics data.

- *Fine mapping of genetic variants*

Once a list of SNPs significantly associated to a trait have been identified, fine mapping can be performed, aiming to find the causal variant for the trait of interest. The general workflow of fine mapping is to first examine the LD structure around identified variants and identify genes that have been mapped to that trait-associated region. Then, several fine-mapping strategies are available, and researchers need to weigh the pro and cons for each strategy and choose the most appropriate one based on context. Next, identified SNPs can be integrated with functional data. Finally, selected SNPs can be evaluated with the wet-lab based functional analysis.

LD in fine-mapping LD complicates fine-mapping, particularly in regions with complex LD patterns. It is challenging to identify causal variants in high-LD regions because SNPs in these areas are highly correlated[49]. Since SNPs in such regions tend to be inherited as blocks, the combinatorial complexity within the study population may be insufficient to disentangle the effects of individual variants[50]. As a result, the outcomes of fine-mapping analyses in high-LD regions should be interpreted with caution, as the identified variant may not be the true causal variant.

Visualize local associations and LD structures Local LD structures and associations mapped to the regions of interests can be visualised using tools such as Haploview[51] and LocusZoom[52]. These approaches are useful for selecting SNPs for follow-up in regions where the LD structure is relatively simple, as the index SNP should, in theory, be close to the causal variant (discussed in detail in Chapter III). However, as these approaches do not consider the joint effect of SNPs, and do not have a numeric measurement of whether SNPs are causal, more informative methods should be used, especially when LD structure is complex[49].

Fine-mapping strategies Commonly used fine-mapping methods include conditional joint analyses, penalised regression models, and Bayesian approaches.
Conditional joint analysis incorporates the lead variant as a covariate in the regression model. If a second associated variant is detected, it is added to the model and the regression is repeated. This iterative process continues until no additional

variants are included or excluded, thereby enabling the identification of all independently trait-associated variants. Details of the model selection procedure will be presented in Chapter III. GCTA-COJO is currently the most widely used tool for conditional and joint analysis[53]. It requires two inputs: GWAS summary statistics and individual-level genotype data, with the latter used as the LD reference. When individual-level data from the original GWAS are unavailable, an external cohort or a publicly available reference panel may be used as the LD reference instead. In GCTA-COJO, estimated joint effects ($\widetilde{\boldsymbol{\beta}}$) and standard error ($\hat{S}_{\widetilde{\boldsymbol{\beta}}}$) are given by:

$$\widetilde{\boldsymbol{\beta}} = \mathbf{B}^{-1}\mathbf{D}\hat{\mathbf{b}}$$

$$\hat{S}_{\widetilde{\boldsymbol{\beta}}} = \sqrt{\sigma_J^2 \mathbf{B}^{-1}}$$

Where, assuming HWE,

$$\mathbf{B} = \mathbf{D}^{1/2}\mathbf{D}_{\mathbf{W}}^{-1/2}\mathbf{W}^{\mathbf{T}}\mathbf{W}\mathbf{D}_{\mathbf{W}}^{-1/2}\mathbf{D}^{1/2}$$

Here, assuming there are *M* SNPs in the model,

- $\mathbf{D}$ is a diagonal matrix with the *i*th element given by $\widehat{D}_i = 2\hat{p}_i\hat{q}_i N_i$. This term accounts for scaling, as in real data the marginal effect sizes correspond to a genotype matrix that is not standardised.
- $\mathbf{W}$ is an *M*x*M* genotype matrix in the reference samples
- $\mathbf{D_W}$ is a diagonal entries of $\mathbf{W^TW}$ with $\mathbf{D_{W(j)}} = \sum_{\mathbf{i}}^{\mathbf{m}} \mathbf{w_{ij}^2}$
- $\hat{\mathbf{b}}$ is an *M*x1 vector of observed GWAS marginal effects

However, the conditional joint analysis tends to eliminate useful variants when the LD structure is complex, because of the nature of forward stepwise selection it uses[54]. Several *penalised regression models* have been developed to address this issue. In brief, penalised regression models perform SNP effects estimation and selection at the same time, with the small effect estimates being shrunk towards zero. However, *penalised regression models* tend to give a sparse model where only one or few SNPs are selected within each region, and causal SNPs can be excluded when the LD structure is complex[49]. In simulation studies, *Bayesian approaches* have shown they can confer advantages over the two methods mentioned above in terms of causal variant selection[55,56]. The general idea of Bayesian approaches can be expressed as[56]:

$$\mathrm{P}\left(\mathbf{c}^*|\hat{\mathbf{b}}\right) = \frac{\mathrm{P}\left(\hat{\mathbf{b}}|\mathbf{c}^*\right)\mathrm{P}(\mathbf{c}^*)}{\sum_{\mathbf{c}\in\mathbf{C}} \mathrm{P}\left(\hat{\mathbf{b}}|\mathbf{c}\right)\mathrm{P}(\mathbf{c})}$$

Here,

- **c** is an *M*x1 indicator vector of causal status, where possible values of c are -1(negative causal effect), 0(no causal effect), 1(positive causal effect). *M* is the number of SNPs.
- **c*** is an *M*x1 vector of particular causal status
- $\hat{\mathbf{b}}$ is an *M*x1 vector of observed marginal effect estimates, as defined previously

The idea is to estimate posterior probability distributions $\mathrm{P}(\mathbf{c}^*|\hat{\mathbf{b}})$, the probability of a particular **c*** given the observed marginal effects, by using the prior probability distributions. Bayesian based approaches have several advantages, for example they can incorporate factors such imputation quality into the model, and posterior probability can be directly compared with each other. However, output of causal variants is inconsistent between different approaches. In regions with very complex LD, the power of Bayesian based approach will also decrease, as multiple causal configurations may yield nearly indistinguishable likelihoods, leading to increased uncertainty in variant prioritisation. Additionally, Bayesian based approaches can be time consuming and results of posterior probability can be hard to interpret. To address these issues, the method SuSiE[57] has been developed by applying stepwise selections in a Bayesian framework. The method also generates credible sets–a set of variables which contain at least one effect variable at a certain probability level, and they prioritised variables within each set based on the probability. Notably, the credible sets are built with distributional assumptions, and methods are validated using simulation based on those assumptions. However, in real life those assumptions can be violated. Studies show the credible sets from a GWAS do not always hold up in a GWAS of larger sample sizes of the same trait[49,58].

In general, all fine mapping strategies are better at identifying causal variants compared to simply taking index SNPs, and each method has their own advantages. However, their performance can be suboptimal in regions with complex LD structure. This is a real issue in biology, as massively parallel reporter assays (MRPAs) showed that genetic effects of each region on the traits are moderated by joint actions of multiple genetic variants[59]. Trans-ancestry meta-analysis and functional analysis may help increase resolution of fine-mapping and provide biological insights of the identified genetic variants.

- *Trans-ancestry meta-analysis*

The basis of using trans-ancestry meta-analysis to improve fine-mapping resolution is under the assumption that causal variants for specific traits are same across

ancestries, and the direction of the association is also the same. Under such assumption, fine-mapping resolution will improve because the LD structure of different ancestries is different. Difference in LD structure result from events including natural selection, genetic drift, and population subdivision. Consequently, non-causal SNPs that are in high LD with true causal SNPs in one ancestry might not be in LD with causal SNPs in another ancestry. Hence, variants tagged in the trans-ancestry meta-analysis are more likely to be true causal variants. The common causal variant assumption generally holds, because common SNPs are ancient and so common causal variants are likely shared across ancestries[60], with the exception of some ancestry-specific rare variants. An example of such variant is a missense mutation in the *GRASP* gene, which relates to asthma only in the Latino population[61]. However, ancestry-specific rare variants explain an insignificant proportion of phenotypic variance and the majority of causal variants are shared across ancestries[61,62]. Research showed that trans-ancestry meta-analysis of European and Asian population will provide little gain in terms of fine mapping resolution, but mixture of European and African population will provide much better resolution because of narrower LD regions in African ancestry[63]. Simulation studies suggest that datasets composed of equal proportions of European and African ancestry samples achieve the greatest fine-mapping resolution compared with other ancestry compositions and relative sample size combinations[41,49]. The lack of data availability in other ancestries is currently the rate-limiting step in trans-ancestry meta-analysis.

- *Functional analysis of identified variants*

Once a list of potential causal variants is identified, the function of the identified variant is of great interest to researchers as it may provide insights into the pathophysiology of diseases, or the biological mechanisms behind certain traits. The interaction between genetic variants and the level of transcription, translation, as well as protein structure, have been studied extensively in past few years.

Genome annotation Some identified genetic variants are located within protein-coding regions. For these variants, several computational tools can be used to assess their potential functional effects. For example, ANNOVAR provides variant annotations such as genomic location, predicted amino-acid changes, and known disease associations[64]. CADD generates a deleteriousness score that estimates the potential impact of a variant[65], while REVEL predicts the pathogenicity of missense variants[66]. Integrating results from multiple annotation methods can improve the accuracy of variant prioritisation in fine-mapping analyses[49]. However, most

identified genetic variants are located in the non-protein-coding regions[6], which commonly have regulatory functions such as histone modification and DNA methylation. They most frequently affect genes in nearby regions but can also have long range effect[67]. The effect of genetic variants in non-coding regions has been the focus of projects such as ENCODE[68] and the Roadmap Epigenomics[69].

Transcriptome Efforts have also been made to increase understanding the associations between genetic variants and gene expression. Expression quantitative trait loci (eQTL) are DNA variants that are significantly associated with gene expression levels, and eQTL database are available as a result of projects such as GTEx[70]. As regulation of gene expression is different in different cell types, cell type specific eQTL analysis (such as the single-cell RNA sequencing analysis) has become popular in recent years with large databases such as the GTEx becoming available for analysis[71]. The eQTL data can also be used in fine-mapping and transcriptome-wide association studies (TWAS) which link GWAS results to gene expression data. Notably, TWAS may identify non-causal genes when the analysed tissue is not relevant to the trait of interest, because some eQTLs are shared across tissues and can produce spurious gene expression–trait associations. Therefore, eQTL data from the most biologically relevant tissue (e.g., brain tissue for psychiatric disorders) should be used[72].

Proteomics pQTLs are genetic variants associated with protein level. Recent advances in affinity-based proteomics technologies have enabled large-scale measurement of protein abundance and have accelerated progress in this field. Large biobanks, including the UK Biobank and the China Kadoorie Biobank, have released datasets measuring the levels of thousands of proteins. However, proteomic studies remain an emerging area, and measured protein levels may be subject to various sources of bias. Epitope effects and cross-reactivities are two major concerns in pQTL studies. Epitope effects refer to condition when mutation in genetic variants will only change the assay binding epitope of protein, with no other function effects. In this situation, the protein abundance will be reported to be low not because of decreased in protein expression but because difference in binding capacities. Another concern is when the selectivity of affinity reagents is insufficient and reported protein abundance will be higher than true abundance here because of the non-specific bindings of other types of proteins. Thus, pQTL analysis should be followed up with validation studies[73].

I.2.3.2 Understanding genetic architectures

Genetic architecture refers to the number of causal variants underlying a trait, the magnitude of their effect sizes, and their distribution across the genome. Key parameters are the proportion of variance attributable to genetic factors (heritability), the proportion of variance associated with genotyped SNPs (SNP-based heritability) and generic correlations between traits. Estimation of these parameters helps make comparisons across traits and planning of trait-specific experimental designs and make predictions about future discoveries as sample sizes increase.

- *Estimation of SNP-based heritability*

Heritability refers to the proportion of phenotypic variance that can be explained by genetic variance. Genetic variance in broad sense heritability (H^2) refers to total genetic variance, which includes non-linear effects such as dominance effects. Narrow sense heritability (h^2) considers only the variance attributable to additive genetic effects. The SNP-based heritability (h^2_{SNP}) refers to the proportion of total phenotypic variance associated with measured common SNPs[74]. The h^2_{SNP} can be estimated using individual genotype data, or GWAS summary statistics.

GREML is a commonly used method for estimating SNP-based heritability using individual-level genotype data. The GREML model assumes the phenotypic variance can be decomposed as[75]:

$$\mathbf{V} = \mathbf{Var}(\mathbf{y}) = \mathbf{A}\sigma^2_g + \mathbf{I}\sigma^2_e$$

where $\mathbf{A} = \mathbf{ZZ}'/M$ is the GRM and $\mathbf{Z}$ is an *N*x*M* standardised genotype matrix (with mean 0 and variance 1). *N* and *M* denote the population size and the number of SNPs, respectively. The vector $\mathbf{y}$ is an *N*x1 phenotype vector for the population and is typically standardised for simplicity. By definition, $h^2_{SNP} = \sigma^2_g/\sigma^2_P$, where σ^2_P is the phenotypic variance in a population. When the phenotype is standardised, $\sigma^2_P = 1$, and it follows that $h^2_{SNP} = \sigma^2_g$.

In practice, it is infeasible to collect genotype and phenotype data for every individual in a population. Consequently, the variance components σ^2_g and σ^2_e must be estimated from the observed data. GREML estimates these parameters using a REML approach, which maximises the restricted Gaussian likelihood of the phenotype vector under the linear mixed model $\mathbf{y} \sim \boldsymbol{N}(\mathbf{X_F b_F}, \mathbf{V})$, while treating the GRM $\mathbf{A}$ as known.

Alternatively, SNP-based heritability can be estimated from GWAS summary statistics. LD score regression (LDSC) is widely used for this purpose because it is computationally efficient. The LDSC model is defined as follows[76]:

$$E[\chi_m^2|l_m] = \frac{Nh_{SNP}^2 l_m}{M} + Na + 1$$

where,

- N is the sample size
- M is the number of SNPs
- a is a measure of confounding bias including cryptic relatedness and population stratification
- l_m denotes the LD score of variant *m*, defined as the sum of squared LD correlations between variant *m* and all SNPs included in the GWAS summary statistics. For example, the LD correlation between SNP *m* and SNP *k* can be calculated as $\mathbf{Z_m}'\mathbf{Z_k}/N$, where $\mathbf{Z_m}$ and $\mathbf{Z_k}$ are *N*x1 standardised genotype vectors for SNP *m* and *k*, respectively. To reduce noise and improve computational efficiency, LD correlations between distant SNPs (defined based on a user-specified threshold, for instance, 1 Mb) are typically shrunk to 0.
- χ_m^2 represents the squared z-statistic for SNP *m* in the GWAS summary statistics

LDSC assumes a polygenic model. The method is termed LD score regression because each SNP in the GWAS summary statistics is treated as a data point, with the GWAS test statistic χ^2 as the response variable and the LD score as the predictor.

Performing a linear regression of χ^2 on LD scores yields a slope equal to $\frac{Nh_{\text{SNP}}^2}{M}$ and an intercept equal to $Na + 1$. From these quantities, one can estimate h_{SNP}^2 as well as quantify inflation in test statistics due to population stratification and cryptic relatedness.

In general, GREML is more accurate than LDSC in the estimation of heritability[77] because LDSC relies on more simplifying assumptions such as a polygenic architecture, and its LD scores are usually calculated from a reference sample. However, GREML is much more computationally intensive and requires individual-level genotype data.

For binary traits, h_{SNP}^2 and its standard error should be converted from the observed scale (h_{obs}^2, which is typically the output from most software) to the liability scale (h_l^2). The liability-threshold model assumes that an individual's risk of disease is determined by an unobserved continuous variable, termed liability, which captures

the combined influence of genetic and environmental factors. This liability is normally distributed in the population, and disease occurs only when an individual's liability exceeds a defined threshold (*t*). Under liability threshold model[78]:

$$h_l^2 = \frac{K(1-K)}{z^2}\frac{K(1-K)}{P(1-P)}h_{obs}^2$$

$$SE_{h_l^2} = \frac{K(1-K)}{z^2}\frac{K(1-K)}{P(1-P)}SE_{h_{obs}^2}$$

Here, *K* denotes the lifetime prevalence of the disease in the population, and *P* represents the proportion of cases in the sample; *z* is the height of the standard normal density function at the liability threshold *t*, where *t* corresponds to the quantile of the standard normal distribution associated with prevalence *K*.

One of the key questions in the study of SNP-based heritability is the "missing heritability" problem. Using human height as an example, pedigree-based studies have estimated heritability at approximately 80%, whereas heritability explained by genome-wide significant SNPs was only about 5% in early GWAS analyses conducted around 2010[75]. This discrepancy led to the long-standing question: *"Where is the missing heritability?"*

Subsequent analyses using the GREML approach showed that approximately 56% of the variance in height can be explained when considering all common SNPs simultaneously[17]. A significant proportion of the remaining missing heritabity can be attributed to rare variants[79,80], and recent WGS studies have shown that the gap between pedigree-based heritability and h_{SNP}^2 has become negligible for at least 15 quantitative traits. However, even with WGS data, the gap persists for some traits. For example, the estimated h_{SNP}^2 for height is approximately 71%, leaving a remaining gap of about 9%[81].

Several explanations have been proposed for the remaining discrepancy. These include non-additive genetic effects, structural variants, incomplete coverage of the reference genome, and potential overestimation of heritability in twin or pedigree studies[81,82]. Additionally, the inclusion of cohorts with binary phenotypes may affect the robustness of SNP-based heritability estimates[79].

- *Estimation of genetic correlation*

Genetic correlation between different traits is an important concept in quantitative genomics. The genetic correlation between trait *x* and trait *y* can be defined as[83]:

$$r_g = \frac{Cov_g(x,y)}{\sqrt{h_x^2 h_y^2}}$$

Here,

- $Cov_g(x,y)$ is the genetic covariance between traits *x* and *y*. Under an additive genetic model, it can be expressed as $\sum_{m=1}^{M} Var(\mathbf{X_m})\, b_{xm} b_{ym}$, where *M* is the total number of SNPs included in the GWAS, $\mathbf{X_m}$ is the genotype vector of SNP *m*, and b_{xm} and $\mathrm{b_{ym}}$ are the marginal effects of SNP *m* on traits *x* and *y*, respectively.
- h_x^2 and h_y^2 denote the SNP-based heritabilities of traits *x* and *y*, respectively

Similar to heritability, commonly used methods to estimate genetic correlation include bivariate GREML[84] and LDSC[76]. In GREML, r_g is estimated using a restricted maximum likelihood framework. In LDSC, the regression equation becomes:

$$E\left[z_{xm} z_{ym} \middle| l_m\right] = \frac{r_p N_s}{\sqrt{N_x N_y}} + \frac{\sqrt{N_x N_y}\, r_g \sqrt{h_x^2 h_y^2}}{M} l_m$$

Here, r_p denotes the phenotypic correlation between the two traits, and z_{xm} and z_{ym} are the GWAS z-statistics for SNP *m* in traits *x* and *y*. Similarly, each SNP is treated as a data point, with l_m as the predictor variable and $z_{xm} z_{ym}$ as the response variable in a linear regression. The slope of this regression can be used to estimate r_g, while the intercept can be used to estimate r_p.

Although LDSC tends to underestimate SNP-based heritability[85], estimates of genetic correlation are generally considered unbiased, as any systematic biases affect both the numerator and denominator proportionally and therefore cancel out[86].

It is notable that, due to the property of covariance, if there are positive genetic correlations in some genomic regions and negative genetic correlations in others, their effects across the genome may cancel out. As a result, the overall genetic correlation between two traits may be close to zero even when strong regional correlations exist. An example of this phenomenon is LDL and HDL where positive and negative regional associations have been found in the absence of global associations[87]. Several regional association methods have been developed to address this issue, such as rho-HESS[87] and LAVA[88].

I.2.3.3 Translation into clinical care

A fascinating aspect of GWAS analysis is its potential to translate into clinical care. Here several commonly used tools are presented, and they are vital in transformation of quantitative genomics studies into clinical areas in terms of disease prevention, diagnosis, and treatment.

- *Polygenic scores and out-of-sample predictions*

PGS is a summary of genetic predisposition/risks of an individual, by weighting effect sizes of SNPs and summing all weighted genetic effects. Polygenic scores should be calculated in a sample independent of the discovery cohort (the cohort used in generating genetic predictors) to avoid overfitting. Covariates such as age, sex, covariates relating to population structure, and important trait-specific confounding are commonly included in the model, and the testing hypothesis is[9]:

$$H_0: Phenotypes \sim Covariates + e$$

$$H_1: Phenotypes \sim PGS + Covariates + e$$

Methods such as ANOVA can be used to assess whether the alternative hypothesis where the PGS is included in the model is significantly different from the null hypothesis.

PGS methods Different methods use different strategies to decide which SNPs will be included in the model and what weight will be giving to that SNP.
The simplest way of calculating PGS is to use the pruning and thresholding (P+T) method, which was first demonstrated in a study of psychiatric disorders as a proof-of-principles for PGS[89]. SNPs are pruned to account for LD and filtered based on certain p-value thresholding (P_T). Selected SNPs will then be used to predict genetic effects in an independent cohort and predicted phenotypes will be compared with real phenotypes to test for associations.

Several other methods exist which better account for LD and include functional annotations, which improve the polygenic prediction accuracy. BayesR is a Bayesian mixture model based approach[90], which can be modelled as[91]:

$$p\left(\beta\middle|\pi,\sigma_g^2\right)=\sum_{k=1}^{K}\pi_k N(\beta|0,\sigma_k^2)$$

Here, the assumption is that the causal effect size β of SNPs come from a *K* components finite mixture.

- $\sigma_1^2 \ldots \sigma_k^2$ are the variance of components 1,…,K
- $\pi_1 \ldots \pi_k$ are the mixture proportion of components 1,…,K

- Default variances of $\sigma_1^2,\ldots,\sigma_k^2$ in the SBayesR GCTB software are 0, $10^{-2}\sigma_g^2$, $10^{-1}\sigma_g^2$, σ_g^2, however for traits such as height, different parameters (e.g. 0, $10^{-4}\sigma_g^2$, $10^{-3}\sigma_g^2$, σ_g^2) will provide better prediction accuracies accounting for the extreme polygenic nature of the traits[92].
- Unknown parameters π, β, σ_e^2, and σ_e^2 are estimated using a Markov chain Monte Carlo (MCMC) approach by sampling from the conditional posterior distribution of each parameter and taking the mean after convergence has been reached.

BayesR can be extended to SBayesR[93] where summary statistics are used together with a LD reference to model the correlation between SNPs. Functional information can be integrated in BayesR model as well, which motivated the development of BayesRC[94] and SBayesRC[95]. Other methods that use different strategies to account for LD are also available such as LDPred2[96].

Prediction accuracies The prediction accuracy of PGS is commonly tested in a cohort independent of the discovery cohort. The predicted phenotype will be compared with the real phenotype and the correlation between them evaluated.
For continuous traits, prediction accuracies can be simply measured by the coefficient of determination (R^2) of linear regression. For binary traits, pseudo-R^2 are commonly used but it should be converted to the liability scale to enable comparison of pseudo-R^2 between different studies (under the liability thresholding model). The formula of conversion is given by[97]:

$$R_{l_{cc}}^2 = \sigma_g^2 = \frac{R_{o_{cc}}^2 C}{1 + R_{o_{cc}}^2 \theta C},$$

$$\text{where } C = \frac{K(1-K)}{\mathcal{Z}^2}\frac{K(1-K)}{P(1-P)} \text{ and } \theta = m\frac{P-K}{1-K}\left(m\frac{P-K}{1-K} - t\right).$$

Here,

- $R_{l_{cc}}^2$ is the R^2 under the liability scale. $R_{o_{cc}}^2$ is the R^2 under the observed scale, which is obtained through linear regression
- K is the lifetime disease risk in the population, and P is the proportion of cases in the case control study
- t is the threshold of the normal distribution curve when the truncated area in the upper tail equals to K, $\mathcal{Z}$ is the height at point t, and $m = \mathcal{Z}/K$

In clinical settings, another commonly used metric to evaluate the performance of risk predictors (such as clinical scores, or PGS in our case) is the area under the

receiver operating characteristic curve (AUC)[98]. The AUC represents the probability that a randomly selected case has a higher predicted risk (i.e., a higher PGS in our context) than a randomly selected control. An AUC of 0.5 indicates no discrimination between cases and controls, whereas larger AUC values indicate better discrimination; an AUC of 1 corresponds to perfect separation. However, even if the out-of-sample prediction R^2 reaches 1, the maximum AUC may still be less than 1, depending on aspects of the underlying genetic epidemiology such as heritability and disease prevalence. Therefore, it is important to establish a relationship between AUC and prediction R^2, which allows us to assess whether a PGS has sufficient discriminative ability for clinical applications (i.e., a sufficiently high AUC), and conversely to infer prediction R^2 from a given AUC. On the liability scale, such a relationship is given by[97]:

$$R^2_{AUC} = \frac{2Q^2}{(m_2 - m)^2 + Q^2 m(m - t) + m_2(m_2 - t)}$$

where $Q = \Phi^{-1}(AUC)$, and $m_2 = -m\,K/(1 - K)$.

Alternatively, another useful classification tool is the net reclassification index. It measures whether individuals are reclassified into more appropriate clinical risk categories when a new predictor (PGS in our case) is added to an existing risk prediction model. Because it assesses how individuals shift across clinically defined risk thresholds, it is often viewed as more directly relevant to clinical decision-making than discrimination metrics (such as AUC) alone. However, caution is advised when three or more risk categories are used, as this may inflate the test statistic. Therefore, risk categories should be chosen based on clinically meaningful thresholds[9,99].

- *Mendelian randomisation*

Mendelian randomisation (MR) uses natural variation of genetic variants, which are fixed at conception, as instrument variables (IV) to infer causality between an exposure and outcomes of interests. It has gained popularity over past few years because it is able to address key challenges in observational studies, including unmeasured confounding and reverse causality[100]. Large online databases and MR tools are available, including MR-Base[101]. Mendelian randomisation is a powerful tool that can not only inform the causality between traits, but also be extended into other areas such as drug repurposing. However, in order for the Mendelian randomisation studies to be valid, several assumptions need to be met, along with proper MR model selection. In this section, basic ideas behind MR, the commonly used MR

design, assumptions in MR, tests for these assumptions, and applications of MR will be reviewed.

Basic ideas behind MR Traditional observational studies are limited by unmeasured confounding and reverse causalities to inform causal relationships between the exposure and the outcome. For example, in observational studies increased high-density lipoprotein (HDL) levels are shown to be related to reduce the risk of coronary artery diseases (CAD)[102]. A key question is whether this relationship is causal, because if so, we can reduce the CAD risk by developing medications that increase HDL level. However, it is difficult to determine if the relationship is causal based on observational studies, because the association between HDL and CAD can be mediated by unmeasured confounding factors, which could increase HDL level and reduce CAD risks at the same time[103]. In addition, reverse causality can also happen in observational studies. For example, one could argue that associations between HDL and CAD is because reduced CAD risk will cause an increase in HDL. The gold standard for testing causal relationships is a randomised controlled trial (RCT), in which participants with similar baseline characteristics are randomly assigned to either a treatment group or a control group, and outcomes are subsequently compared between the two groups. Well-designed RCTs are informative but are time and resource intensive. Additionally, RCTs can be unethical in certain situations, especially when exposures cause harm. The basic idea of MR is to use genetic variants as an instrumental variable (IV) to represent the exposure. As a result of Mendel's first and second law, in MR, nature randomly assign participants into the two groups in RCT. As the genotype is randomly assigned and fixed at conception, unmeasured confounding and reverse causality is less of an issue in MR design[100].

Design of MR Extended from the original one-sample MR model, several study designs have been developed and gained popularity because the simplicity of the design.

- One sample and two sample MR

One-sample MR represents the original MR design (**Figure I.2**), in which genetic variants are used as instrumental variables to proxy an exposure, with both the exposure and the outcome measured in the same sample. If genetic variants associated with an exposure (for example, variants identified in GWAS as influencing smoking behaviour) are also associated with an outcome of interest (such as lung

cancer), this provides evidence consistent with a causal effect of the exposure (e.g., smoking) on the outcome (e.g., lung cancer risk).

The rationale is that genetic variants are randomly allocated at conception. As a result, individuals carrying genetic variants that increase the likelihood of a particular exposure (e.g., smoking) can be viewed as analogous to an "exposed" group, while those without such variants form a control group. Differences in the outcome (e.g., lung cancer incidence) between these groups can therefore be examined, mimicking the random assignment of treatments in a RCT.

The design of two-sample MR is similar to that of one-sample MR, with the exception that the measurements of exposures and outcomes are obtained from different samples. The advantage of two sample MR is that it allows the use of GWAS summary statistics from different studies which are time and cost efficient, and statistical power in two sample MR is higher. It is notable that two sample MR

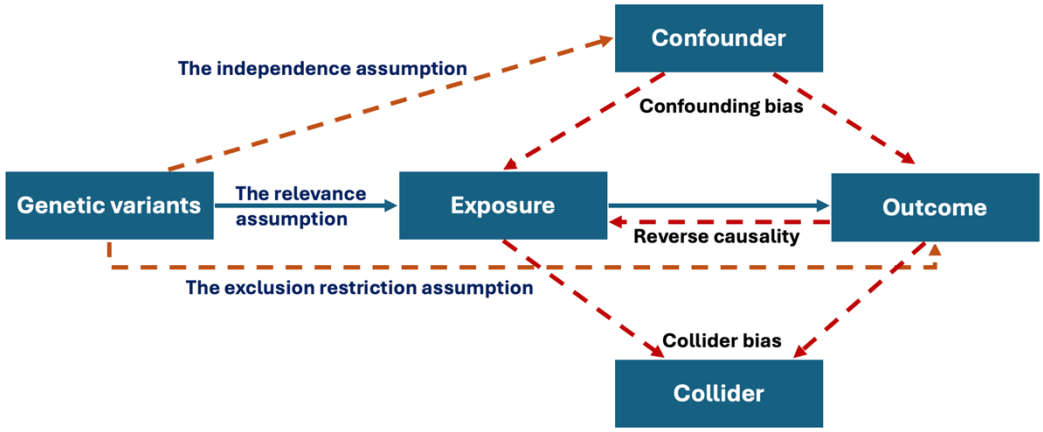


Figure I.2. Conceptual diagram of potential sources of bias in observational studies and the assumptions underlying Mendelian randomisation. Arrows in the figure indicate the direction of causality (e.g., A → B means A causally affects B), and solid arrows represent the hypothesised causal pathway of interest in MR. In observational studies, confounding bias occurs when a confounder causes changes in both exposure and outcome, and reverse causality occurs when the outcome causes changes in the exposure. Both can generate significant associations between exposure and outcome even when the exposure is not causal for the outcome. Collider bias occurs when both exposure and outcome cause changes in a collider; the existence of a collider alone does not induce spurious association, but conditioning on the collider can generate such associations. In MR, genetic variants serve as instrumental variables for the exposure and must be associated with the exposure (the relevance assumption). The orange dashed lines represent violations of the independence assumption (genetic variants associated with confounders) or the exclusion restriction assumption (genetic variants affecting the outcome through pathways other than the exposure), which would invalidate MR inference.

assumes the underlying populations are same and overlapping samples will bias the result towards the existence of exposure-outcome associations[104].

- Bidirectional MR

Extending from two sample MR where GWAS summary statistics are used to identify variants related to exposures and outcomes, bidirectional MR tests the causal relationship in both directions. Using the example above, bidirectional MR will test both the HDL as exposure, CAD as the outcome model, and the CAD as the exposure, HDL as the outcome model.

- Two-step MR, multivariable MR and factorial MR

Several other MR designs have been developed in an effort to better model the properties of MR under different circumstances (causal and confounding relationships). Two-step MR focuses on determining whether there is an intermediate trait that causally mediates the effect between exposures and outcomes. Factorial MR measures the combined causal effects of two or more risk factors, which require individual level data[100]. Multivariable MR accounts for the horizontal pleiotropic effects where a single genetic variant can be significantly associated with multiple traits. Multivariable MR considers multiple exposures jointly, and integrated multiple exposures in a single model to estimate individual causal effect of those exposures. An example of a multivariable MR analysis is the investigation of the causal relationship between HDL and CAD, where multivariable MR showed no causal relationship between HDL and CAD, despite univariable MR showed the opposite. In this case, RCT result is more consistent with multivariable MR[100,105].

Assumptions in MR Three important assumptions in MR are the relevance assumption, the independence assumption, and exclusion restrictions assumption (**Figure I.2**). All three assumptions need to be met in order the MR results to be valid[104].

- The relevance assumption

The association between genetic variants used as instrumental variables and the exposure needs to be a true association. Otherwise results will be biased toward rejecting null hypothesis in one sample MR and towards the null in two sample MR[100]. In the single sample MR, the rule of thumb is that F-statistics should be larger than 10 in order for the variant to be a valid IV[104]. In two sample settings, variants that can be used as IV are those reach the genome-wide significance threshold in

GWAS and replicable in independent samples. Research found combining multiple SNPs into an aggregate score can be a more powerful IV than using a single SNP[106].

- The independence assumption

Genetic variants used as IV should not be associated with confounders of risk factor-outcome associations. It is not possible to prove this assumption, but it can be checked by scanning the relationship between genetic variants and known confounders[100].

- The exclusion restrictions assumption

The only association between genetic variants used as IV and the outcome should be mediated through the exposure. Again, this is hard to prove, but several methods exist to detect the existence of such associations, and attempt to derive the causal relationship by adjusting for these associations[100].

Other considerations in MR Pleiotropic effects, where a single genetic variant is associated with multiple traits, are common in real settings and may invalidate MR if such pleiotropy violates the exclusion restriction assumption described above. Several methods have been developed to account for the pleiotropic effects. The median estimators method, combining data from multiple genetic variants into a single estimate with weighted median estimators, will provide consistent results when up to half of IVs are invalid[107]. MR Egger regression, which adopts a tool used in estimating the small sample bias of the meta-analysis, can detect pleiotropic effects and provide a causal effect estimate that are robust to the violation of IV assumptions[108]. Causal estimates resulting from MR Egger analysis are less precise likely due to a power penalty[104]. Other tools are also available for the detection of pleiotropic effects, such as the HEIDI test used with summary based Mendelian randomisation (SMR)[109] and MR-PRESSO[110]. Analysis of the pleiotropic effects should involve multiple methods outlined above and results should be compared. It is important to note that the original MR framework was designed to exploit Mendelian segregation within families. Mendelian segregation does not strictly operate at the population level. However, to overcome the lack of within-family data, MR methods based on population-level data have been developed to approximate the causal inference attainable in within-family designs.

Applications of MR Besides informing causality described above, an exciting application of MR is that it can be used in a cost-effective way to evaluate potential drug-disease relationships. An exemplary way of using MR in drug target prediction is

to use eQTL and pQTL as the exposure, and disease status as outcome. A causal relationship between certain gene/protein expressions and disease status, informed by MR, will allow us to find potential drug target to treat disease. For example, if increased gene expression is found to be causally linked to a disease, a drug that found to decrease the expression of such gene can potentially be the cure for the disease. Several caveats exist though in the drug development using MR, which is reviewed elsewhere[111,112]. Information on the gene target of medications can be found on large databases such as DrugBank[113] and chEMBL[114].

I.3 Expanding genomic association analyses

Conventional GWAS largely rely on SNP array technology, which focus on individual SNPs and small-scale genomic changes such as single-nucleotide substitutions or short insertions and deletions. However, the genome can also undergo larger-scale structural changes. These structural variations range from changes of a few base pairs, such as microsatellites, to sub-kilobase events including inversions, deletions, and copy number variants, and extend to larger alterations such as chromosomal inversions or whole-chromosome deletions or duplications[115]. With the increasing availability of whole-genome sequencing data—particularly from long-read sequencing platforms—the ability to detect structural variants has improved substantially.

I.3.1 Sequencing technique and genome assembly

Sequencing reads generated by modern platforms are essentially strings of nucleotides. To reconstruct the complete genome of a sample, these reads must be assembled, much like piecing together a puzzle, in a process known as *de novo* assembly. However, this approach is costly and time-consuming. Alternatively, because most human sequences are conserved across individuals, a reference genome can be established, onto which sequencing reads can then be mapped. The Human Genome Project (HGP) marked the first successful effort to build a complete human reference genome. This endeavour spanned 13 years, involved international collaboration, and relied on the accurate but time-intensive Sanger sequencing technique[116].

Since the HGP, sequencing technologies have advanced substantially. A major breakthrough was the development of high-throughput sequencing, exemplified by platforms such as the Illumina NovaSeq 6000. This short-read sequencing technology has enabled whole-genome sequencing at scale, facilitating the sequencing of nearly

one million individuals worldwide[117,118]. However, short-read sequencing, typically producing reads of ~150 bp, has limited ability to detect larger structural variants. This limitation motivated the development of long-read sequencing technologies, with Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT) as leading representatives[119]. Studies have shown that nearly 50% of deletions and 80% of insertions are missed by Illumina short-read sequencing, particularly variants in the 50 bp to 2 kb range[119,120]. By contrast, PacBio HiFi sequencing achieves >99% accuracy with read lengths exceeding 10 kb, while ONT can generate ultra-long reads of up to 2.3 Mb, albeit at slightly lower accuracies (87–98%)[119]. These ultra-long reads are particularly valuable for resolving highly repetitive genomic regions, such as centromeres and segmental duplications.

By leveraging these technological advances, researchers have recently been able to generate complete telomere-to-telomere (T2T) reference genomes, such as the CHM13 assembly[121].

I.3.2 Variant calling

Raw WGS data must undergo two major stages, including mapping and alignment followed by variant calling, before it can be used in quantitative genomic analyses. During the mapping and alignment stage, raw reads from the sequencing platform are first trimmed to remove adaptor sequences and subjected to quality control to exclude low-quality reads[122]. The processed reads are then aligned to a reference genome using software such as *bwa-mem*[123], *bowtie2*[124], or *minimap2*[125] (the latter being particularly suitable for long-read sequencing). Reads can map to the reference genome with varying confidence; for example, repetitive sequences are widespread in the genome, and a read may map to multiple locations, leading to lower mapping quality scores. Quality control summaries for mapped data can be generated using tools such as *Picard CollectQualityYieldMetrics* [126] or *samtools flagstat*[127], which provide metrics useful for filtering in downstream analyses. The most commonly used integrated pipelines that incorporate these steps are the *GATK-based pipeline* and the *DRAGEN-based pipeline*[128].

Both the *GATK* and *DRAGEN* pipelines also provide toolkits for variant calling. For small insertions or deletions (indels), commonly used tools include *GATK HaplotypeCaller*, *DRAGEN*, and deep-learning approaches such as *DeepVariant* [129]. A key advantage of WGS-based variant calling over imputation lies in its ability to detect rare variants. For larger structural variants (SVs), the *GATK-SV* pipeline integrates multiple callers, including *GATK gCNV* [130], *MELT* [131], and *Manta*[132].

Similarly, *Illumina* developed *DRAGEN SV*[128], which builds upon the same logic as Manta but incorporates algorithmic improvements, such as refined parameter tuning, to achieve higher accuracy. Optimised for the Illumina platform, the *DRAGEN* pipeline demonstrates superior precision and recall compared with the *GATK* pipeline[122]. This advantage motivated its adoption for processing the UK Biobank's WGS data of nearly half a million individuals[117].

Beyond these general pipelines, specialised tools and algorithms have been developed for specific tasks. For example, in detecting rare CNVs in protein-coding regions, haplotype-informed approaches have been applied under the assumption that comparing individuals with shared haplotype backgrounds improves CNV detection compared with treating all samples as independent[133]. Another example comes from mitochondrial genomics: in estimating mitochondrial copy number or identifying somatic mutations (i.e. heteroplasmy), researchers have reconstructed individual-specific consensus sequences rather than relying solely on the reference genome. This approach has been shown to improve mitochondrial DNA coverage, particularly in African haplogroups[134].

I.3.3 Limitation of current approaches

Accurate variant detection depends critically on the quality of the reference genome. However, even the most recent versions of the reference genome remain imperfect. In WGS, a contig represents a continuous DNA sequence without gaps. In an idealised scenario, a complete reference genome for a homogeneous population would contain only 25 contigs (one for each of the 22 autosomes, plus the two sex chromosomes and mitochondrial DNA). By contrast, GRCh38 comprises 998 contigs across the 24 chromosomes[119], including regions that were unsequenceable, such as highly repetitive sequences near telomeres, as well as sequences whose precise chromosomal locations remain unknown.

Long-read sequencing has substantially improved contiguity by reducing the number of contigs and eliminating gaps. The CHM13 human genome assembly provides a gapless reference genome; however, it represents only a single complete human haplotype and therefore does not adequately capture human genomic diversity[121]. The Human Pangenome Reference Consortium (HPRC) has addressed this limitation by developing a graph-based genome, which allows multiple haplotypes to be represented simultaneously[135]. While this is a significant advance, widespread adoption remains limited: long-read sequencing, which underpins pangenome

generation, remains relatively expensive, and even the HPRC samples may not fully represent all human haplotypes.

As short-read sequencing continues to dominate due to cost and is largely mapped to GRCh38[136], it may be some time before long-read sequencing reaches biobank-scale and reference genomes fully capture human haplotype diversity. Consequently, current analyses must be conducted with careful consideration of these limitations, recognising the constraints imposed by available reference genomes and sequencing technologies.

I.4 Thesis overview

In the previous sections, the state-of-the-art in quantitative genomics analyses were reviewed. Interestingly, the implementation of these analyses has raised some unexpected questions. During the meta-analysis of GWAS summary statistics within the PGC for major depressive disorder, it was observed that SNP-based heritability tends to decline as the number of cohorts included in the meta-analysis increases, and that out-of-sample prediction in certain cohorts can exceed the SNP-based heritability—which is the theoretical upper bound. Chapter II addresses these observations by deriving the statistical theory underlying these phenomena, demonstrating that they can be explained by heterogeneity in study design and/or differences in genetic ancestry.

Importantly, such heterogeneity should not be regarded solely as a challenge, but rather as a factor that must be appropriately considered, and one that can provide unique opportunities for advancing the understanding of human genetic architecture. Building on this insight, Chapter III introduces a novel tool for integrating information from multi-ancestry GWAS summary statistics. This approach shows an improved ability to detect independent associations compared with single-ancestry analyses of equivalent sample size.

In addition to analysing genetic sequence itself, information generated during the sequencing process can also provide valuable insights relevant to human disease. Chapter IV explores this concept, focusing on the potential utility of by-products of genetic sequencing, such as mitochondrial copy number, and evaluates their relevance to human diseases including mood disorders.

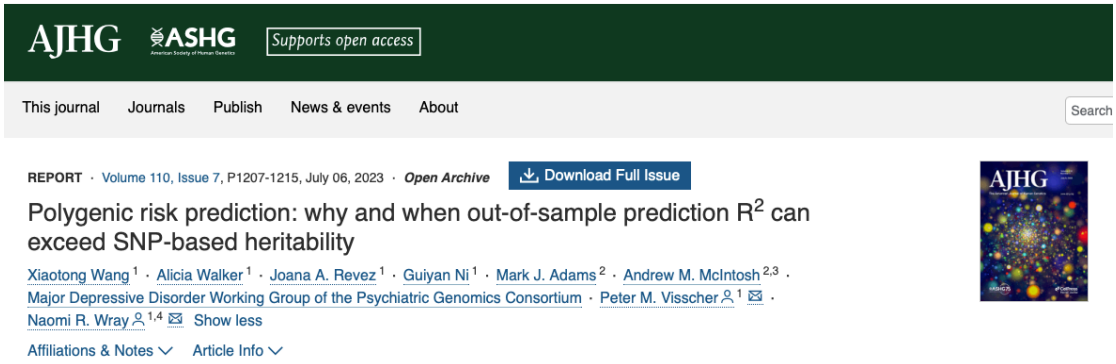
Although many of the motivating observations and analytic needs originated in the study of psychiatric disorders, the theoretical principles and methodological innovations described in this thesis are broadly applicable across the spectrum of human disease. In light of these findings, Chapter V discusses the implications and limitations of the results within the wider context of quantitative genomics, reflecting on their contribution to the broader understanding of genetic architecture and outlining several promising directions for future research.

II

Polygenic Risk Prediction: why and when out-of-sample prediction R^2 can exceed SNP-based heritability

This chapter is adapted from the following published paper, with parts of the analysis updated to reflect the availability of a new data release from the PGC MDD Working Group:

Wang, X., Walker, A., Revez, J. A., Ni, G., Adams, M. J., McIntosh, A. M., ... & Wray, N. R. (2023). Polygenic risk prediction: why and when out-of-sample prediction R2 can exceed SNP-based heritability. *The American Journal of Human Genetics*, *110*(7), 1207-1215.



"A theory is just a model of the universe, or a restricted part of it, and a set of rules that relate quantities in the model to observations. " - Stephen Hawking

II.1 Introduction

Complex human traits or diseases are polygenic[137]. Trait-associated alleles can be identified in genome-wide association studies (GWAS). Polygenic scores (PGS), estimates of the genetic contribution to a trait or disease liability for individuals, are calculated as an aggregate score of associated variants (with weights derived from GWAS results). The coefficient of determination (R^2) is a key statistic to evaluate the efficacy of PGS. R^2 is the proportion of phenotypic variance explained by the PGS in a "target" cohort independent of the GWAS used to identify risk alleles and estimate their effect sizes. By definition, the SNP-based heritability (h^2_{SNP}, the proportion of total phenotypic variance attributable to all common SNPs[138]) is the upper limit of the out-of-sample prediction R^2. The difference between h^2_{SNP} and R^2 is attributed to measurement errors of SNP effects which decrease as sample sizes increase[139]. h^2_{SNP} can be estimated from individual-level genotype data using methods such as GREML[140] implemented in software such as GCTA[141]. Increased power for GWAS is achieved through meta-analysis of GWAS summary statistics from multiple cohorts, and methods to estimate h^2_{SNP} from summary statistics are available. LD Score Regression (LDSC)[142] is an example of such method, commonly used in the field in large part due to its computational efficiency[85].

In practice, a decrease in estimates of h^2_{SNP} is often noted as the number of cohorts included in the GWAS meta-analysis increases, until the estimate reaches a plateau. For example, in the GWAS of major depression (MD), the h^2_{SNP} of a cohort with ~18,000 samples (9,041 cases and 9,381 controls) was 0.21 (standard error (s.e.) 0.021)[143], while in a subsequent GWAS meta-analysis of more than half million samples (135,458 cases and 344,901 controls), the h^2_{SNP} estimate declined to 0.087 (s.e., 0.004)[144]. Similar trends were also observed in educational attainment[145] and Alzheimer's disease[146]. At the same time, as opposed to the standard narrative, out-of-sample prediction R^2 can sometimes approximate or even exceed the h^2_{SNP} estimated from the GWAS meta-analysis. For example, in studies of educational attainment, h^2_{SNP} of years of education was 0.122 (s.e. 0.003), but the out-of-sample prediction R^2 in the National Longitudinal Study of Adolescent to Adult Health (Add

Health) cohort was 0.158, with the lower limit of 95% confidence interval (C.I.) of 0.143[145,147].

The aim of our study is to provide the theory to explain these observations. Previously, de Vlaming *et al*.[148] demonstrated that heterogeneity in genetic effects across cohorts attenuates the statistical power of GWAS, i.e., the empirical power from a GWAS meta-analysis is less than the power from a single-cohort GWAS of the same sample size. Their conclusions thus focused on the reduced performance of PGS from meta-analysis compared to expectation from a sample of equal size constructed under idealised conditions of equal h^2_{SNP} and genetic correlations (r_g) between cohorts of 1 (both conditions are expected if all cohorts are random samples of the same population and with phenotype measured in the same way). Using similar principles, we derive an equation for the expected value of the SNP-based heritability of the meta-analysis GWAS (h^2_{ma}) as a function of cohort-specific SNP-based heritabilities and between-cohort genetic correlations and show this explains the observed decrease in h^2_{ma} estimates as the number of cohorts in a GWAS meta-analysis increases. We show the validity of our derivations using simulation and empirical data. Building on the work from de Vlaming *et al.*[148], Dudbridge[149] and Daetwyler *et al.*[150], we derive theoretical conditions when out-of-sample prediction R^2 can exceed h^2_{ma} (it will only occur when the SNP based heritability of the target sample is greater than h^2_{ma}), and test our theory with major depression and educational attainment data sets (**Figure II.1**).

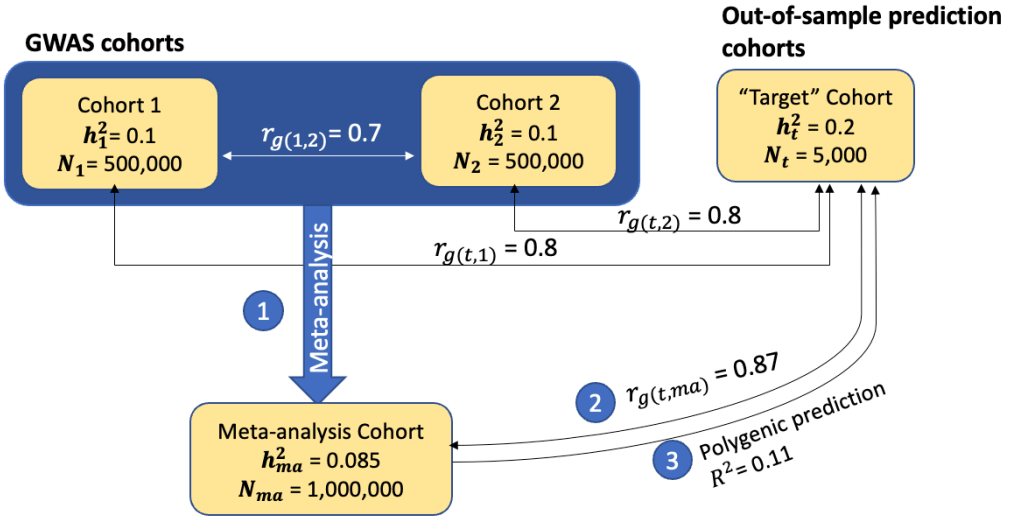


Figure II.1 A schematic illustration of the study in a simplified scenario. In this made-up scenario, there are two large GWAS cohorts, each with a sample size of 500,000 and SNP-based heritability of 0.1. After meta-analysis, the results are used to generate genetic predictors in an independent cohort, the "target" cohort. (1) The h_{ma}^2 can be expressed as a function of per-cohort h_{SNP}^2, the between-cohorts genetic correlations, and meta-analysis weights (equation II.4) or directly estimated from the meta-analysis summary statistics (these two estimates should be consistent). (2) Similarly, the genetic correlation between the meta-analysed cohort and the target sample ($r_{g(ma,t)}$) can be calculated from equation II.5 (which should be consistent with that estimated from summary statistics of the "target" and the meta-analysis cohort using LD Score Regression). (3) From equation II.6, $h_t^2 r_{g(ma,t)}^2 > h_{ma}^2 + \frac{M_e}{N_T}$, the out-of-sample prediction R^2 (0.11) is greater than the SNP-based heritability in the meta-analysis cohort (0.085) that is used to generate genetic predictors.

II.2 Method

II.2.1 Derivation of SNP-based heritability (h^2_{SNP}) in the meta-analysis cohort and conditions when out-of-sample prediction R^2 will exceed h^2_{SNP}

II.2.1.1 Assumptions and notation

We assume, without loss of generality, that phenotypes and genotypes of SNPs are centred and standardised to have unit variance, and that the GWAS data represent M_T markers in total (which is a function of the minor allele frequency threshold used in the GWAS). If there is linkage disequilibrium (LD) correlations between the markers, then we define the effective number of markers (M_e) as[151]:

$$M_e = \frac{M_T^2}{\sum_{k=1}^{M_T}\sum_{j=1}^{M_T} r_{jk}^2}$$

Here, r_{jk}^2 is a standard measurement of the LD between the SNP *j* and SNP *k* in the study[152].

Suppose there are *C* cohorts in a GWAS meta-analysis (subscript *ma*) and a single target cohort (subscript *t*) for out-of-sample prediction. We define the SNP-based heritability (h_{SNP}^2) as the proportion of total phenotypic variance captured by the M_e SNPs. To make the notation simpler, we denote h_{SNP}^2 of the *i*-th cohort in the meta-analysis as h_i^2 (which refers to the true SNP-based heritability of the base population from which cohort i is sampled) and h_{SNP}^2 of the target cohort as h_t^2.

Let β_{ik} be random effects representing the true effect size of SNP *k* in cohort *i*, with $\hat{\beta}_{ik}$ the estimated effect size of SNP *k* in cohort *i*. If the SNP effects are estimated with ordinary least squares,

$$\hat{\beta}_{ik} = \beta_{ik} + e_{ik};$$

$$Var(\hat{\beta}_{ik}) \approx Var(\beta_{ik}) + \frac{1}{N_i}$$

Assume, without loss of generality, that $\beta_{ik} \sim N\left(0, \frac{h_i^2}{M_e}\right)$.[153,154] It follows that

$Cov(\beta_{ik}, \beta_{jk}) = \frac{r_{g(i,j)} h_i h_j}{M_e}$, with $r_{g(i,j)}$ the genetic correlation between the populations from which cohorts *i* and *j* were sampled.

As noted previously, our derivations build on the work of de Vlaming *et al.*[148], Daetwyler *et al.* [150] *and* Dudbridge [149]. de Vlaming *et al.* used simulation to show that violations of the underlying model assumptions have little impact on final results. We use the same model assumptions and similarly the impact of violating assumptions is expected to be negligible.

II.2.1.2 Meta-analysis and the SNP-based heritability from meta-analysis GWAS

In a meta-analysis, the estimated effect size of SNP *k* is

$$\hat{\beta}_{ma,k} = \sum_{i=1}^{C} w_i \hat{\beta}_{ik}$$

where w_i is the weight applied to the *i*-th cohort with expected value (here, $\sum_{i=1}^{C} w_i = 1$)

$$E(\hat{\beta}_{ma,k}) = \sum_{i=1}^{C} w_i \, \beta_{ik}$$

and variance

$$Var(\hat{\beta}_{ma,k}) = Var(\sum_{i=1}^{C} w_i \, \hat{\beta}_{ik})$$

$$= \sum_{i=1}^{C} w_i^2 Var(\hat{\beta}_{ik}) + \sum\nolimits_{i=1}^{C} \sum\nolimits_{j=1}^{C} w_i w_j Cov(\hat{\beta}_{ik}, \hat{\beta}_{jk})$$

$$= \sum\nolimits_{i=1}^{C} \frac{w_i^2}{N_i} + \frac{\sum_{i=1}^{C} \sum_{j=1}^{C} w_i w_j r_{g(i,j)} h_i h_j}{M_e}$$

where the first term of $Var(\hat{\beta}_{ma,k})$ is the variance due to sample size and the second term is the (co)variance between true effects. N_T is the total sample size in the meta-analysis (Note: if all sample sizes are equal to N_T/C, all heritabilities equal h^2_{SNP} and all genetic correlations are 1, then $Var(\hat{\beta}_{ma,k}) = 1/N_T + h^2_{SNP}/M_e$, as expected).

The SNP-based heritability in the meta-analysis cohort is,

$$h^2_{ma} = M_e \, Var(E(\hat{\beta}_{ma,k})) = Var(\sum\nolimits_{i=1}^{C} w_i \, \beta_{ik}) = \sum\nolimits_{i=1}^{C} \sum\nolimits_{j=1}^{C} w_i \, w_j r_{g(i,j)} h_i h_j \quad \text{(II.1)}$$

For the purpose of our study we have defined h^2_{ma} as a population parameter. However, we note that its definition depends on the specific cohorts and their sample sizes that contributed to the GWAS meta-analysis.

The covariance between the meta-analysis estimate and the true value in a cohort (*t*) that was not part of the meta-analysis is,

$$\begin{aligned} Cov(\hat{\beta}_{ma,k}, \beta_{tk}) &= Cov(\beta_{ma,k}, \beta_{tk}) \\ &= Cov(\sum\nolimits_{i=1}^{C} w_i \, \beta_{ik}, \beta_{tk}) \\ &= \frac{\sum_{i=1}^{C} w_i r_{g(i,t)} h_i h_t}{M_e} \end{aligned}$$

The genetic correlation between the meta-analysis cohort and the target cohort is,

$$r_{g(ma,t)} = M_e \frac{cov(\beta_{ma,k}, \beta_{t,k})}{h_t h_{ma}} = \frac{\sum_{i=1}^{C} w_i r_{g(i,t)} h_i}{\sqrt{\sum_{i=1}^{C} \sum_{j=1}^{C} w_i w_j r_{g(i,j)} h_i h_j}} \quad \text{(II.2)}$$

II.2.1.3 R^2 from out-of-sample prediction

For individual *u* in the out-of-sample "target" cohort *t*, their predicted phenotype, i.e. their polygenic score, is

$$\hat{y}_{u,t} = \sum_{k=1}^{M_e} X_{u,k}\,\hat{\beta}_{ma,k}$$

With $\mathrm{Var}(\hat{y}_{u,t}) = M_e \mathrm{Var}(\hat{\beta}_{ma,k}) = M_e \sum_{i=1}^{C} \frac{w_i^2}{N_i} + \sum_{i=1}^{C}\sum_{j=1}^{C} w_i\, w_j r_{g(i,j)} h_i h_j$

$Var(y_{u,t}) = 1$, as we assumed a standarised phenotype, and
$Cov(\hat{y}_{u,t}, y_{u,t}) = \sum_{i=1}^{C} w_i r_{g(i,t)} h_i h_t = h_t \sum_{i=1}^{C} w_i r_{g(i,t)} h_i$

Hence,

$$R^2 = \left(\sum_{i=1}^{C} w_i r_{g(i,t)} h_i h_t\right)^2 / \left[M_e \sum_{i=1}^{C} \frac{w_i^2}{N_i} + \sum_{i=1}^{C}\sum_{j=1}^{C} w_i\, w_j r_{g(i,j)} h_i h_j\right]$$

$$= h_t^2 \left(\sum_{i=1}^{C} w_i r_{g(i,t)} h_i\right)^2 / \left[M_e \sum_{i=1}^{C} \frac{w_i^2}{N_i} + \sum_{i=1}^{C}\sum_{j=1}^{C} w_i\, w_j r_{g(i,j)} h_i h_j\right]$$

$$= h_t^2 r_{g(ma,t)}^2 h_{ma}^2 / \left[M_e \sum_{i=1}^{C} \frac{w_i^2}{N_i} + \sum_{i=1}^{C}\sum_{j=1}^{C} w_i\, w_j r_{g(i,j)} h_i h_j\right]$$

(Note, under the same equality assumptions in terms of sample size, heritabilities and genetic correlations as before, $R^2 = h_{SNP}^4/(M_e/N_T + h_{SNP}^4)$, which is identical to the standard expression from e.g., Daetwyler and others[150]).

II.2.1.4 Condition when out-of-sample prediction R^2 will exceed the h^2_{SNP}

An observation that $R^2 > h_{ma}^2$ occurs when :

$$h_t^2 r_{g(ma,t)}^2 > \sum_{i=1}^{C}\sum_{j=1}^{C} w_i\, w_j r_{g(i,j)} h_i h_j + M_e \sum_{i=1}^{C} \frac{w_i^2}{N_i} = h_{ma}^2 + M_e \sum_{i=1}^{C} \frac{w_i^2}{N_i} \quad \text{(II.3)}$$

Note, when $h_t^2 = h_{ma}^2$, and $r_{g(ma,t)}^2 = 1$, the inequality will never hold because $M_e \sum_{i=1}^{C} \frac{w_i^2}{N_i} > 0$. It indicates that the SNP-based heritability is the theoretical upper limit of the out-of-sample prediction R^2 under idealised conditions.

II.2.1.5 The fixed-effect inverse variance meta-analysis model

The fixed-effect inverse-variance meta-analysis is commonly used in the meta-analysis of GWAS from different cohorts. The approach assumes that each cohort is a random draw from the same underlying population, so that SNP effects and the

number of effective SNPs M_e are the same. Therefore, SNP-based heritabilities in each cohort are assumed to be the same and equal to the SNP-based covariances $Cov(\beta_i, \beta_j)$ and hence the genetic correlations between cohorts are assumed to be 1. The weight of the i-th cohort in this setting is given by:

$$w_i = \frac{\frac{1}{SE_i^2}}{\sum \frac{1}{SE_i^2}} \approx \frac{N_i}{N_T}$$

Substituting this into the equations above gives

$$h^2_{ma} = \frac{1}{N_T^2} \sum_{i=1}^{C} \sum_{j=1}^{C} N_i \, N_j r_{g(i,j)} h_i h_j \qquad \text{(II.4)}$$

$$r_{g(ma,t)} = \sum_{i=1}^{C} N_i \, r_{g(i,t)} h_i \, / \sqrt{\sum_{i=1}^{C} \sum_{j=1}^{C} N_i \, N_j r_{g(i,j)} h_i h_j} \qquad \text{(II.5)}$$

The condition when out-of-sample prediction R^2 will exceed the SNP-based heritability become:

$$h_t^2 r^2_{g(ma,t)} > \frac{1}{N_T^2} \sum_{i=1}^{C} \sum_{j=1}^{C} N_i \, N_j r_{g(i,j)} h_i h_j + \frac{M_e}{N_T} \qquad \text{(II.6)}$$

Notably, the assumption of the fixed-effect meta-analysis are violated when heterogeneities in heritabilities exist and when between-cohort genetic correlations are less than 1. The purpose of deriving h^2_{ma} under the fixed-effect IVM model is to reflect and explain observations based on practices used in current studies. In practice, genetic correlations between cohorts are estimated to be less than 1 for which the key consequence is that we will observe h^2_{SNP} estimates decrease as the number of cohorts being meta-analysed increases. Usually the h^2_{SNP} estimate from the meta-analysis plateaus such that it does not decrease further when additional data sets are added. The drop in h^2_{SNP} estimates contributes to the observation that the out-of-sample prediction R^2 can exceed the h^2_{SNP} of the (meta-analysed) cohort from which predictors were generated.

II.2.1.6 Special considerations in binary traits

For binary traits such as MD, the estimated SNP-based heritability on the observed risk scale (subscript *obs*) needs to be converted to the unobserved liability scale (subscript *l*) to account for the scale of heritability and ascertainments bias. We converted the heritability and its standard error to the liability scale using the formula from Lee et al. study[155]:

$$h_l^2 = \frac{K(1-K)}{z^2}\frac{K(1-K)}{P(1-P)}h_{obs}^2 \quad \text{(II.7)}$$

$$SE_{h_l^2} = \frac{K(1-K)}{z^2}\frac{K(1-K)}{P(1-P)}SE_{h_{obs}^2} \quad \text{(II.8)}$$

K is the disease lifetime prevalence, and *z* is the height of the normal curve at the threshold for the quantile of the normal distribution of proportion *K*. *P* is the proportion of cases in the sample. Additionally, for binary traits, sample sizes in all equations above should be replaced by the effective sample sizes (N_{eff}):

$$N_{eff} = 4N_{case}N_{control}/(N_{case}+N_{control})$$

II.2.2 Simulating impacts of heterogeneities in cohort-specific heritability and between-cohorts genetic correlations on the h^2_{ma}

To illustrate how heterogeneities in heritability and genetic correlations affect h^2_{ma}, we simulated h^2_{SNP} of 100 cohorts and r_g between these 100 cohorts under different assumptions as follows:

II.2.2.1 Same underlying h^2_{SNP} for all cohorts; same underlying r_g for all pairs of cohorts

Under this assumption, we assume the true SNP-based heritability is the same for all cohorts, and differences in h^2_{SNP} estimations are due to sampling variation alone. At the same time, we assume the true r_g for all pairs of cohorts are the same and differences are due to sampling variation as well.

To simulate h^2_{SNP} of 100 cohorts under this setting, we randomly drew 100 values from the uniform distribution:

Uniform [max(0, $h^2_{SNP-true}$- 679/N),min(1, $h^2_{SNP-true}$ + 679/N)]

Here, the $h^2_{SNP-true}$ denotes the true underlying h^2_{SNP} (we arbitrarily chose 0.2, 0.5, and 0.8 as $h^2_{SNP-true}$), which is unmeasurable in practice. N is the sample size. Visscher et al. [156] previously showed the GREML standard error of h^2_{SNP} estimation can be accurately approximated by $\sqrt{2M_e}/N$. To make sure the $h^2_{SNP-true}$ lies within the 95% confidence interval (C.I.) of the simulated h^2_{SNP}, and to ensure simulated h^2_{SNP} falls inside [0, 1], we set max(0, h^2_{SNP}- 1.96$\sqrt{2M_e}/N$), and min(1, h^2_{SNP}+ 1.96$\sqrt{2M_e}/N$) to be the lower and the upper limit of the uniform distribution. The effective number (M_e) of SNPs on a standard GWAS chip in the European population is around 60,000.[157]

Similarly, to simulate genetic correlations between cohort *i* and *j*, we randomly drew a value from the uniform distribution:

Uniform [max(0, r_{g-true} – 1.96*SE($\hat{r}_{g-true}$),min(1, r_{g-true}+ 1.96*SE($\hat{r}_{g-true}$)]

Here, r_{g-true} is the true underlying genetic correlations between all pairs of cohorts. Following Visscher et al. [156] study, the standard error of the estimated genetic correlation is given by

$$SE(\hat{r}_{g-true}) = \sqrt{\frac{r_{g-true}^2(N_i^2h_i^4+N_j^2h_j^4)+2h_i^2h_j^2N_iN_j}{4x10^{-5}h_i^4h_j^4N_i^2N_j^2}}$$

Here, N_i and N_j are the sample size of cohort *i* and *j*, respectively; h_i^2 and h_j^2 are the true SNP-based heritability of cohort *i* and *j* respectively.

II.2.2.2 Same underlying h^2_{SNP} for all cohorts; different underlying r_g between pairs of cohorts

Under this assumption, the true underlying r_g is different between cohorts. Therefore, differences in genetic correlations between cohorts are due to both differences in the underlying r_g, and sampling variation. Simulations under this assumption are similar to *II.2.2.1*, except that the r_{g-true} is now sampled from a uniform distribution, with the mean arbitrarily set to be 0.2, 0.5, and 0.8, and the range arbitrarily set to be 0.2.

II.2.2.3 Different underlying h^2_{SNP} for all cohorts; same underlying r_g between pairs of cohorts

Under this assumption, the true underlying h^2_{SNP} is different. Therefore, differences in cohort-specific SNP-based heritability are due to both differences in the underlying true $h^2_{SNP-true}$, and sampling variation. Simulations under this assumption are similar to *II.2.2.1*, except that the $h^2_{SNP-true}$ is now sampled from a uniform distribution, with the mean set to be 0.2, 0.5, and 0.8, and the range around the mean set to be 0.2.

II.2.2.4 Different underlying h^2_{SNP} for all cohorts; different underlying r_g between pairs of cohorts

Under this assumption, the true underlying h^2_{SNP} and r_g are both different. We randomly sampled r_{g-true} and $h^2_{SNP-true}$ from a uniform distribution, with the

mean set to be 0.2, 0.5, and 0.8, and the range set to be 0.2 around the mean. All other parts of the simulations followed procedures outlined in *II.2.2.1*

We simulated sample sizes of each cohort under four settings:
(1) Sample sizes of all cohorts are drawn from Uniform [50,000 , 100,000]
(2) Sample sizes of all cohorts are drawn from Uniform [5,000 , 10,000]
(3) Sample sizes are in an arithmetic sequence, with the smallest sample size being 1,000, the largest sample size being 100,000, and the common difference between terms being 1,000
(4) Sample sizes are in an arithmetic sequence, with the smallest sample size being 100, the largest sample size being 10,000, and the common difference between terms being 100
We repeated simulations under each assumption 100 times, ordered the cohort based on their sample size (the largest cohort being meta-analysed first), and calculated the h^2_{ma} with the formula (II.4).

II.2.3 Validate equation derivations with empirical data

We use real data to test whether the equations we derived provide valid estimates of SNP-based heritability in the meta-analysis cohort and genetic correlations between the target and the meta-analysis cohort. We compared the h^2_{ma} and the $r_{g(ma,t)}$ estimated by equations (II.4) and (II.5), with those directly estimated from the meta-analysis summary statistics.

II.2.3.1 Estimation of cohort-specific SNP-based heritability and between cohort genetic correlations

GWAS summary statistics for major depression (MD) of 34 cohorts[158] were obtained from the Major Depressive Disorder Working Group of the Psychiatric Genomics Consortium (PGC). Access to the dataset can be obtained via submitting a research proposal to the PGC (https://pgc.unc.edu/for-researchers/data-access-committee/data-access-information/). Summary statistics were generated by the PGC team following the RICOPILI pipeline[16]. No further QC was performed in our analysis. We used LDSC regression[142] to estimate cohort-specific h^2_{SNP} and between-cohort r_g. The analysis requires an LD reference panel to account for correlations between SNP estimates, and here we used the European ancestry cluster of the 1000 Genome projects, and only HapMap3 SNPs were retained. It is notable that although genetic correlation estimates from LDSC are unbiased[159], the heritability estimate in LDSC is known to be

downward biased[85]. However, LDSC is still a popular tool for estimating SNP-based heritability in practice because of its simplicity and computational efficiency. To reflect what is done in real practice, we also chose LDSC as the tool to estimate SNP-based heritability (this does not impact conclusions but allows for comparison with published papers). After getting LDSC estimates from binary trait GWAS that are made on the observed scale (h^2_{obs}, which means the estimates reflect in part the proportion of cases in the sample, *P*), we converted the heritability and its standard error to the liability scale using the equation (II.7) and (II.8), and we set *P* to be 0.15 in line with the common practice for MD. Once we converted the heritability to the liability scale, we removed cohorts with heritability estimates smaller than 0 or larger than 1. As a result of small effective sample sizes, r_g estimates are unavailable for certain pairs (NA estimates are generated reflecting negative SNP-based heritability estimates, for which the square root is not estimable). As equations (II.4) and (II.5) require valid r_g input, we built up an r_g matrix, and calculated the total number NAs of each row (corresponding to each cohort). We removed the cohort with the largest number of NAs, and when two or more cohorts had an equal number of NAs, we removed the cohort with the smallest effective sample sizes. We repeated the above processes iteratively until no NAs were left in the r_g matrix. In total, 21 out of 34 cohorts were left after applying the above criteria.

II.2.3.2 Calculating the h^2_{ma} and the $r_{g(ma,t)}$ directly from the meta-analysis summary statistics

We conducted four sets of meta-analyses, varying the order in which studies were added to the meta-analysis. We conducted analyses where MD cohorts were entered by sample size in ascending (smallest study first) and descending (largest study first) orders of sample size. We conducted two further analyses where MD cohorts were ordered randomly. For each ordering, cohorts were entered and meta-analysed by adding one cohort at a time. We used METAL[45] to perform meta-analyses under the fixed-effect IVM assumption. With the 21 resulting meta-analysis summary statistics (the first resulting meta-analysis is in fact the GWAS summary statistics of the first cohort being meta-analysed), we estimated their h^2_{ma} using LDSC [142], resulting in 21 h^2_{ma} estimates. Next we estimated $r_{g(ma,t)}$ using standard practice, we held out the last cohort being meta-analysed, and meta-analysed the remaining 20 cohorts one at a time, and estimated genetic correlations with the LDSC[142], which gave 20 $r_{g(ma,t)}$ estimates.

II.2.3.3 Calculating the h^2_{ma} and the $r_{g(ma,t)}$ with the derived equations

Similarly, to get h^2_{ma} and $r_{g(ma,t)}$ estimates, we meta-analysed adding one cohort at a time following the same orders as outlined in *II.2.3.2*. But this time, instead of directly calculating h^2_{ma} and $r_{g(ma,t)}$ from summary statistics, we calculated them using the cohort-specific heritability and between cohort genetic correlations with our derived equations (II.4) and (II.5).

II.2.4 Comparing out-of-sample predictions R^2 with h^2_{ma}

To show that out-of-sample prediction R^2 can be higher than the h^2_{ma} under the condition outlined by the inequality (II.6), we chose a binary trait (major depression, MD) and a continuous trait (Educational attainment, EA) where cohort-specific heritabilities and between-cohort genetic correlations were available to us, as proof-of-principle examples.

II.2.4.1 Major depression

We obtained access to the results of the leave-one-cohort out (LOO) analysis in the PGC MD wave 3 analysis[158,160]. In brief, for the 48 cohorts (of the 49 cohorts in the MDD49 set) with individual-level genotype data, one cohort at a time was left out as the target sample, and the remaining 47 cohorts were meta-analysed together with the 27 additional cohorts whose summary statistics level data are available. Polygenic risk scores (PRS) were generated in the left-out target sample with SNP weights derived from the LOO meta-analysis summary statistics using the SBayesR method[161]. The h^2_t of the left-out sample, h^2_{ma} of the LOO meta-analysis, and the $r_{g(ma,t)}$ between the left-out sample and the remaining cohorts were calculated using LDSC and HapMap3 SNPs as described above. The effective number (M_e) of SNPs on a standard GWAS chip in the European population is set to be 60,000 [157]. The total effective sample sizes of the LOO meta-analyses are the sum of the effective sample sizes of individual cohorts in the LOO meta-analyses. We estimated the h^2_{ma} and $r_{g(ma,t)}$ directly from the summary statistics but found close agreement with estimates using equations (II.4) and (II.5). We removed cohorts whose h^2_t estimate was unavailable because of small sample sizes. We then plotted either cohorts whose h^2_t estimates lie between 0 and 1 (**Figure II.4**), or all cohorts with no restrictions on h^2_t (**Supplementary Figure II.5**)

II.2.4.2 Educational attainment

For educational attainment (EA), we obtained per-cohort heritability and between cohort genetic correlations for all pairs of cohorts from the supplemental file of Lee et al. study[145]. In their paper, they calculated between cohort r_g for cohorts with h^2_{SNP} estimates between 0 and 1 (47 out of 71 cohorts). We built up an r_g matrix and removed cohorts with the largest number of NAs (as for MD above, when two or more cohorts have an equal number of NAs, we removed the cohort with the smallest effective sample sizes). We repeated the above processes iteratively until no NAs were left in the r_g matrix. There are 35 out of 47 cohorts left for further analysis.

In Lee *et al.* study, they conducted out-sample prediction analyses in two cohorts (the Add Health and the HRS) with predictors generated from LOO meta-analyses of 70 cohorts. However, as summary statistics of LOO meta-analyses are not publicly available, we are unable to estimate h^2_{ma} and $r_{g(ma,t)}$ directly from the meta-analysis summary statistics. As shown in the **Figure II.2**, h^2_{ma} estimates are expected to plateau with an increased number of cohorts. Therefore, we can use the h^2_{ma} of the meta-analysis of all 71 cohorts to represent the h^2_{ma} of the LOO meta-analysis of 70 cohorts as these two estimates should be almost identical. In fact, we calculated the $h^2_{ma}$with equations (II.4) using per-cohort h^2_{SNP} and between cohorts r_g of 35 selected cohorts, and found h^2_{ma} estimated (h^2_{ma}= 0.132, 95% C.I. [0.121,0.143]) from those 35 cohorts (N = 671,314) are not significantly different from h^2_{ma} estimated (h^2_{ma}= 0.122, 95% C.I. [0.116,0.128]) from the meta-analysis of all 71 cohorts (N=1,131,881). Similarly, to test our theory, we need to assume the $r_{g(ma,t)}$ calculated with equation (II.5) using 34 selected cohorts (we left the Add Health or the HRS cohort out) is not significantly different from the $r_{g(ma,t)}$ between the left-out cohort and all 70 remaining cohorts. The effective number (M_e) of SNPs is also set to be 60,000.[157]

II.3 Results

II.3.1 Trends in SNP-based heritability in meta-analysis (h^2_{SNP}) with increasing meta-analysed cohorts under different genetic correlation and h^2_{SNP} settings

Our derivations require recognition of analytical approaches taken in practice and the assumptions of the underlying true model on which they depend, which contrast to an alternative model that likely operates when bringing together real data sets. An

underlying assumption of standard GWAS meta-analyses is that each contributing GWAS cohort is a random sample from a homogenous idealised population, such that the "true" SNP effects and the "true" h^2_{SNP} of each cohort are the same, and that the "true" genetic correlations between cohorts are 1. Hence, the assumption is that differences in estimated h^2_{SNP} between cohorts, and genetic correlations less than 1 simply reflect statistical sampling. However, with real data these assumptions may be violated. For example, genetic ancestry between cohorts can be different even when labelled as the same continent-based ancestry, and/or there may be differences in experimental settings and/or measured phenotypes[162]. Notably, GWAS results derived from population-based databases may use 'proxy' phenotypes in place of formal clinical phenotypes. For example, genetic understanding of major depressive disorder has been facilitated by use of data sets that record major depression 'proxy' phenotypes (e.g. From different UK Biobank data fields, multiple major depression phenotypes have been derived and significant variabilities in h^2_{SNP} have been reported in these phenotypes, and genetic correlations are significantly less than 1[163]). The expected value of the parameter h^2_{ma} can be expressed as a function of cohort-specific SNP-based heritability (i.e., true h^2_{SNP} of the "population" from which the cohort is sampled) and between-cohorts genetic correlations (i.e., the true genetic correlations between the cohort populations) (see **section II.2.1** above). Here "population" reflects genetic ancestry, phenotype definition and sampling frame of the phenotype.

To illustrate how heterogeneities in r_g and h^2_i will affect h^2_{ma}, and to explain empirical observations, we use simulations (as described in **II.2.2**) to investigate the impact of varying r_g and h^2_{SNP} on estimates of h^2_{ma}. To reflect common practice, the meta-analysis weights are determined under the fixed-effect IVM model. As shown in **Figure II.2** (sample sizes between 5,000 and 10,000) and **Supplementary Figures II.1-II.3** (other sample sizes), when r_g and h^2_{SNP} vary, the h^2_{ma} drops initially with an increasing number of cohorts being meta-analysed, but eventually reach a plateau. Note that in a single simulation, h^2_{ma} can both increase and decrease with an increasing number of cohorts, but the average over simulations shows the clear trend to a decreased plateau value (**Supplementary Figures II.4**). The overall trends are consistent across different $h^2_{SNP,}$ $r_{g,}$ and sample size assumptions, and the main difference is the increased standard error with increased heterogeneity or decreased sample sizes.

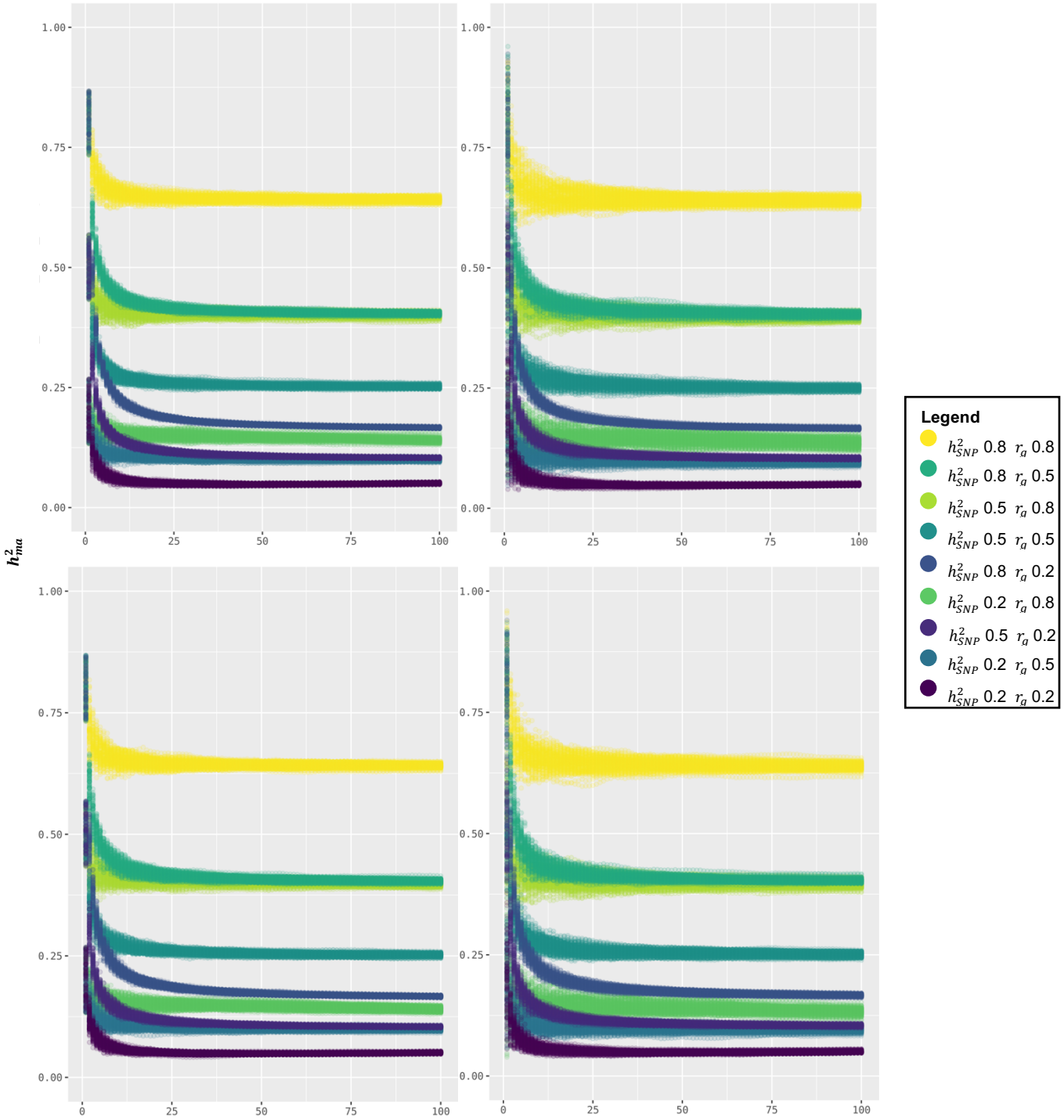


Figure II.2. Estimates of h^2_{ma} as the number of cohorts being meta-analysed increases under different r_g and h^2_{SNP} simulation settings. X-axes are the number of cohorts being meta-analysed, and y-axes are the estimates of h^2_{ma}. Each cohort being meta-analysed is simulated to have a sample size between 5,000 and 10,000, and the cohort with the largest effective sample size was meta-analysed first. (A) True r_g and h^2_{SNP} are the same in all cohorts so variation between simulation replicates represents sampling variation. (B) True r_g are the same but h^2_{SNP} are different between cohorts (C) True h^2_{SNP} are the same but r_g are different between cohorts. (D) True r_g and h^2_{SNP} are all different in each cohort. Each scenario (depicted by colour) has been repeated 100 times and so the graph shows the variation across replicates. Supplementary Figures II.1-II.3 show similar simulation results but for different sample sizes.



Figure II.3. Comparison of h^2_{ma} and $r_{g(ma,t)}$ estimated by the derived equations with those directly estimated from the meta-analysis summary statistics. Left column: 21 major depression cohorts were meta-analysed, adding one cohort at a time, in different orders. The first meta-analysis results are simply the GWAS summary statistics of the first cohort. **Right column:** We held out the last cohort being meta-analysed and calculated genetic correlations between the left-out cohort and each of 20 remaining meta-analyses. X-axes are the number of cohorts included in the meta-analysis. In general, h^2_{ma} (first column) and the $r_{g(ma,t)}$ (second column) estimated with the formulae are consistent with those directly estimated from the meta-analysis summary statistics. Notably, when the sample size of the left-out cohort is small, $r_{g(ma,t)}$ estimated with two methods could have some insignificant differences because of large standard errors.

II.3.2 Comparison of h^2_{ma} and genetic correlations derived from equations vs. estimates from meta-analysis summary statistics

To show empirical validity of our derivations, we obtained GWAS summary statistics of 21 cohorts for major depression[158] (**section II.2.3.1**). The h^2_{SNP} estimates of these cohorts, along with genetic correlations between them, were estimated using LDSC[142], following standard practice of the PGC. We randomly ordered the 21 selected cohorts, and meta-analysed (using IVM and the software METAL[45]) adding one cohort at a time. With the 21 resulting meta-analysis summary statistics we estimated their h^2_{ma} using LDSC [142] and show good agreement with estimates of h^2_{ma} from equation (II.4) (**Figure II.3**) Notably, in some meta-analyses addition of a cohort generates an increase in the magnitude of h^2_{ma}, both estimated from the data and predicted from equation (II.4). However, the clear trend is a decrease in the estimated h^2_{ma} as more cohorts are added, a reflection of the estimated genetic correlations being less than 1. To demonstrate consistent estimates of $r_{g(ma,t)}$ estimated directly and from equation (II.5) we held out the last cohort being meta-analysed, estimated genetic correlations between the meta-analysed sample and target samples with the LDSC[142] (20 resulting $r_{g(ma,t)}$ estimates) (**Figure II.3**). We repeated these analyses with different orders of the cohorts and show our derivations are valid regardless of this (**Figure II.3**).

II.3.3 Empirical investigation using data on major depression and educational attainment

To show that out-of-sample prediction R^2 can be higher than the estimated h^2_{ma} under the conditions outlined by the inequality in equation (II.6) above, we chose a binary trait (major depression, MD[164]) and a continuous trait (educational attainment, EA[145]) where cohort-specific SNP-based heritabilities and between-cohort genetic correlations were available, as proof-of-principle examples.

For MD (see methods in **II.2.4.1**), we obtained access to results of leave-one-cohort out (LOO) analysis [158,160]. In brief, for the 48 cohorts with individual-level genotype data, one cohort at a time was left out and the remaining cohorts meta-analysed, together with 27 additional cohorts whose GWAS summary statistics (but not individual-level data) were available. PGS were calculated for all individuals in the left-out sample with SNP weights derived from the LOO meta-analysis summary statistics using the SBayesR method with default settings[165]. The h^2_t of the left-out sample, h^2_{ma} of the LOO meta-analysis, and the $r^2_{g(ma,t)}$ between the left-out

sample and the remaining cohorts were estimated using LDSC and HapMap3 SNPs. For the left-out cohort, we only retained those h^2_{SNP} estimates that are greater than 0 and smaller than 1 (**Figure II.4**) or retained all cohorts whose h^2_{SNP} are available (**Supplementary Figure II.5**). Results show that the derived inequality (equation II.6) is consistent with empirical results.

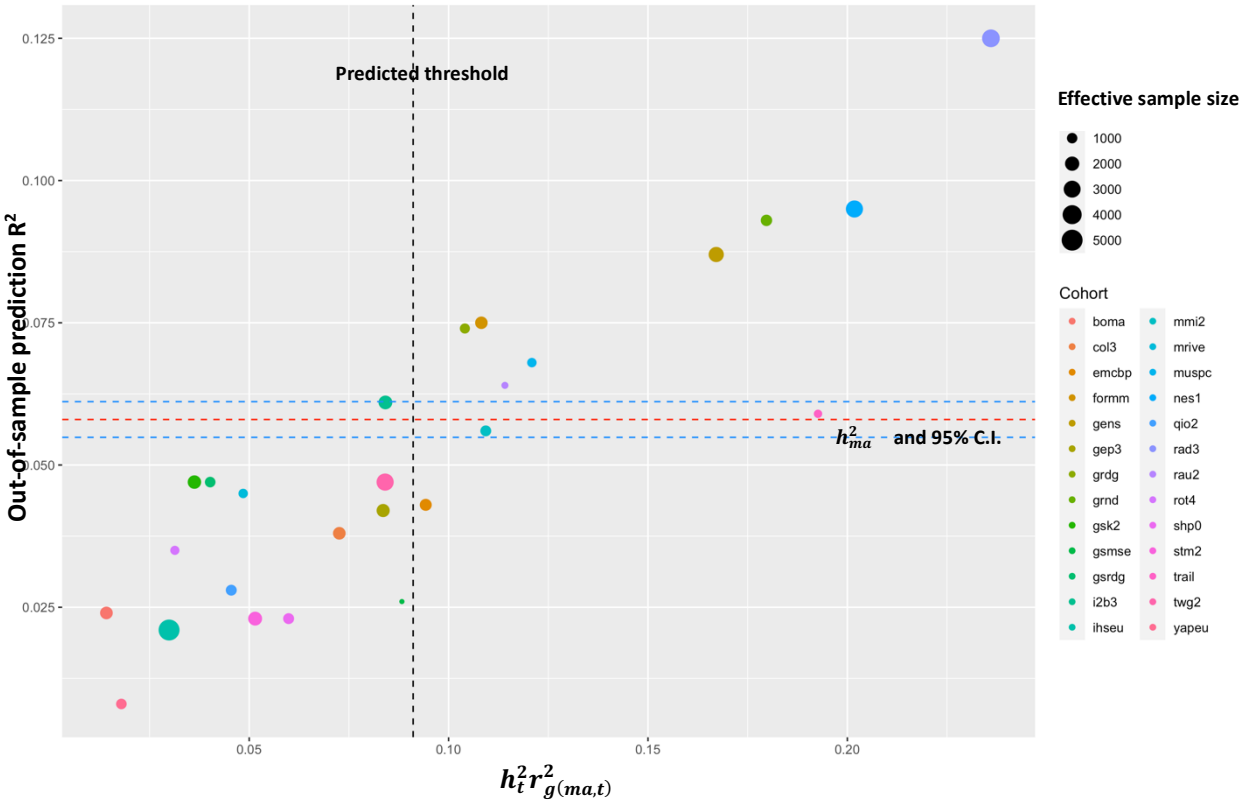


Figure II.4 Out-of-sample prediction R^2 on the liability scale for cohorts being predicted into and h^2_{ma} of the leave-one-cohort-out (LOO) meta-analysis for major depression datasets. In this figure, x-axis is the product of h^2_t (SNP-based heritability of cohorts being predicted into on the liability scale) and $r^2_{g(ma,t)}$ (Squared genetic correlations between the left-out target cohort we predicted into, and the LOO meta-analysis cohort used in generating the predictors); y-axis is the out-of-sample prediction R^2 on the liability scale. Each dot represents the target cohort (whose estimated h^2_t is between 0 and 1, sizes of dots are in proportion to the effective sample size). The horizontal line denotes h^2_{ma} (black dotted line) and its 95% C.I.(blue dotted line). (Although we left a different cohort out each time, the h^2_{ma} and 95% CI remain unchanged because the sample size of the left-out cohort is small when compared with the total sample size of meta-analysis.) Our mathematical derivations show that when the h^2_{SNP} in the cohort being predicted into is higher than the predicted threshold (vertical line), out-of-sample prediction R^2 will exceed the h^2_{ma} in the meta-analysis used to generate the predictor. Small cohorts have estimates (depicted by small dots) with large standard errors.

For EA, we obtained cohort-specific SNP-based heritability estimates and between cohort genetic correlations for all pairs of cohorts from the supplemental file of Lee

et al. study[145]. We removed cohorts with heritability estimates smaller than 0 or larger than 1, or not available (see method in **II.2.4.2**). 35 out of 71 cohorts met the criteria. We conducted analyses as for MD, but LOO PGS results were only available for 2 cohorts. The empirical results again agree with our derived expectation (**Supplementary Figure II.6**).

II.4 Discussion

In conclusion, in this study, we demonstrate that the h^2_{ma} can be expressed in terms of per-cohort h^2_{SNP}, between-cohort genetic correlations, and meta-analysis weights (which are a function of the sample sizes under the commonly used fixed effect IVM model). Under idealised conditions where between-cohorts genetic correlations are all equal to 1, and all cohorts have a common SNP-based heritability, the out-of-sample prediction R^2 will always be smaller than SNP-based heritability (smaller because of error associated with estimates of SNP effect sizes). The difference between h^2_{SNP} and R^2 will tend to be 0 with an infinitely large sample size. However, when h^2_{SNP} and r_g heterogeneities exist, h^2_{SNP} estimates made from GWAS meta-analysis results will decrease as the numbers of meta-analysed cohorts increases (equation II.4) until reaching a plateau, and the out-of-sample prediction R^2 can be greater than SNP-based heritability (equation II.6). Notably, a key assumption of the fixed-effect meta-analysis is that true underlying effect sizes of SNPs are the same for each cohort, and the experimental settings and measured phenotype are the same[162]. These assumptions do not always hold, especially when population-based databases are used where phenotypes may be 'proxy' phenotypes. With the knowledge of how between cohort heterogeneity can impact SNP-based heritability estimates it may be relevant to select cohorts that represent the focal trait (e.g., clinically measured major depressive disorder for MD, or years of education (as opposed to the proxy trait of attended college (yes/no) for EA), and treat other cohorts as genetically correlated traits (i.e., an MTAG analysis)[166].

III

Multi-ancestry conditional and joint analysis (Manc-COJO) applied to GWAS summary statistics

" A thousand streams—born of separate sources—
Find their unity in the sea. " – from the Huainanzi, a classic Daoist-inspired text dated back to the 2nd century BC

III.1 Introduction

GWAS have revolutionised understanding of the genetic architecture of complex traits[167]. Typically, the GWAS paradigm tests for association between a trait and SNPs, one SNP at a time. However, associations detected in GWAS are often not independent because of linkage disequilibrium (LD)[167] between SNPs. In principle, multivariate linear regression analysis can be applied to GWAS data to identify secondary SNP associations, which are SNP associations detected after conditioning on more significantly associated SNPs. However, such analyses are computationally demanding and are not possible if individual-level data are not available[168-170]; modern GWAS are often a meta-analysis of summary statistics from contributing studies. To address this challenge, the conditional and joint (COJO) multiple-SNP analysis of GWAS summary statistics was developed and implemented into the GCTA software[170]. The algorithm requires an LD matrix to model the correlation structure between SNPs which the software generates from user-provided individual-level genotype data. The GCTA-COJO algorithm was developed with an implicit assumption that the GWAS was conducted in a single ancestry, and is hereafter called Sanc-COJO (**S**ingle-**anc**estry **Co**nditional and **Jo**int analysis).

GWAS are now commonly conducted as meta-analyses of GWAS across samples with diverse ancestries[1,171,172]. Ancestry groups differ in allele frequencies and LD patterns because of demographic and evolutionary processes such as random drift[173]. A key question when implementing post GWAS analyses such as COJO is how to model the correlation structure between SNPs in a multi-ancestry GWAS setting. Some studies have repurposed Sanc-COJO to analyse multi-ancestry GWAS data either by using LD matrices derived from European ancestry cohorts[172,174], as these dominate (typically >80%) current GWAS meta-analyses[175], or by constructing LD matrices from reference samples that match the ancestry composition in the meta-analysis[176,177]. Other approaches include conducting COJO analysis separately within each ancestry[178,179] or performing COJO analysis in one ancestry and attempting replication in other ancestries[180,181]. However, to our knowledge these *ad hoc* approaches have not been formally evaluated, and optimal strategies require further consideration. Meanwhile, although Sanc-COJO was not originally designed as a fine-mapping tool, its ease of use has led many researchers to incorporate COJO into their fine-mapping pipeline[182,183]. However, interpretation of COJO-based independent associations as being causal requires further investigation.

In this chapter, we introduce **m**ulti-**anc**estry **co**nditional and **jo**int analysis (Manc-COJO), which extends the capabilities of stepwise conditional and joint association analysis to multi-ancestry datasets. Through extensive simulations and real-world data applications, we demonstrate that Manc-COJO improves the detection of independent trait-associated loci and reduces false positive discoveries. An underlying assumption of our method (supported by an increasing body of evidence) is that causal variants are largely shared across ancestries and have similar effect sizes across ancestries[184-187]. However, we demonstrate that Manc-COJO is robust to violations of this assumption. To address scenarios involving ancestry-specific causal variants[188,189], we propose application of a follow-up analysis: multi-ancestral data-informed single ancestry analysis (Manc-COJO --MDISA), that applies a COJO analysis within a single ancestry while accounting for SNPs identified by Manc-COJO.

III.2 Methods

In this section, we first review the conceptual assumptions underlying the COJO models. We then detail our algorithms and outline the step-by-step model selection process. Finally, we demonstrate the validity of the algorithms implemented in our method through theoretical derivations.

III.2.1 Concepts and assumptions

In GWAS and post-GWAS analyses, there are three SNP effect size concepts: the true causal effect size (β_c), the marginal effect size (b_m), and the effect size from the joint analysis of SNPs (β_J).

III.2.1.1 Modelling true effect

The true causal effect size (β_c) is the effect of true causal SNPs on the phenotypes.

We make three assumptions about true causal effects:

(1) β_c of SNPs act on the phenotype additively, which means that for a specific SNP, each copy of the alternative allele adds β_c to the phenotype.

(2) β_c of different SNPs are independent, which means the effect of a specific SNP will not affect the effect of another SNP. Notably, although SNP-substitution effects are assumed to be independent, the LD between different SNPs is **NOT** independent. For example, when two variants in nearby

locations confer a beneficial effect for natural selection, they will compete with each other for fixation and thus create a huge negative LD between these two loci until recombination breaks down the LD between them (Hill-Robertson Interference[190,191]), which takes many generations to occur, as there is only ~1 cross-over per chromosome per generation[192].

(3) The number of true causal variants, and their effect sizes on **standardised** phenotypes, are identical in different ancestries. In other words, under an additive model, if a variant is causal in one ancestry, it must be causal in another ancestry. Although allele frequencies of the same SNP in different ancestries can be different (potential causes of such differences include different selective pressure in different ancestries or other evolutionary forces like genetic drift and population bottlenecks, or the same mutation happened at different time in different ancestries[193]), their effect sizes are the same. (We later relax this assumption by including ancestry-specific variants).

We assume an underlying generative model that phenotypes of **all** individuals in a population (size *N*) depend on causal SNP effects according to the following equation:

$$\mathbf{y} = \mathbf{X}_C\boldsymbol{\beta}_C + \mathbf{e}_C \qquad \text{(III.1)}$$

where

- $\mathbf{y}$ is a *N*x1 vector of **standardised** residuals of phenotypes (after regressing out all covariates)
- $\mathbf{X_C}$ is an *N*x*C* **mean-centred** matrix of *C* genotypes of causal variants shared across ancestries
- $\boldsymbol{\beta}_\mathbf{C}$ is a *C*x1 vector of true causal effects
- $\mathbf{e_C}$ is a *N*x1 vector of error terms, which is normally distributed under N(0,σ_e^2)

However, the genomic location and effect sizes of true causal SNPs are not always observable for several reasons. For example, not all true causal SNPs are genotyped in the SNP-array (imputed) data. Even in whole genome sequencing data, there are still gaps (for example, regions near the centromere and telomere[194]). LD adds complexity in identifying the true causal variants, especially when there are multiple causal variants within a region of correlated SNPs. In following sections, we will model what is observable in real practice, and link it to the unoberservable true causal effects.

III.2.1.2 Modelling marginal effects in GWAS analysis

In GWAS analysis, one SNP at a time is tested for association with the phenotype. Here, we commence the derivation by assuming access to the genetic data of every individual in the population; accordingly, all parameters are treated as population parameters.

Let $\mathbf{x_m}$ be an *N*x1 mean-centred vector of genotypes for the SNP *m* to be tested in the GWAS analysis, and b_m to be the marginal effect size of the SNP *m*. As the genotype and error are independent (i.e. $\mathbf{x'_m e_C} = 0$), $\mathbf{x'_m y} = \mathbf{x'_m x_C \beta_C}$. It follows that,

$$b_m = (\mathbf{x'_m x_m})^{-1} \mathbf{x'_m y} \qquad \text{(III.2)}$$

Now, let p_m to be minor allele frequency of the SNP *m*, $q_m = 1 - p_m$. Based on the definition of binomial variance, and under the Hardy-Weinberg equilibrium (HWE), $\mathbf{x'_m x_m} = 2p_m q_m N$. Let us denote $D_m = 2p_m q_m N$.

Applying equation (III.1) into equation (III.2), we get

$$b_m = \frac{\mathbf{x'_m X_C \beta_C}}{D_m} = \sum_{c=1}^{C} r_{m,c} \beta_c \sqrt{\frac{p_c q_c}{p_m q_m}} \qquad \text{(III.3)}$$

where, $r_{m,c}$ denotes the (signed) LD correlation between SNP *m* and the *c*-th causal SNP. In other words, the value of true marginal effect size (b_m) for the SNP *m* in the GWAS analysis, can be viewed as a weighted sum of true effect size of all causal SNPs, where the weights are the LD correlations between the SNP *m* and causal SNPs, adjusted for minor allele frequencies.

lII.2.1.3 Modelling joint effects

In a joint analysis, any set of j SNPs can be fitted together, yielding the corresponding true joint effects (assuming we have genetic data for every individual in the population). The major benefit of considering SNPs jointly is to account of the LD correlation between these SNPs. Joint effects of *j* SNPs can related to the true marginal effects of those *j* SNPs and the LD correlation between those *j* SNPs:

$$\boldsymbol{\beta_J} = \mathbf{D_J^{-\frac{1}{2}} R_J^{-1} D_J^{\frac{1}{2}} b_M} \qquad \text{(III.4)}$$

Here,

- $\boldsymbol{\beta_J}$ is an *Jx1* vector of true **joint** effect sizes when fitting those *J* SNPs are considered jointly in a model
- $\mathbf{b_M}$ is a *Jx1* vector of true **marginal** effective sizes
- $\mathbf{R_J}$ is a *JxJ* true LD correlation matrix between *J* SNPs that are fitted jointly in the model

- $\mathbf{D_J}$ is a *JxJ* diagonal matrix with *j*-th elements being $D_{j,j} = 2p_j q_j N$

Applying equation (III.3) into equation (III.4), gives

$$\boldsymbol{\beta}_{\mathbf{J}} = \mathbf{D}_{\mathbf{J}}^{-\frac{1}{2}} \mathbf{R}_{\mathbf{J}}^{-1} \mathbf{r}_{\mathbf{JC}} \mathbf{D}_{\mathbf{C}}^{\frac{1}{2}} \boldsymbol{\beta}_{\mathbf{C}} \qquad \text{(III.5)}$$

where,

- $\mathbf{D_J}$ is a *JxJ* diagonal matrix with j-th elements being $D_{j,j} = 2p_j q_j N$
- $\mathbf{R_{JC}}$ is a *JxC* true LD correlation matrix between J SNPs that are fitted jointly in the model and *C* causal SNPs
- $\mathbf{D_C}$ is a *CxC* diagonal matrix with c-th elements being $D_{c,c} = 2p_c q_c N$
- $\boldsymbol{\beta}_{\mathbf{c}}$ is an *Cx1* vector of **true** effect sizes of *C* causal SNPs

Now, let

- $\mathbf{V_J}$ be a *JxJ* diagonal matrix with j-th elements being $V_{j,j} = 2p_j q_j$
- $\mathbf{V_C}$ be a *CxC* diagonal matrix with c-th elements being $V_{c,c} = 2p_c q_c$

Equation (III.5) can be simplified to:

$$\boldsymbol{\beta}_{\mathbf{J}} = \mathbf{V}_{\mathbf{J}}^{-\frac{1}{2}} \mathbf{R}_{\mathbf{J}}^{-1} \mathbf{r}_{\mathbf{J,C}} \mathbf{V}_{\mathbf{C}}^{\frac{1}{2}} \boldsymbol{\beta}_{\mathbf{C}} \qquad \text{(III.6)}$$

When all C true causal SNPs are fitted in a model, $\mathbf{R_J} = \mathbf{r_{J,C}}$, $\mathbf{V_J} = \mathbf{V_C}$. It follows that $\boldsymbol{\beta}_{\mathbf{J}} = \boldsymbol{\beta}_{\mathbf{C}}$. In other words, when all true causal SNPs are included jointly in a model and all the assumptions described above are satisfied, the true joint SNP effects will equal the true causal SNP effects.

III.2.1.4 Modelling conditional effects

Having modelled the marginal and joint effects, we can now define an additional quantity: the conditional effect. Given a set of SNPs (set S), we fit these SNPs jointly into a model as in (III.4), and obtain the residual:

$$\mathbf{y}_{\mathrm{S}} = \mathbf{y} - \mathbf{X}_{\mathrm{S}} \boldsymbol{\beta}_{\mathrm{S}}$$

For each SNP i not included in the model, the conditional effect of SNP i given set S, denoted $b_{i|\boldsymbol{\beta}_S}$, is essentially the marginal effect of SNP i in the following regression:

$$\mathbf{y}_{\mathrm{S}} = \mathbf{x_i} \mathrm{b}_{\mathrm{i}|\boldsymbol{\beta}_{\mathrm{S}}} + \mathbf{e_S}$$

Intuitively, the conditional effect represents the remaining contribution of SNP i to the phenotype after adjusting for the effects of SNPs in set S.

The conditional effect $b_{i|\boldsymbol{\beta}_S}$ can be expressed as follows:

$$b_{i|\boldsymbol{\beta}_S} = b_i - V_i^{-\frac{1}{2}} \mathbf{r}_{i,S} \mathbf{V}_S^{\frac{1}{2}} \boldsymbol{\beta}_S$$

$$= b_i - V_i^{-\frac{1}{2}} \mathbf{r}_{i,S} \mathbf{R}_S^{-1} \mathbf{V}_S^{\frac{1}{2}} \mathbf{b}_S \qquad \text{(III.7)}$$

where,

- $\mathbf{r}_{i,S}$ is a 1 x *s* vector of LD correlations between SNP *i* and each of *s* SNPs already selected set *S* included in the model
- $\boldsymbol{\beta}_S$ is a *s* x 1 vector of true **joint effects** of the *s* SNPs in set *S* in the model
- $\mathbf{R}_S^{-1}$ is a *s* x *s* matrix of LD correlations between s SNPs that are already in the model
- $\mathbf{b}_S$ is a s x 1 vector of true **marginal effects** of *s* SNPs in the model
- b_i is the true marginal effect of the SNP *i*
- $\mathbf{V}_S$ is a *s* x *s* diagonal matrix with *s*-th elements $V_s = 2p_s q_s$
- $V_i = 2p_i q_i$ (Assuming HWE)

III.2.1.5 Assumption of a homogenous LD structure

It is important to note that all derivations above implicitly assume a homogeneous LD structure within the population. Consider a scenario in which a population (denoted by the subscript *com*) is not homogeneous, but consists of two sub-populations (denoted by subscripts *1* and *2*) **com**bined together. Each sub-population is internally homogeneous, yet they differ in allele frequencies at the same SNPs, resulting in distinct correlation structures between the sub-populations. In such scenario, the genotypic variance of SNP *i* in the combined population ($V_{com,i}$) will no longer be $2p_{com,i}q_{com,i}$, but rather

$$V_{com,i} = 2p_{1,i}q_{1,i}w_1 + 2p_{2,\mathrm{i}}q_{2,\mathrm{i}}w_2 + 4w_1w_2(p_{1,\mathrm{i}} - p_{2,\mathrm{i}})^2 \qquad \text{(III.8)}$$

Where w_1 and w_2 are the proportion of sub-population 1 and 2 in the combined population, respectively. In other words, the Hardy–Weinberg equilibrium (HWE) assumption is violated when the population is not homogeneous.

Let $V_{a,i}$ be the genotypic variance of SNP *i* in the population *a*, where $V_{a,i} = 2\mathrm{p}_{a,i}q_{a,i}$, and $r_{\mathrm{a},i,k}$ be the (signed) LD correlation between SNP *i* and SNP *k*. The LD correlation between SNP *i* and SNP *k* in the combined population, can be expressed as:

$$r_{\mathrm{com},i,k} = w_1 r_{1,\mathrm{i},k}\sqrt{\frac{V_{1,i}V_{1,k}}{V_{\mathrm{com},i}V_{\mathrm{com},k}}} + w_2 r_{2,i,k}\sqrt{\frac{V_{2,i}V_{2,k}}{V_{\mathrm{com},i}V_{\mathrm{com},k}}} + \frac{4w_1w_2(p_{1,\mathrm{i}}-p_{2,\mathrm{i}})(p_{1,\mathrm{k}}-p_{2,\mathrm{k}})}{\sqrt{V_{\mathrm{com},i}V_{\mathrm{com},k}}} \qquad \text{(III.9)}$$

It is clear that $r_{\mathrm{com},i,k} = r_{1,i,k}$ will hold for all SNP pairs i and k only in a homogeneous population (i.e. $w_1 = 1, w_2 = 0$, $V_{1,i} = V_{\mathrm{com},i}$, and $V_{1,k} = V_{\mathrm{com},k}$).

III.2.2 Addressing the Challenges of Real-World Data

In an idealised scenario, where genetic data are available for all individuals in a population, the population is homogeneous, and all model assumptions hold, fitting all causal SNPs jointly would recover the true causal effect sizes. In practice, such conditions never exist, and some assumption violations are beyond our control. Therefore, the aim of our algorithm (and, arguably, all COJO- and fine-mapping–type algorithms) is to identify a set of SNPs that approximates the true causal variants—in both number and genomic location—while controlling for manageable sources of error and recognising inherent limitations of the model. To achieve this aim, we can implement an extended version of conditional and joint analyses in each iteration of the algorithm. As in Sanc-COJO[170], LD correlations between SNPs can be estimated from a reference sample, with allele frequencies and marginal effect sizes available from GWAS (meta-analysis) summary statistics. However, when using summary statistics several additional quality control procedures are needed to make sure results are comparable to scenarios where individual level genotype data are available.

III.2.2.1 Considerations in dealing with real-world data

- *Per-SNP Sample size*

In real data sets, per-SNP sample size can be significantly different between SNPs. Within a single data set differences in sample size between SNPs occur because of insufficient SNPs in a region to allow accurate imputation, which in turn may reflect genotype missingness. In GWAS meta-analyses different sub-studies may contribute different SNPs which can generate very large differences in sample sized between SNPs. We suggest exclusion of SNPs to avoid extreme differences in per-SNP sample sizes.

- *Phenotypic variance*

In our model we assumed phenotypic variance to be 1 across studies. However, this may not be true in real practice (even when traits are described as standardised this is often prior to accounting for covariates). Therefore, as in Sanc-COJO, the phenotypic variance in each study is estimated from GWAS summary statistics. For each SNP i, the estimate of $\mathbf{y}'\mathbf{y}$ can be estimated from:

$$\widehat{\boldsymbol{y}}'\widehat{\boldsymbol{y}} = \widehat{D}_i\hat{S}_i^2(n_i - 1) + \widehat{D}_i\hat{b}_i^2$$

where $\hat{S}_i^2$ is the squared standard error of the marginal effect, $\hat{b}_i$ is observed marginal effect, n_i is the per-SNP sample size of SNP i (i.e. not the population size), and $\widehat{D}_i = 2\hat{p}_i\hat{q}_i$.
Phenotypic variance (V_P) can then be estimated by:

$$\widehat{V}_P = \widehat{D}_i\hat{S}_i^2 + \frac{\widehat{D}_i\hat{b}_i^2}{n_i - 1} \qquad \text{(III.10)}$$

In our analyses, we took the median of $\widehat{V}_P$ across all SNPs in the GWAS summary statistics as the estimate of the phenotypic variance.

- *Effective sample size*

As discussed in the Sanc-COJO paper[170], when data sets contributing to the GWAS results include relatives, the reported sample size will be larger than the effective sample size. In our analyses we estimated the per-SNP sample size for any SNP i using the equation provided in the Sanc-COJO paper[170]:

$$\hat{n}_i = \frac{\widehat{V}_P - 2\hat{p}_i\hat{q}_i\hat{b}_i^2}{2\hat{p}_i\hat{q}_i\hat{s}_i^2} + 1 \qquad \text{(III.11)}$$

- *Choice of reference sample to estimate LD correlation matrix*

In Section III.2.1.5, we examined the assumption of a homogeneous LD structure within a GWAS cohort. When GWAS summary statistics are obtained from meta-analyses of multiple cohorts, or when they are combined with an external LD reference cohort (i.e., out-of-sample LD), as in Sanc-COJO, an additional layer of complexity is introduced. For the analysis to yield reliable outcomes, all cohorts contributing to the meta-analysis within a single inferred ancestry should exhibit a broadly similar LD structure (i.e., heterogeneity may exist between GWAS meta-analysis cohorts but not within each cohort). Furthermore, it is essential to ensure that the LD structure of the GWAS cohort and that of the LD reference cohort are sufficiently similar to produce robust results.

Ideally, the LD estimated from the same sample used to generate the GWAS should also be employed in COJO analyses. For this reason, we support sharing of LD matrix together with GWAS summary statistics thereby facilitating information sharing when access to raw data is restricted. When the use of out-of-sample LD is unavoidable, its selection may be informed by prior knowledge or by approaches such as PC analysis, which lie beyond the scope of this work.

In addition, we recommend testing for differences in allele frequency between the reference and GWAS cohorts. The allele frequency of SNP *i* in a cohort

of *N* individuals can be modelled as a binomial distribution: $\hat{p}_i \sim \frac{1}{2N}\mathrm{Bin}(2N, p)$. When the sample size is large enough, it can be approximated by a normal distribution. $\hat{p}_i \sim \frac{1}{2N}\mathrm{N}(2Np, 2Npq) \sim \mathrm{N}\left(p, \frac{pq}{2N}\right)$. Hence, we suggest removing SNPs that are significantly different at specified threshold (for example, P = 5x10^{-8}) between the GWAS (subscript *GWAS*) and reference data (subscript *ref*) sets based on following z-score:

$$z = \frac{\hat{p}_{ref} - \hat{p}_{GWAS}}{\sqrt{\frac{\hat{p}_{ref}\hat{q}_{ref}}{2n_{ref}} + \frac{\hat{p}_{GWAS}\hat{q}_{GWAS}}{2n_{GWAS}}}} \qquad \text{(III.12)}$$

Another important consideration is the size of the reference sample. The estimation error of the Pearson correlation, assuming both variables are normally distributed, can be approximated by $\frac{1-r^2}{\sqrt{n-3}}$[195]. However, as genotypes are not normally distributed, we cannot get a closed-form solution for the estimation error of r^2. Empirical evidence suggest that reference sample sizes > 5,000 will generate relatively accurate LD correlation estimates[170].

III.2.2.2 Overview of single ancestry COJO algorhithm

In our study, we re-derived the equations used in the original GCTA-COJO (i.e. Sanc-COJO) so that the LD correlation matrix could be used as the LD reference, in addition to individual-level genotype data.

In the first iteration of the Sanc-COJO algorithm, the SNP with the smallest p-value is selected into the model. Conditional on this SNP, the conditional z-statistics of all other SNPs (taking SNP i as an example) are then calculated as:

$$z_{i|\hat{\boldsymbol{\beta}}_S} = \frac{\hat{b}_i - \hat{V}_i^{-\frac{1}{2}}\hat{\mathbf{r}}_{i,S}\,\hat{\mathbf{R}}_S^{-1}\hat{\mathbf{V}}_S^{\frac{1}{2}}\hat{\mathbf{b}}_S}{\hat{S}_i} \qquad \text{(III.13)}$$

where,

- $\hat{\mathbf{r}}_{i,S}$ is a 1 x *S* vector of **observed** LD correlations between SNP *i* and each of the *S* SNPs already selected into the model (in the first iteration, only 1 SNP is in the model)

- $\widehat{\mathbf{R}}_{\boldsymbol{S}}$ is an *S* x *S* matrix of **observed** LD correlations between S SNPs that are already in the model (in the first iteration, this reduces to a 1×1 matrix with the value equal to 1)
- $\hat{\mathbf{b}}_{\boldsymbol{S}}$ is an S x 1 vector of **observed marginal effects** of *S* SNPs in the model
- $\widehat{\boldsymbol{\beta}}_{\boldsymbol{S}}$ is an S x 1 vector of **observed joint effects** of *S* SNPs in the model
- $\hat{b}_i$ is the **observed** marginal effect of the SNP *i*
- $\widehat{\mathbf{V}}_{\boldsymbol{S}}$ is an *S* x *S* diagonal matrix with *s*-th elements of the diagonal $\widehat{V}_s = 2\hat{p}_s\hat{q}_s$ (assuming HWE)
- $\hat{V}_i = 2\hat{p}_i\hat{q}_i$ (assuming HWE)
- As in GCTA-COJO, to avoid over-fitting, a conservative approach is used to get the conditional p-value, where $\hat{S}_i$ denotes the standard error of the observed marginal effect of SNP *i*.

After getting the z-statistics for each of the remaining SNPs, the SNP with the smallest p-value is selected into the model (i.e. the SNP with largest z^2). Afterwards, a joint analysis is performed as follows:

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{S+1}} = \widehat{\mathbf{V}}_{\boldsymbol{S+1}}{}^{-\frac{1}{2}}\widehat{\mathbf{R}}_{\boldsymbol{S+1}}{}^{-1}\widehat{\mathbf{V}}_{\boldsymbol{S+1}}{}^{\frac{1}{2}}\hat{\mathbf{b}}_{\boldsymbol{S+1}} \qquad \text{(III.14)}$$

Here, the definition of $\widehat{\boldsymbol{\beta}}_{\boldsymbol{S+1}}$, $\hat{\mathbf{b}}_{\boldsymbol{S+1}}$, $\widehat{\mathbf{R}}_{\boldsymbol{S+1}}$, and $\widehat{\mathbf{V}}_{\boldsymbol{S+1}}$, are the same as defined in equation (III.13), except that now there are S+1 SNPs in the model. The estimation error of joint effects of S+1 SNPs can be expressed as:

$$\text{Var}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{S+1}}) = \widehat{\sigma}_J^2\widehat{\mathbf{D}}_{\boldsymbol{S+1}}{}^{-\frac{1}{2}}\widehat{\mathbf{R}}_{\boldsymbol{S+1}}{}^{-1}\widehat{\mathbf{D}}_{\boldsymbol{S+1}}{}^{-\frac{1}{2}} \qquad \text{(III.15)}$$

Where $\widehat{\mathbf{D}}_{\boldsymbol{S+1}}$ is a (s+1)x(s+1) diagonal matrix with *s*-th elements $\widehat{D}_s = 2\hat{p}_s\hat{q}_s\hat{n}_s$, and $\hat{n}_s$ is the effect sample sizes of SNP s (as defined in section III.2.2.1). Here, the residual variance of the joint analysis, can be estimated with the equation:

$$\widehat{\sigma}_J^2 = \widehat{V}_P - \hat{\mathbf{b}}_{\boldsymbol{S+1}}{}'\widehat{\mathbf{V}}_{\boldsymbol{S+1}}\widehat{\boldsymbol{\beta}}_{\boldsymbol{S+1}}$$

The vector of z-scores for the joint SNP effects for these s SNPs is given by

$$\mathbf{z}_{\boldsymbol{S+1}} = \text{diag}(\text{Var}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{S+1}}))^{-\frac{1}{2}}\widehat{\boldsymbol{\beta}}_{\boldsymbol{S+1}} \qquad \text{(III.16)}$$

In this single ancestry COJO algorithm, in each iteration, SNPs whose joint p-value fall above certain significance threshold are removed. The above conditional and joint analysis steps are repeated until no SNP can be removed or added. In our analyses we set the significance threshold for selection as p-value < 5×10^{-8}.

III.2.2.2 Extending COJO analysis into trans-ancestry settings

The single-ancestry COJO algorithm outlined in the previous section can be extended into a trans-ancestry setting. By leveraging diverse LD structures from different ancestries, we expect to gain better resolution in tagging true causal variants (shared across ancestries), and hence in estimating the number of independent associations.

- *Overview of the Manc-COJO algorithm*

The trans-ancestry COJO algorithm assumes GWAS summary statistics of two or more ancestries which have been processed through quality control steps described in Section III.2.2.2. The trans-ancestry COJO algorithm is summarised in seven steps and we have implemented it in the new *Manc-COJO* software. The software requires a predefined genome-wide significance threshold (we used 5×10^{-8}) and a pre-set collinearity threshold between SNPs (i.e. maximum LD r^2, we used 0.9 by default).

Step 1: Conduct an inverse-variance meta-analysis (IVM) of GWAS summary statistics across all ancestries. The SNP with the smallest meta-analysis p-value is selected for inclusion in the model.

Step 2: SNPs above the pre-set collinearity threshold with any SNPs already included in the model in any ancestry are removed in all ancestries.

Step 3: Within each ancestry, a conditional analysis is performed and the conditional effects of each SNP are calculated using Equation (III.13).

Step 4: An IVM of the conditional effects of each SNP is then conducted and the SNP with the smallest meta-analysis p-value is selected for inclusion in the model.

Step 5: A joint analysis is conducted within each ancestry using equation (III.14).

Step 6: An IVM of the joint effects of SNPs included in the model is performed, and SNPs whose meta-analysis p-value for joint effects exceeds the pre-set threshold are removed from the model.

Step 7: Steps 2 through 6 are repeated until no SNP can be added to or removed from the model.

- *Overview of the Manc-COJO --MDISA algorithm*

To identify additional associations within each ancestry that may represent ancestry specific variants, we developed an algorithm complementary to Manc-COJO, termed the Manc-COJO --MDISA. In MDISA, a subset of SNPs is fixed in the model prior to analysis. For identifying ancestry specific variants, these fixed SNPs are those identified by the Manc-COJO model, which assumes shared causal variants across ancestries. The framework can also be applied in other contexts; for example, it can incorporate variants supported by external evidence such as wet lab experiments.

Once the fixed SNPs are incorporated into the model, steps 2 through 7 described in the previous section are iteratively repeated until no additional SNPs can be added or removed. Importantly, MDISA employs two distinct p-value thresholds: a primary threshold for newly added SNPs (default: $5x10^{-8}$) and a secondary threshold for fixed SNPs (set to 1 in our analyses, but user-adjustable).

All newly added SNPs must satisfy the more stringent primary threshold, and any candidate SNP whose inclusion would reduce a fixed SNP's significance below the secondary threshold is rejected. In contrast, fixed SNPs are retained in the model regardless of their statistical significance, as they are supported by prior evidence.

If a more stringent secondary threshold is applied to fixed SNPs (for example $5x10^{-8}$), then no additional SNPs will be identified if the fixed SNPs are not all significant at the start of the analysis. Therefore, the choice of the secondary threshold represents a balance between maximising discovery and limiting false positive findings.

- *Optional R^2 parameter*

In variable selection algorithms, in addition to the approach implemented in the Sanc-COJO model, another commonly used strategy is to consider changes in the adjusted R^2 (Coefficient of determination) of model. Using summary statistics and an LD reference, the R^2 of the multiple linear regression can be approximated as:

$$\hat{R}^2 = \frac{\hat{\mathbf{b}}_{\boldsymbol{S}}{}'\widehat{\mathbf{V}}_{\boldsymbol{S}}\widehat{\boldsymbol{\beta}}_{\boldsymbol{S}}}{\hat{V}_P}$$

where $\hat{\mathbf{b}}_{\boldsymbol{S}}, \widehat{\mathbf{V}}_{\boldsymbol{S}}, \widehat{\boldsymbol{\beta}}_{\boldsymbol{S}}$ are as defined above. The estimated adjusted R^2 of the model can then be calculated as:

$$\hat{R}^2_{adj} = 1 - \frac{(1 - \hat{R}^2)(\hat{n} - 1)}{\hat{n} - s - 1}$$

Where *s* is the number of SNPs in the model, $\hat{n}$ is the effective sample size as defined in equation (III.11).

In the Manc-COJO algorithm, it is possible to use changes in $\hat{R}^2_{adj}$ as an *optional* selection criterion for researchers who prefer a more conservative model in regions with extremely complex LD structure. Researchers can specify thresholds that require the $\hat{R}^2_{adj}$ of the model to increase by a certain percentage during both forward and backward selection in *both* ancestries for a SNP to be included in the model.

For example, setting the forward-selection $\hat{R}^2_{adj}$ threshold to 0.1 means that, in each iteration, after including a SNP, the $\hat{R}^2_{adj}$ of the model must be at least 1.1 times the value from the previous iteration in both ancestries. Conversely, setting the backward-selection threshold to −0.1 means that, if adding a new SNP causes other SNPs to become insignificant and be dropped, then after excluding those SNPs, the $\hat{R}^2_{adj}$ of the resulting model must remain at least 0.9 times the value from the previous iteration for the new SNP to be retained. Otherwise, the algorithm reverts to the model from the last iteration and continues screening other SNPs for inclusion. By default, the $\hat{R}^2_{adj}$ thresholds for both forward and backward selection are set to −1, meaning that the selection criterion based on $\hat{R}^2_{adj}$ is not applied.

- *The benefit of trans-ancestry analyses*

By leveraging differences in LD structures across ancestries, Manc-COJO may enhance the ability to detect independent associations. To illustrate this, we begin with the single-ancestry analysis.

Under the infinitesimal model, and given the standardised phenotypic variance, the estimation error in GWAS for SNP m ($\hat{S}_m$) can be approximated by $\frac{1}{\sqrt{2\hat{p}_m\hat{q}_m\hat{n}_m}}$. In GWAS analyses, a common simplifying assumption is that the genotype matrix is fixed, even though it typically represents only a sample of the population. Consequently, we can approximate $\frac{1}{\sqrt{2\hat{p}_m\hat{q}_m\hat{n}_m}} \approx \frac{1}{\sqrt{2p_mq_mn_m}}$ (Note, this assumption again underscores the importance of selecting an appropriate LD reference, as discussed previously, and highlights the need to understand the limitations of each model arising from its underlying assumptions.).

Therefore for SNP m, the association z-score in GWAS is,

$$z_m = \frac{\hat{b}_m}{\hat{s}_m} = \frac{b_m + e_m}{\hat{s}_m} = e_m\sqrt{2p_m q_m n_m} + \sum_{c=1}^{C} r_{m,c}\beta_c\sqrt{2p_c q_c n_m}$$

$$= \sum_{c=1}^{C} r_{m,c}\beta_c\sqrt{2p_c q_c n_m} + \varepsilon_m \qquad \text{(III.17)}$$

Where $\varepsilon_m = e_m\sqrt{2p_m q_m n_m}$, and it follows that $\varepsilon_m \sim \mathrm{N}(0,1)$

When there is only one causal SNP within an LD block, equation (III.17) can be simplified to

$$z_m = r_{m,c}\beta_c\sqrt{2p_c q_c n_m} + \varepsilon_m \qquad \text{(III.18)}$$

where $\beta_c\sqrt{2p_c q_c}$ is a constant across all SNPs (note $r_{m,c}$, β_c, p_c, and q_c here are population parameters). In other words, in the absence of genotype missingness—so that the sample sizes of all SNPs are **equal**—and when causal SNP effect sizes are **large**, it is relatively straightforward to identify the SNP closest to the causal SNP (i.e., the SNP with the highest $|r_{m,c}|$). However, even in this idealised scenario, if per-SNP sample sizes differ due to genotype missingness, or if there are multiple tightly linked variants, the true causal SNP may not correspond to the largest $|z_m|$, even when it is genotyped.

When there is more than one causal SNP in a region of correlated SNPs, it is evident from equation (III.4) that the most significant SNP may not be the SNP closest to causal SNPs owing to effects such as genetic masking[196] (which refers to when trait increasing alleles are present on different haplotype backgrounds). Recognising that genomic regions are likely to habour multiple causal SNPs motivates the joint analysis approach.

In multi-ancestry meta-analysis, the z-score of SNP *i* ($z_{i,manc}$), derived from an inverse-variance meta-analysis of summary statistics from each ancestry, can be expressed as follows:

$$z_{i,manc} = \sum_{anc}\sqrt{\frac{D_{i,anc}}{D}}\, r_{i,c,anc}\beta_c\sqrt{D_{c,anc}} + e \qquad \text{(III.19)}$$

Here,

- $z_{i,manc}$ is the z-score of the SNP *i* in the trans-ancestry analysis IVM
- $\mathrm{D_{i,anc}} = 2p_{i,anc}q_{i,anc}n_{anc}$, where $p_{i,anc}$ is the MAF of the SNP *i* in a specific ancestry
- $D = \sum_{anc} D_{i,anc}$
- $r_{i,c,anc}$ is the LD correlation between SNP *i* and the causal SNP *C* in a specific ancestry
- β_c is the causal effect, which is assumed to be identical across ancestries

- Assuming the estimation error of effect sizes are independent across ancestries, it follows $e \sim \mathrm{N}(0,1)$

Note that in the equation (III.19), $p_{i,anc}$,$r_{i,c,anc}$, $q_{i,anc}$, and β_c are all population parameters. When there are no genotype missingness, $\beta_c\sqrt{D_{c,anc}}$ is constant for each ancestry. Assuming causal variants are shared across ancestries, $r_{i,c,anc}$ will reach its maximum (*i.e.*, 1) in all ancestries when the SNP *i* is either the causal SNP or in complete LD with causal SNP in all ancestries. Given that LD structures differ across ancestries, extending COJO algorithms across multiple ancestries can narrow the LD window, thereby improving resolution in identifying the true causal variants. However, as discussed previously, when per-SNP sample sizes vary, causal effect sizes are small, or causal SNPs are rare (i.e., low MAF, resulting in $D_{c,anc}$ values), the SNP selected by the COJO algorithm may not correspond to the SNP closest to the causal variant, due to the influence of the error term.

When there is more than one causal SNP in a region, equation (III.19) can be further extended to

$$z_{i,manc} = \sum_{anc}\sqrt{\frac{D_{i,anc}}{D}}\sum_{c=1}^{C} r_{i,c,anc}\beta_c\sqrt{D_{c,anc}} + \varepsilon \qquad \text{(III.20)}$$

Now the expression gets very complicated. Nonetheless, simulations show that the trans-ancestry COJO algorithm performs better than the single ancestry analysis of the same sample size in terms of finding independent associations, as shown in the results sections.

III.2.3 Simulations

We conducted a series of simulations to compare the trans-ancestry COJO analysis with the single-ancestry COJO analysis of the same sample size with multiple criteria.

III.2.3.1 Simulation settings

- Simulation Cohort

We used real genotype data from chromosome 22 of UK Biobank participants to establish our simulation framework. From all unrelated individuals (coefficient of relatedness ≤ 0.05, as calculated using GCTA) of inferred European ancestry (*N* ~ 340,000) and African ancestry (*N* = 6,901)—defined based on published ancestry projections[197]—we selected five cohorts. The African cohort (AFR1) comprises 6,901

unrelated participants of inferred African ancestry. European cohort 1 (EUR1) and cohort 2 (EUR2) each consist of 6,901 randomly selected, unrelated individuals of inferred European ancestry, with no overlap or relatedness between EUR1 and EUR2. We then combined EUR1 and EUR2 to form a cohort named EUR1_EUR2 (*N* = 13,802). Additionally, we randomly selected 13,802 unrelated European participants, independent of EUR1_EUR2, for out-of-sample prediction (named EURpred), and selected 2,000 African participants for out-of-sample prediction in African ancestry (named AFRpred), independent of AFR1. The sample sizes of EUR1 and EUR2 were matched to AFR1 because previous studies have shown that combining equal-sized European and African ancestry cohorts maximises power for detecting causal associations in multi-ancestry meta-analyses[198,199].

- Partitioned the genome

Previously, Berisa et al. partitioned the genome and defined LD-independent blocks separately for African and European ancestries[200], assuming that LD correlations between SNPs in distinct LD-independent blocks are negligibly different from zero. In this study, we first partitioned chromosome 22 separately for the African and European cohorts (N=33 and N=23 blocks, respectively). The blocks were partially overlapping, and N=52 overlapping sections were extracted (**Supplementary Figure III.1**). Within each of these 52 overlapping blocks, we calculated allele frequencies for all SNPs (using PLINK 1.9 --freq) in the AFR1, EUR1, EUR2, EURpred and AFRpred cohorts, retaining only SNPs with a minor allele frequency (MAF) >= 0.01 across all five cohorts. Blocks with fewer than 1,000 SNPs remaining after this filtering were excluded, resulting in 38 blocks for use in simulation.

- Simulation of Causal SNPs

Within each of the 38 selected blocks, we randomly selected 1, 5, 15, or 20 SNPs to be causal variants. The effect sizes (β_c) of these variants were simulated to be identical across ancestries, ensuring that the SNP heritability ($h_c^2 = 2p_c q_c \beta_c^2$) of each causal SNP c ranged between 0.5% and 1.5% in all European and African cohorts. However, heritability values could differ across ancestries owing to varying minor allele frequencies for the same SNP. To avoid collinearity, LD r^2 between all pairs of SNPs selected to be causal were required to be ≤ 0.8 in both ancestries. For each combination of block and number of causal variants, we repeated the simulation 50 times, using different sets of causal variants and effect sizes in each iteration. This resulted in 7,600 scenarios, accounting for 38 blocks, 4 different numbers of causal SNPs, and 50 variations of causal SNP sets and effect sizes.

- Simulation of Phenotypes

We simulated phenotypes for individuals within each cohort as

$$\mathbf{y} = \mathbf{X}_C\boldsymbol{\beta}_C + \mathbf{e}_C \quad \text{(III.21)}$$

Here,

- $\mathbf{y}$ is a *N* x 1 vector of simulated phenotypes.
- $\mathbf{X}_C$ is a *N* x *C* genotype, with entries encoded as 0, 1, or 2, corresponding to the number of alternative alleles at each SNP. Missing values in $\mathbf{X}_C$ are imputed as the mean value of genotypes of that SNP in that specific cohort.
- $\boldsymbol{\beta}_C$ is a *C* x 1 vector of simulated effect sizes as described in the previous section, and $\mathbf{e}_C$ is a *N* x 1 error vector, and each element of the vector is assumed to be drawn from a normal distribution N ~ (0, var($\mathbf{X}_C\boldsymbol{\beta}_C$)*($\frac{1-h^2}{h^2}$)), where, $h^2 = \sum_c 2p_c q_c \beta_c^2$ (β_c is *c*-th element of vecter $\boldsymbol{\beta}_C$) is the total SNP-based heritability.

We simulated phenotypes within each of six cohorts (AFR1, EUR1, EUR2, EUR1_EUR2, EURpred, AFRpred).

- GWAS simulation, and benchmark against existing method

For each of the 7,600 scenarios, we conducted a GWAS in four cohorts (EUR1, EUR2, EUR1_EUR2 and AFR1) using PLINK 2.0 (--glm hide-covar cols=+a1freq). Subsequently, we performed an inverse-variance meta-analysis of summary statistics from EUR1 and AFR1 using custom *R* code, naming the resulting cohort EUR1_AFR1. We then compared the outputs of three GWAS and model combinations: EUR1_EUR2 + Sanc-COJO, EUR1_AFR1 + Sanc-COJO, and EUR1 with AFR1 (as two separate inputs) + Manc-COJO. Default settings were applied across all scenarios for both Sanc-COJO and Manc-COJO, with in-sample LD used throughout. For the AFR1_EUR1 cohort, in-sample LD was calculated from the merged genotypes from AFR1 and EUR1 using PLINK 1.9 (--bmerge).

- Model performance measurement metrics

To assess model performance, we established several criteria to evaluate Manc-COJO outputs. First, to compare Manc-COJO outputs with true effects, we recorded how many COJO-selected SNPs were simulated as true causal SNPs and the proportion of COJO-selected SNPs that were causal. This criterion is justified because the primary objective of Manc-COJO is to construct a model with SNPs that have independent associations. However, in some instances, rather than selecting true causal SNPs, nearby SNPs in high LD with the true causal SNPs may be included in the final selected set, which still aligns with the goal of the method. To account for such

scenario, we calculated two additional measures as indicators for power and false positive discovery: we first build a *CxM* LD correlation matrix, where *C* is the number of true causal SNPs, and *M* is the number of COJO SNPs. For each of *M* rows, we pick the highest value, and then calculate the mean of this group of highest values. This can be viewed as an indicator for power, as it measures for each causal SNP the highest LD correlation to COJO SNPs. In an idealised situation, where all causal SNPs have been correctly identified, this number should be 1. Similarly, for each of *M* COJO SNPs, we calculated the mean of the highest value from each column, and subtracted the mean from 1. The resulting value provides an indicator of false-positive discoveries: it equals 0 when all COJO SNPs are causal SNPs and approaches 1 when COJO SNPs are less correlated with the causal SNPs. LD correlation matrices were calculated from both EUR1_EUR2 and AFR1 cohort.

Beyond comparisons with the ground truth, we also assessed out-of-sample prediction accuracies in independent European and African ancestry cohorts using joint SNP-effect sizes of COJO SNPs. Although the COJO algorithm is not designed to maximise out-of-sample risk prediction, comparing out-of-sample prediction using COJO selected SNPs is a criterion that can be applied in real data analysis as well as in simulation. We randomly divided the 13,800 EURpred individuals into 23 groups of 600, calculated polygenic scores for each group ($\mathbf{y}_{\mathrm{Pred}} = \mathbf{X}_{Pred}\boldsymbol{\beta}_{COJO}$), and fitted a linear model in R using lm() to obtain the adjusted coefficient of determination (R^2). We then computed the mean and standard error of these 23 regression coefficients. Similarly, we randomly divided the 2,000 AFRpred individuals into 20 groups of 100, and calculated the adjusted coefficient of determination.

III.2.3.2 Simulation sensitivity analysis

Section III.2.3.1 outlined our simulation strategies under idealised conditions, where the ancestry composition maximises power to detect causal associations, all causal SNPs are genotyped, causal effects are identical across ancestries, and the LD structure in the reference panel matches that of the cohort used for GWAS. However, in practice, these assumptions are often violated. To assess the impact of such violations, we conducted a series of simulations in which one assumption was altered at a time.

We first considered a scenario where European samples constitute the majority of the cohort. All simulation settings remained identical to those described in Section I, except that the EUR1 sample size was increased to 69,010. For all simulations in this scenario, in-sample LD reference panels were used. This means that the LD reference

for EUR1 was also larger than in Section I. As a result, the input for Manc-COJO was the GWAS summary statistics from EUR1 (*N* = 69,010) and AFR1 (*N* = 6,901). For single-ancestry Sanc-COJO, the input was GWAS from EUR1_EUR2 (*N* = 75,911). For multi-ancestry Sanc-COJO with ancestry-matched LD reference, the input was the meta-analysis of EUR1 and AFR1 GWAS, and the LD reference panel was the merged EUR1 and AFR1 cohort. We also examined an additional setting: multi-ancestry Sanc-COJO using a European-only LD reference panel. In this case, the input GWAS was the same meta-analysis of EUR1 and AFR1, but the LD reference was EUR1_EUR2.

In a second simulation scenario we dropped the assumption that all causal SNPs are genotyped. To achieve this, we kept all simulation settings identical to those described in Section I, with the sole exception that we restricted the analysis to HapMap3 SNPs (using gcta --extract and Manc-COJO --extract).

In the third simulation scenario we dropped the assumption of the in-sample LD reference panel. To do this, we downloaded genotype data from the 1000 Genomes Project and randomly sampled two sets of unrelated individuals (N = 503) from European participants, as defined by the 1000 Genomes Project, to serve as the LD reference for EUR. Similarly, we selected N = 503 African individuals as the LD reference for AFR. In the simulations, we ensured that all simulated variants were present in the genotype data from the 1000 Genomes Project.

In the final simulation scenario we allowed SNP effect sizes to differ across ancestries, although retaining direction of association. As in previous settings, all simulation parameters matched those in Section III.2.3.1, except that we relaxed the requiring identical effect sizes across ancestries. Instead, we sampled the true effect sizes independently in the two ancestries, while ensuring that the direction (sign) of effect was consistent.

III.2.3.3 Simulation for Manc-COJO--MDISA

Simulations in Sections III.2.3.1 and III.2.3.2 assumed shared causal SNPs across ancestries. In this study, we developed Manc-COJO --MDISA to account for scenarios where ancestry-specific causal variants exist, and tested it in simulation. In our simulations, 60% of the causal SNPs in each block were shared across ancestries, while 20% were specific to each of the two ancestries. For example, in a region with 10 simulated causal SNPs, 6 were shared, 2 were European-specific, and 2 were African-specific. We simulated scenarios with either 5, 10, or 20 causal SNPs per region. The effect sizes of shared causal SNPs were identical across ancestries.

Phenotype simulation followed the same procedure as described in Section I.4. In total, we conducted simulations in 38 blocks, across 3 different numbers of causal SNPs, with 50 replicates per combination, resulting in 5,700 simulated scenarios.

III.2.4 Real data analysis

To evaluate the performance of our algorithm using real-world data, we obtained ancestry-specific GWAS summary statistics of European and African ancestries for lipid traits and height. Specifically, we used data for four lipid traits—high-density lipoprotein (HDL), low-density lipoprotein (LDL), total cholesterol (TC), and log-transformed triglyceride levels (logTG)—from the GLGC consortium[201] and for height from the GIANT consortium[1].

III.2.4.1 Data quality control

We used GWAS summary statistics that excluded participants from the UK Biobank. Only common SNPs (minor allele frequency, MAF $\geq$ 0.01) present in both ancestries and bi-allelic SNPs were retained for analysis. Further filtering was conducted to retain only those SNPs that were genotyped in the UK Biobank across both ancestries and had a MAF > 0.01 among unrelated African participants (N = 6,901; genetic relatedness <5%). Following quality control, the median per-SNP sample sizes and the number of SNPs included for each trait were as follows: HDL (N_{AFR} = 90,804, N_{EUR} = 881,859, M = 8,134,172 SNPs); LDL (N_{AFR} = 87,759, N_{EUR} = 837,509, M = 8,131,191 SNPs); TC (N_{AFR} = 92,554, N_{EUR} = 924,769, M = 8,128,415 SNPs); logTG (N_{AFR} = 89,467, N_{EUR} = 858,015, M = 8,138,699 SNPs); Height (N_{AFR} = 104,384, N_{EUR} = 1,143,908, M = 1,169,127 SNPs).

III.2.4.2 Data input

For each trait, the inputs to the Manc-COJO algorithm consisted of two separate GWAS summary statistics (one for each ancestry) and two corresponding LD reference panels: a randomly sampled set of 69,010 unrelated European individuals and 6,901 unrelated African individuals from the UK Biobank. Default parameters were applied: a window size of 10Mb (assuming SNPs more than 10Mb apart to be independent), a collinearity threshold of 0.9, and a significance threshold of 5×10^{-8}, consistent with the original Sanc-COJO protocol.

We benchmarked Manc-COJO against Sanc-COJO applied to both trans-ancestry and single-ancestry datasets. For the trans-ancestry analysis, the input was the inverse-variance meta-analysis of the aforementioned European and African GWAS summary statistics from GIANT (for height) and GLGC (for lipid traits). The LD reference was either a merged cohort of 69,010 European and 6,901 African ancestry individuals, or a merged cohort of 69,010 European and 6,901 European ancestry individuals (using PLINK's –bmerge function). For the single-ancestry analysis, new GWAS were conducted on 100,000 unrelated European UK Biobank participants for height and 90,000 for each of the lipid traits. Covariate adjustment followed the original GIANT and GLGC protocols: for lipid traits, covariates were age, age-squared, batch, sex, assessment centre, and 20 genetic PCs. For individuals on cholesterol-lowering medications (Field IDs p6153_i0 and p6177_i0), LDL (LDL_{adj} = LDL/0.7) and TC (TC_{adj} = TC/0.8) levels were adjusted. For height, age at assessment, batch, sex, assessment centre, and 20 genetic PCs were included. Analyses were sex-stratified: residuals were computed separately for males and females and then pooled. Subsequently, we conducted an IVM of new GWAS summary statistics derived from UKB European participants and aforementioned European-specific GWAS, excluding UKB data, from the GLGC (lipid traits) or GIANT (Height) consortia. The LD reference for single-ancestry Sanc-COJO was a merged cohort of 69,010 and 6,901 unrelated European UK Biobank individuals. In general, we ensured that the input GWAS summary statistics involved approximately equal sample sizes and identical SNP sets, and LD reference panel sizes are comparable across methods for fair comparison (**Supplementary Table III.1**).

Given that samples from GIANT and GLGC were heavily skewed towards European ancestry (approximately 10:1 ratio of European to African participants after excluding UK Biobank samples), we conducted a down-sampling analysis to examine whether closer ancestry balance (approximately 1:1) would enhance detection of independent associations. In this analysis, we again applied three methods: Manc-COJO on multi-ancestral data, Sanc-COJO on multi-ancestral data (Sanc-COJO->Manc), and Sanc-COJO on single-ancestry data. For Manc-COJO, inputs comprised African-specific GWAS from GLGC or GIANT and new European-specific GWAS conducted using the UK Biobank samples (N = 100,000 for height and 90,000 for lipid traits) as previously described. For multi-ancestry Sanc-COJO, input GWAS were obtained via IVM of African and European-specific GWAS results used in Manc-COJO. The LD reference panel combined 6,901 unrelated European and 6,901 unrelated African participants. For single-ancestry Sanc-COJO, GWAS summary statistics were generated from N=180,000 for lipid traits and N=200,000 unrelated UKB European

samples for height, following the same analytical pipeline outlined in full-size analysis. The corresponding LD reference comprised 13,802 unrelated Europeans, excluding individuals used in the GWAS (**Supplementary Table III.1**).

III.2.4.3 Out-of-sample prediction

To obtain the weights for polygenic prediction and ensure a fair comparison across all methods, we first identified the list of SNPs selected by the Manc-COJO and Sanc-COJO models. These SNPs were then jointly re-estimated using the GCTA --cojo-joint command, based on the summary statistics and LD reference panels specified in **Supplementary Table III.1** (for full-sample analyses) and **Supplementary Table III.2** (for down-sampled analyses). After obtaining the joint effect sizes for each model, polygenic scores (PGS) were computed using the PLINK --score command.

To evaluate out-of-sample prediction performance in European cohorts, we included 47,066 unrelated European participants for lipid traits and 48,658 for height, none of whom were included in previous GWAS analyses. Prediction into African cohorts was assessed using 6,901 unrelated UKB African participants for height and 3,000 for lipid traits. Residuals for all traits were computed following the same covariate adjustments described for the GWAS analyses.

To estimate the prediction error, we randomly divided 47,066 unrelated UK Biobank European participants for lipid traits and 48,658 for Height into 50 groups, and 6,901 unrelated UK Biobank African participants for Height and 3,000 for lipid traits into 20 groups, with one group having more individuals if the sample size was not evenly divisible. Within each group, a simple linear regression (residual ~ PGS + error) was performed, and the out-of-sample R^2 was calculated as the coefficient of determination. The final R^2 estimate was averaged across groups for each trait and ancestry, and the standard error of the estimate was computed.

To compare genomic regions where the Manc-COJO and Sanc-COJO models identified different numbers of SNPs, we partitioned the genome into LD-independent blocks based on a previously published framework[187]. Note that the LD blocks used here differ from those in the simulation study, as we adopted a more conservative definition to reduce the likelihood of false-positive findings in the real data analysis. Specifically, we used a block definition that yields fewer but wider blocks. This reduces the risk that independent associations identified by different approaches are assigned to adjacent blocks purely due to block partitioning, when in

fact they reflect the same underlying causal signal. This does not matter in the simulation setting, as we know exactly where the causal variants are and only need to ensure that the same genomic regions are analysed across ancestries. For each individual LD block, the out-of-sample prediction accuracy of each COJO model (Manc-COJO or Sanc-COJO) was estimated by first removing the COJO-selected SNPs within that block, recalculating the joint effect sizes using the remaining SNPs, and then computing the difference in prediction accuracy between the full model and the reduced model.

III.3 Results

III.3.1 Overview of Manc-COJO

Manc-COJO is a forward and backward stepwise conditional and joint analysis method that extends the current Sanc-COJO algorithm into multi-ancestry settings. Illustrated in **Figure III.1** for two ancestries, Manc-COJO utilises GWAS summary statistics and LD reference matrices from each ancestry as input. Manc-COJO is equivalent (when an in-sample LD matrix is used) to a stepwise forward and backward variable selection algorithm applied to individual-level genotypic data. Briefly, in each iteration, the top-associated SNPs are regressed out from the original phenotype within each ancestry, and the regression residuals serve as the new phenotype for a subsequent GWAS within each ancestry (conditional analysis, conditioned on the top-associated SNPs). A new SNP is then selected for inclusion in the model based on p-values from the meta-analysis of GWAS of each individual ancestry (forward selection). Following selection of a new SNP, all previously chosen SNPs, together with the newly selected SNP, are fitted jointly in a multiple linear regression model using the original phenotype within each ancestry (joint analysis). The newly selected SNP is retained as a COJO-SNP based on the p-value in the meta-analysis of joint-effects and adjusted R^2 (coefficient of determination; an optional criterion) of the joint model within each ancestry. If its inclusion renders previously included SNPs insignificant in the meta-analysis of joint effects, those insignificant SNPs are removed, and the process continues (backward selection). This iterative process repeats until no further SNPs can be added or excluded from the model. Notably, Manc-COJO software has improved computational efficiency over the Sanc-COJO implementation in GCTA and has been developed with an option to take LD matrices as reference input (as opposed to individual level data to build the LD matrices).

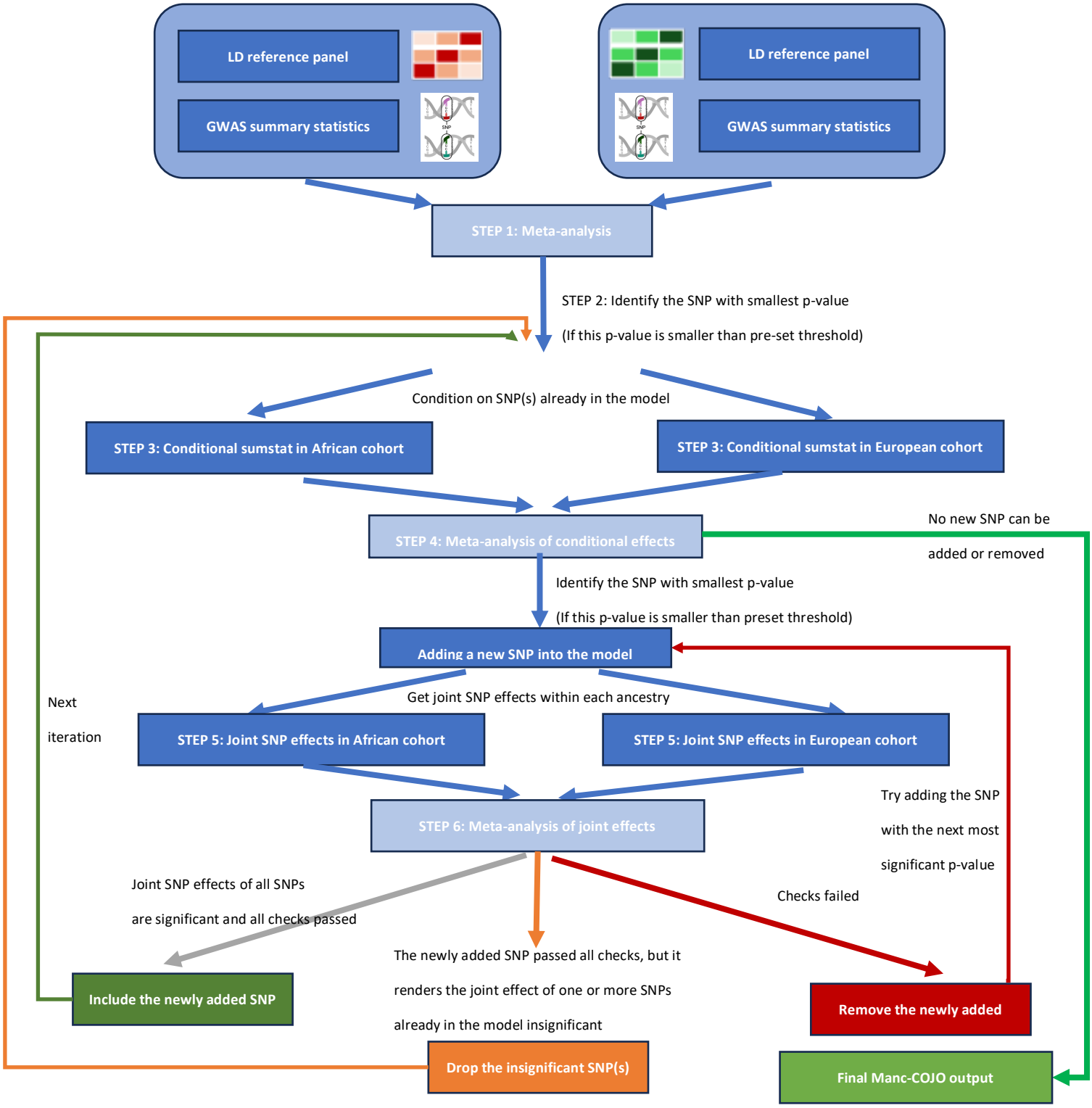


Figure III.1. A schematic overview of the Manc-COJO algorithm. To facilitate the integration of multi-ancestry data, Manc-COJO is a a forward and backward stepwise conditional and joint analysis method that extends the current Sanc-COJO algorithm into trans-ancestry settings. Manc-COJO utilises GWAS summary statistics and LD references from different ancestries as input (illustrated here with two ancestries), enabling the use of LD matrices as reference inputs to support the global sharing of accurate LD structures.

III.3.2 Simulation study results

We compared the model output (hereafter referred to as "COJO SNPs") from three model applications: 1) Manc-COJO applied to GWAS summary statistics from two ancestries (European and African) each of sample size (N=6,900) 2) Sanc-COJO applied to the same meta-analysed GWAS summary statistics, denoted as Sanc-COJO->Manc 3) Sanc-COJO applied to European ancestry GWAS of the total sample size (N=6,900*2) (**Table III.1**). Simulations used real genotype data from the UK biobank, hence sample sizes were limited by participant numbers of African ancestry. Simulations used chromosome 22 divided into 38 LD blocks[200] with 1, 5, 10, or 20 randomly chosen causal SNPs assigned into each LD blocks. COJO SNPs were compared to the simulation ground truth (hereafter referred to as "causal SNPs") using various criteria (see **Methods**) across 7,600 simulated scenarios with results summarised in **Figure III.2**. Trends within individual LD blocks are generally consistent with the pooled results (**Supplementary Figure III.2**).

Under the idealised simulation conditions (i.e., well-powered GWAS, all causal SNPs genotyped, true SNP effects equal in both ancestries, and LD structure derived from the GWAS cohort), analyses incorporating multi-ancestry data (i.e. Manc-COJO and Sanc-COJO->Manc) identify a greater number of causal SNPs (**Figure III.2a,b**), exhibit superior statistical power (**Figure III.2c,e**), and generate fewer false-positive discoveries (**Figure III.2d,f**) compared to analyses of equivalent sample sizes within a single ancestry (i.e., Sanc-COJO). Contrasting the two multi-ancestry approaches, the median numbers of causal SNPs tagged by Manc-COJO versus Sanc-COJO->Manc and power are similar, but Manc-COJO produces fewer false-positive discoveries (**Figure III.2d,f**) which becomes more evident when the simulated trait is highly polygenic (*i.e.*, 10 or 20 causal SNPs within a block).

Color-coding	Analysis description	GWAS data	LD reference
Manc-COJO	Manc-COJO on multi-ancestry data	Multi-ancestry GWAS; separate ancestry GWAS are needed	Separate LD reference for each ancestry
Sanc-COJO	Sanc-COJO on single ancestry data with European-only LD reference	Single ancestry EUR GWAS	EUR LD reference
Sanc-COJO -> Manc	Sanc-COJO on multi-ancestry data with multi-ancestry LD reference	Multi-ancestry GWAS meta-analysis (with known proportions of the different ancestries)	Single LD reference constructed from genotype data with known proportions of each ancestry
Sanc-COJO -> Manc/EURref	Sanc-COJO on multi-ancestry data with European-only LD reference	Multi-ancestry GWAS meta-analysis	EUR LD reference

Table III.1. Summary of implemented methods. This table summarises the methods compared in this study with color-coding matched to box plots used in Figures and Supplementary Figures. It clarifies the GWAS inputs and LD reference panels used for each method

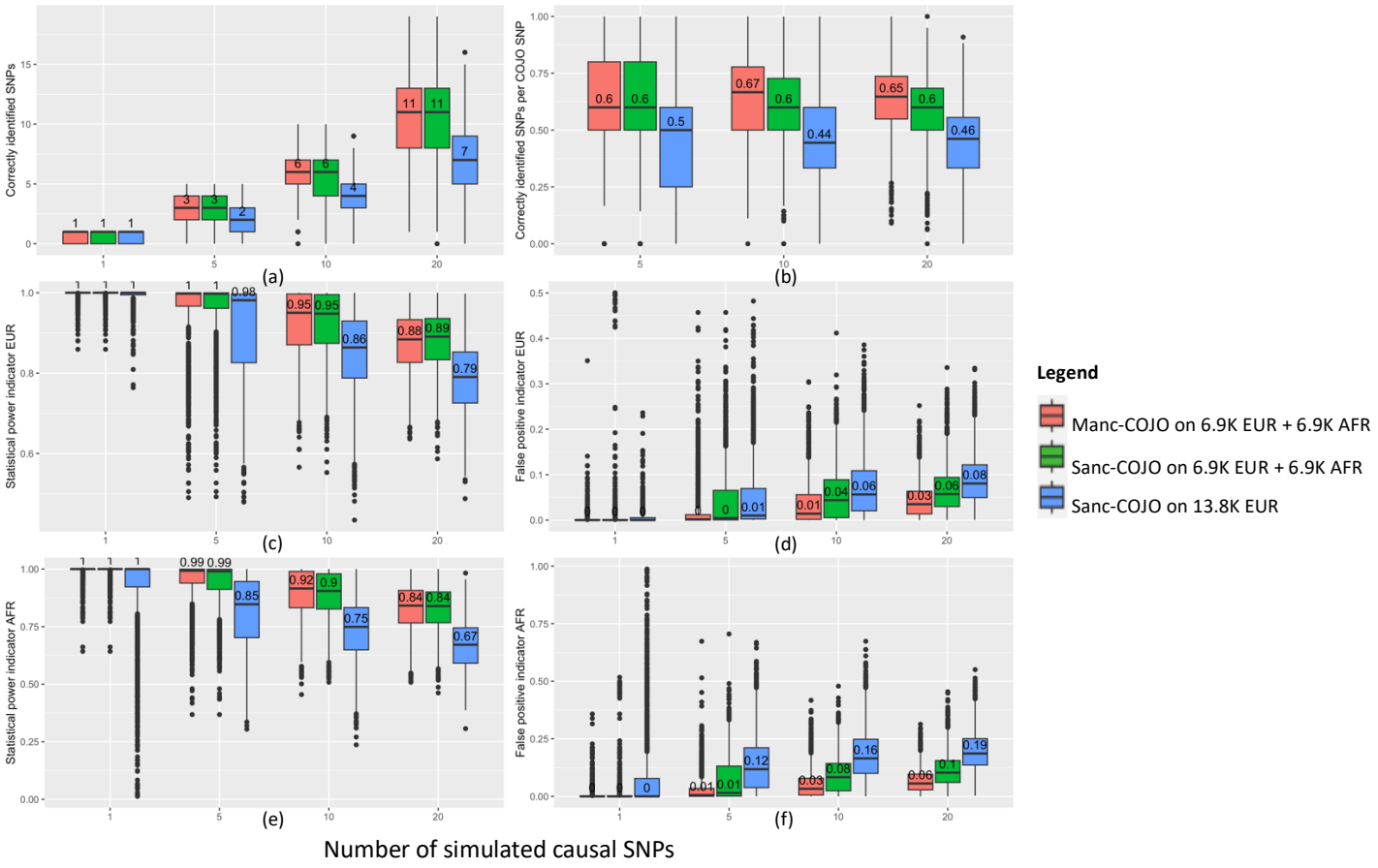


Figure III.2. Manc-COJO on average identifies a similar number of causal SNPs, but exhibits superior power, and shows fewer false-positive discoveries in simulations. This figure summarises simulation results of 7,600 scenarios, and illustrates the performance of Manc-COJO in identifying causal SNPs, showcasing superior power and fewer false-positive discoveries. The algorithm was applied to both a simulated European (EUR) and a simulated African (AFR) cohort, each with a sample size of 6,900. The comparisons include Sanc-COJO applied to either 13,800 Europeans or a meta-analysis of 6,900 Europeans with 6,900 Africans. In all cases, in-sample LD was used. The x-axis represents the number of simulated causal SNPs. (a) The y-axis denotes the number of causal SNPs correctly identified by each algorithm. (b) The y-axis denotes the proportion of SNPs identified by the COJO algorithm that are causal SNPs. (c) Instead of identifying the causal SNP, the COJO SNPs maybe located in nearby locations. We identified the COJO-SNP with the highest LD correlation (in European cohort) to each causal SNP and averaged these correlations across all causal SNPs. Higher value indicates better power in identifying causal variants. (d) For each COJO SNP, we assessed identified the closest true causal SNP, calculated the (European) LD between them andaveraged these LDs across COJO SNP instances, providing an indicator of false-positive discoveries (*i.e.*, instances where SNPs not in LD with causal SNPs were incorrectly identified will result in lower averaged LD-correlation). Higher value indicates higher false positive discoveries. (e)(f) Power and false positive indicators as in (c) and (d), except that African LD is used in calculations. Numbers in each plots are the median value of each group.

Out-of-sample prediction using the SNPs selected by a COJO model is an objective measure to compare validity of the SNPs selected by the different models. Despite Sanc-COJO resulting in a model with fewer true causal SNPs and more false-positive discoveries than Manc-COJO (and Sanc-COJO->Manc) from GWAS of the same total sample size, its out-of-sample prediction performance is comparable to that of Manc-COJO when predicting into an independent cohort of the same ancestry (**Figure III.3**). Hence, Sanc-COJO SNPs have good out-of-sample prediction accuracy within an independent sample sharing a similar LD structure to the GWAS samples. However, Sanc-COJO displays relatively low cross-ancestry portability, because the single ancestry COJO SNPs tag true causal SNPs less accurately than a multi-ancestry GWAS of the same total sample size which exploits differences in LD correlation between SNPs across different ancestries[197].

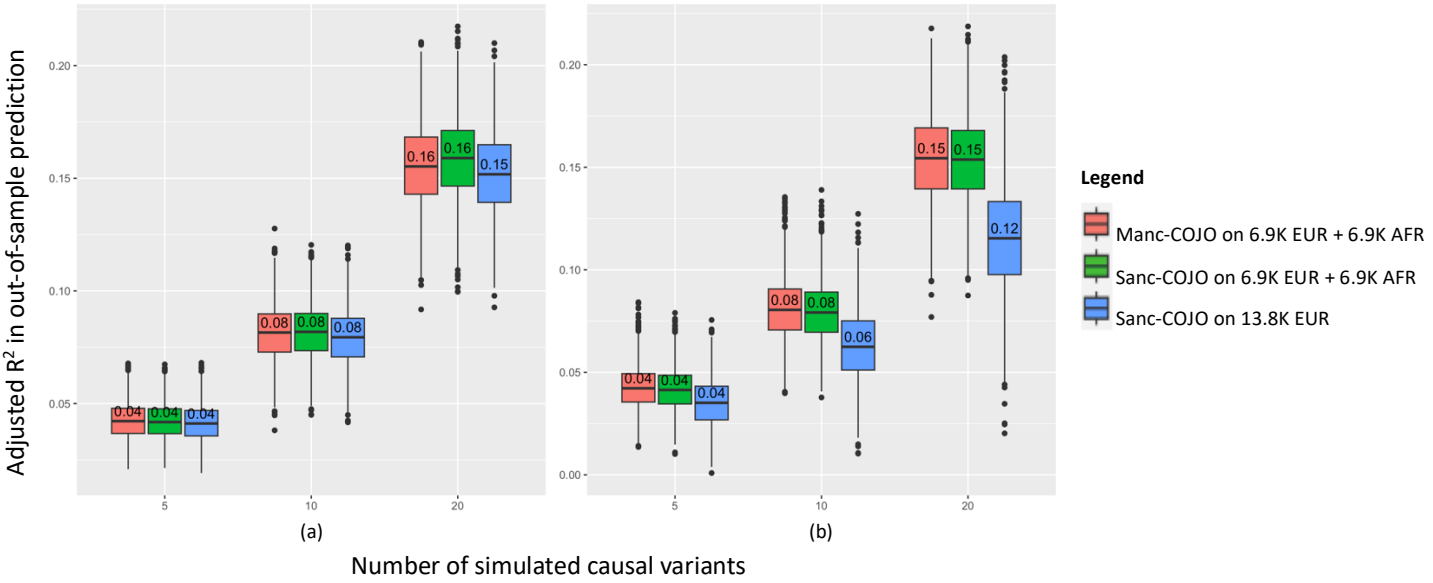


Figure III.3. Application of Sanc-COJO to single-ancestry simulation yields a model with good out-of-sample prediction within the same ancestry but relatively low cross-ancestry portability in simulations. This figure summarises simulation results from 7,600 scenarios, illustrating the out-of-sample prediction performance of different COJO methods. SNP effect sizes were generated under the simulation settings presented in **Figure III.2** and used to predict outcomes in an independent European cohort (*N* = 13,800) and an African cohort (*N* = 2,100). The x-axis represents the number of simulated causal SNPs. (a) The y-axis indicates the adjusted R^2 when predicting into the European cohort. (b) The y-axis indicates the adjusted R^2 when predicting into the African cohort. Numbers in each plot represent the median value for each group. When Sanc-COJO is applied to single-ancestry data, the resulting model demonstrates out-of-sample prediction accuracy into an independent sample similar to Manc-COJO, yet it exhibits limited portability across ancestries.

III.3.3 Simulation sensitivity analysis

In practice, model assumptions are often violated. To investigate the impact of such violations, we conducted a series of simulations in which one assumption was altered at a time. As expected, the performance of all models declined under non-

idealised conditions. Nonetheless, Manc-COJO consistently demonstrated greater robustness compared to the Sanc-COJO-based methods.

The scarcity of multi-ancestry data remains a significant challenge to date. We simulated a scenario in which the European cohort sample size was increased ten-fold to 69,010. Compared with Sanc-COJO-based methods, Manc-COJO, on average, identifies a greater number of causal SNPs, exhibits superior power, and produces fewer false-positive discoveries in simulations (**Figure III.4**). This demonstrates that, even when sample sizes are heavily skewed towards the European cohort,

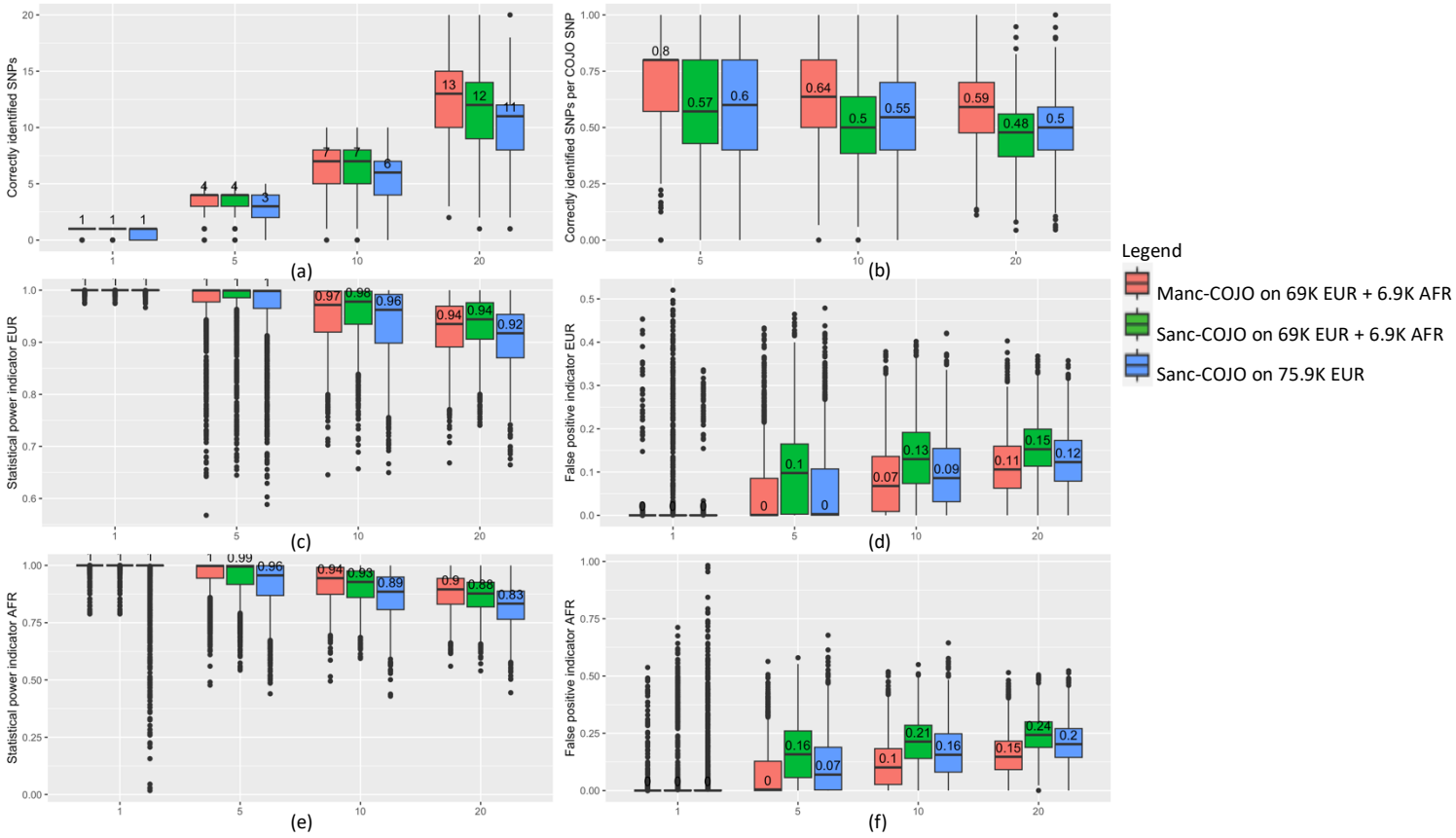


Figure III.4. **Manc-COJO demonstrates superior performance in simulations with uneven sample sizes across ancestries.** This figure summarises simulation results from 7,600 scenarios. We applied our algorithm to a simulated European cohort (*N* = 69,000) and a simulated African cohort (*N* = 6,900), comparing these results with Sanc-COJO applied to either 75,900 Europeans or a meta-analysis of 69,000 Europeans and 6,900 Africans. In-sample LD was used in all scenarios. The x-axis represents the number of simulated causal SNPs. (a) The y-axis indicates the number of causal SNPs correctly identified by each algorithm. (b) The y-axis represents the proportion of identified SNPs that are causal. (c) As COJO algorithms may identify SNPs (termed COJO-SNPs) near the causal SNP rather than the causal SNP itself, we identified the COJO-SNP with the highest LD correlation (in the European cohort) to each causal SNP and averaged these correlations across all causal SNPs; higher values reflect greater power in detecting causal variants. (d) For each COJO-SNP, we calculated its LD (in the European cohort) with the nearest true causal SNP and averaged these values across all COJO-SNPs, providing an indicator of false-positive discoveries (i.e., lower averaged LD correlations indicate instances where SNPs not in LD with causal SNPs were incorrectly identified); higher values suggest more false-positive discoveries. (e)(f) Power and false-positive indicators as in (c) and (d), respectively, but calculated using African LD. Numbers in each plot represent the median value for each group. This figure illustrates that Manc-COJO, on average, identifies a greater number of causal SNPs, exhibits superior power, and produces fewer false-positive discoveries when sample sizes across ancestries are unequal.

incorporating data from the African cohort still enhances the ability to detect independent associations with fewer false-positive discoveries. Notably, the application of Sanc-COJO→Manc—even when using an LD matrix constructed from the same individuals included in the analysis (i.e. in-sample mixed-ancestry LD)—results in a higher number of false-positive discoveries. Specifically, the proportion of COJO-selected SNPs that are truly causal is 0.57, 0.50, and 0.48 for regions with 5, 10, and 20 causal SNPs, respectively. These values are not only lower than those obtained with Manc-COJO (0.80, 0.64, and 0.59 for the same scenarios) but also lower than those from single-ancestry Sanc-COJO (0.60, 0.55, and 0.50 for the same scenarios).

Causal SNPs are not always directly genotyped in practice. To simulate this scenario, we retained the same settings as under the idealised condition but restricted the GWAS analysis to HapMap3 SNPs. All three methods performed reasonably well in European ancestry, with the resulting models identifying SNPs at similar LD distances from the true causal variants. However, Manc-COJO produced fewer false-positive findings. The false-positive indicator was consistently lower for Manc-COJO compared with Sanc-COJO→Manc: when calculated using LD from European ancestry samples, the values were 0.03 vs 0.04 for regions with 1 causal SNP, 0.09 vs 0.11 for 5 SNPs, 0.11 vs 0.14 for 10 SNPs, and 0.13 vs 0.16 for 20 SNPs. Using LD from African ancestry samples, the corresponding values were 0.17 vs 0.18, 0.21 vs 0.25, 0.23 vs 0.27, and 0.25 vs 0.30. Single-ancestry Sanc-COJO showed comparable power to both Manc-COJO and Sanc-COJO→Manc when evaluated with European ancestry LD, but—as expected—displayed reduced power and higher false-positive rates when assessed with African ancestry LD (**Supplementary Figure III.3**).

In our modelling framework, we initially assumed that the true effect sizes of causal SNPs were identical across ancestries. To assess robustness under non-idealised scenarios, we relaxed this assumption in simulation, while keeping all other parameters consistent with the idealised scenarios. In all simulated conditions, the effect sizes of the causal SNPs varied across ancestries; however, the direction of association remained consistent (see **Methods**). Under these settings, Manc-COJO consistently outperformed both Sanc-COJO and Sanc-COJO->Manc with in-sample, mixed-ancestry LD reference by identifying a greater number of causal SNPs, demonstrating higher statistical power, and yielding fewer false positive discoveries (**Supplementary Figure III.4**).

Finally, we simulated scenarios with imperfect LD references. The simulation settings were identical to those of the idealised scenarios, with the sole exception that, instead of using in-sample LD, the LD references were derived from the 1000 Genomes Project. For this, we randomly selected *N* = 503 unrelated European individuals and *N* = 503 unrelated African individuals as a reference sample (see **Methods**). Generally, the trends remain consistent with the idealised scenario: Manc-COJO or Sanc-COJO->Manc identify a greater number of causal SNPs and exhibit superior statistical power compared to Sanc-COJO, while among the two multi-ancestry approaches, Manc-COJO generates fewer false-positive discoveries. However, compared to the scenario using in-sample LD, the ability to detect independent associations diminishes across all methods. For example, the median number of true causal SNPs identified by Manc-COJO decreases from 6 to 5 and from 11 to 8 when there are 10 and 20 causal SNPs in a region, respectively. Concurrently, the median false-positive indicator rises sharply, increasing from 0 to 0.01, 0.01 to 0.09, and 0.03 to 0.19 when there are 5, 10, and 20 causal SNPs in a region, respectively (**Supplementary Figure III.5**). This underscores the benefit of using in-sample LD.

Our final sensitivity analysis mimicked studies which had sample sizes biased towards European population and so used the Sanc-COJO->Manc approach but used European-only LD references (hereafter referred to as Sanc-COJO->Manc/EURref)[172,174]. In simulations, when European ancestry dominates the sample (10:1 ratio), the approach performs reasonably well, achieving higher power in both ancestries and making fewer false positive discoveries in African ancestry, compared to Sanc-COJO applied to European-ancestry GWAS of the same total sample size (Manc-COJO is still superior). When the ancestry balance is equal (1:1 ratio), the use of a European-ancestry LD reference becomes ineffective, leading to marked inflation in false-positive discoveries, particularly in African ancestry samples (**Supplementary Figure III.6**).

III.3.4 Manc-COJO--MDISA

All simulations described above assumed shared causal SNPs across ancestries. We developed a complementary approach Multi-Ancestry Data-Informed Single Ancestry Analysis (Manc-COJO --MDISA) which uses multi-ancestry information to increase power to identify variants which may have ancestry-specific effects. We designed a series of simulated scenarios in which 60% of the simulated SNPs in a block are common across ancestries, while 20% are specific to each of the two ancestries (see

Methods). Both multi-ancestry algorithms (Manc-COJO and Sanc-COJO->Manc) detect more shared causal variants than Sanc-COJO of the same total GWAS sample

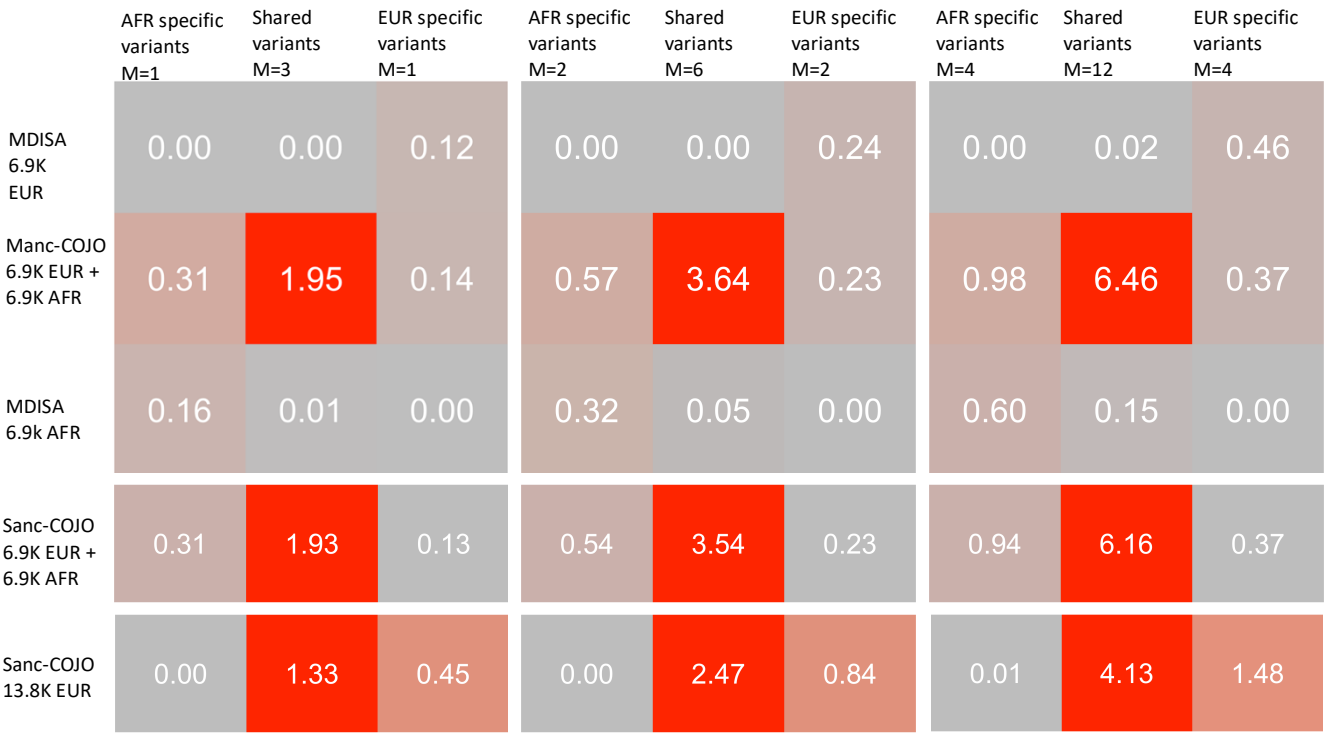


Figure III.5. Manc-COJO identifies a greater number of shared causal SNPs in scenarios where ancestry-specific trait-associated variants exist, while Manc-COJO --MDISA recovers some ancestry-specific associations. This figure summarises 5,700 simulated scenarios across 38 blocks on chromosome 22, where 60% of the simulated SNPs in each block are common across ancestries and 20% are specific to each of the two ancestries. Column labels indicate the number of simulated SNPs (M) and whether they are shared across ancestries or ancestry-specific, while row labels correspond to the analyses performed. The numbers in the boxes are the mean number, where the mean values are calculated by pooling results across 38 blocks and 1,900 simulation replicates for each specified number of causal SNPs (5, 10, or 20). Consistent with prior simulations, we applied Manc-COJO to a simulated European cohort (N = 6,900) and a simulated African cohort (N = 6,900). Building on the Manc-COJO output, we subsequently applied Manc-COJO --MDISA within each ancestry. Comparisons include Sanc-COJO applied to either 13,800 Europeans or a meta-analysis of 6,900 Europeans and 6,900 Africans, with in-sample LD utilised in all cases. Numbers in the plot represent the mean number of causal SNPs correctly identified by each algorithm within each category, with a colour gradient reflecting these values (red for higher numbers, grey for lower). Both multi-ancestry algorithms—Manc-COJO and Sanc-COJO applied to multi-ancestry data—detect more shared causal variants than a single-ancestry approach with the same sample size, with Manc-COJO outperforming Sanc-COJO on multi-ancestry data. By applying Manc-COJO --MDISA to the Manc-COJO output, we are able to recover some additional ancestry-specific associations.

size, with Manc-COJO outperforming Sanc-COJO->Manc. Applying Manc-COJO --MDISA to the Manc-COJO output recovers around 10% - 20% additional ancestry-specific causal variants (**Figure III.5**).

III.3.5 Real data analysis

In real data analysis, since the true causal SNPs and effect sizes are unknown, we use out-of-sample prediction accuracy from the COJO SNPs and model size as the primary

criteria for comparing Sanc-COJO and Manc-COJO models. We analysed summary statistics from a GWAS meta-analysis of HDL, LDL, TC, and logTG including approximately 1.1 million samples (1 million EUR + 0.1 million AFR for multi-ancestry models, and 1 million EUR + 0.1 million EUR for single ancestry models; **Supplementary Table III.1**). We selected these traits because they were previously used to evaluate the multi-ancestry fine-mapping approach, MEsusie[202]. and the

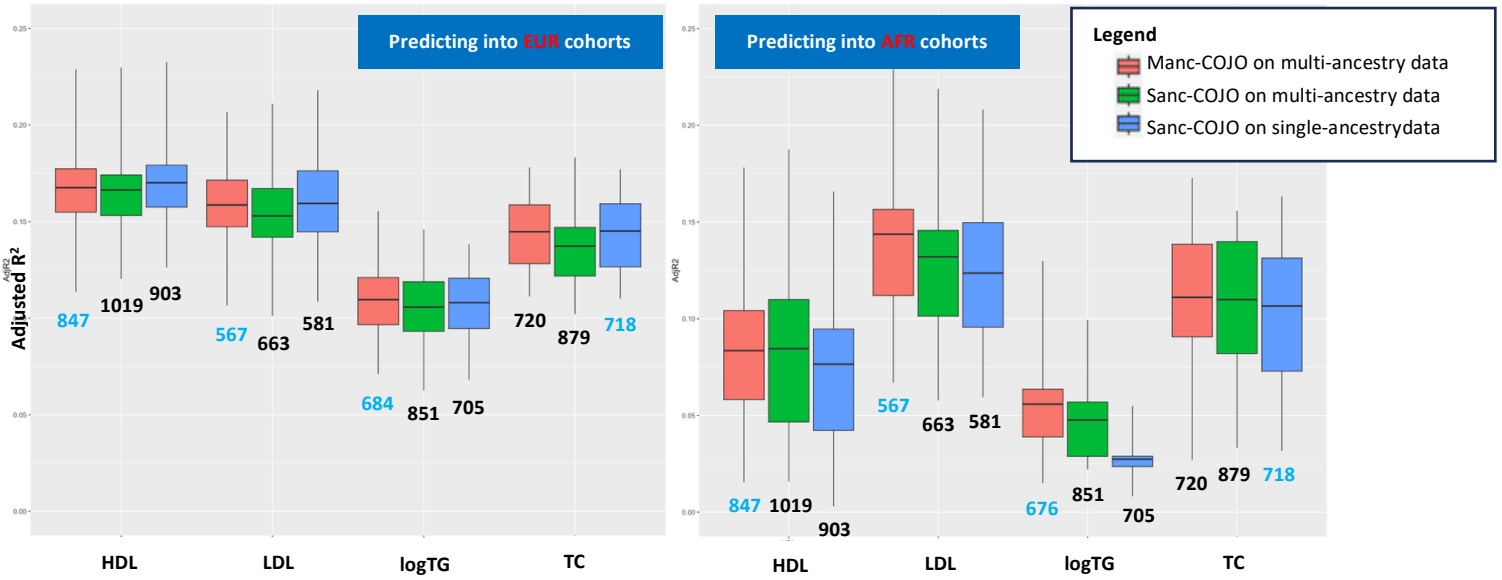


Figure III.6. Manc-COJO selected fewer COJO-SNPs with superior trans-ancestry portability compared to the single-ancestry Sanc-COJO model. This figure summarises the out-of-sample prediction accuracies of Manc-COJO on multi-ancestry data, Sanc-COJO on multi-ancestry data, and Sanc-COJO on single-ancestry data for HDL, LDL, logTG, and TC when predicting into independent European and African cohorts. Numbers below each boxplot are the number of independently associated SNPs outputted from the algorithms, and the model with the smallest number of SNPs is highlighted in blue for each trait. (i) Manc-COJO input: European-specific GWAS summary statistics from GLGC (N ~ 0.9M), combined with African-specific GWAS from GLGC (N ~ 90,000). All GWAS summary statistics excluded UK Biobank participants. The LD reference panels consisted of 69,010 unrelated European individuals and 6,901 unrelated African individuals from the UK Biobank. (ii) Sanc-COJO on multi-ancestry data input (Sanc-COJO->Manc): Meta-analysis of the European-specific and African-specific GWAS summary statistics described above. The LD reference panel was a merged set of the unrelated European and African individuals from the UK Biobank (69,010 Europeans and 6,901 Africans). (iii) Sanc-COJO on single-ancestry data input: European-specific GWAS used in i) meta-analysed with European-specific GWAS of additional independent UK Biobank participants (N = 90,000) using pipelines consistent with the original GLGC studies. The LD reference panel merged the 69,010 unrelated Europeans with an additional independent sample of 6,901 unrelated Europeans from the UK Biobank. To ensure fair comparisons across methods, we harmonized the GWAS summary statistics to have identical SNP sets and approximately equal sample sizes, and we kept identical LD reference panel sizes across methods.

GWAS have both European and African ancestry GWAS that exclude the UK biobank, allowing the UK biobank samples to be used for evaluation of the COJO SNPs selected from each model.

We first assessed the prediction accuracy of single ancestry Sanc-COJO- and Manc-COJO-derived predictors in an independent sample of 47,066 individuals of European ancestry. The prediction R^2 (and standard error) of the Manc-COJO model was 0.168

(0.022), 0.159 (0.021), 0.145 (0.019), and 0.110 (0.018) for HDL, LDL, TC, and logTG, respectively. The R^2 of the Sanc-COJO were similar: 0.170 (0.022), 0.159 (0.023), 0.145 (0.020), and 0.108 (0.017) for the same traits. When applied to an independent AFR cohort of 3,000 individuals, the Manc-COJO model consistently gave higher R^2 values (although not statistically different given the sample size available) than Sanc-COJO: 0.084 (0.044) vs 0.077 (0.047) for HDL, 0.144 (0.053) vs 0.124 (0.041) for LDL, 0.111 (0.046) vs 0.106 (0.042) for TC, and 0.0559 (0.031) vs 0.0274 (0.013) for logTG. Notably, the number of independently associated SNPs identified by Manc-COJO was generally lower than that of Sanc-COJO model: 847 vs 903 for HDL, 567 vs 581 for LDL, 720 vs 718 for TC, and 684 vs 705 for logTG (**Figure III.6**). In summary, Manc-COJO achieved comparable predictive accuracy in the EUR sample and superior performance in the AFR sample, while selecting fewer or an equal number of SNPs, suggesting a lower rate of false-positive inclusions relative to the Sanc-COJO model.

To assess whether a more balanced ancestry composition (~1:1) in GWAS sample sizes would enhance the detection of independent associations, we down-sampled the EUR GWAS to match the sample size of the AFR GWAS (**Supplementary Table III.2**). Given the reduced GWAS sample sizes the R^2 values in all analyses were naturally lower. In EUR cohorts, the out-of-sample R^2 (and standard error) for Manc-COJO models were lower than that of Sanc-COJO models: 0.125 (0.023) vs 0.134 (0.020) for HDL, 0.137 (0.022) vs 0.144 (0.024) for LDL, 0.126 (0.022) vs 0.130 (0.020) for TC, and 0.076 (0.013) vs 0.083 (0.015) for logTG. Nonetheless, in African cohorts, Manc-COJO models demonstrated consistently better trans-ancestry portability, achieving higher out-of-sample R^2 values than the Sanc-COJO models: 0.090 (0.047) vs 0.073 (0.032) for HDL, 0.140 (0.048) vs 0.128 (0.045) for LDL, 0.121 (0.048) vs 0.116 (0.047) for TC, and 0.046 (0.027) vs 0.012 (0.0080) for logTG. Notably, as the ancestry balance approached parity, the difference in model size between the methods became more pronounced: Manc-COJO consistently yielded more compact models with fewer selected SNPs—140 vs 182 for HDL, 117 vs 141 for LDL, 138 vs 157 for TC, and 97 vs 144 for logTG (**Supplementary Figure III.7**). Overall, while Sanc-COJO models achieved slightly higher predictive accuracy in EUR cohorts under the down-sampled setting Manc-COJO models still maintained reasonable accuracy in European cohorts and outperformed Sanc-COJO models in African cohorts, using up to 30% fewer SNPs. This suggests that a more balanced ancestry composition may enhance Manc-COJO's ability to detect independent associations.

In previous attempts to adapt Sanc-COJO for multi-ancestry GWAS, an *ad hoc* approach has employed LD matrices estimated from samples whose ancestry composition matches that of the GWAS sample, with the choice of LD matrix largely driven by empirical considerations[176,177]. Using this approach, Sanc-COJO->Manc models yielded slightly lower—but generally comparable—prediction accuracies in both EUR and AFR cohorts across full-sample and down-sampled analyses. However, the resulting models were substantially larger. For example, in the down-sampled analysis, the number of SNPs selected by the Sanc-COJO->Manc model was 261 for HDL, 215 for LDL, 218 for logTG, and 256 for TC—nearly twice the size of the corresponding Manc-COJO models (**Figure III.6; Suppplementary Figure III.7**). Despite this increased complexity, predictive performance was marginally lower compared to Manc-COJO across all traits and scenarios. These findings suggest that Sanc-COJO->Manc, an *ad hoc* application of Sanc-COJO to multi-ancestry meta-analysis, tends to include a higher number of false-positive SNPs relative to the more parsimonious Manc-COJO approach.

Researchers have also attempted to apply Sanc-COJO to meta-analyses of multi-ancestry GWAS using EUR-only LD reference panels, particularly in studies where EUR populations dominate the sample sizes. The results aligned with observations from our simulations. When the European-to-African (EUR: AFR) sample size ratio was 10:1, this approach was reasonably effective— Sanc-COJO->Manc/EURref produced models with comparable out-of-sample prediction accuracy in both European and African cohorts, and the model sizes remained similar. However, as the ancestry balance shifted to a 1:1 ratio, this approach was no longer viable. The number of SNPs selected by Sanc-COJO->Manc/EURref model increased substantially: 277 for HDL, 295 for LDL, 330 for logTG, and 165 for TC, approximately twice the size of the corresponding Manc-COJO models (**Figure III.7**). Moreover, for all traits except TC, these model sizes also exceeded those of Sanc-COJO->Manc models (which constructed with ancestry-proportion-matched LD reference panels), suggesting a higher rate of false-positive SNP inclusion.

III.3.6 Ancestry-specific variants in real data analysis

For the Manc-COJO model, we further conducted Manc-COJO --MDISA analysis (using ~7 million SNP) within the 1 million EUR and 0.1 million AFR cohorts separately, to examine ancestry-specific associations. The changes in out-of-sample prediction accuracy were minimal in both EUR and AFR ancestries (**Supplementary Figure III.8**). In AFR, compared with the Manc-COJO model, the increase in prediction

R^2 (Number of ancestry specific SNP) was 0.002 (15), 0.002 (22), -0.0004 (23), and

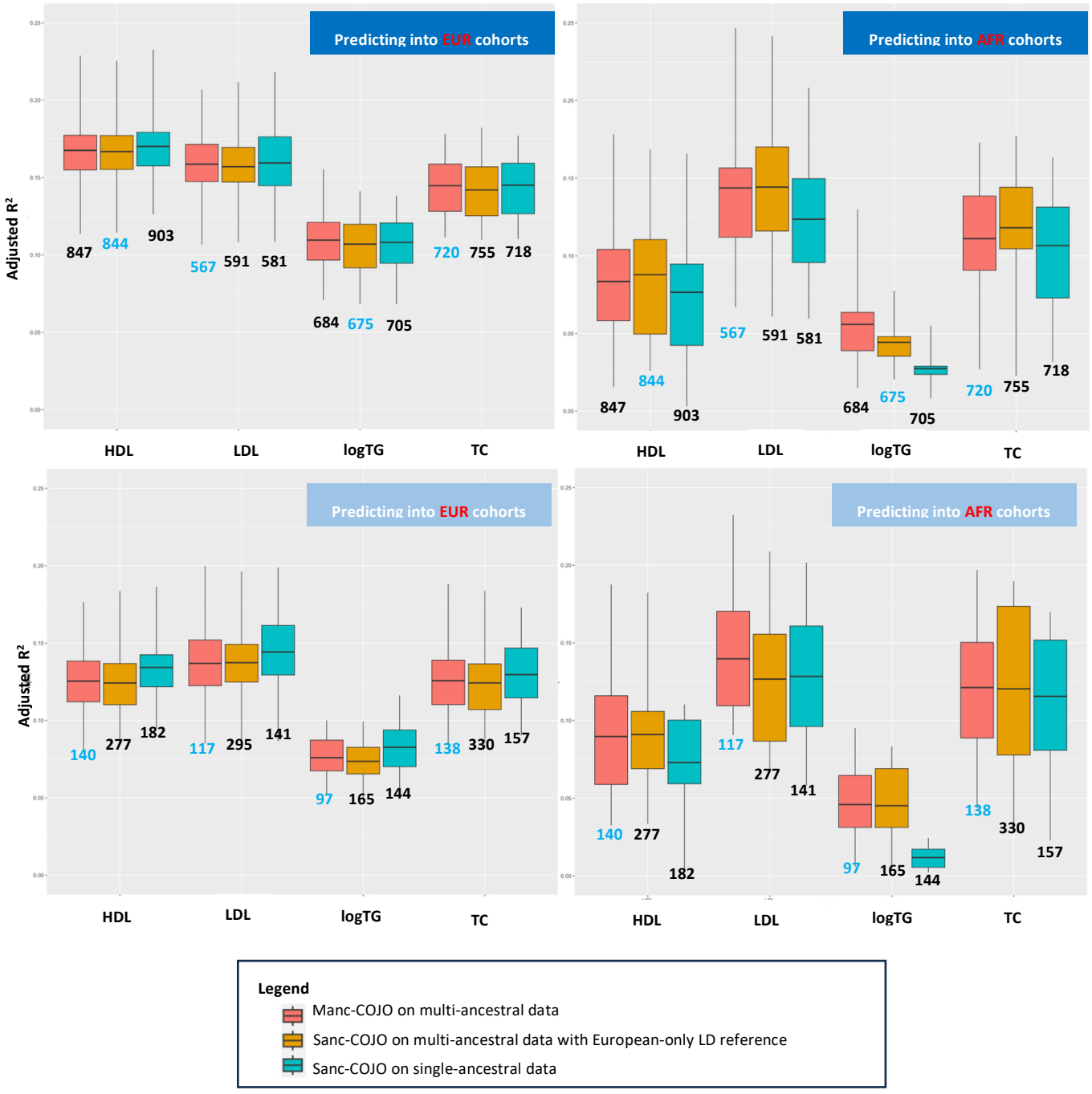


Figure III.7. Applying Sanc-COJO to meta-analyses of multi-ancestry GWAS using European-only LD reference panels is only effective when European populations dominate the sample. Published studies have applied Sanc-COJO to multi-ancestry GWAS meta-analyses using European-only LD reference panels, particularly when European individuals comprised the majority of the sample. In this figure, we apply Sanc-COJO using a European-only LD reference under two scenarios: a European:African (EUR:AFR) sample size ratio of 10:1 (top row) and 1:1 (bottom row), predicting into both an independent European cohort (left column) and an African cohort (right column). The results mirrored our simulation findings. When Europeans dominated the sample (10:1), the approach performed reasonably well. However, when the ancestry balance shifted to 1:1, the strategy became ineffective. The number of SNPs selected by Sanc-COJO increased markedly, indicating a higher rate of false-positive inclusions and reduced model parsimony.

0.005 (9) for HDL, LDL, TC, and logTG, respectively. In EUR, the corresponding values were −0.002 (55), −0.001 (28), 0.001 (24), and 0.002 (38). These results indicate that most of the prediction R^2 can be explained by a model based on common causal SNPs assumptions, and that some SNPs identified by MDISA may represent false-positive discoveries.

III.3.7 SNP density

All analyses above included approximately 7 million SNPs. However, sometimes in published GWAS the SNP density for summary statistics is lower. To evaluate the performance of the Manc-COJO model under such conditions, we repeated our analysis on 1.1 million samples (1 million EUR and 0.1 million AFR for the multi-ancestry analysis, or 1.1 million EUR for the single-ancestry analysis) of lipid traits, but restricting the analysis to 1.1M HapMap3 SNPs. We also included height, because the published multi-ancestry GWAS for this trait is limited to HapMap3 SNPs[1]. Out-of-sample prediction R^2 for Manc-COJO vs Sanc-COJO model when predicting into a European cohort: 0.160 (0.023) vs 0.162 (0.022) for HDL, 0.144 (0.022) vs 0.147 (0.021) for LDL, 0.135 (0.020) vs 0.135 (0.019) for TC, 0.102 (0.018) vs 0.106 (0.017) for logTG (as expected all values are lower than for the 7M SNP version), and 0.296 (0.018) vs 0.299 (0.018) for height. However, the Manc-COJO model retained better trans-ancestry portability, with higher prediction accuracies when predicting into an African cohort: 0.085 (0.044) vs 0.075 (0.041) for HDL, 0.094 (0.041) vs 0.081 (0.033) for LDL, 0.091 (0.042) vs 0.067 (0.033) for TC, 0.028 (0.018) vs 0.024 (0.013) for logTG, and 0.096 (0.017) vs 0.079 (0.021) for height. As for the 7M SNP analyses, the model sizes of Manc-COJO were smaller or roughly equal to those of Sanc-COJO models: 714 vs 807 for HDL, 475 vs 501 for LDL, 545 vs 606 for TC, 594 vs 629 for logTG, and 3,206 vs c for height. Again consistent with the analyses using ~7 million SNPs, Sanc-COJO->Manc/EURref remained comparable to Manc-COJO at a EUR:AFR sample size ratio of 10:1. In contrast, Sanc-COJO->Manc (uses an LD reference matched to ancestral proportions) included substantially more SNPs without improving prediction accuracy, suggesting the inclusion of false positives (**Supplementary Table III.3**).

III.3.8 Choice of LD window size

The default LD window for both the Manc-COJO and Sanc-COJO algorithms is 10 Mb, meaning that any pair of SNPs separated by more than 10 Mb is assumed to be LD-independent (i.e., have an LD correlation of 0). While this assumption holds for most

SNP pairs, it can break down in regions with long-range LD, such as the Major Histocompatibility Complex (MHC), where a 10 Mb window may be insufficient.

To address this, we also applied the Manc-COJO algorithm to the full-sample analysis using a 20 Mb LD window. When predicting into the European population, the prediction accuracies for the 20 Mb versus 10 Mb windows were 0.167 (0.022) vs 0.168 (0.022) for HDL, 0.159 (0.021) vs 0.159 (0.021) for LDL, 0.145 (0.019) vs 0.145 (0.019) for TC, and 0.109 (0.019) vs 0.110 (0.018) for logTG. For the African population, the prediction accuracies were 0.0850 (0.047) vs 0.0835 (0.044) for HDL, 0.141 (0.051) vs 0.144 (0.053) for LDL, 0.110 (0.047) vs 0.111 (0.046) for TC, and 0.0546 (0.027) vs 0.0559 (0.031) for logTG with 20 Mb and 10 Mb windows, respectively. Across lipid traits, the number of SNPs included in the model was generally reduced with the larger window size: 831 vs 847 for HDL, 567 vs 567 for LDL, 715 vs 720 for TC, and 676 vs 684 for logTG when comparing the 20 Mb to the 10 Mb window.

Notably, for HDL, applying a 20 Mb window resulted in a model with 16 fewer SNPs while maintaining similar prediction accuracy in Europeans and achieving improved prediction accuracy in Africans. In particular, using a 20 Mb window led the Manc-COJO model to identify 7 fewer SNPs within the block on chromosome 6 (base pairs 31,428,014–32,682,664, GRCh37 build) alone, highlighting the presence of long-range LD in this region.

Overall, while the default 10 Mb window performs well in most settings, assuming that LD in both ancestries decays to near zero beyond this range, certain regions and traits may benefit from alternative window sizes (with increase computing time). We therefore recommend that researchers consider testing different LD windows in regions known to exhibit long-range LD and/or large effect sizes. For regions with extremely complex LD structures, such as the MHC, as discussed in the main text, more specialized analytical tools may be required. In future study designs, window size selection could also be informed by recombination metrics such as genetic distance in centimorgans.

III.3.9 Choice of R^2 parameter

To evaluate the impact of applying different $\mathrm{R}^2_{\mathrm{Adj}}$ thresholds (relative to the default setting, in which the R^2_{Adj}threshold was set to −1 for both forward and backward selection), we tested two scenarios in both simulation and real-data analyses: (1)

R^2_{Adj} in both forward and backward selection was set to 0. In this setting, for a SNP to be included in the model, the R^2_{Adj} of the new model after forward and backward selections must be higher than that of the previous model in both ancestries. (2) R^2_{Adj} in forward selection was set to 0, and in backward selection set to −1. In this scenario, we imposed a threshold only for forward selection, requiring an increase in R^2_{Adj} to add a SNP, while allowing any decrease in R^2_{Adj} during backward selection. We performed the simulation with a EUR:AFR ratio of 10:1 to match the ancestry proportions used in the real-data analysis.

In general, applying a more stringent R^2_{Adj} threshold reduces the false positive inclusion of non-causal SNPs in the model, but at the cost of reduced power to include causal SNPs, especially when the number of causal SNPs increases (**Supplementary Figure III.9**).

The trend observed in the real-data analysis is consistent with the simulation results. Applying an R^2_{Adj} threshold of 0 in both forward and backward selection substantially reduced model size compared to the default setting with no R^2_{Adj}restriction. Specifically, model sizes decreased from 847 to 774 for HDL, 567 to 512 for LDL, 720 to 679 for TC, and 684 to 620 for logTG.

Applying an R^2_{Adj} threshold generally (though not always) led to a slight drop in out-of-sample prediction accuracy in both ancestries but produced a much more compact model. Of note, for TC, using an R^2_{Adj} threshold of 0 in both directions reduced the out-of-sample prediction accuracy in European ancestry from 0.145 (0.019) to 0.143 (0.020), but increased prediction accuracy in African ancestry from 0.111 (0.046) to 0.117 (0.043) with a model that included 6% fewer SNPs (**Supplementary Table III.4**).

Overall, while the default setting without an R^2_{Adj} threshold works well in most scenarios, adding such a threshold can help reduce false-positive discoveries in regions with highly complex LD structures. The optimal approach may depend on the specific genetic architecture of each trait.

III.3.10 Insights into Manc-COJO and Sanc-COJO outcomes

In order to understand the factors contributing to differences in model size between the Manc-COJO and Sanc-COJO models we partitioned SNPs identified in the full-sample analysis (using lipid GWAS traits, 7M SNPs, 1 million EUR + 0.1 million AFR for

Manc-COJO and 1 million EUR + 0.1 million EUR for Sanc-COJO) into LD-independent blocks as defined previously [187]. We then compared the number of SNPs included in the Manc-COJO model with those in the single-ancestry Sanc-COJO model within each LD-independent block (**Figure III.8**). For most LD blocks (specifically, 66.4% of blocks for HDL, 69.5% for LDL, 70.7% for TC, and 66.3% for logTG) the Manc-COJO and Sanc-COJO models identified an identical number of independent associations. The main reason the Manc-COJO model is generally smaller than the Sanc-COJO model is that in many LD blocks, the Sanc-COJO model selects one or occasionally two SNPs, whereas the Manc-COJO model does not select any SNPs in those blocks. Specifically, this pattern accounts for 52 SNPs across 49 blocks for HDL, 46 SNPs across 43 blocks for LDL, 38 SNPs across 34 blocks for TC, and 50 SNPs across 48 blocks for logTG (**Figure III.8**). These SNPs identified by the Sanc-COJO model make up approximately 5–7% of all SNPs in the model and are not rare variants, with mean minor allele frequencies of 0.27 for HDL, 0.28 for LDL, 0.26 for TC, and 0.28 for logTG in the European population. However, they collectively explain only a negligible proportion of phenotypic variance. The sum of out-of-sample prediction accuracies (**Methods**) for these SNPs when predicting into European cohorts was 0.00098 for HDL, 0.00055 for LDL, 0.0013 for TC, and 0.00037 for logTG. Prediction accuracies in African cohorts were even lower, at 0.00023, −0.0017, 0.00077, and −0.0012 for HDL, LDL, TC, and logTG, respectively. The fact that this large set of SNPs consists of common variants yet explains only a negligible or even negative proportion of phenotypic variance suggests that at least some of the independent associations tagged by the Sanc-COJO model in these blocks are more likely to represent false-positive signals rather than true causal SNPs with small effects.

Notably, despite Manc-COJO identifying fewer COJO SNPs than Sanc-COJO genome-wide, there are a few blocks where the Manc-COJO model identified markedly more SNPs than the Sanc-COJO model (**Supplementary Figure III.10-III.11**). In general, these blocks have complex LD structures. The most prominent example is the block located between base pairs 31,428,014 and 32,682,664 (GRCh37 build) on chromosome 6, encompassing part of the Major Histocompatibility Complex (MHC) region. Here, Manc-COJO tagged 8 and 12 more SNPs than Sanc-COJO for HDL and LDL, respectively. However, despite the identification of multiple independent associations in this region, the out-of-sample prediction accuracies of both models for these SNPs were near zero when predicting into cohorts of either ancestry (**Supplementary Table III.5**). While the lack of difference in prediction may be a reflection of low power given the number of SNPs involved and sample size, these results may reflect the inherent difficulty of detecting independent associations in

such regions. The MHC region is well known for its complex recombination patterns and multiple haplotypes[203,204]; even the reference genome itself may not reliably represent its structure[205]. Consequently, the additional independent associations identified in this region may be artefacts arising from differences in LD structure between or even within populations. See **Supplementary Table III.5** and **Supplementary Figure III.11** for other examples. While the Manc-COJO model generally performs well, analyses in regions such as the MHC may require dedicated tools and further investigation when larger sample sizes are available for non-European ancestries. When applying Manc-COJO in these regions, we recommend using larger LD windows and/or more stringent parameter settings, as discussed in the III.3.8 and III.3.9.

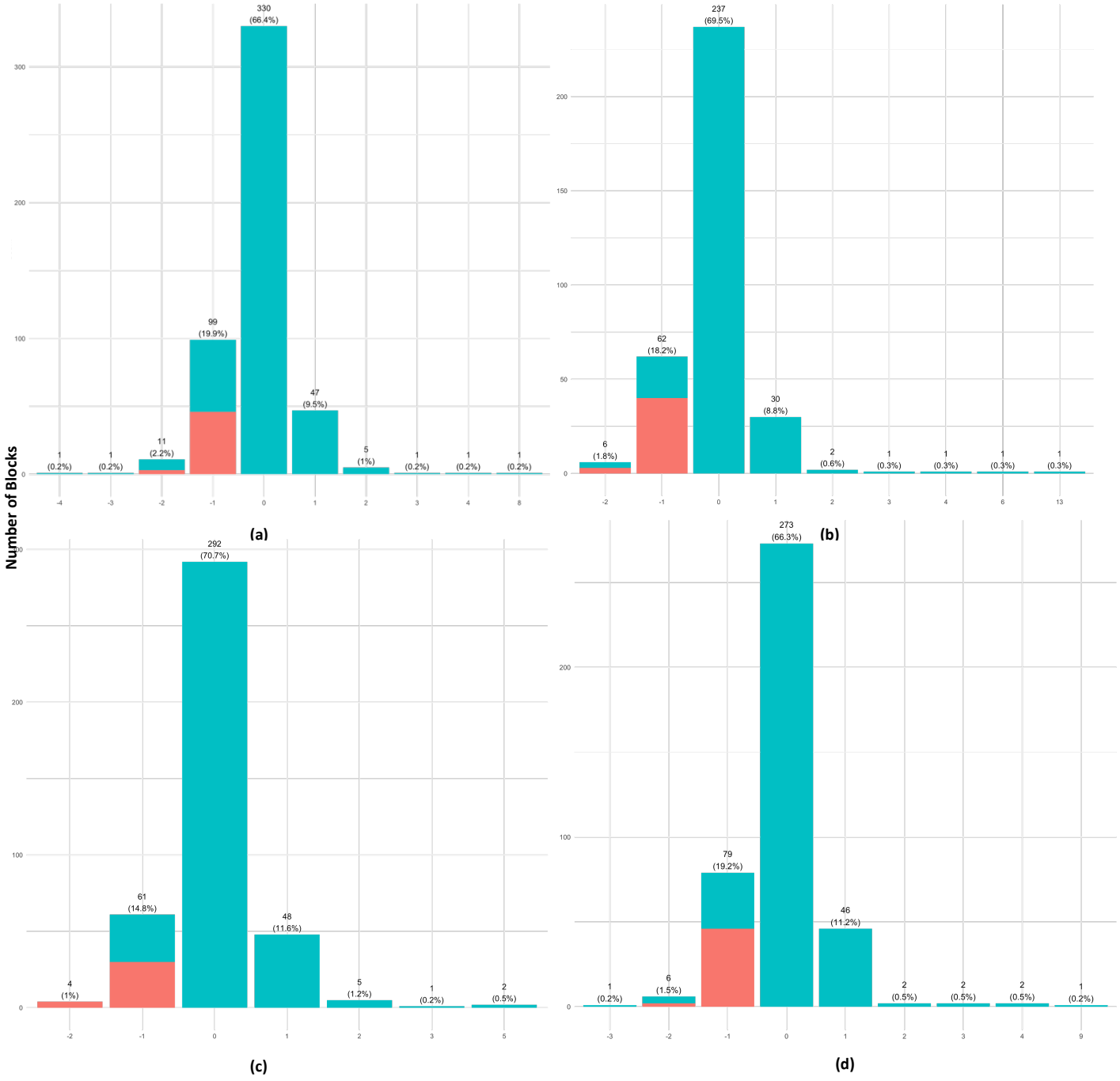


Figure III.8. Comparison of Manc-COJO and Sanc-COJO outputs. This figure compares the number of COJO SNPs selected the Manc-COJO model (using GWAS input of 1 million EUR + 0.1 million AFR) with the single-ancestry Sanc-COJO model (using GWAS input of 1 million EUR + 0.1 million EUR) for (a) HDL, (b) LDL, (c) TC, and (d) logTG. COJO-selected SNPs in both models were partitioned into LD-independent blocks. The x-axis shows the difference in SNP counts between the models (number of Manc-COJO SNPs minus number of Sanc-COJO SNPs), where a positive value indicates that Manc-COJO identified more SNPs than Sanc-COJO. The y-axis represents the number of blocks. The number shown above each histogram bar is the count of blocks, with the percentage in parentheses indicating the proportion of all blocks corresponding to that x-value. The red sections of the histograms highlight blocks where the Manc-COJO model identified zero SNPs while the Sanc-COJO model identified more than one SNP.

III.4 Discussion

We developed a novel multi-ancestry algorithm to perform conditional and joint analysis of GWAS summary statistics. We demonstrate through extensive simulations that Manc-COJO consistently outperformed Sanc-COJO applied on both single- and multi- ancestry data by correctly identifying, on average, more independent associated variants and reducing false positive associations. Moreover, Manc-COJO was robust to scenarios when sample sizes are biased, causal SNPs are not all genotyped, true SNP effects are different within each ancestry, and LD references are imperfect. We applied Manc-COJO to four polygenic lipid traits (which have genetic architectures that include large effect size associations) and validate SNPs selected by Manc-COJO using out-of-sample prediction and observed Manc-COJO yielded more parsimonious models with improved trans-ancestry portability compared to single ancestry Sanc-COJO model of the same sample size. In real data applications, the differences between Manc-COJO and Sanc-COJO models in terms of variance explained by their COJO SNPs were not statistically significant. The lack of statistical significance likely reflects power (sample sizes of non-European GWAS are currently much smaller than for European GWAS, 10-fold for the lipid traits studied here, and non-European sample sizes for out-of-sample prediction are also limited). Nonetheless, the pattern of results, replicated across the traits studied is consistent with expectations from well-powered simulations. The use of Manc-COJO applied to cross-ancestry GWAS meta-analyses will become increasingly relevant as samples sizes for non-European GWAS increase over time. Moreover, our results provide insights likely relevant to other post-GWAS methods that rely on GWAS summary statistics and LD references.

A key result from our analyses is that Manc-COJO achieves comparable predictive accuracy in European ancestry cohorts and markedly better portability in African ancestry cohorts, while selecting fewer SNPs. This indicates that the number of independent associations detected using single-ancestry data may be overestimated and tagging of true causal variants is improved through break down of correlation structure in the genome which can be exploited even when the proportion of African ancestry represented in the GWAS is relatively modest (*e.g.*, 10%).

Previous efforts to adapt Sanc-COJO for identification of independent associations in multi-ancestry GWAS meta-analyses (such as using LD matrices from European-ancestry individuals in predominantly European ancestry datasets or building a merged LD references from data that mirror the ancestry composition of the GWAS)

are reasonable but face important limitations. Using LD estimated in European ancestry samples performs adequately when the EUR:AFR sample size ratio is around 10:1, but it leads to a surge in false positives as this ratio approaches 1:1. Pinpointing the exact threshold where performance deteriorates is challenging (and likely varies by chromosomal region and genetic architecture). Likewise, constructing LD reference panels that match the ancestry proportions of the GWAS can work well under idealised conditions (e.g., in-sample LD, balanced ancestral representation, and sufficient power) but becomes prone to inflated false discoveries when these assumptions are violated. One potential contributing factor to inflated false discoveries is that all COJO models assume Hardy–Weinberg equilibrium in their calculations; however, mixed-ancestry LD reference panels may violate this assumption. Moreover, to ensure accurate LD estimation, each ancestry in the reference panel typically requires at least 5,000 individuals[170]. In imbalanced scenarios (e.g., 10:1 ancestry ratio), this dramatically increases the required panel size, creating significant computational burdens. In contrast, Manc-COJO avoids these complications by treating LD reference panels independently across ancestries, making their construction more straightforward and scalable.

Here, we focused on individuals of inferred European and African ancestry. However, the Manc-COJO methodology is readily extendible, and our software supports the inclusion of an arbitrary number of ancestries—limited only by computational and memory resources. While expanding the analysis to incorporate more than two ancestries is theoretically straightforward, it may introduce additional complexities that require careful methodological and empirical evaluation. Addressing these challenges lies beyond the scope of the current study and will be the subject of future work.

This study has several limitations, some inherent to the nature of the algorithm, others attributable to data quality and the scarcity of real-world data. First, we made several assumptions: true causal SNPs act additively on the phenotype, causal SNPs are independent, and the number of true causal variants, as well as their effect sizes on standardised phenotypes, are identical across ancestries. While these are standard assumptions they can be violated in various ways. For instance, dominant or recessive SNP effects may exist[206] and/or may interact with other SNPs[207], but these can only be identified in the context of large SNP effects. Previous research has shown that the additive effect model explains most genetic variation in complex traits[208,209].

Second, although an increasing number of studies provide empirical evidence that causal SNPs are shared across ancestries with similar effect sizes[184-187], ancestry-specific causal variants have also been reported[188,189]. To address this, we developed a complementary method, Manc-COJO --MDISA, to accommodate such scenarios. In real-data applications the additional ancestry-specific NPs identified did not explain variance in out-of-sample prediction. There are several reasons why marginal effects of causal SNPs may not be significant, including, but not limited to, inadequate sample size and power, masking effects of nearby SNPs with large effects, or the causal SNP not being genotyped and not tagged by a nearby SNP. Thus, differences in the significance of marginal effects alone cannot confirm the existence of ancestry-specific causal variants. Although larger sample sizes in future from non-European ancestries will help provide some clarity, our simulations show, with regard to joint effects, that even when a true causal SNP is genotyped, it may not necessarily be included in the model. Therefore, the presence of ancestry-specific variants in a joint model should not be taken as definitive evidence, as it may simply reflect artefacts from variable selection algorithms compounded by power/sampling error. When not all causal variants are genotyped, the scenario becomes even more complex due to differing LD structures across ancestries; for example, LD in African ancestry tends to decay faster along chromosomes than in other ancestries[210]. In such cases, the apparent existence of ancestry-specific variants might result solely from varying rates of LD decay surrounding true causal variants. Similar principles apply to claims of ancestry-specific effect sizes. In studies combining summary statistics from different sources, varying covariates may be fitted, meaning differences in effect sizes could arise purely from discrepancies in analysis pipelines. Therefore, at least in the short-term, bioinformatics-based claims of ancestry specific variants should be substantiated by direct functional evidence.

With this in mind, we developed Manc-COJO --MDISA not only to identify additional ancestry-specific variants but also to serve a broader purpose. Traditionally, bioinformatics-based analysis acts as an upstream step for wet-lab studies. However, Manc-COJO --MDISA establishes a “two-way street”: causal variants confirmed in wet-lab analyses can be incorporated into the Manc-COJO model—even if they are neither the most significant variants (due to sampling error) nor reach significance thresholds (due to inadequate sample sizes)—and will not be excluded by subsequent variable selection algorithms. Integrating wet-lab evidence back into the model may enhance our ability to detect independent associations. As a side note, in single-ancestry MDISA, if no SNPs are fixed in the model, it is equivalent to Sanc-

COJO but can use an LD matrix as input (as opposed to individual level data used to generate an LD matrix).

Throughout this chapter, we frequently refer to true causal variants and benchmark simulation outcomes against them, raising the question of whether COJO should be considered a fine-mapping tool. Simulations show that, in an idealised scenario where all causal variants are genotyped, no pairs of causal variants exhibit LD above 0.8, and studies are well-powered, Sanc-COJO identifies 40% of causal variants for less polygenic traits (5 causal SNPs in a region) and 16% for more polygenic traits (20 causal SNPs in a region). Manc-COJO increases this to 60% and 40%, respectively, by leveraging differing LD structures across ancestries. However, the assumptions (common to all fine-mapping methods) that all causal variants are genotyped (or imputed) and studies are well-powered are rarely met in practice. With whole-genome sequence data, where nearly all genetic variants are genotyped (though gaps remain, especially with short-read sequencing[194]), many variants are in complete or near-complete LD, necessitating indefinitely large sample sizes to completely resolve LD in numerous regions. Thus, Manc-COJO outcomes should be interpreted as a model of variants proximal to true causal variants that are nearly independently associated. In other words, while fine-mapping methods aim to pinpoint (using credible sets) the exact causal variant, Manc-COJO seeks to construct a model with nearly independent variants, where tagging variants close to causal ones suffices.

In summary, our study underscores the importance of collecting data from diverse ancestries, particularly as GWAS for some traits approach saturation[1]. It is essential to allocate resources to underrepresented populations during data collection efforts. With multi-ancestral data, approaches like Manc-COJO can better resolve LD and refine candidate variants for experimental validation. Given that LD blocks are smallest in African ancestry cohorts, GWAS in African ancestry cohorts will be the most cost-effective for optimisation of Manc-COJO SNPs as well as for fine-mapping methods.

IV

Refined mitochondrial DNA copy number (mtCN) calling pipeline reveals nuclear genetic regulation of mtCN in blood and its association with mood disorders

"Details make perfection, and perfection is not a detail." – Leonardo da Vinci

IV.1 Introduction

Mitochondria are semi-autonomous cellular organelles hosting 37 genes in humans[211]. The number of mitochondria within a cell varies by cell-type and physiological state. Together they serve as the biological powerhouse of cells, with both their quality and quantity playing critical roles in cellular metabolism and disease but regulated at least in part by the nuclear genome[212]. Mitochondrial copy number (mtCN) estimated from DNA copy number can serve as a proxy and an indicator for mitochondrial function[134].

A link between mood disorders and mitochondrial function has been studied motivated by changes in cell metabolism and energy requirements in mood disorders[213]. Studies have reported significant associations between mtCN and mood disorders, including major depressive disorder (MDD)[214,215] and bipolar disorder (BIP)[216-218], suggesting potential biological pathways such as stress response mechanisms[6]. However, these findings are inconsistent in the direction of association and indicate that the relationship may depend on disease state[214-219]. Moreover, many of these studies have been limited by relatively small sample sizes and/or a lack of adjustment for important covariates, such as blood cell composition and the timing of sample collection. These factors can have a substantial impact on analytical outcomes—for instance, mtCN levels are known to increase during the daytime and winter months[134], potentially due to heightened energy demands. As such, changes in mtCN may merely reflect seasonal variation in mood disorder incidence (mood disorders more common in winter[220,221]) or chronotype differences (eveningness is associated with an increased risk of mood disorders[222], which may lead affected participants to attend the assessment centre at a later time of day compared to controls). Furthermore, several blood cell parameters— including total white blood cell count, as well as neutrophil, monocyte, and specific lymphocyte subsets[223,224]— have been reported to be elevated in mood disorders. In such cases, alterations in mtCN may be attributable to shifts in blood cell composition rather than to a direct pathological effect of the disorder.

Although mtCN derived from whole-genome sequencing (WGS) data is often regarded as the gold standard[225], substantial inconsistencies exist across different mtCN calling pipelines[134,226,227]. At first consideration, estimation of mtCN from WGS data seems straightforward: it is commonly calculated as twice the ratio of the mean coverage of mitochondrial DNA (mtDNA) to that of autosomal DNA (**Methods**), thereby correcting the mtDNA haploid read count for the total diploid read count of

the individual sample. In practice, however, mtCN estimation pipelines differ in how coverage is computed. Differences between pipelines arise from choice of the genomic regions selected for calculation of mapped bases. For autosomal DNA, some studies use reads from the entire autosomal genome[226], whereas others restrict the analysis to specific subsets to account for potential structural variations—such as copy number variations—that may influence read depth across genomic regions[215]. For mtDNA, some pipelines use the full mitochondrial genome[226], whereas others exclude regions affected by nuclear mitochondrial DNA segments (NUMTs)[134,227]—mitochondrial DNA segments that have become incorporated into the human nuclear genome over evolutionary time—which can create ambiguity in read alignment between mitochondrial and nuclear sequences. Other differences occur in how mapped reads are defined and filtered; while some pipelines include all mapped reads[226], others apply quality control measures such as excluding low-quality reads[3] and adjusting for GC content bias[227].

The primary driver of this methodological heterogeneity is the absence of a definitive ground truth: direct quantification of mitochondrial content is difficult, particularly in blood samples containing a mixture of cell types. Without such a reference, it is difficult to determine which estimation pipeline is more accurate. Here, rather than directly calling mtCN from different pipelines, we systematically evaluate their underlying rationale by applying them to scenarios with known expected ratios. These include: (1) the ratio of sex chromosome to autosomal coverage, (2) individuals with whole-chromosome duplications, and (3) coverage of ultra-conserved sequences (UCS)[228]—regions highly conserved across species and depleted of single nucleotide variants (SNVs) or structural changes[229]. We also investigate potential confounders and, based on our findings, propose a refined mtCN calling and quality control pipeline designed to meet the analytical requirements of large biobank datasets. This approach may also offer valuable insights for other studies involving copy number variation detection. Finally, we conduct a series of statistical genetic analyses to explore the relationship between mtCN and mood disorders. To our knowledge, this represents the largest study to date examining the association between mtCN and mood disorders.

IV.2 Methods

IV.2.1 Cohorts

1000 Genomes Project (1000G)

The 1000 Genomes Project provides open-access WGS data without restrictions on use. In its final phase 3 release, WGS data were generated for 3,202 individuals, consisting of 2,504 unrelated participants and 602 members of family trios. Detailed sequencing protocols have been reported by the 1000G Consortium[230]; briefly, DNA libraries were prepared using the TruSeq DNA PCR-Free High Throughput Kit and sequenced on the Illumina NovaSeq 6000 platform, producing 150 bp paired-end reads. In our study, we analysed all unrelated individuals (N=90) labelled as GBR to ensure consistency with the geographical origin of the UK Biobank samples.

As part of the phase 3 structural variant initiative, the structural variation group of the 1000G sequenced approximately 50 individuals with long-read technology at the time our study was initiated. High-molecular-weight DNA was extracted from lymphoblastoid cell lines, libraries were prepared, and sequencing was carried out on the Pacific Biosciences (PacBio) Sequel II platform. Since then, long-read datasets have expanded substantially, with additional samples generated using both PacBio and Oxford Nanopore technologies[231]; however, only the individuals available at the start of our analysis were included in our study. Details of the samples analysed, and their ancestral backgrounds are provided in the **Supplementary Materials**.

UK Biobank

The UK Biobank is a large-scale cohort study that has collected comprehensive genetic, health, and lifestyle information from approximately 500,000 participants across the United Kingdom to facilitate research on the determinants of complex diseases. In late 2023, WGS data for all participants were released to researchers. Sequencing was performed on the Illumina NovaSeq 6000 platform, employing three distinct protocols—deCODE, Sanger Vanguard, and Sanger Main—within the sequencing and processing pipeline. Detailed descriptions of the WGS processing pipeline are provided elsewhere[232].

IV.2.2 Calculation of mtCN

Coverage can have different meanings depending on context. In this paper, we define the coverage of a genomic region (C) as:

$$C = \frac{M}{G} \qquad \text{(IV.1)}$$

where M is the total number of bases mapped to the region, and G is the length of the region. The total number of bases mapped can be calculated as:

(IV.2)

$$M = \sum_{i=1}^{N} L_i = NL$$

Where, L_i is the length of i-th read, N is the total number of reads, and L is the mean read length. Substituting formula (2) into formula (1) gives:

$$C = \frac{LN}{G} \quad \text{(IV.3)}$$

This is the Lander–Waterman equation commonly used in previous research[233]. Following prior work[215],mtCN can be defined as:

$$mtCN = 2\frac{C_T}{C_R} \quad \text{(IV.4)}$$

where C_T and C_R are the coverage for the target (mitochondrial) and reference (autosomal) regions, respectively. The factor of two accounts for the fact that the mitochondrial genome is haploid, in contrast to the diploid autosomal genome. Assuming the average read length is identical in the target and reference regions—which is reasonable within a single library and sequencing platform, as reads in a lane are uniform in length determined by sequencing cycles—the equation can be expressed as:

$$mtCN = 2(\frac{N_T}{G_T} / \frac{N_R}{G_R}) \quad \text{(IV.5)}$$

where N_T and G_T are the number of mapped reads and genome length for the target region, and N_R and G_R are the corresponding values for the reference region.

The same pipeline can be readily repurposed for other copy-number variation or aneuploidy detection. For example, the copy number of chromosome X can be estimated by comparing the sequencing coverage of chromosome X (target region) with that of the autosomes (reference region).

IV.2.3 Assessing coverage bias

We obtained the coordinates of ultra-conserved sequences (UCS) defined by Bejerano et al. [228] and converted them to Human Genome Assembly GRCh38 coordinates using UCSC *LiftOver* [234]. Each UCS was partitioned into 100 bp bins by rounding coordinates to the nearest multiple of 100; for example, a UCS spanning positions 38–302 was represented by two 100 bp bins: 100–200 and 200–300. GC content for each bin was calculated using the *bedtools nuc* function [235], and bins with GC content below 30% or above 70% were excluded. For mappability, we used the

pre-calculated file *k100.Umap.MultiTrackMappability.bw* from UCSC[236] and retained only bins with a mappability score of 1.

For short-read WGS data from the 1000G and UKB, which are already aligned to GRCh38, we counted bases with mapping quality >30 when assessing regional coverage bias using *samtools bedcov*[237], which by default excludes unmapped reads (sequencing reads that could not be aligned to the reference genome), secondary alignments (when a read matches multiple positions in the genome with similar alignment scores, most aligners choose one position as the primary alignment but still record the other possible matches as secondary alignments), duplicates (multiple reads that appear to originate from the same DNA fragment), or reads failing quality control (QC). For the 1000G long-read PacBio data, reads were aligned to GRCh38 (*GRCh38_full_analysis_set_plus_decoy_hla.fa*) using *pbmm2* (https://github.com/PacificBiosciences/pbmm2), and only HiFi reads (highly accurate long reads) were retained. Total bases mapped to each 100 bp UCS bin were calculated using the same protocol as in short-read sequencing.

Mapped base counts were obtained for each individual across three independent cohorts: (i) unrelated UKB participants of genetically inferred European ancestry[197] (relatedness ≤ 0.05, defined by GCTA[238] ; N = 347,512; hereafter referred to as the unrelUKB-EUR cohort), which were sequenced exclusively on the Illumina short-read platform; (ii) UK participants from the 1000G (N = 90; hereafter the 1000G-GBR cohort, unrelated to UKB participants), also sequenced on the Illumina short-read platform; and (iii) samples from the 1000G that were sequenced on both the PacBio long-read and Illumina short-read platforms (N = 47; non-overlapping with the 1000G-GBR cohort; hereafter the 1000G-LR cohort).

To assess coverage bias across UCS, the total number of mapped reads for each UCS bin was rank transformed within each individual (and calculated for each individual in all cohorts). Subsequently, within each cohort, the mean rank for each UCS bin was calculated and compared using linear regression in R (*lm()* function).

To assess coverage bias at the chromosome level, we used *samtools idxstats*[237] to obtain the total number of mapped reads for each chromosome in each individual. Relative coverage was calculated using equation (IV.5), omitting the factor of 2 since we are comparing between diploid genomes. Here, the target region corresponds to each chromosome, while the reference region comprises all autosomes combined.

The genome length used was the mappable length (rather than the full chromosome length), as provided in the second column of the *idxstats* output.

IV.2.4 Comparing different copy number calling pipelines

In counting the number of mapped reads, different filtering strategies have been implemented in previous studies. Here, we used either chromosome X (chrX) or chromosome Y (chrY) as the target region and autosomes as the reference region to evaluate various copy number calling pipelines in scenarios where the ground truth is known.

We assessed several pipelines, including an Idxstats-based pipeline, which extracts the number of mapped reads from the output of *samtools idxstats*[237]. This approach counts all mapped reads without additional quality control filtering and implicitly assumes that the average read length is the same in both the target and reference regions.

We also evaluated a flagstat-based pipeline, in which coverage is calculated using equation (IV.3) with outputs from *samtools flagstats*[237]. In this pipeline, singletons(paired-end reads where only one read of the pair is successfully mapped to the reference genome), reads with discordant mates (both reads map**,** but their alignments are inconsistent with the expected pairing configuration), and duplicated reads (multiple sequencing reads that originate from the same original DNA fragment) are excluded as previously recommended[134]. The flagstat-based method requires either the same implicit assumption regarding average read length in target and reference regions or an external estimate of average read length, for example using *GATK CollectQualityYieldMetrics*[126].

Notably, *GATK CollectQualityYieldMetrics*[126] can also be used directly to estimate copy number. In GATK, the total number of mapped reads corresponds to the number of paired reads in *samtools flagstat*. By default, GATK excludes secondary and supplementary alignments (defined as above) and reports bases above specified mapping quality thresholds. In our study, we applied two GATK-based pipelines using mapping quality scores of 20 and 30, respectively.

It is important to note that the reference genome used in both the 1000G and UKB datasets is the version of GRCh38 that includes alternative haplotypes, decoy sequences, Epstein-Barr virus (EBV), and HLA regions. To avoid inflating the total

number of bases, we extracted autosomes and target chromosomes (chrX and chrY) separately using *samtools view*[237] before running *GATK CollectQualityYieldMetrics*[126].

Additionally, for chrX, we implemented a *UCS-based approach* (as UCS are absent on chrY). In this pipeline, the target region consisted of UCS on chrX with perfect mappability and moderate GC content (between 30% to 70%), while the reference region included all UCS on autosomes with perfect mappability and moderate GC content. The number of mapped bases was calculated using *samtools bedcov*[237], which excludes unmapped reads, secondary alignments, duplicates, or reads failing quality control, with a mapping quality threshold of 30. This represents the most stringent quality control pipeline in our study.

Finally, for both chrX and chrY, we implemented a mixed approach in which coverage of the target region was estimated using portions of chrX or chrY with perfect mappability, though not necessarily ultra-conserved, and total mapped bases on the target region were calculated using *samtools bedcov*. Coverage of the reference genome was calculated from all autosomes using *GATK CollectQualityYieldMetrics*, with a mapping quality threshold of 30 applied to both target and reference regions. This mixed approach represents a relatively stringent pipeline, designed to address the absence of UCS on chromosomes such as chrY and to maximise the length of both target and reference regions. We focused on chrX and chrY because the homologous regions between these chromosomes provide a useful model for understanding the impact of NUMTs on mtCN calling.

IV.2.5 Calling mtCN

In the UKB, we estimated mtCN using both the idxstats-based approach (least stringent) and the UCS-based approach (most stringent) proposed here and found that results were nearly perfectly correlated for most individuals. For downstream analyses, we used the idxstats-based approach due to its smaller sampling variation, but outliers were removed when autosome or mitochondrial coverage differed substantially between the two approaches, as this typically indicates a high number of reads with low mapping quality. Additionally, individuals lying outside 4 SDs of the mean in several WGS QC metrics—including Freemix verifyBamID (i.e. the fraction of contaminating DNA), proportion of mapped reads, and coverage—were excluded. The 4 SD threshold was chosen to balance sample size with data quality.

Covariates

For both phenotypic association and GWAS analyses, covariates were grouped into three categories:

1. **Basic covariates**: age, age², sex, sex × age, sex × age², assessment center, 10 genetic PCs, sequencing provider, and shipment batch number.
2. **Time-related covariates**: Fasting time, hour, month, and year of sample collection.
3. **Blood composition covariates**: red blood cells (RBCs), nucleated RBCs, reticulocytes, high light scatter reticulocytes, basophils, eosinophils, lymphocytes, monocytes, neutrophils, and platelets counts. Absolute counts (rather than percentages) were used in all analyses to avoid collinearity among covariates, as percentages are derived by normalising counts to the total blood cell count.

Following prior research, we excluded individuals sampled in 2006[134]. Sample collection time was defined by the collection of the first tube, and individuals with samples collected at or before 7 am or after 9 pm were excluded. Hour and month of collection were modelled as categorical variables. Finally, individuals with any blood count lying outside 4 SDs of the mean were removed.

IV.2.6 Definition of MDD and BIP cases and control

In analysis of association between mtCN and MDD or BIP, we first only limit the analysis to unrelated individuals of inferred European ancestry as defined previously. Within this unrelated cohort of individual, we get scaled residuals of each of four models below:

1. Raw model: log(mtCN)
2. Basic model: log(mtCN) ~ Basic covariates
3. Time model: log(mtCN) ~ Basic covariates + Time-related covariates
4. Complete model: log(mtCN) ~ Basic covariates + Time-related covariates + Blood composition covariates

Definition of different MDD phenotype are based on a previous study[239]. In brief, we included eight MDD phenotypes: (i) seen by a general practitioner for nerves, anxiety, tension, or depression (GPpsy); (ii) seen by a psychiatrist for nerves, anxiety, tension, or depression (Psypsy); (iii) probable recurrent major depression or a single probable major depressive episode (DepAll); (iv) self-reported depression (SelfRepDep); (v) ICD-10–defined depression (ICD10Dep); (vi) DSM-V–defined major

depression (LifetimeMDD); (vii) recurrent major depression (MDDRecur); and (viii) seen by a GP for depression without cardinal symptoms (GPNoDep). In addition to these definitions, we introduced an additional phenotype, current depression (MDDCurrent), defined as ICD10Dep cases in which the first recorded occurrence of MDD was within 30 days of the blood sample collection for sequencing.

BIP phenotypes were defined based on UK Biobank data fields, combining field 20122 (Bipolar disorder status) and field 130892 (Date F31 first reported). We defined five BIP phenotypes: (i) Bipolar I and (ii) Bipolar II, both based on field 20122 (Bipolar disorder status); (iii) ICD-10–defined bipolar disorder (ICD10BIP), based on field 130892 (Date F31 first reported); (iv) current bipolar disorder (BIPCurrent), defined as ICD10BIP cases in which the first recorded occurrence of BIP was within 90 days of blood sample collection for sequencing; and (v) *combined bipolar disorder* (BIPCombined), comprising all bipolar definitions described above.

We applied a more stringent definition for controls, in additional to not diagnosed with any type of MDD or BIP, we require participants to explicitly answer “no” to question “Seen doctor (GP) for nerves, anxiety, tension or depression”(UK Biobank data field 2090) and no mental health condition ever diagnosed by a professional (UK Biobank data field 29000).

IV.2.7 Sensitivity analysis

ICD10 diagnostic data were retrieved from the UK Biobank. ICD10 codes were simplified to their base form (3-character codes), and the diagnosis of every base code was determined for each participant. The association between diagnostic codes and mitochondrial read counts was investigate using two approaches: (1) correlation with the ICD10 chapter-specific burden of diagnostic codes, and (2) association testing of individual base codes. For the ICD10 chapter-level analysis, base codes were mapped to their respective ICD10 chapter, and the number of diagnostic codes per chapter was counted for each participant. The association with mitochondrial count was then tested in a logistic regression model. For the base-code level analysis, only base codes assigned to at least 50 participants were included. Each of these codes was tested for association with scaled residual of mitochondrial read count in a logistic regression model. Scaled residual used here is identical to what used in analysis of MDD and BIP.

IV.2.8 GWAS of mtCN and downstream analysis

We first fitted a linear model to get the scaled residual from the model:

log(mtCN) ~ Basic covariates + Time-related covariates + Blood composition covariates

We included all individuals of inferred European ancestry who passed quality control (N= 413,946), including related participants, and performed GWAS using *GCTA fastGWA*[240], a mixed linear model requiring a sparse GRM as input. Independent association signals were identified using *GCTA COJO*[170]. In addition, we conducted a gene-based test using *mBAT-combo*[196]. All tools were run with default parameters. Using the resulting list of genes, we performed GO functional enrichment and KEGG pathway analyses with the R packages *clusterProfiler*[241], *enrichplot*, *org.Hs.eg.db*, *msigdbr*, and *ReactomePA*[242].

For genetic correlation analyses, we used the latest publicly available GWAS summary statistics of MDD and BIP from European populations (excluding 23andMe) and applied *LDSC*[243] with HapMap3 SNPs and default settings.

IV.2.9 Single cell RNAseq analysis

To identify the cell types through which genetic variants may influence mtCN, we performed single-cell RNA sequencing (scRNA-seq) analyses following the pipeline described in a previous publication[244]. In brief, we integrated scRNA-seq data from three databases comprising human brain tissue[245], bone marrow[246], and multiple non-brain tissues[247]. Trait–cell type associations were assessed using three complementary methods: Cepo[248], sc-DRS[249], and MAGMA[250]. The resulting p-values from these methods were combined using the Cauchy combination test to derive overall trait–cell type association p-values. Multiple testing was addressed using Bonferroni correction, with a significance threshold of $0.05/171 = 2.9 \times 10^{-4}$, where 171 represents the total number of tests performed.

IV.3 Results

IV.3.1 Regional and chromosomal coverage bias

Accurate estimation of mtCN requires assessing whether sequencing coverage is uniform across the genome. In an ideal scenario with evenly distributed coverage, the mean depth of any autosomal or mitochondrial region without structural variation could be considered representative of overall autosomal or mitochondrial coverage, respectively. Nonetheless, earlier studies have reported coverage bias, especially in genomic regions characterised by extreme GC content—which can lead to reduced and inconsistent coverage across samples[251] —or by low mappability, as

seen in telomeric and centromeric regions enriched with repetitive sequences that hinder accurate read alignment[236].

In our analysis, we found that coverage bias persists even after correcting for GC content and mappability differences. To illustrate this bias, we first extracted the mean coverage of ultra-conserved regions (N = 481 in total; N = 457 from autosomes and N = 24 from chromosome X) from the unrelUKB-EUR cohort (N = 347,512; see **Methods**). These ultra-conserved regions were selected because they are highly stable across individuals and species, with minimal single nucleotide variants (SNVs) or structural changes. Therefore, any variation in coverage within these regions is likely attributable to technical artefacts rather than genome changes. We partitioned each UCS into 100 bp bins and restricted the analysis to bins with moderate GC content (30–70%) and perfect mappability scores (score = 1; see **Methods**) to minimise confounding effects. This filtering left 565 bins across 366 UCS regions for further analysis. For each individual, we applied a rank transformation to the number of reads mapped to each 100 bp bin across the 565 bins. In a scenario where coverage is uniform across the genome and any variation arises purely from random sampling, the mean rank across individuals for each of the 565 UCS bins should be similar. However, we observed a clear deviation from this expectation, with an uneven distribution of mean ranks (**Figure IV.1a**), implying that some UCS regions have consistently more reads than others.

To further support that the pattern of coverage unevenness is more than expected by chance, we find that the pattern of read coverage of UCS regions is specific to the sequencing platform, providing evidence for technical bias. When the mean rank of each UCS bin in unrelUKB-EUR cohort was compared with those from the 1000G-GBR cohort, a statistically significant coefficient of determination was observed ($R^2 = 0.66$, $P < 2 \times 10^{-16}$; **Figure IV.1b**). This concordance was evident despite the samples having been generated and processed in different institutions and followed different protocols, as both datasets were produced using the Illumina NovaSeq 6000 platform. In a similar manner, the 1000G-LR cohort—irrespective of whether sequencing was carried out at the Jackson Laboratory or the University of Washington—also demonstrated strong agreement in UCS bin rankings ($R^2 = 0.62$, $P < 2.2 \times 10^{-16}$; **Figure IV.1c**).

In contrast, the mean UCS bin rankings exhibited substantial variation when the same individuals were sequenced on different platforms. To illustrate this, we examined individuals from the 1000G-LR cohort who had also undergone sequencing using the

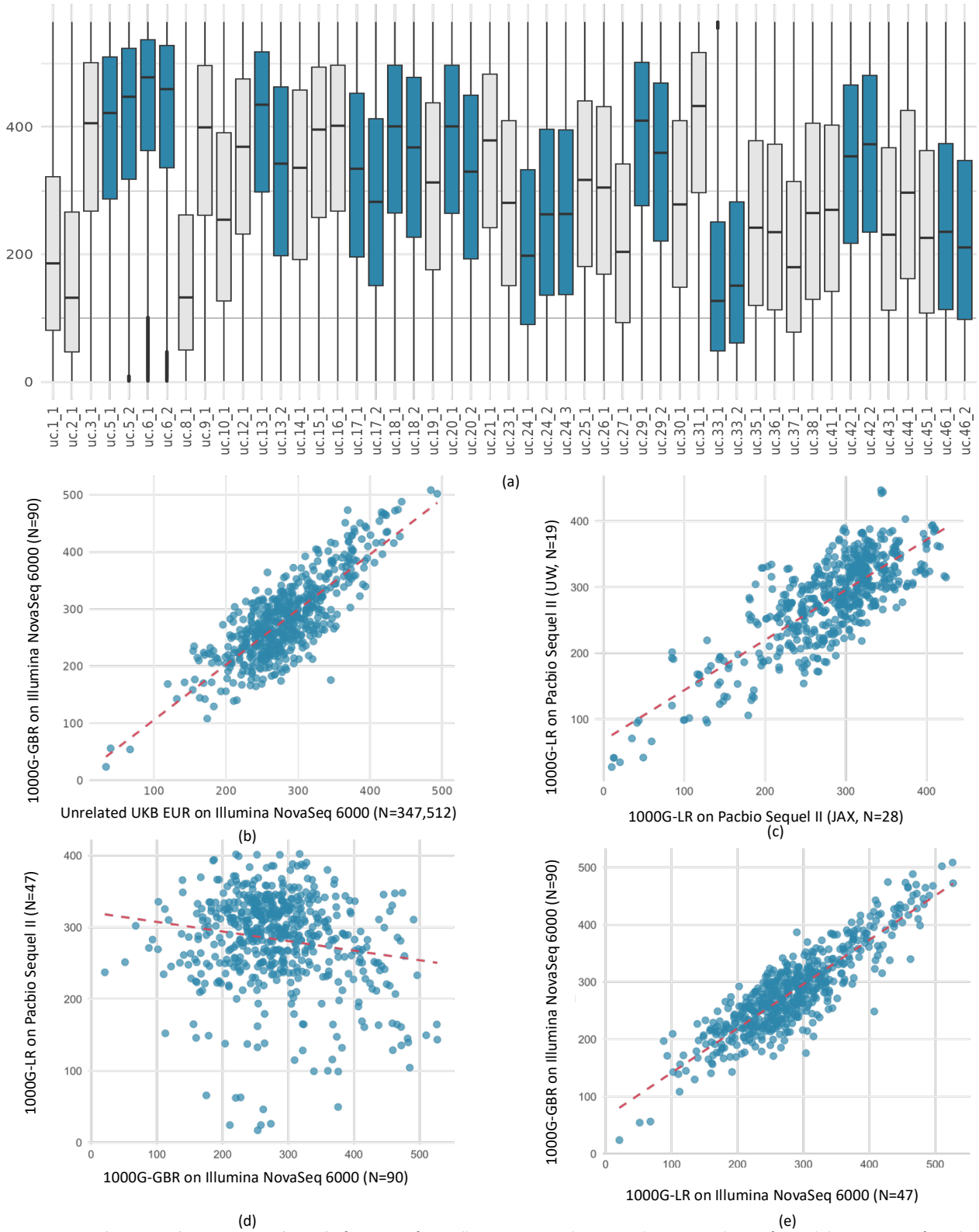


Figure IV.1. Regional coverage biases exist and are platform specific. To illustrate regional coverage bias, we rank-transformed the coverage of 100 bp bins of ultra-conserved sequences (UCS) within each individual and then calculated the mean value for each UCS element within each cohort, including unrelUKB-EUR (sequenced on Illumina NovaSeq 6000), 1000G-GBR (sequenced on Illumina NovaSeq 6000), and the 1000G-LR cohort (sequenced on both Illumina NovaSeq 6000 and PacBio Sequel II). (a) Boxplot of mean rankings on chromosome 1 in the UKB unrelated EUR cohort. Each column on the x-axis represents a 100 bp UCS bin, and columns coloured in blue indicate instances where more than one bin exists within the same ultra-conserved region. The y-axis denotes the mean ranking. (b–e) Each dot represents a 100 bp UCS bin, and its coordinates correspond to the mean ranking in the indicated cohort and sequencing platform, as specified in the axis titles. The sample numbers (N) in each genotyping instance are provided on the x and y axes.

short-read Illumina NovaSeq 6000 platform. For these same set of individuals, the coefficient of determination of UCS bin rankings between the two platforms was markedly lower ($R^2 = 0.027$, $P = 7.9 \times 10^{-5}$; **Figure IV.1d**).

The lack of concordance in UCS bin rankings between platforms cannot be attributed to the smaller sample size of the 1000G-LR cohort. When data from these same individuals were generated on the Illumina NovaSeq 6000 platform, their UCS bin rankings displayed a high coefficient of determination with those of the 1000G-GBR samples sequenced on the same platform ($R^2 = 0.77$, $P < 2.2 \times 10^{-16}$; **Figure IV.1e**). It is noteworthy that participants in 1000G-LR cohort were not exclusively of European ancestry (**Supplementary Table IV.1**). Thus, the observed strong correlation in UCS bin rankings suggests that differences in coverage profiles are, at least in part, driven by the sequencing platform itself rather than being solely a consequence of genetic background.

In addition to sequencing platform–specific regional coverage biases, we also observed biases at the whole-chromosome level. Notably, the direction and magnitude of these biases became more consistent across both the Illumina NovaSeq 6000 and PacBio long-read sequencing platforms at the chromosomal level, irrespective of the sequencing facility (**Supplementary Figure IV.1**).

IV.3.2 Impact of coverage bias and QC pipelines

Despite the presence of coverage bias at both regional and chromosomal levels, the ratio of coverage between target and reference regions remains informative. A clear example is the comparison of coverage between sex chromosomes (target regions in this case) and autosomes (reference regions in this case). In both the 1000G-GBR and unrelUKB-EUR cohorts, the observed mean coverage ratios of sex chromosomes to autosomes were largely consistent with theoretical expectations. In the UK Biobank, the mean (SD) chromosome X–to–autosome ratio was 0.512 (0.0039) in male participants (N = 159,520) and 0.998 (0.0095) in female participants (N = 187,994). Similarly, in the 1000G-GBR cohort, the corresponding values were 0.500 (0.0038) for males (N = 46) and 0.950 (0.049) for females (N = 44) (**Supplementary Figure IV.2**). Notably, even in individuals with whole-chromosome duplications (**Supplementary Figure IV.3**), such as trisomy 21—an extreme form of structural variation—the sex chromosome–to–autosome coverage ratio remained stable. This robustness is observed despite the inflation of autosomal coverage caused by chromosome duplication (**Figure IV.2**), owing to the law of large numbers.

Although the overall sex chromosome–to–autosome ratios were in line with expectations, slight deviations from the idealised values were observed (i.e. an expected chrX-to-autosome ratio of 1.0 in females and 0.5 in males, and a chrY-to-autosome ratio of 0.0 in females and 0.5 in males). The discrepancies were particularly evident for chromosome Y. In the unrelUKB-EUR cohort, the mean (SD)

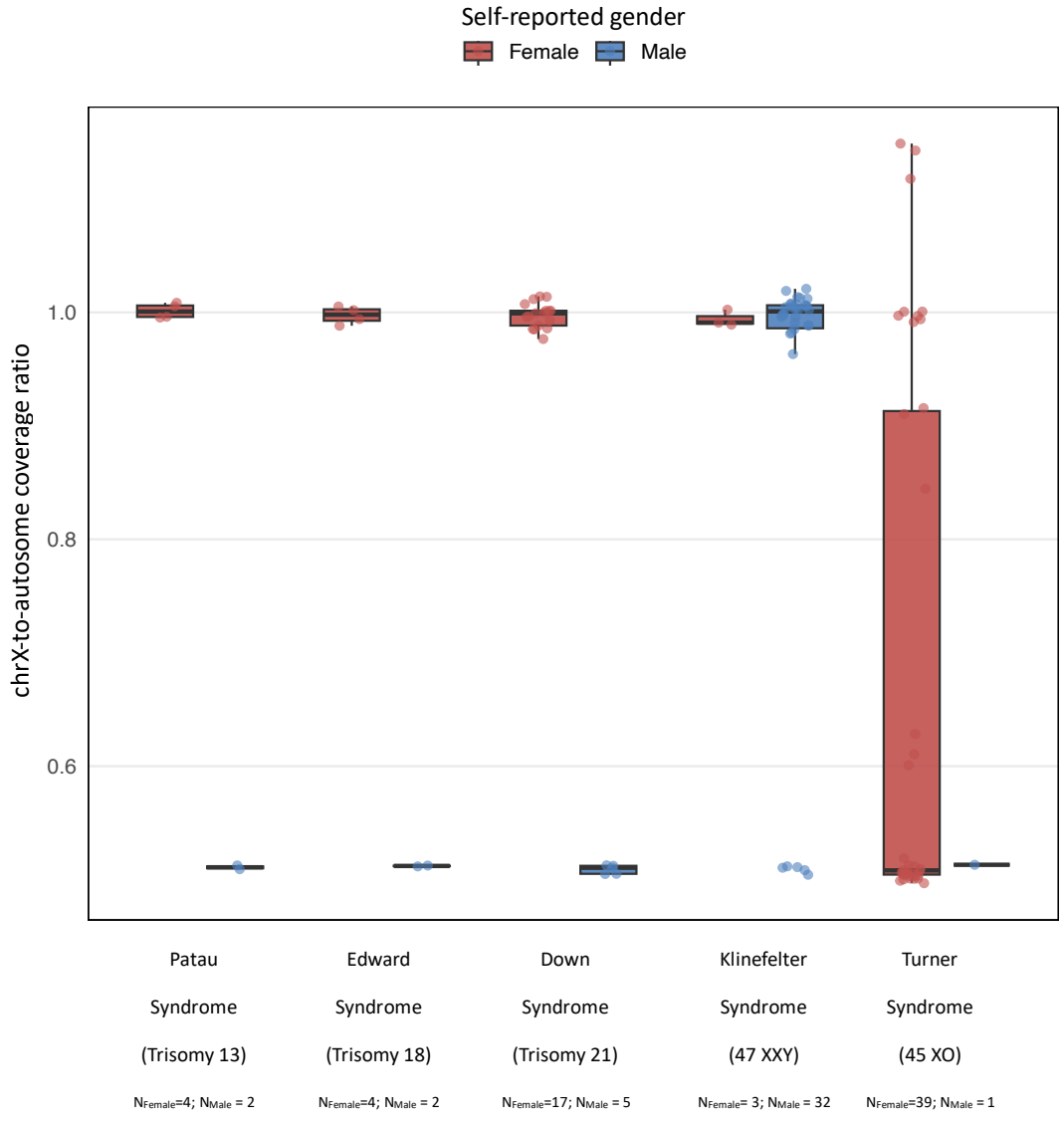


Figure IV.2. Coverage ratio remains informative even in the presence of whole-chromosome duplications. Despite the presence of coverage biases at both regional and chromosomal levels, the ratio of coverage between targeted and reference sequences remains informative. In the figure, the y-axis represents the coverage ratio between chromosome X and autosomes, while the x-axis indicates the status of known aneuploidy. Column colours denote sex, as shown in the legend. Even in individuals with extreme forms of structural variation, such as trisomy 21, the chrX-to-autosome ratio generally conforms to expectations, despite the inflation of autosomal coverage caused by the additional chromosome. The expected chrX-to-autosome ratio is ~1.0 in biological females and ~0.5 in biological males, except in cases of Klinefelter syndrome (expected ratio of ~1.0) and Turner syndrome (expected ratio of ~0.5), regardless of self-identified gender. As reported in previous studies, the observed continuum in Turner syndrome cases is most likely due to mosaicism, where only a subset of cells is affected, or to partial chromosomal duplication.

chrY-to-autosome ratio was 0.409 (0.033) in males and 0.0561 (0.0036) in females. In the 1000G-GBR cohort, the respective values were 0.413 (0.044) and 0.0943 (0.0082) (**Supplementary Figures IV.4a and 4b**). These deviations likely arise from homologous regions between sex chromosomes (e.g. PARs) causing ambiguous mapping, as well as technical factors such as contamination and alignment artefacts.

As all sex chromosome–to–autosome ratio analyses above included all mapped reads without filtering, this raises the question of whether applying more stringent quality control procedures would yield coverage ratios more closely aligned with theoretical expectations. To examine this, we excluded duplicate reads, secondary alignments, reads failing quality control, and those with a mapping quality score below 30 across all chromosomes. In addition, analysis of chrX and chrY was restricted to regions with perfect mappability (**Methods**). In the 1000G-GBR cohort, the mean (and s.d.) chrX-to-autosome ratio under this stringent pipeline was 0.462 (0.012) for males and 0.898 (0.053) for females. The corresponding chrY-to-autosome ratios were 0.493 (0.051) for males and 0.0465 (0.0030) for females. Interestingly, applying stricter quality control moved the chrX-to-autosome ratio further from the expected values, whereas the chrY-to-autosome ratio shifted closer to expectation. This highlights the persistent technical limitations of current read-alignment protocols. Notably, different quality control strategies used in previous studies—such as filtering by mapping quality, removing secondary or duplicate reads, or incorporating GC content and mappability—produce highly correlated results. (**Figure IV.3; Supplementary Figure IV.5**).

In the sex chromosome–to–autosome ratio analyses above, both target and reference sequences exceeded 50 Mb. To test whether the ratios remain reliable for shorter sequences, we calculated chrX-to-autosome ratios using individual 100 bp UCS bins on chrX as target sequences. In all cases, ratios were significantly higher in females (**Figure IV.4a** and **4b**), but variability increased, with a median standard deviation of 0.10 in males and 0.15 in females. This demonstrates that the method remains informative for short regions, though with reduced precision.

IV.3.3 Strategies in calling mtCN

For most UK Biobank participants, coverage estimates from different mapping strategies remained highly correlated. For each autosome and the mitochondrial genome, we compared coverage calculated from all mapped reads without quality control filtering to coverage calculated using stringent QC strategies (see **Methods**).

High coefficient of determination was observed for both autosomes ($R^2 = 0.894$; $p < 2.2 \times 10^{-16}$) and mitochondria ($R^2 = 0.969$; $p < 2.2 \times 10^{-16}$). We identified N = 4,147

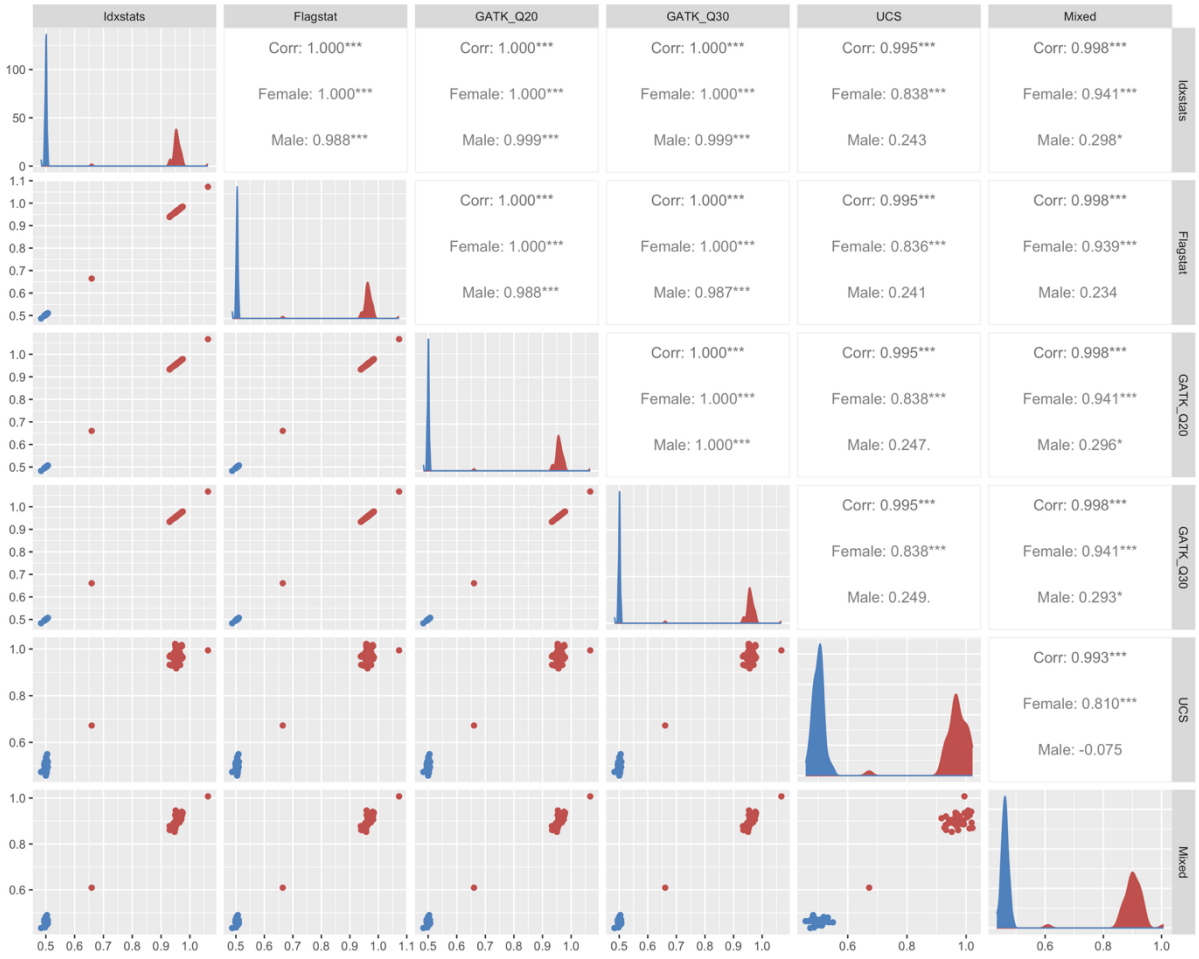
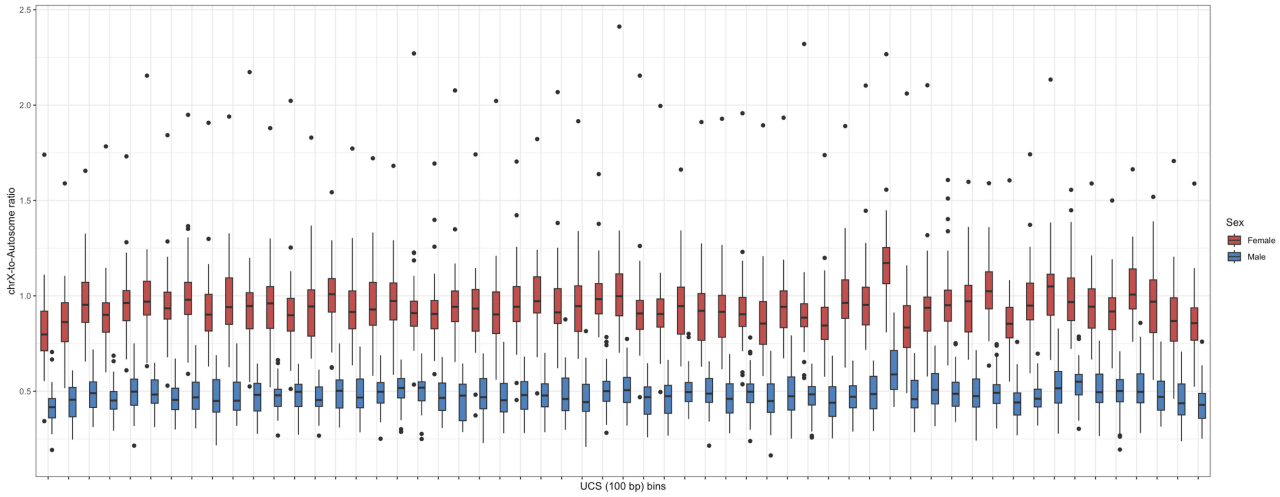
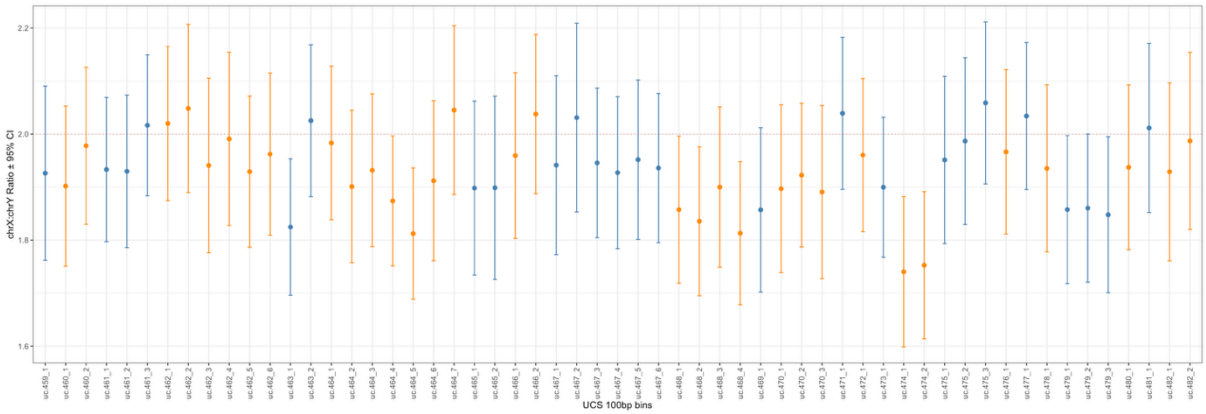


Figure IV.3. Impact of different pipelines on the chrX-to-autosome ratio. In this figure, we applied methodologies used in previously described pipelines to calculate the chromosome X–to–autosome coverage ratio. Six approaches were tested: Idxstat-based (least stringent), Flagstat-based, GATK-based (MAPQ > 20), GATK-based (MAPQ > 30), UCS-based (most stringent), and a mixed approach. Detailed descriptions of each method are provided in the Methods section. Colours correspond to biological sex (red = female; blue = male). Row and column labels indicate each method: the lower triangle shows pairwise comparisons of inferred ratios between methods (each point represents an individual from the 1000G-GBR cohort), the diagonal displays density plots of the inferred ratios, and the upper triangle presents correlation coefficients from linear regression (combined and stratified by sex). Overall, results from the different approaches were highly correlated and close to the expected values. However, as illustrated in the UCS and mixed approaches, when the target and reference regions became shorter, estimation variability increased, resulting in weaker correlations with approaches that used entire chromosomes as target and reference regions. Nevertheless, the estimated ratios remained close to the expected values (a chrX-to-autosome coverage ratio of approximately 0.5 in males and 1 in females).



(a)



(b)

Figure IV.4. Coverage ratios remain informative for shorter target regions despite larger variability. To test whether coverage ratios remain reliable for shorter sequences, we calculated chrX-to-autosome ratios using individual 100 bp UCS bins on chrX as target sequences. (a) Boxplot of chrX-to-autosome ratios. Each column corresponds to a 100 bp UCS bin on chrX, with colours indicating biological sex. (b) For each 100 bp bin, we compared the chrX-to-autosome ratio (and its 95% confidence interval) between females and males, treating the two groups as independent and applying a t-test. The red dotted line indicates the expected ratio of 2. Alternating colours were used to highlight bins belonging to the same UCS. In all UCS bins, the ratio was significantly greater than 1, indicating higher chrX-to-autosome ratios in females compared with males. These results demonstrate that the method remains informative even for short regions, though with greater variability.

and 1,118 outliers (defined as lying outside 3SD of the mean; **Supplementary Figure IV.6**) for autosomes and mitochondria, respectively. Notably 958 of the 1,118 mitochondrial outliers also classified as autosomal outliers. After removing all outliers, coverage calculated from the stringent QC pipeline (UCS-based) became nearly perfectly correlated with that from the least stringent pipeline (Idxstat-based; **Supplementary Figure IV.7**).

Compared with the other participants, these outliers possessed markedly greater total sequence data (112.41 Gb vs. 97.52 Gb; $p < 2.2 \times 10^{-16}$) and displayed higher levels of sample contamination—quantified through both read haplotypes (0.0087% vs. 0.0068%; $p < 2.2 \times 10^{-16}$) and ancestry-agnostic approaches (0.10% vs. 0.080%; $p < 2.2 \times 10^{-16}$)—as well as increased non-reference discordance (0.18% vs. 0.12%; $p < 2.2 \times 10^{-16}$). These findings indicate that the elevated sequencing yield (i.e. the total volume of data successfully produced) may stem from technical artefacts, such as overloading during library preparation, rather than genuine biological variation. Taken together, these findings indicate that previously employed QC strategies are reasonable and produce highly comparable coverage estimates. When the target sequence is sufficiently long, the inferred coverage proportion remains robust even in the presence of large structural variants, such as whole-genome duplications. Nonetheless, our analyses suggest it is important to remove outliers, as defined by comparisons between UCS-based (most stringent) and Idxstat-based (least stringent), since they serve as indicators of technical artefacts.

In our pipeline of calling mtCN, we removed outliers as described above, and we also removed outliers based on several additional WGS QC metrics (see **Method**). For our analyses we only retained the European participants N = 413,946 that passed the quality control, with a mean mitochondrial copy number of 33.9 with a right-skewed distribution (**Supplementary Figure IV.8**).

IV.3.4 Association between mtCN and mood disorders

In the UK Biobank, several definitions of MDD exist, which are correlated but not identical[239]. We therefore compared scaled raw log(mtCN), scaled residuals after adjusting for basic covariates (age, sex, 10 PCs, etc.; **Methods**), scaled residuals after further adjusting for time, month, and year of blood sample collection, and scaled residuals after additionally adjusting for blood composition. mtCN values were log-transformed to correct for skewness. Significant differences in mtCN between MDD cases and controls were observed in raw counts for eight out of nine of the depression phenotypes, with five showing significantly lower count and three a

significantly higher count. Including basic/time covariates for certain definitions of depression gave no clear pattern of attenuating or enhancing association. However, associations for all nine depression definitions became non-significant once blood composition was included as a covariate. This suggests that differences in mtCN MDD cases and controls largely reflect variation in blood cell composition rather than a direct disease effect (**Supplementary Table IV.2**; **Figure IV.5**).

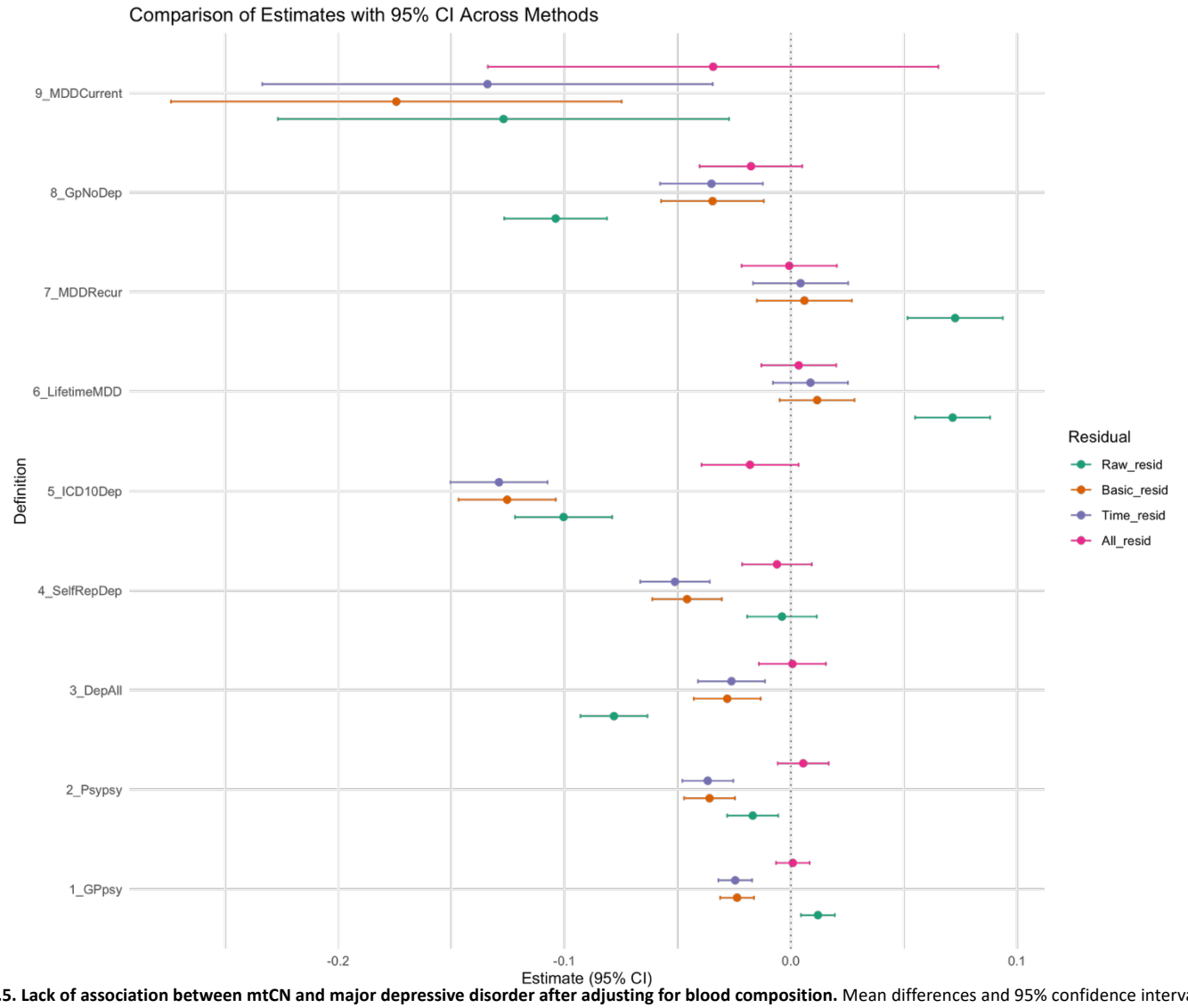


Figure IV.5. Lack of association between mtCN and major depressive disorder after adjusting for blood composition. Mean differences and 95% confidence intervals (CIs) for scaled residuals of log(mtCN) comparing cases and controls, derived from models with varying covariate adjustments. Cases were defined using multiple established criteria, while controls were selected based on stricter definitions (see Methods). Analyses used two-sample t-tests of case/control status on mtCN. Colours indicate model type: Raw model – log(mtCN); Basic model – scaled residuals from log(mtCN) adjusted for basic covariates; Time model – scaled residuals adjusted for basic and time-related covariates; Complete model – residuals adjusted for basic, time-related, and blood composition covariates. Case definitions: GPpsy – seen by a general practitioner for nerves, anxiety, tension, or depression; Psypsy – seen by a psychiatrist for the same conditions; DepAll – probable recurrent or single major depressive episode; SelfRepDep – self-reported depression; ICD10Dep – ICD-10–defined depression; LifetimeMDD – DSM-V–defined major depression; MDDRecur – recurrent major depression; GPNoDep – seen by a GP for depression without core symptoms; MDDCurrent – ICD-10–defined depression with first occurrence within 30 days of blood sampling. After adjustment for blood composition covariates, all significant associations disappeared, indicating that mtCN differences among depression cases likely reflect variations in blood composition.

Similarly, for bipolar disorder (BIP), the association between mtCN and disease status was no longer significant after adjusting for blood composition (**Supplementary Table IV.3**; **Figure IV.6**).

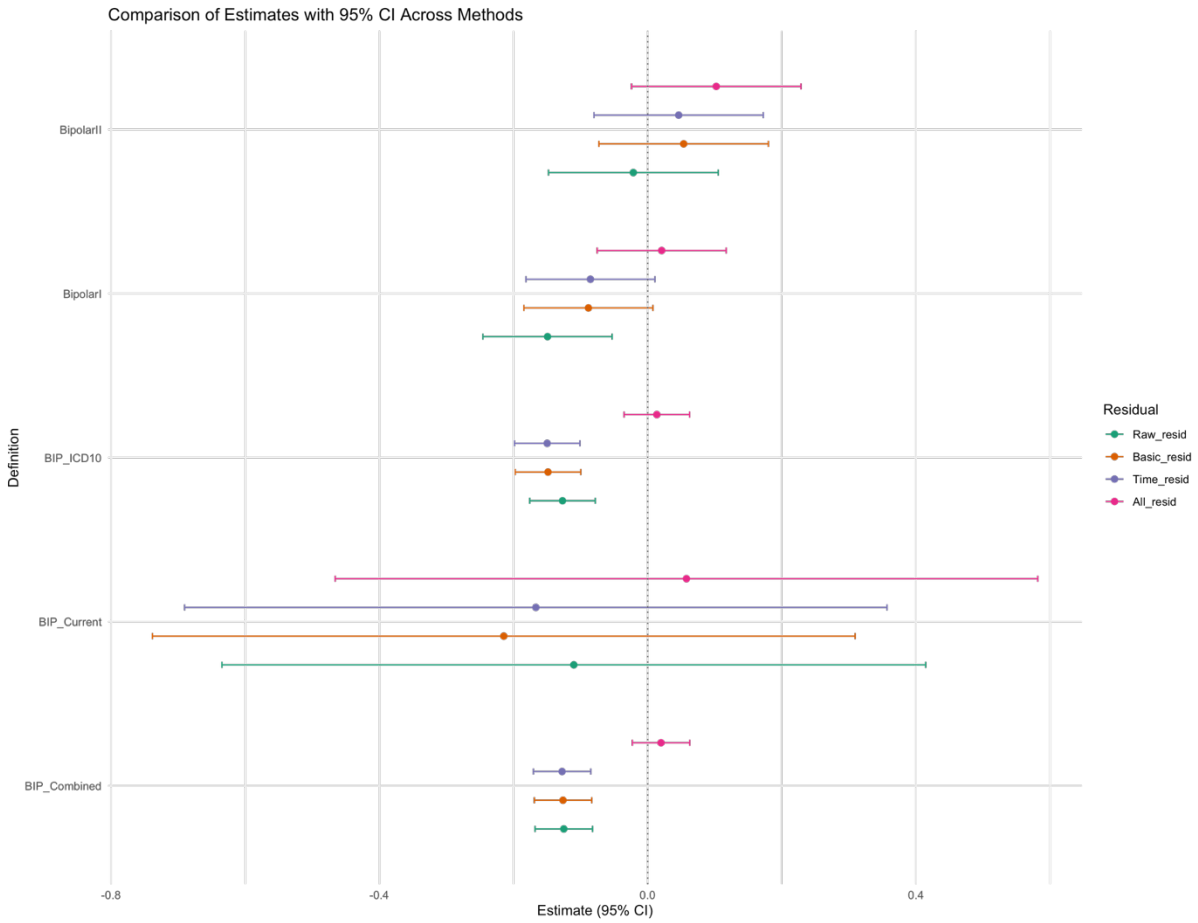


Figure IV.6. Lack of association between mtCN and bipolar disorder after adjusting for blood composition. This figure presents the mean difference and 95% confidence intervals (CI) in the scaled residuals of mtCN from models fitted with different sets of covariates (models identical to Figure IV.6), comparing bipolar disorder cases and controls (see case definition in Method). After adjusting for blood composition covariates, all significant associations disappeared, suggesting that the apparent changes in mtCN are primarily attributable to differences in blood composition. It should be noted, however, that the analysis remains underpowered to detect associations in bipolar disorder owing to the limited number of cases.

IV.3.5 Sensitivity analysis

To ensure that the lack of association between mtCN and mood disorders (MDD and BD) was not an artefact of our analytical pipeline, we conducted a sensitivity analysis to verify whether the pipeline could detect established associations reported in previous research. Specifically, we tested associations between mtCN and all ICD-10 codes that passed quality control and further grouped these codes into disease categories (**Methods**). As with mood disorders, several significant associations were observed after adjusting for basic covariates, or basic covariates plus sampling time, but most became non-significant after adjusting for blood composition. Interestingly,

in some cases the direction of association reversed. For example, the apparent lower mtCN in UKB participants recorded as having peripheral artery disease (PAD), compared to those without, before adjusting for blood composition, likely reflects differences in blood cell composition. After accounting for this, mtCN is higher in those recorded as having PAD—consistent with previous reports finding significantly elevated leukocytes in PAD patients and the impact of blood cell composition on the direction of associations[252] (**Figure IV.7**). Overall, these results support that the absence of association between mtCN and mood disorders is unlikely to be due to methodological artefacts.

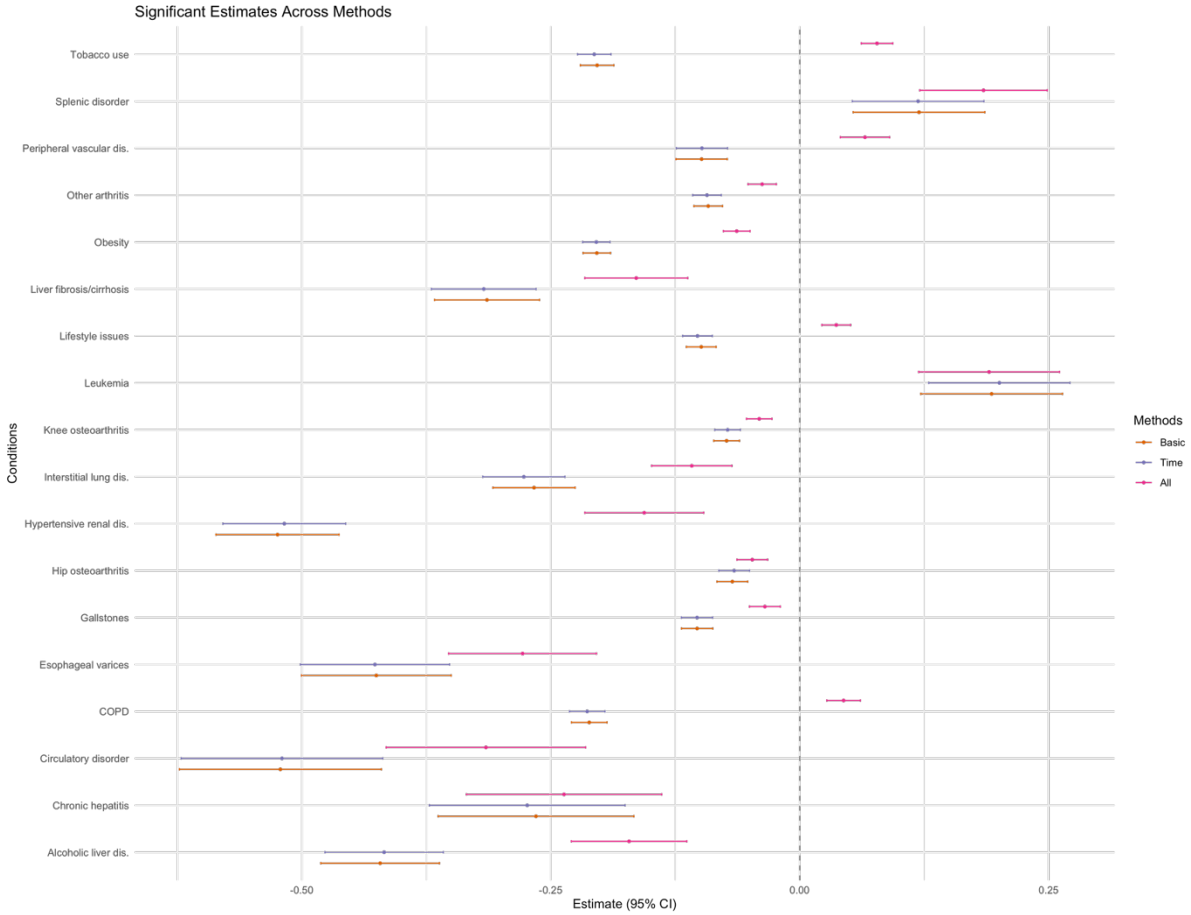


Figure IV.7. Sensitivity analysis confirms that the analytical pipeline detects established associations reported in previous research. To ensure that the lack of association between mtCN and mood disorders (MDD and BIP) was not an artefact of the analytical pipeline, we conducted a sensitivity analysis to test whether the pipeline could recover previously reported associations. Specifically, we examined associations between scaled residuals from different models of mtCN (models identical to Figure IV.6) and all ICD-10 codes that passed quality control. This figure displays health conditions that were significant after Bonferroni correction for multiple testing ($p < 1x10^{-5}$), ordered by significance level.

IV.3.6 Genetic control over mtCN

We performed a GWAS of mtCN after adjusting for all covariates described above, and the direction of association between covariates and mtCN are consistent with previous report (**Supplementary Table 4**). A total of 162 independent association signals were identified across the autosomes, broadly consistent with results from a recently published study also using the UKB[130] despite minor differences in the quality control pipeline (**Method**). The SNP-based heritability of mtCN was estimated at 0.056 (0.006). No significant genetic correlation was detected between mtCN and either MDD (P = 0.45) or BIP (P = 0.068), whereas, as expected, a strong genetic correlation was observed between MDD and BIP (rg= 0.51, P = 1.1 x 10^{-103}).

In a gene-based analysis, 772 genes were significantly associated with mtCN after correction for multiple testing (see **Method**). These genes were enriched in biological processes such as positive regulation of cell–cell adhesion, haemostasis, and immune functions including MHC class II assembly (**Figure IV.8**). They were also enriched in immune-related pathways, including co-inhibition mediated by PD-1 (**Figure IV.9**).

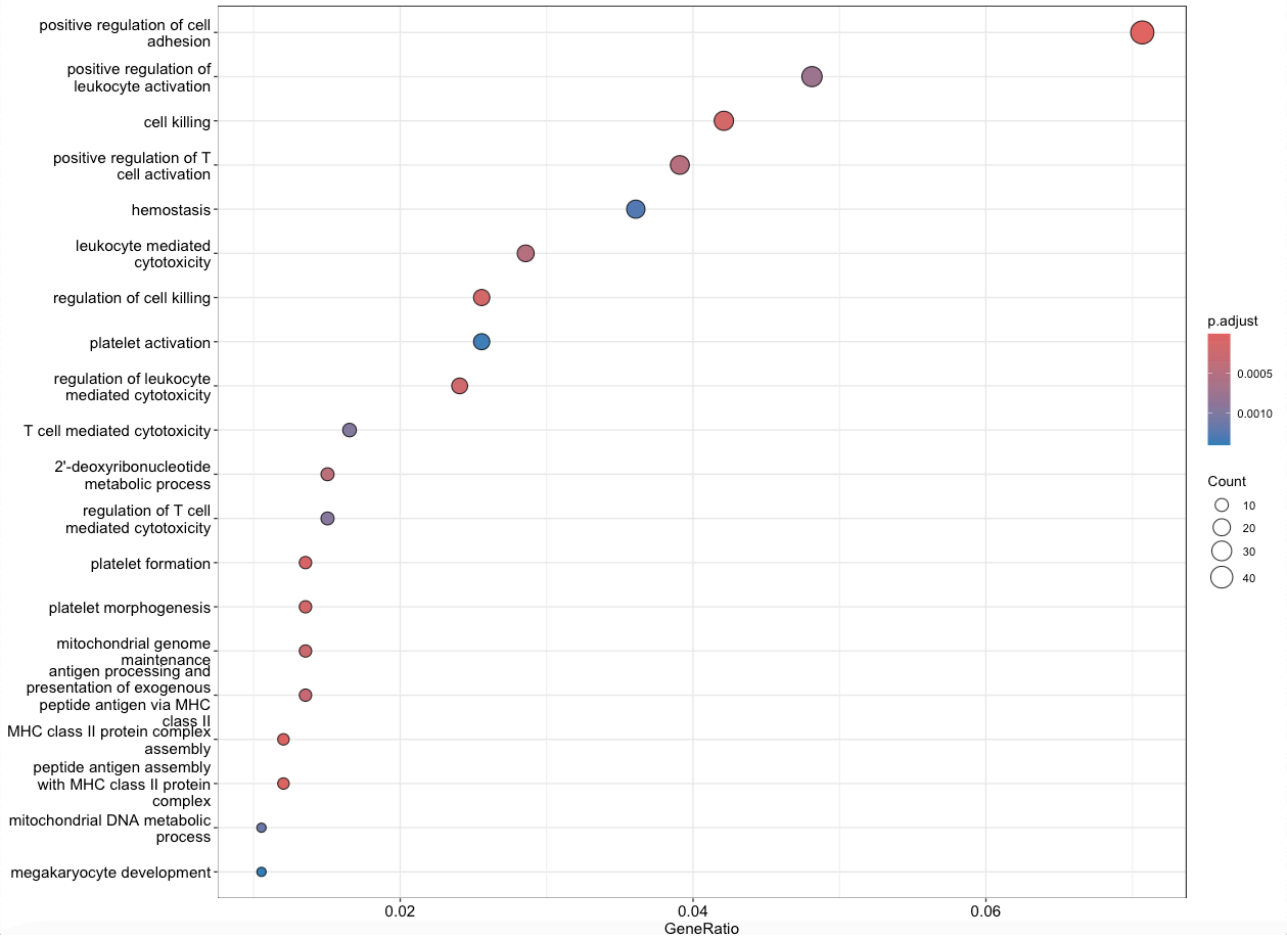


Figure IV.8. GO functional enrichment analysis for genes associated with mtCN in blood. In a gene-based analysis, we identified 772 genes significantly associated with mtCN after correction for multiple testing. To investigate their functional relevance, we performed a GO functional enrichment analysis. These genes were enriched in biological processes such as positive regulation of cell–cell adhesion, haemostasis, and immune-related functions, including MHC class II assembly.

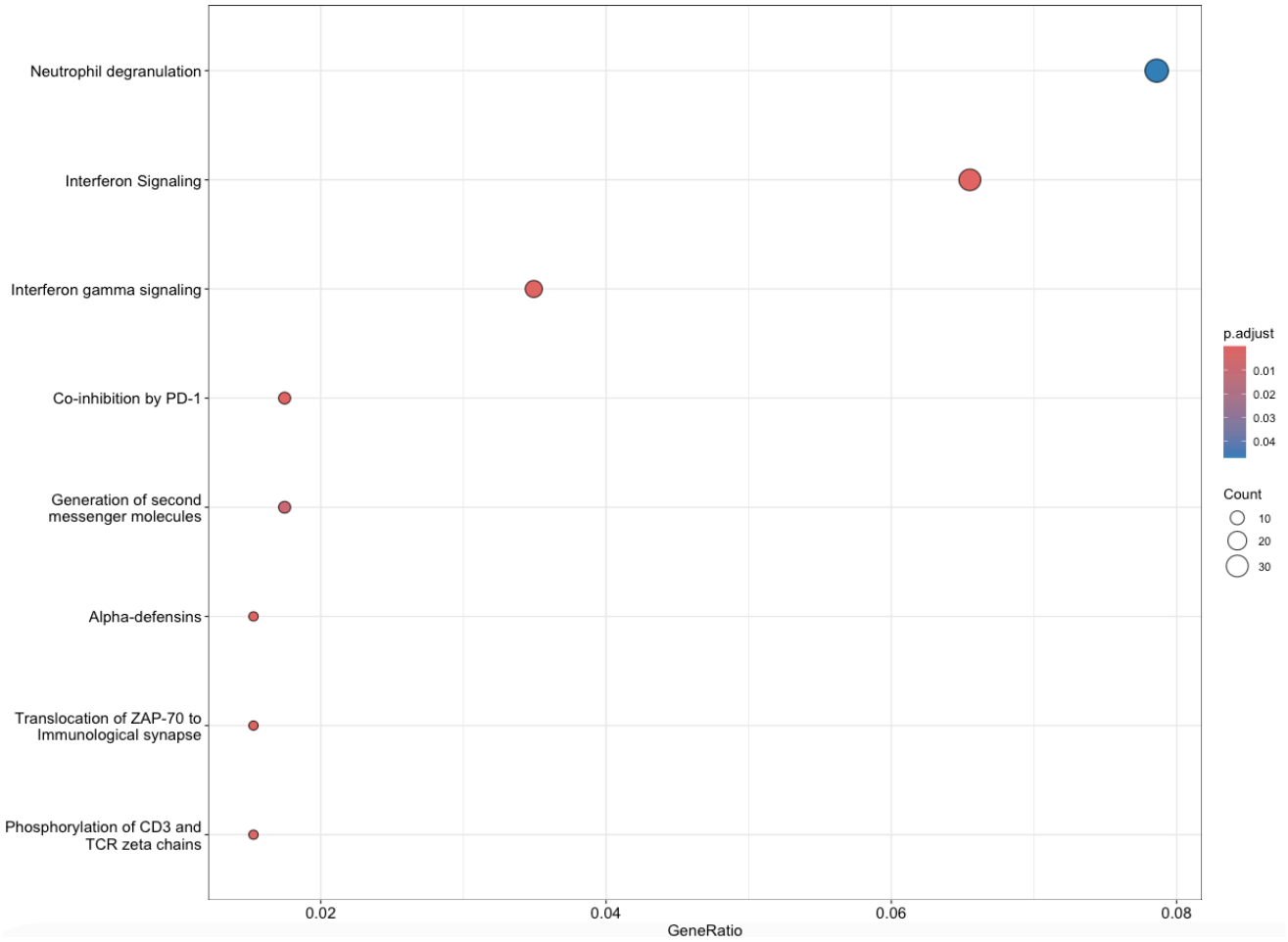


Figure IV.9. KEGG pathway analysis for genes associated with mtCN in blood. We tested the 772 genes significantly associated with mtCN in blood for enrichment in biological pathways using KEGG enrichment analysis. The results showed that these genes were enriched in pathways related to processes such as Interferon signaling and co-inhibition by PD-1.

IV.3.7 Cell types in blood and brain that are associated with mtCN

Although mtCN in blood showed no direct association with MDD or BIP, the observed links between white blood cell traits and MDD, together with the enrichment of mtCN-associated genes in immune pathways, suggest that mtCN regulation may be relevant to mood disorders in other tissues, such as the brain. To investigate this possibility, we analysed single-cell expression profiles to identify cell types in which genetic variants influence mtCN (see **Methods**). After Bonferroni correction, significant associations were found between microglia in the brain tissue ($P = 4.24 \times 10^{-5}$), monocytes in the bone marrow ($P = 4.45 \times 10^{-5}$), as well as circulating memory B cells ($P = 2.25 \times 10^{-4}$). Interestingly, microglia are resident immune cells of the central nervous system and have been shown to play important roles in both MDD[253], and BIP[254].

IV.4 Discussion

In this study, we systematically evaluated different mtCN calling strategies and made several observations relevant not only to mtCN estimation but also to copy number variation (CNV) pipelines more broadly. First, we found that regional coverage is uneven across the genome even in the absence of structural variation, and that such biases are platform specific. At the chromosomal level, chromosomes 9, 13, 14, 15, and 22 showed consistently lower coverage across sequencing platforms. This is unsurprising, as chromosomes 13, 14, 15, and 22 are all acrocentric chromosomes whose short arms contain genes encoding ribosomal RNA (rRNA) and are enriched for satellite (i.e. repetitive) sequences that are difficult to map[255]. Chromosome 9 is also known to contain a large number of satellite sequences[256]. These findings highlight the need for caution when comparing coverage across genomic regions: increased coverage at one locus does not necessarily reflect greater copy number variation relative to others, and such biases must be accounted for. When calculating autosomal coverage, we recommend including all autosomes, as coverage bias can outweigh the effects of even extreme structural variants, such as whole-chromosome duplications.

Despite these biases, coverage ratios between target and reference regions remain informative, as illustrated by the detection of sex chromosome copy number through comparison of coverage on chrX or chrY (target regions) with that of the autosomes (reference regions). Variability decreases when larger genomic regions are considered, consistent with the law of large numbers—for example, using the entire autosomal genome as a reference reduces the influence of individual chromosomal events. At finer scales, such as 100 bp bins on chrX, coverage ratios remain useful, albeit with increased noise.

Nevertheless, coverage ratios can be systematically biased even in large samples. For example, inferred chrY dosage in females deviates significantly from zero. These biases arise not only from biological or technical sources individually but also from their interaction. Biologically, homologous regions shared across chromosomes—such as PARs on chrX and chrY—cause ambiguous read mapping, while repetitive or low-complexity sequences further complicate assignment. Other biological processes, such as pregnancy and the presence of fetal DNA in maternal blood, may also interfere (although this is unlikely to feature in the UKB blood samples given the ages of recruitment). Technically, sample contamination, alignment artefacts, and limitations of the reference genome or sequencing technology itself all contribute.

Importantly, these biases persist even after stringent filtering, suggesting fundamental constraints of current approaches. Thus, while coverage ratios are informative, they should be interpreted with an awareness of their limitations. Across different calling strategies, we found that QC strategies implemented in previous studies yield highly correlated results, consistent with prior reports[227]. This suggests that results are generally reliable provided that target and reference regions are handled consistently. However, we recommend applying outlier-detection checks by comparing coverage estimates obtained using a more stringent approach (e.g. the UCS-based method) with those from a less stringent one (e.g. the Idxstats-based method), and excluding individuals with inconsistent coverage estimates to minimise technical artefacts and potential batch effects.

Our analyses indicate that previously reported associations between blood-derived mtCN and mood disorders[214-219] are more likely to reflect differences in blood cell composition and/or the timing of sample collection, rather than a direct biological relationship. A limitation of our study is that most mood disorder diagnoses were lifetime diagnoses, with many cases identified through questionnaires administered several years after participants attended the assessment centre. Although we attempted to identify current cases by comparing the date of the first recorded ICD-10 code for MDD or BIP with the date of blood sample collection for whole-genome sequencing, this approach remains less robust than those used in clinical studies, such as Cai et al.[215], where cases were current and hospitalised. Nonetheless, many previous studies were unable to adjust for cell type composition owing to limitations in sample collection. Given the substantial influence of cell type composition on mtCN—similar to its effect on numerous other traits—it is unsurprising that significant associations between mtCN and MDD were detected before adjusting for this factor. As shown in our analyses, raw mtCN exhibited significant associations with MDD, sometimes in opposing directions depending on the case definition, but these associations became non-significant once blood cell composition was accounted for.

As mtCN varies considerably across tissues—with mean values differing by up to 50-fold between tissue types[257] —it remains possible that mtCN is relevant to mood disorders in other tissues. Notably, genes associated with mtCN in blood are enriched near loci related to microglia-specific gene expression in the brain. Given the crucial role of microglia in both MDD and BIP, mtCN in brain tissue warrants further investigation and may represent a potential therapeutic target.

V

Discussion

V.1 Summary of Findings

The overarching aim of this thesis is to develop and apply new methodologies to address key challenges in quantitative genomics and to advance our understanding of the genetic architecture of human disease.

To achieve this, the thesis introduces methods and theoretical frameworks spanning the full spectrum of GWAS—including pre-GWAS preparation, GWAS analysis, and post-GWAS analytical pipelines. In addition, with the growing availability of WGS data, the thesis highlights the potential of WGS to uncover changes including copy number variation that may not be adequately captured by traditional GWAS approaches.

Since the publication of the first GWAS in 2005[258] and the release of the full genotyping cohort from the UK Biobank in 2017[259], the field has witnessed remarkable progress. Within a relatively short period, numerous quantitative genomics tools and methodologies, as reviewed in this work, have been developed and shared with the broader research community. This rapid evolution has greatly enhanced both our understanding of human genetic architecture and the translation of genomic findings into clinical practice.

While it is exciting to embrace emerging technologies and novel methods, it is just as important to pause from time to time, reflect on the path we have taken, ask whether we used each tool in the right context, recognise their limitations, and consider if there might be better ways to tackle the same scientific challenges.

Observations in Chapter II serve as an example of this point. In the meta-analysis of multiple GWAS summary statistics from different study cohorts, the fixed-effect inverse-variance meta-analysis is the most commonly used model. However, this approach rests on the assumption that the true effect sizes are identical across all cohorts. When cohort heterogeneity exists—for example, differences in ancestry, phenotype definition, or covariates included in the analysis—this assumption no longer holds. In such cases, it is unsurprising to observe both a reduction in SNP-based heritability in the meta-analysed cohort and the seemingly paradoxical finding that out-of-sample prediction R^2 in some cohorts can exceed the theoretical upper limit.

However, the aim of Chapter II is not to suggest that we should avoid collecting heterogeneous samples. On the one hand, heterogeneity is inevitable; on the other, it can in fact enhance our understanding of human genetic architecture. The focus, therefore, should be on applying the most appropriate model whenever possible, while also recognising and highlighting the limitations inherent in any model that is used.

To further illustrate this point, in Chapter III we developed a new software tool designed to improve the identification of independent associations in GWAS summary statistics by leveraging differences in LD structure across ancestries. Both simulation studies and real-world data analyses demonstrated that this approach provides greater power and reduces false discovery rates compared with existing methods. These findings highlight the value of incorporating data from diverse genetic backgrounds.

Beyond the analysis of genetic sequences themselves, the growing availability of WGS offers opportunities to extract valuable insights from the sequencing process itself. Chapter IV explores this concept by focusing on one such by-product—mitochondrial copy number. In this chapter, we developed an additional quality control pipeline for WGS data and evaluated different approaches for mitochondrial copy number estimation, including the role of key covariates. With these improvements, we demonstrated that previously reported associations between mitochondrial copy number in blood and major depressive disorder were more likely attributable to changes in blood cell composition in affected individuals. In contrast, mitochondrial copy number remained significantly elevated in cardiovascular and cancer patients under the revised pipeline, suggesting either altered energy demand or pathogenic processes with potential as therapeutic targets. Together, these findings underscore the vast potential of WGS, not only in terms of sequence data but also in the wealth of information generated during the sequencing process.

V.2 Limitations

All models and discoveries have limitations, and the methods and findings presented in this thesis are no exception. As discussed above, it is important to recognise these limitations. Some can, in fact, serve as valuable directions for future development. In addition to the limitations discussed within each chapter, this section considers the limitations of each chapter from a broader and more fundamental perspective.

V.2.1 Limitations in the theory explaining the impact of heterogeneity on h^2_{SNP} and out-of-sample prediction R^2

The derivations presented in Chapter II are based on the infinitesimal model, which assumes that each genetic variant contributes an equally small proportion of the total genetic variance. For highly polygenic traits, such as depression and educational attainment discussed in the text, this framework provides a plausible explanation for patterns observed in empirical data. However, the extent to which these results generalise to traits with lower levels of polygenicity, or to traits in which some common variants exert large effects, remains uncertain and may be limited.

Several approaches could potentially improve the generalisability of the theoretical framework. Rather than assuming that each SNP explains an equal proportion of the genetic variance, it may be more realistic to assume that effect sizes follow a mixture distribution. For example, models such as LDpred2 adopt a "spike and slab" model in which the distribution of causal effect sizes consists of a point mass at zero combined with a normal distribution representing non-zero effects[96]. Similarly, SBayesR models SNP effects as a mixture distribution containing a point mass at zero together with several normal distributions with different variances[165]. These approaches allow for a more flexible representation of genetic architecture and may better accommodate traits in which per-SNP heritability vary substantially across loci.

Further theoretical and empirical work will be required to extend and refine these models. Empirical analyses depend on the availability of detailed summary statistics for each trait across multiple cohorts and, ideally, access to individual level data in order to improve the estimation of h^2_{SNP} and r_g. However, access to such data is often limited. This constraint highlights the importance of developing frameworks that facilitate secure data sharing and international collaboration while maintaining privacy and ensuring the protection of sensitive information.

V.2.2 Limitations in the Manc-COJO algorhithm

As described in Section III.2.1, the method assumes additive genetic effects, shared causal SNPs and effect sizes across ancestries, and independence among SNP effects. These assumptions can be violated in multiple ways, as discussed in Chapter III. Moreover, even when the assumptions hold, the method cannot always resolve independent associations in certain regions, depending on the underlying genetic architecture.

For example, consider a region containing one causal SNP and $M-1$ non-causal SNPs in complete LD with it (**Supplementary Figure V.1a**). Such situations are not uncommon, especially with WGS data where higher SNP density increases the likelihood of including causal variants, but also inflates the number of variants in strong or perfect LD with them. Under such conditions, the probability of correctly identifying the true causal variant is approximately $1/M$. Moreover, for a substantial fraction of loci, even when only a single causal SNP is present and multi-ancestry data are available, fully resolving LD remains infeasible with realistic sample sizes (**Supplementary Figure V.1b**). This underscores a fundamental limitation of the Manc-COJO algorithm in fully disentangling LD structure and helps explain the commonly observed high false-positive inclusion rates across many fine-mapping methods[260].

Recognising the limitations inherent in the Manc-COJO model, we developed Manc-COJO --MDISA—not only to address ancestry-specific variants, but more importantly, to allow the incorporation of causal variants validated through wet-lab analyses. We therefore extended the framework from Rummel and Ziller et al.[261] and propose an idealised pipeline for future genetic research that is cost-effective and emphasises health equity across ancestries (**Supplementary Figure V.2**). Resources should be allocated to underrepresented ancentry groups during data collection. Outcomes from statistical genetic analyses should not only guide functional follow-ups but also inform future data collection strategies. With trans-ancestral data, methods like Manc-COJO can break down LD and narrow down candidates for wet-lab validation. Conversely, wet-lab findings should refine statistical models, enhancing the detection of independent associations.

V.2.3 Limitation in mitochondria copy number analyses and WGS analyses

In our WGS analysis, we found that each stage—from sample preparation and sequencing through quality control, reference alignment, and variant calling—introduces its own limitations and potential sources of bias. We observed both regional coverage biases that differ by platform and broader, chromosome-wide biases that are consistent across sequencing technologies. Notably, the recurrent depression of coverage on chromosomes 13, 14, 15, 21, and 22 is not coincidental: these are acrocentric chromosomes whose short arms contain multiple copies of rRNA genes. Because these regions consist of highly repetitive sequences that are

challenging to map unambiguously, sequencing reads are frequently misaligned or excluded, leading to apparent reduced depth.

Another clear example of mapping artifacts arises when examining sex-to-autosome coverage ratios. In the UK Biobank dataset, the mean chrY:autosome coverage ratio in females is approximately 0.05. While a true biological signal may explain this in rare cases (e.g., the presence of fetal Y DNA during pregnancy), the more likely explanation for the majority of female participants is mis-mapping, due to homologous regions shared between chrY, chrX, and autosomes.

Together, these observations highlight the interplay between sequencing chemistry, reference genome completeness, and mapping algorithms in shaping observed coverage. While adjusting for such biases is beyond the scope of this thesis, they underscore the considerable room for future improvement in sequencing technologies and analytical pipelines.

Another important limitation of this study is that for many diseases, associations between mtCN and disease status likely reflect lifetime risk rather than disease state at the time of sampling. This limitation arises because the available data do not capture how frequently a disease occurs or how long it persists. For conditions such as depression and bipolar disorder, we attempted to infer disease status at the time the blood sample was taken. However, the number of individuals who could be confidently classified in this way was small due to limited information in the Biobank records, highlighting the difficulty of inferring contemporaneous disease status from these data.

Moreover, we rely on WGS data derived from bulk blood samples to infer mtCN. Such samples contain a mixture of intracellular DNA from multiple cell types as well as cell-free DNA (cfDNA). As different cell types have varying mtCN, estimates derived from bulk tissue complicate the interpretation of the results. In particular, it is difficult to determine whether observed mtDNA changes are attributable to specific tissues or cell types, or whether they reflect a general change across all cells. Ideally, this limitation could be addressed using single-cell or at least tissue-level WGS data. Sequencing such samples would not only improve mtCN estimation but also enable the quantification of somatic mutations in different cell types and the evaluation of their impact.

For example, when examining the coverage ratio between individual chromosomes and the autosomal average, most participants cluster around the mean; however, approximately 0.3–1% (depending on the threshold) show pronounced elevations or reductions in the ratio for a specific chromosome relative to the autosomes. Notably, most of these individuals have no recorded congenital aneuploidy or chromosomal abnormalities, are non-cancer participants, and were not excluded during quality control, suggesting that these deviations are unlikely to be explained solely by sample contamination. One plausible explanation is clonal haematopoiesis—an age-related process characterised by the accumulation of somatic mutations and structural alterations in haematopoietic stem cells cells[262]. These observations highlight the potential scale of somatic changes and the promise of single-cell WGS, which is currently limited by both resource and technical constraints.

V.3 Future directions

The ultimate goal of quantitative genomics, from my perspective, is to address three fundamental questions: why biological variation exists in its current form, what the consequences and benefits are of understanding these underlying causes, and how these questions can be rigorously investigated. In terms of future directions, it is essential to think about where the current gaps lie in answering these questions, whether these gaps could be addressed in an idealised setting with unlimited resources, and, if so, how we can approach these questions in practice with limited resources while approximating such ideal conditions.

We have already made substantial progress in understanding biological variation, both from the perspective of biological pathways and from an evolutionary standpoint; however, significant gaps remain. One of the main reasons is that GWAS is, at its core, an association test that identifies relationships between genetic variants and phenotypes rather than establishing causality. Moreover, even the associations reported by GWAS are neither entirely true nor complete. As discussed, such association tests can be influenced by multiple sources of bias. Although we attempt to mitigate these effects, for example by fitting covariates, the process of covariate adjustment can itself introduce bias and generate spurious associations. Furthermore, most study designs, including ours, rely on simplifying assumptions. While these assumptions facilitate statistical modelling and computational efficiency, and can explain a large proportion of variation, they inevitably constrain the generalisability of the methods. For instance, under an additive genetic model, important factors such as gene–environment interactions, gene–gene interactions,

and structural variation are typically ignored, despite their significant biological relevance.

Despite these limitations of GWAS, substantial effort has been directed toward deriving functional and causal insights from association signals, using both statistical approaches in the dry lab and experimental approaches in the wet lab. However, these approaches are constrained not only by the limitations of the input data but also by the nature of the methods themselves. Methods such as MR rely on assumptions that cannot be directly tested, while the generation of multi-omics data can introduce additional sources of bias, for example through epitope effects in proteomic data, as discussed previously. Furthermore, wet-lab experiments are not only resource-intensive but also vulnerable to confounding influences, including the choice of cell line and even the sex of the experimenter[263].

One of the core issues underlying the observations above is that the relationships between environmental factors, genotypic changes, and phenotypic outcomes are extraordinarily complex. These elements interact in complicated and often non-linear ways, such that even with the most careful study designs, it is not possible to identify or properly adjust for all confounding influences across all traits.

In an idealised scenario, in which an indefinitely large number of individuals have been genotyped without error, and comprehensive phenotypic and environmental data capturing all possible sources of variation have been collected, it would in principle be possible to systematically perturb individual factors while holding all others constant and observe the corresponding responses. Under such conditions, one could imagine constructing an extremely complex network capable of explaining the majority of causal relationships between genetic variation and phenotypes.

Apparently, this scenario is fundamentally unattainable. The human genome comprises approximately 3 Gb of sequence, with four possible nucleotides at each locus, in addition to numerous forms of structural variation. When coupled with the vast number of environmental exposures, this results in an astronomical number of possible combinatorial states. Consequently, even if every individual in the world were genotyped, the available data would remain far too sparse to adequately sample this combinatorial space.

However, this recalls the discussions of the 1990s, when the strongest human chess player was defeated by Deep Blue, a supercomputer specifically designed for chess.

At the time, both scientists and chess players argued that such success was possible because chess has a relatively limited combinatorial space, whereas the game of Go presents an astronomical number of possible states. It was therefore widely believed that, at least in the foreseeable future, no supercomputer would be capable of defeating the strongest human Go players.

Yet, two decades later, the development of AlphaGo overturned this assumption[264]. No human player can now surpass these artificial intelligence systems—not because we have acquired sufficient computational power to exhaustively evaluate all possible combinations, but because it has become clear that such exhaustive enumeration is not necessary to achieve our objectives. Some of the principles underpinning AlphaGo have been, and can be, translated to our efforts to understand the mechanisms of biological variation.

As in the game of Go, where not all moves are legal, not all genetic combinations are viable. In fact, many are lethal, and individuals carrying such combinations would not survive to adulthood. Consequently, current GWAS approaches often treat loci as equal (unless augmented with functional annotation), which does not fully reflect underlying biological constraints. In the future, it may be useful to stratify genomic regions based on evolutionary conservation and mutational consequences. Highly constrained or invariant regions, where mutations are more likely to have severe functional effects, could be modelled collectively as haplotypes rather than as independent variants. Notably, conservation does not strictly correspond to regions where mutations are highly deleterious. Perturbation studies in many ultraconserved regions have shown surprisingly mild effects[265]. Therefore, although these ultraconserved regions could also be modelled as a whole, further research to explain this apparent discrepancy would be valuable. Overall, this strategy could reduce the combinatorial search space and alleviate the burden of multiple testing in GWAS.

At the other extreme, certain genomic regions are highly complex, such as the MHC locus, while others exhibit elevated rates of somatic variation. Immune-related regions, for instance, are particularly dynamic and responsive to environmental stimuli. Likewise, chromosome Y is inherently unstable and can undergo progressive loss in association with environmental exposures (e.g., smoking) and aging [266]. Conventional WGS variant calling pipelines are primarily designed to identify consensus germline variants and often filter out signals that deviate from this expectation. As a result, somatic variation, although present in WGS data, is

frequently overlooked or removed during standard processing. In addition, most existing studies of somatic variation focus on tumor samples, with comparatively limited investigation in nonmalignant tissues. Therefore, there is a need for analytical strategies that can effectively incorporate these complex and highly variable regions without substantially increasing the multiple testing burden.

Similar to AlphaGo, not all legal moves occur with equal probability. In genotype data, it is well established that LD leads to certain allele combinations occurring more frequently than expected by chance. A substantial portion of this thesis therefore focuses on evaluating the impact of LD. However, most current LD measures rely on the implicit assumption that LD structure is homogeneous within a population when estimating correlation coefficients. As a result, emerging or localised LD patterns that have not yet become widespread may be underrepresented or overlooked. This limitation is particularly relevant in admixed populations, where reliance on in-sample LD becomes essential. Future work should aim to model LD more flexibly and to exploit this structure to further reduce the dimensionality of genotype data.

Similarly, phenotypes are often not independent. For example, traits such as leg length and standing height are correlated, as are conditions such as bipolar disorder and depression. Environmental variables also exhibit dependence; for instance, temperature and air pollution levels may be correlated, as colder regions often have less industrial activity and consequently lower pollution. Future research should therefore focus on incorporating these correlation structures to reduce the dimensionality of phenotypic and environmental data.

By reducing the dimensionality of genotype and phenotype data, and concentrating on the most informative phenotypic subsets within the domain of interest, future studies may be better positioned to integrate a comprehensive set of genetic variants alongside relevant phenotypic and environmental factors within a unified modelling framework. Such a framework would be better suited to capturing interactions and non-linear relationships. As discussed, the approach adopted here is largely informed by a machine learning perspective, in which model development is inherently iterative. Accordingly, wet-lab and dry-lab evidence in the future should be integrated in a cyclical manner, as outlined in the previous section. In the longer term, this process may become increasingly automated, with elements of experimental work and data collection carried out using robotic systems.

It is anticipated that such an iterative framework will accelerate our understanding of biological variation without requiring explicit enumeration of all possible combinatorial states. Furthermore, with the growing availability of data sources such as electronic health records and imaging data, artificial intelligence may increasingly be applied to integrate multi-modal datasets and construct a “digital self”, enabling more accurate prediction of health outcomes and treatment responses.

In recent years, the field of artificial intelligence has already advanced considerably and has become increasingly integrated with genomics. One example is AlphaGenome, which employs deep learning to model genetic variation and its functional consequences[267]. However, it is important to recognise that the outputs of deep learning models are inherently constrained by the quality and completeness of their training data. Current implementations of AlphaGenome, for instance, are aligned to GRCh38.p13, which is not gapless. Even the more recent telomere-to-telomere (T2T) reference genome (e.g. T2T-CHM13) remains limited, as it represents only a single haplotype.

Genomic regions characterised by repetitive sequences, such as those near telomeres and centromeres, as well as regions with substantial somatic or structural variation, as discussed above, are still not fully captured by existing reference genomes or current deep learning approaches. Addressing these limitations may require the development of individual-level reference genomes, followed by variant calling within groups of individuals sharing similar haplotypic backgrounds.

The limitations of these well-established methods highlight that, alongside adopting emerging technologies and novel approaches, it is equally important to step back and critically evaluate existing quatititive genomics tool from both methodological and engineering perspectives. Many software tools in this field were developed over a decade ago, and advances in modern programming frameworks and computational infrastructure now offer considerable scope for optimisation. Revisiting and re-engineering these tools could substantially improve efficiency, thereby reducing the time and cost associated with large-scale analyses.

V.4 Conclusion

In summary, this thesis advances quantitative genomics by developing and applying new methodologies to tackle central challenges in the study of complex traits and

diseases. While many projects were initially motivated by psychiatric disorders, the principles developed are broadly applicable across human traits and diseases.

A recurring theme is the influence of cohort heterogeneity. Differences in ancestry, phenotype definition, and study design can distort analyses, as illustrated in Chapter II, where polygenic prediction accuracies exceeded theoretical bounds. Yet heterogeneity also creates opportunities: Chapter III introduced Manc-COJO, a trans-ancestry method that leverages differences in linkage disequilibrium to improve power and reduce false positives, while Manc-COJO (--MDISA) extended this framework to incorporate experimentally validated causal variants.

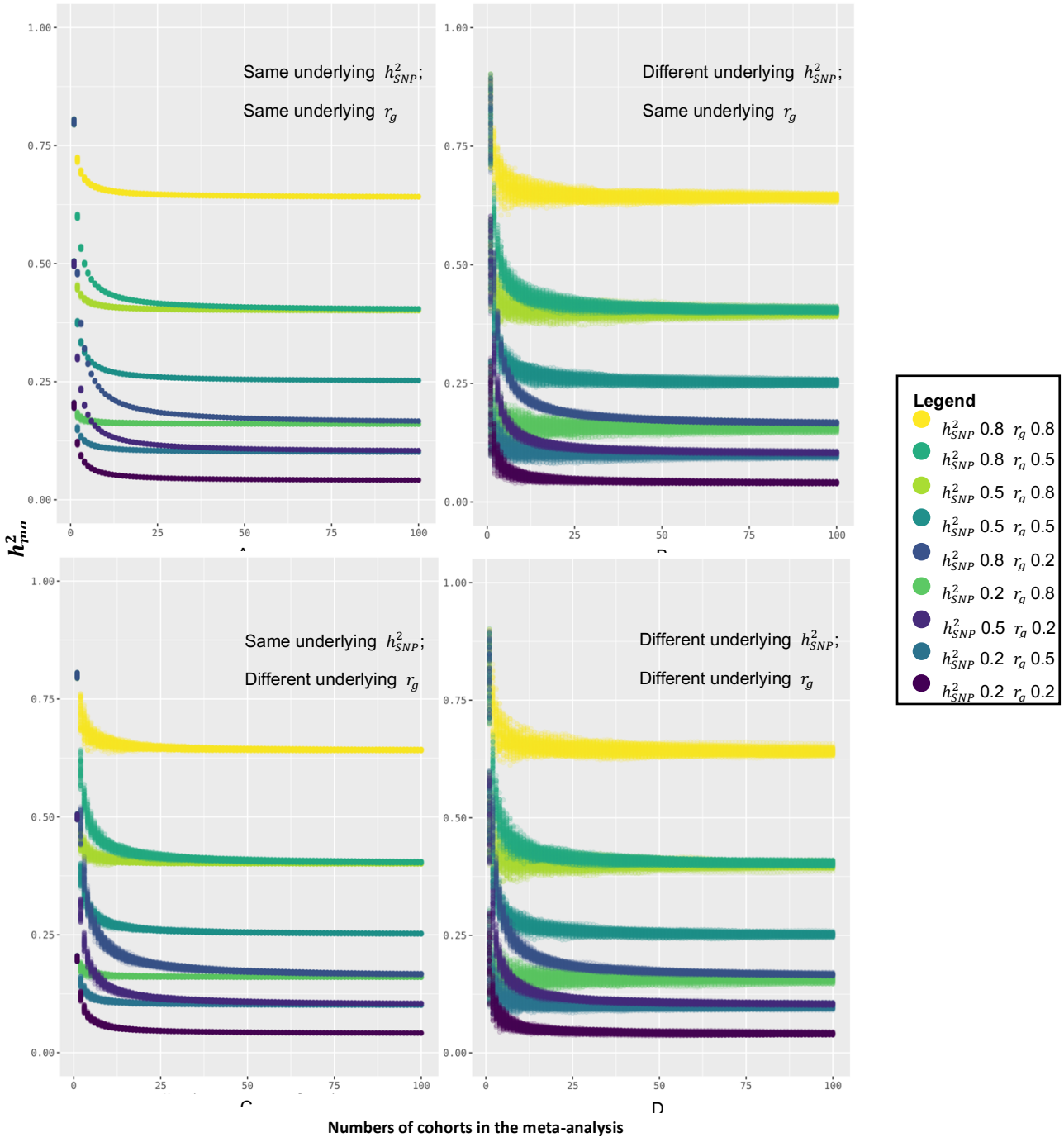
Beyond sequence variation, Chapter IV highlighted the value of secondary measures from whole-genome sequencing. By refining pipelines for mitochondrial copy number estimation, we showed that associations with depression were likely confounded, whereas robust signals in cardiovascular disease and cancer point to potential biological relevance.

Together, these studies demonstrate that every stage of genomic research—from sequencing to modelling—carries limitations, yet each also offers unique opportunities. Recognising and addressing these constraints is essential for progress. Overall, this thesis contributes methodological innovations and conceptual frameworks that deepen our understanding of the genetic architecture of complex human traits and diseases.

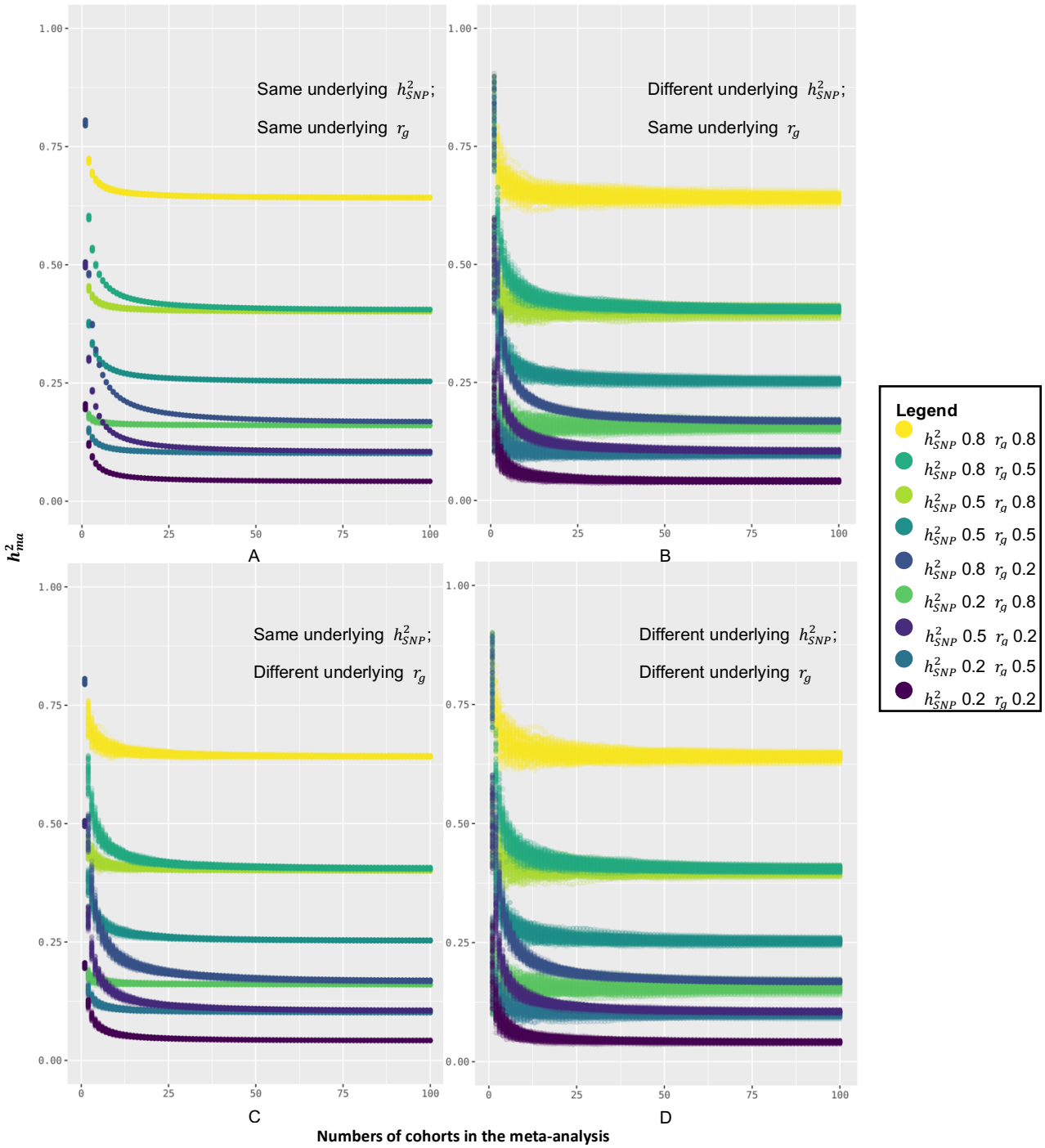
A

Appendix

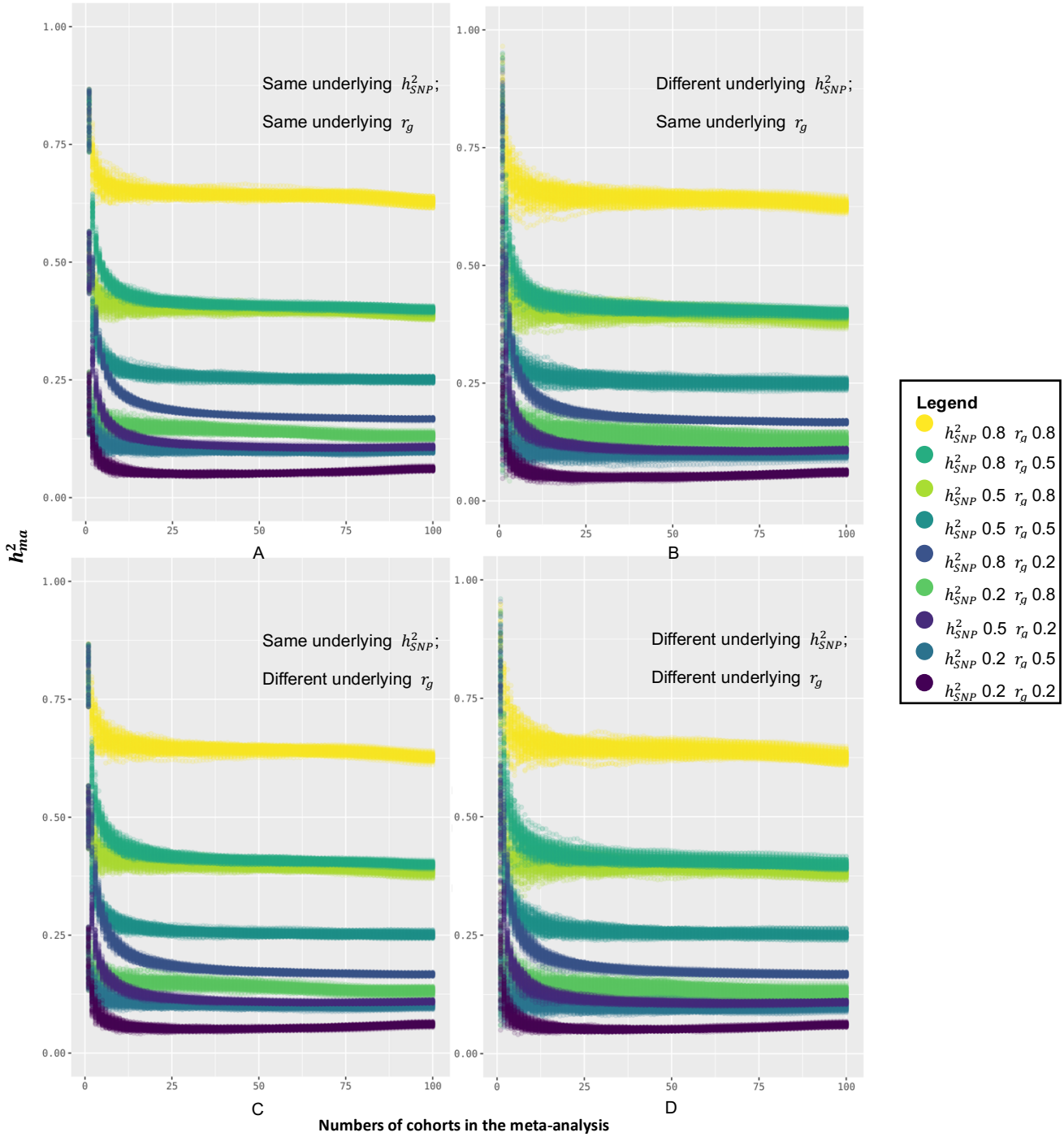
A1. Supplementary files for Chapter II



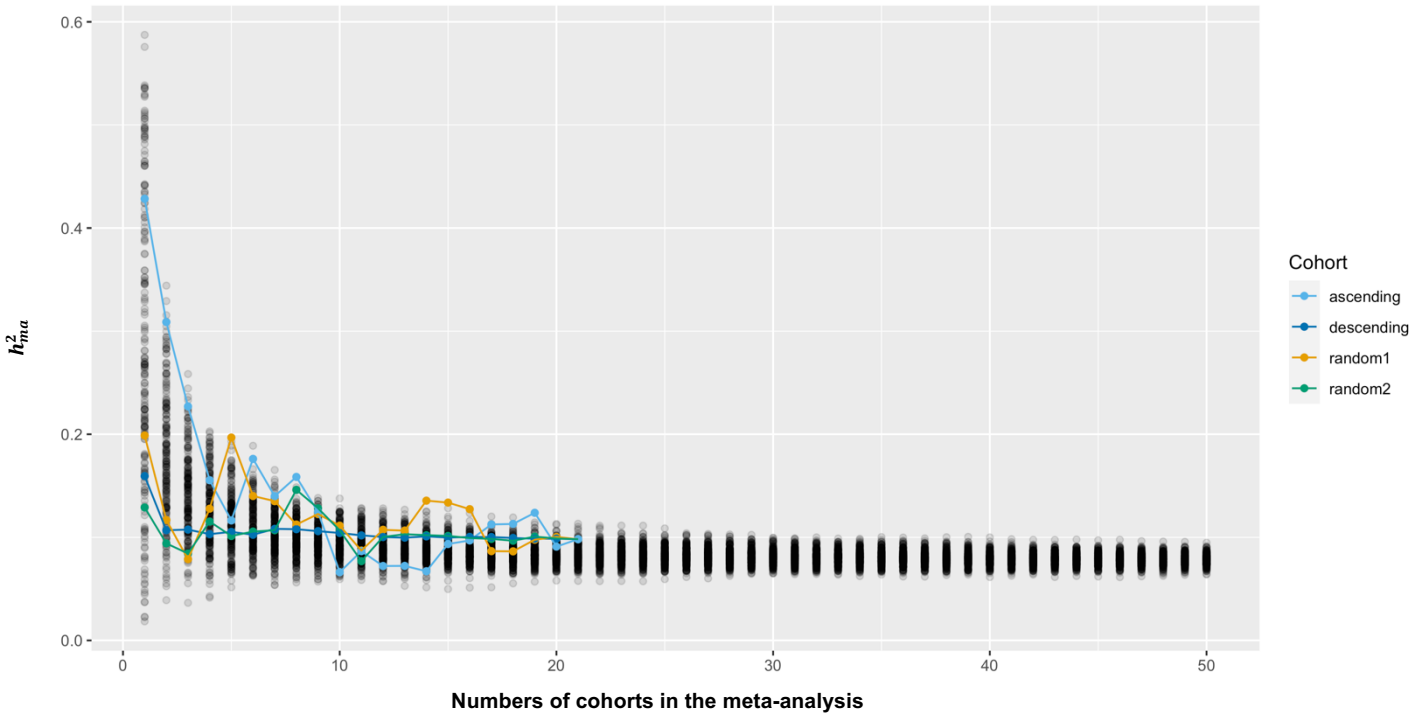
Supplementary Figure II.1 Estimates of h^2_{ma} with an increased number of cohorts being meta-analysed under different r_g and h^2_{SNP} settings, when sample sizes of each cohorts are relatively large and variable. In this figure, x-axes are the number of cohorts being meta-analysed, and y-axes are the h^2_{ma}. Each cohort being meta-analysed is simulated to have a sample size between **50,000** to **100,000**, and the cohort with the largest effective sample size was meta-analysed first. (A) True r_g and h^2_{SNP} are the same in all cohorts so variation between simulation replicates represents sampling variation. (B) True r_g are the same and h^2_{SNP} are different between cohorts. (C) True h^2_{SNP} are the same but r_g are different between cohorts (D) True r_g and h^2_{SNP} are all different in each cohort. Each scenario (depicted by colour) has been repeated 100 times and so the graph shows the variation across replicates.



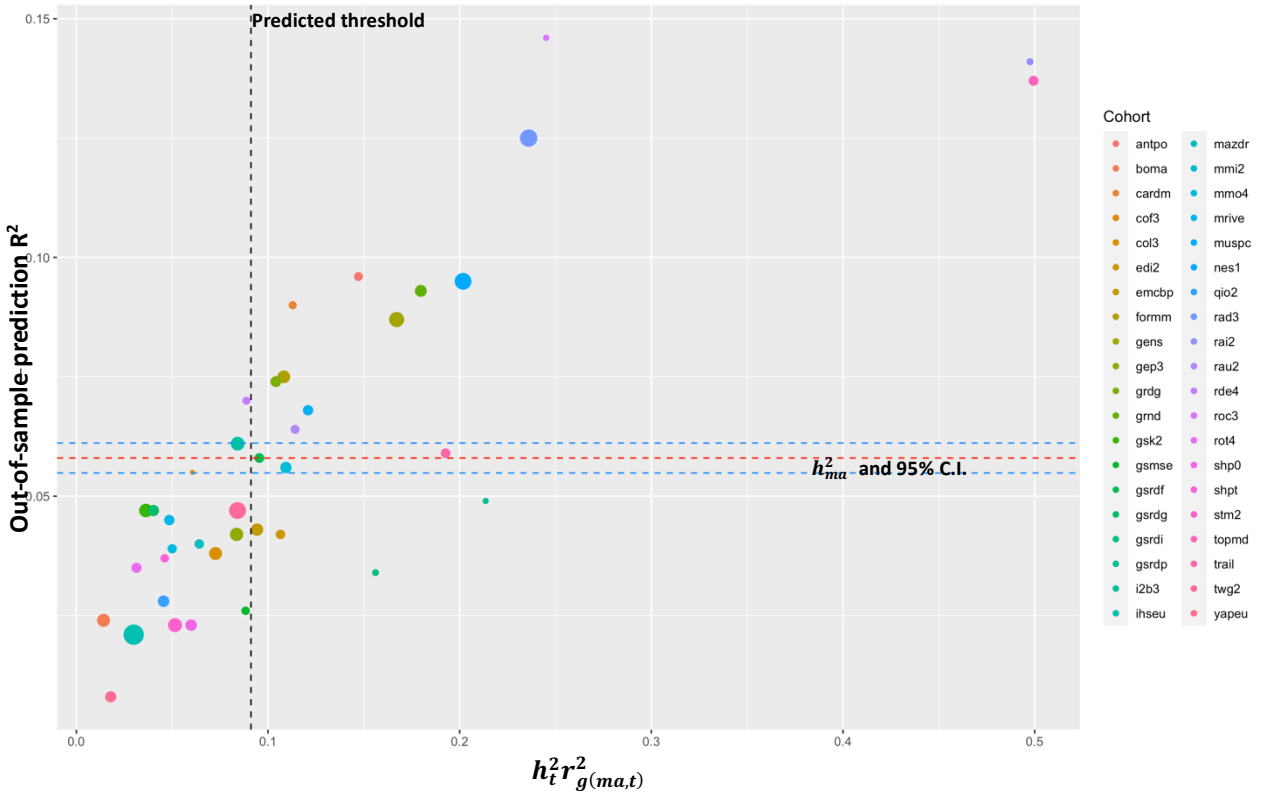
Supplementary Figure II.2 Estimates of h^2_{ma} with an increased number of cohorts being meta-analysed under different r_g and h^2_{SNP} settings, when sample sizes of each cohorts are relatively large and constant. In this figure, x-axes are the number of cohorts being meta-analysed, and y-axes are the h^2_{ma}. Each cohort being meta-analysed is simulated to have a sample size in **an arithmetic sequence from 1,000 to 100,000**, and the cohort with the largest effective sample size was meta-analysed first. A) True r_g and h^2_{SNP} are the same in all cohorts so variation between simulation replicates represents sampling variation. (B) True r_g are the same and h^2_{SNP} are different between cohorts. (C) True h^2_{SNP} are the same but r_g are different between cohorts (D) True r_g and h^2_{SNP} are all different in each cohort. Each scenario (depicted by colour) has been repeated 100 times and so the graph shows the variation across replicates.



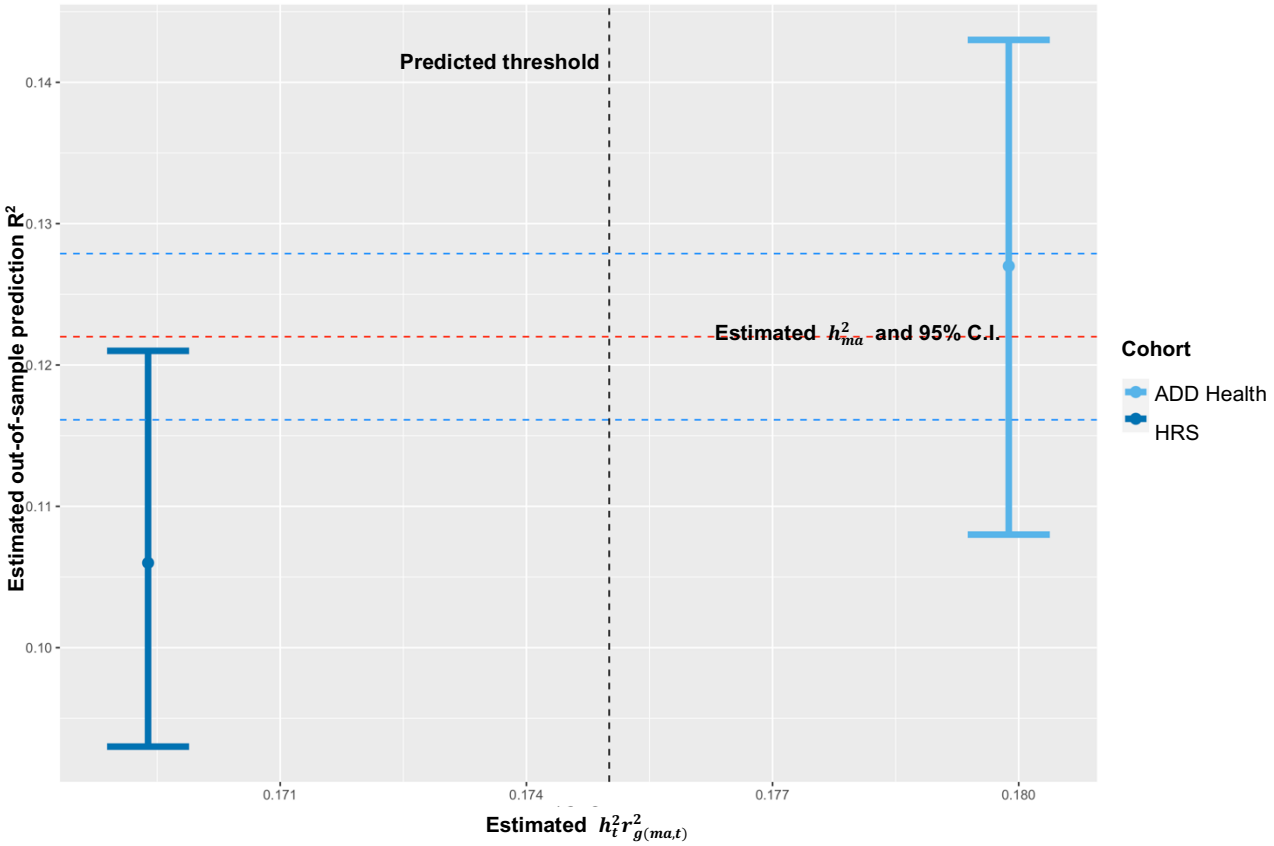
Supplementary Figure II.3 Estimates of h^2_{ma} with an increased number of cohorts being meta-analysed under different r_g and h^2_{SNP} settings, when sample sizes of each cohorts are relatively small and variable. In this figure, x-axes are the number of cohorts being meta-analysed, and y-axes are the h^2_{ma}. Each cohort being meta-analysed is simulated to have a sample size in **an arithmetic sequence from 100 to 10,000**, and the cohort with the largest effective sample size was meta-analysed first. A) True r_g and h^2_{SNP} are the same in all cohorts so variation between simulation replicates represents sampling variation. (B) True r_g are the same and h^2_{SNP} are different between cohorts. (C) True h^2_{SNP} are the same but r_g are different between cohorts (D) True r_g and h^2_{SNP} are all different in each cohort. Each scenario (depicted by colour) has been repeated 100 times and so the graph shows the variation across replicates.



Supplementary Figure II.4 h^2_{ma} can fluctuate with increasing number of cohorts, but the average over simulations shows the clear trends to a plateau. In this figure, black dots represent simulation results as for Figure II.3 and Supplementary Figure II.1-II.3, and the y-axis is the h^2_{ma}. Each cohort being meta-analysed is simulated to have a sample size in **an arithmetic sequence from 100 to 10,000**, and the cohort with the largest effective sample size was meta-analysed first. In our simulation, true h^2_{SNP} range from 0.1 to 0.5 (which reflects what we observed in MD). At the same time, we set the mean of true r_g to be 0.26 (we got this number by averaging all pairs of r_g we observed in the PGC datasets) and set the range of true r_g to be 0.5 (arbitrary but conservative as the range of observed r_g is larger than 0.5). The coloured dots represent the change in estimated h^2_{ma} with an increasing number of cohorts being meta-analysed (in PGC datasets), as shown in Figure 2. This Figure shows that depending on the order of meta-analysis, h^2_{ma} could fluctuate, i.e. does not always decrease, with an increasing number of cohorts. However, when we repeat the meta-analysis under the same h^2_{SNP} and r_g settings a hundred times, the overall trend is clear - h^2_{ma} estimates and s.e. of h^2_{ma} estimates decrease as the number of cohorts being meta-analysed increases until reaching a plateau.



Supplementary Figure II.5 Empirical investigation of equation 6 for Major Depression (all cohorts). In this figure, the x-axis is the product of h_t^2 (SNP-based heritability of cohorts being predicted into on the liability scale) and $r^2_{g(ma,t)}$(squared genetic correlations between the LOO meta-analysis cohort used to generate the PGS, and the left-out target cohort where PGS are calculated); the y-axis is the out-of-sample prediction R^2 on the liability scale calculated in the target cohort. Each dot represents the target cohort (whose estimated h_t^2 is between 0 and 1, sizes of dots are in proportion to the effective sample size). The horizontal line denotes h^2_{ma} (black dotted line) and its 95% C.I. (blue dotted line). (Although we left a different cohort out each time, the h^2_{ma} and 95% CI remain unchanged because the sample size of the left-out cohort is small when compared with the total sample size of meta-analysis.) Our derivations show that when the h^2_{SNP} in the cohort being predicted into is higher than the predicted threshold (vertical line), out-of-sample prediction R^2 will exceed the h^2_{ma} in the meta-analysis used to generate the predictor. Compared to the Figure 4, we removed restrictions on the per-cohort SNP-based heritability, and as a result included 7 more cohorts in the analysis. In general, the trend is still exactly what we expected. Note the error associated with x and y axis values are likely correlated because both depend on the sample size of the target cohort.



Supplementary Figure II.6 Empirical investigation of equation 6 for Educational Attainment. The x-axis is the product of h^2_t (estimated SNP-based heritability of the target cohorts) and $r^2_{g(ma,t)}$(squared estimated genetic correlations between the leave-one-cohort-out meta-analysis cohort used to generate the PGS, and the left-out target cohort where PGS are calculated); the y-axis is the out-of-sample prediction R^2 calculated in the left-out target cohort. Each dot represents a target cohort and error bars represent the 95% C.I. of the estimated out-of-sample prediction R^2 in each target cohort. The horizontal (red dotted) line denotes h^2_{ma} and its associated 95% C.I. (blue dotted line). The vertical (black dotted) line denotes the predicted threshold for $h^2_t r^2_{g(ma,t)}$ when out-of-sample prediction R^2 is expected to exceed h^2_{ma}. The empirical results illustrated for 2 cohorts are consistent with this derivation.

A2. Supplementary files for Chapter III

Model-fitting				
Scenarios	**Methods**	**Input GWAS for EUR**	**Input GWAS for AFR**	**LD reference**
Multi-ancestry Analysis	Manc-COJO	N ~ 1 million EUR GWAS from GLGC (Lipid traits) or GIANT (Height)	N ~ 0.1 million AFR GWAS from GLGC (Lipid traits) or GIANT (Height)	N ~ 69,000 UKB EUR participants and N ~ 6,900 UKB AFR participants (separately)
	Sanc-COJO->Manc : Sanc-COJO on multi-ancestry data with multi-ancestry LD reference			N ~ 69,000 UKB EUR paricipants and N ~ 6,900 UKB AFR participants (Merged into a single cohort)
	Sanc-COJO->Manc/EURref: Sanc-COJO on multi-ancestry data with European-only LD reference			N ~ 75,900 UKB EUR paricipants (As a single cohort)
Single-ancestry Analysis	Sanc-COJO on single ancestry data with European-only LD reference	N ~ 1 million EUR GWAS from GLGC (Lipid traits) or GIANT (Height) plus N ~ 0.1 million EUR GWAS from UKB	N/A	N ~ 75,900 UKB EUR paricipants (As a single cohort)
Generating weight for polygenic prediction				
EUR weight	Output from all COJO model	N ~ 1 million EUR GWAS from GLGC (Lipid traits) or GIANT (Height) plus N ~ 0.1 million EUR GWAS from UKB	N/A	N ~ 75,900 UKB EUR paricipants
AFR weight	Output from all COJO model	N/A	N ~ 0.1 million AFR GWAS from GLGC (Lipid traits) or GIANT (Height)	N ~ 6,900 UKB AFR participants

Supplementary Table III.1. Summary of full sample analyses. This table summarises the GWAS inputs and LD reference panels used for each method in full sample analyses (Total sample size ~ 1.1 million).

Model-fitting				
Scenarios	**Methods**	**Input GWAS for EUR**	**Input GWAS for AFR**	**LD reference**
Multi-ancestry Analysis	Manc-COJO	N ~ 100,000 EUR GWAS from UKB	N ~ 100,000 AFR GWAS from GLGC (Lipid traits) or GIANT (Height)	N ~ 6,900 UKB EUR paricipants and N ~ 6,900 UKB AFR participants (separately)
	Sanc-COJO->Manc : Sanc-COJO on multi-ancestry data with multi-ancestry LD reference			N ~ 6,900 UKB EUR paricipants and N ~ 6,900 UKB AFR participants (Merged into a single cohort)
	COJO->Manc/EURref: Sanc-COJO on multi-ancestry data with European-only LD reference			N ~ 13,800 UKB EUR paricipants (As a single cohort)
Single-ancestry Analysis	Sanc-COJO on single ancestry data with European-only LD reference	N ~ 200,000 EUR GWAS from UKB	N/A	N ~ 13,800 UKB EUR paricipants (As a single cohort)
Generating weight for polygenic prediction				
EUR weight	Output from all COJO model	N ~ 200,000 EUR GWAS from UKB	N/A	N ~ 13,800 UKB EUR paricipants (As a single cohort)
AFR weight	Output from all COJO model	N/A	N ~ 100,000 AFR GWAS from GLGC (Lipid traits) or GIANT (Height)	N ~ 6,900 UKB AFR participants

Supplementary Table III.2. Summary of down-sample analyses. This table summarises the GWAS inputs and LD reference panels used for each method in down-sample analyses (Total sample size ~ 0.2 million).

Size of models (i.e., number of COJO SNPs):

Trait	Manc-COJO	Single ancestry Sanc-COJO	Sanc-COJO->Manc: Multi-ancestry Sanc-COJO with ancestral proportion matched LD reference	Sanc-COJO->Manc/EURref: Multi-ancestry Sanc-COJO with European-only LD reference
HDL	714	807	897	725
LDL	475	501	575	493
TC	594	629	699	628
logTG	545	606	709	557
Height	3206	3195	2430	3284

Prediction R^2 (and s.e.) in EUR:

Trait	Manc-COJO	Single ancestry Sanc-COJO	Multi-ancestry Sanc-COJO with ancestral proportion matched LD reference	Multi-ancestry Sanc-COJO with European-only LD reference
HDL	0.160 (0.023)	0.162 (0.022)	0.155 (0.022)	0.158 (0.023)
LDL	0.144 (0.022)	0.147 (0.021)	0.138 (0.022)	0.145 (0.022)
TC	0.135 (0.020)	0.135 (0.019)	0.126 (0.019)	0.132 (0.021)
logTG	0.102 (0.018)	0.106 (0.017)	0.101 (0.016)	0.101 (0.017)
Height	0.296 (0.018)	0.299 (0.018)	0.250 (0.020)	0.296 (0.020)

Prediction R^2 (and s.e.) in AFR:

Trait	Manc-COJO	Single ancestry Sanc-COJO	Multi-ancestry Sanc-COJO with ancestral proportion matched LD reference	Multi-ancestry Sanc-COJO with European-only LD reference
HDL	0.0846 (0.045)	0.0745 (0.041)	0.0835 (0.045)	0.0873 (0.043)
LDL	0.0944 (0.041)	0.0806 (0.033)	0.0964 (0.047)	0.102 (0.035)
TC	0.0910 (0.042)	0.0674 (0.033)	0.0855 (0.038)	0.0874 (0.040)
logTG	0.0280 (0.018)	0.0240 (0.013)	0.0287 (0.016)	0.0270 (0.017)
Height	0.0960 (0.017)	0.0786 (0.021)	0.0737 (0.013)	0.0988 (0.016)

Supplementary Table III.3. Manc-COJO and Sanc-COJO performances with HapMap3 SNPs. This table summarises the model size and prediction accuracies in European cohorts and African cohorts for various Manc-COJO and Sanc-COJO models using only HapMap3 SNPs.

Size of models (i.e., number of COJO-SNPs):

Trait	Manc-COJO (R^2_{Adj} threshold - Forward selection: -1; Backward selection: -1)	Manc-COJO (R^2_{Adj} threshold - Forward selection: 0; Backward selection: -1)	Manc-COJO (R^2_{Adj} threshold - Forward selection: 0; Backward selection: 0)
HDL	847	774	774
LDL	567	516	512
TC	720	678	679
logTG	684	618	620
Height	3206	3195	2430

Prediction R^2 (and s.e.) in EUR:

Trait	Manc-COJO (R^2_{Adj} threshold - Forward selection: -1; Backward selection: -1)	Manc-COJO (R^2_{Adj} threshold - Forward selection: 0; Backward selection: -1)	Manc-COJO (R^2_{Adj} threshold - Forward selection: 0; Backward selection: 0)
HDL	0.168 (0.022)	0.164 (0.023)	0.164 (0.023)
LDL	0.159 (0.021)	0.157 (0.021)	0.157 (0.021)
TC	0.145 (0.019)	0.143 (0.020)	0.143 (0.020)
logTG	0.110 (0.018)	0.105 (0.018)	0.105 (0.018)
Height	0.296 (0.018)	0.287 (0.020)	0.286 (0.020)

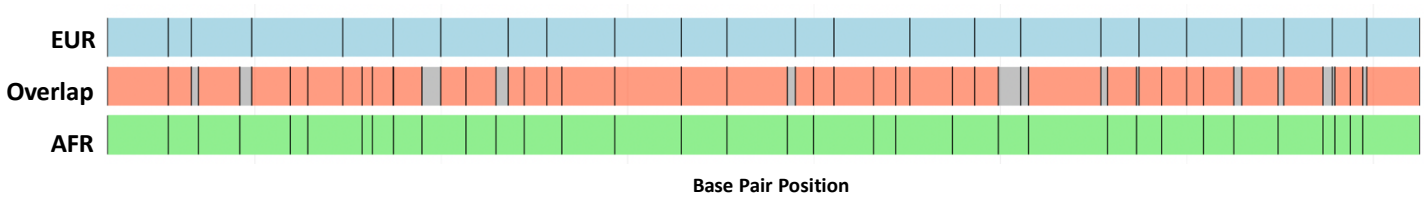
Prediction R^2 (and s.e.) in AFR:

Trait	Manc-COJO (R^2_{Adj} threshold - Forward selection: -1; Backward selection: -1)	Manc-COJO (R^2_{Adj} threshold - Forward selection: 0; Backward selection: -1)	Manc-COJO (R^2_{Adj} threshold - Forward selection: 0; Backward selection: 0)
HDL	0.0835 (0.044)	0.0837 (0.047)	0.0837 (0.047)
LDL	0.144 (0.053)	0.141 (0.051)	0.141 (0.050)
TC	0.111 (0.046)	0.117 (0.042)	0.117 (0.043)
logTG	0.0559 (0.031)	0.0550 (0.033)	0.0532 (0.031)
Height	0.0960 (0.017)	0.0935 (0.018)	0.0941 (0.018)

Supplementary Table III.4. Comparison of Manc-COJO performance using different adjusted R^2 thresholds in forward and backward selections. This table summarises the model size and prediction accuracies in European cohorts and African cohorts by applying various adjusted R^2 threshold settings in Manc-COJO models. Note Due to data availability, the analysis of height in this table includes only HM3 SNPs.

Trait	LD Blocks (Chr: Start BP - End BP)	Manc-COJO Model Sizes	Sanc-COJO Model Sizes	Prediction Accuracies Into EUR cohorts (Manc-COJO/ Sanc-COJO)	Prediction Accuracies Into AFR cohorts (Manc-COJO/ Sanc-COJO)
HDL	6: 31,428,014-32,682,664	12	4	-0.00041 (0.0032)	-0.0012 (0.0036)
				-0.00025 (0.0032)	-0.00035 (0.0020)
	16: 56,347,201-58,536,624	31	27	0.038 (0.031)	0.032 (0.017)
				0.039 (0.031)	0.030 (0.019)
LDL	1: 53,116,453-56,413,125	23	19	0.0042 (0.011)	0.0098 (0.0070)
				0.0034 (0.011)	0.0055 (0.0098)
	6: 31,428,014-32,682,664	15	2	0.0012 (0.0038)	0.00051 (0.0029)
				0.0011 (0.0038)	0.00052 (0.0016)
	19: 45,202,640-46,771,398	26	20	0.062 (0.039)	0.087 (0.037)
				0.061 (0.039)	0.082 (0.033)
TC	1: 53,116,453-56,413,125	18	13	0.0036 (0.0085)	0.087 (0.037)
				0.0028 (0.039)	0.082 (0.033)
	19: 45,202,640-46,771,398	25	20	0.040 (0.020)	0.051 (0.022)
				0.041 (0.020)	0.053 (0.022)
logTG	11: 116,106,324-117,541,718	34	25	0.023 (0.022)	0.018 (0.014)
				0.022 (0.022)	0.0043 (0.0072)
	13: 113,609,290-115,109,853	6	2	0.00025(0.0038)	0.00094 (0.0020)
				-0.000014(0.0038)	0.00047 (0.00082)
	15: 57,743,507-60,474,503	7	3	0.0016(0.0053)	0.0015 (0.0028)
				0.0011(0.0054)	0.0015 (0.0019)

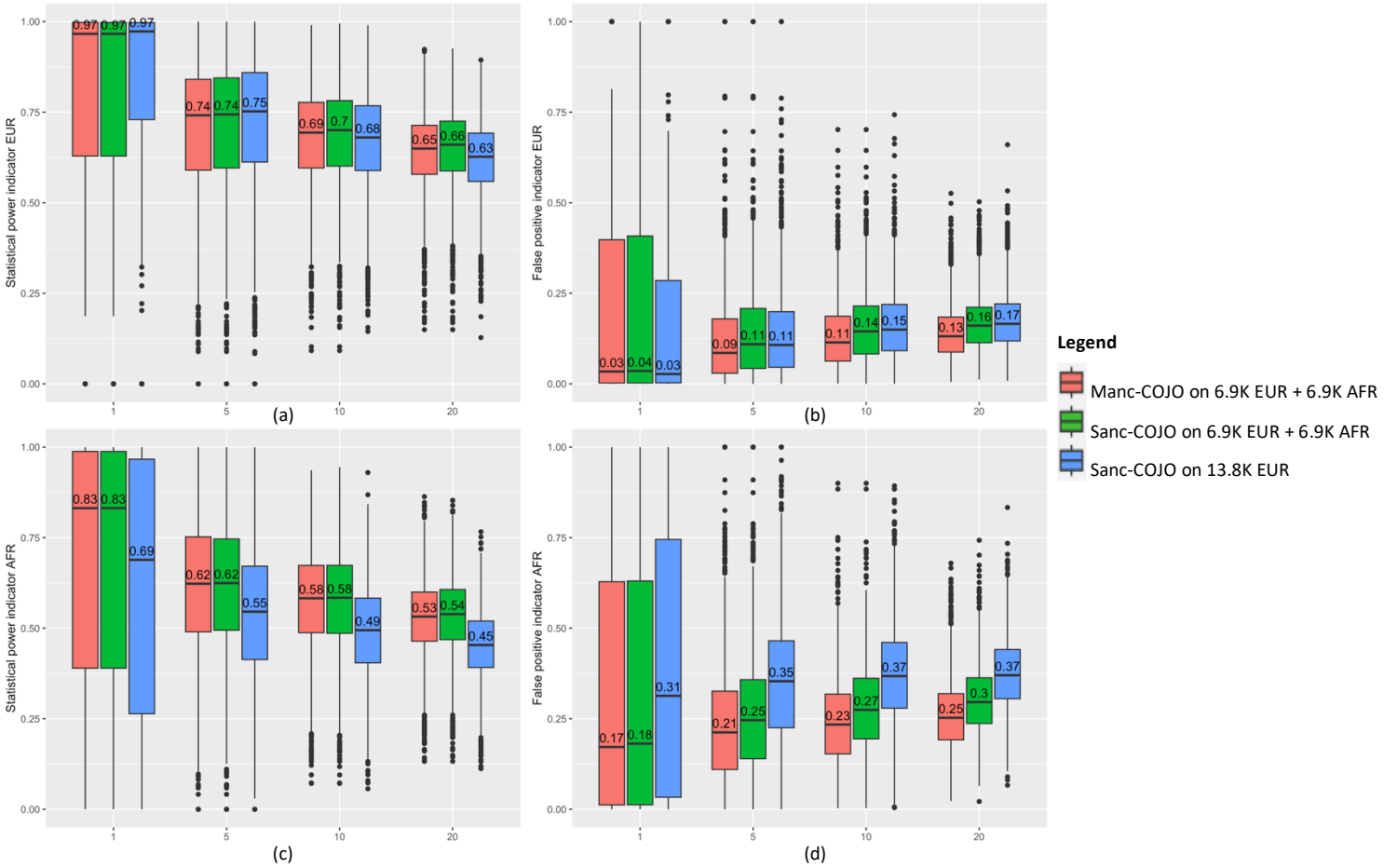
Supplementary Table III.5. Summary of LD blocks where the Manc-COJO model identified more than three additional independent associations compared with the single ancestry Sanc-COJO model. This table summarises all EUR LD blocks in which the Manc-COJO model identified more than three additional independent associations relative to the single ancestry Sanc-COJO model. Such instances were rare, occurring in fewer than 1% of all blocks analysed across the lipid traits. Out-of-sample prediction accuracies into both EUR and AFR cohorts for predictors derived from each model are also reported. The target cohorts used for prediction are the same as those employed in the full-sample analysis.



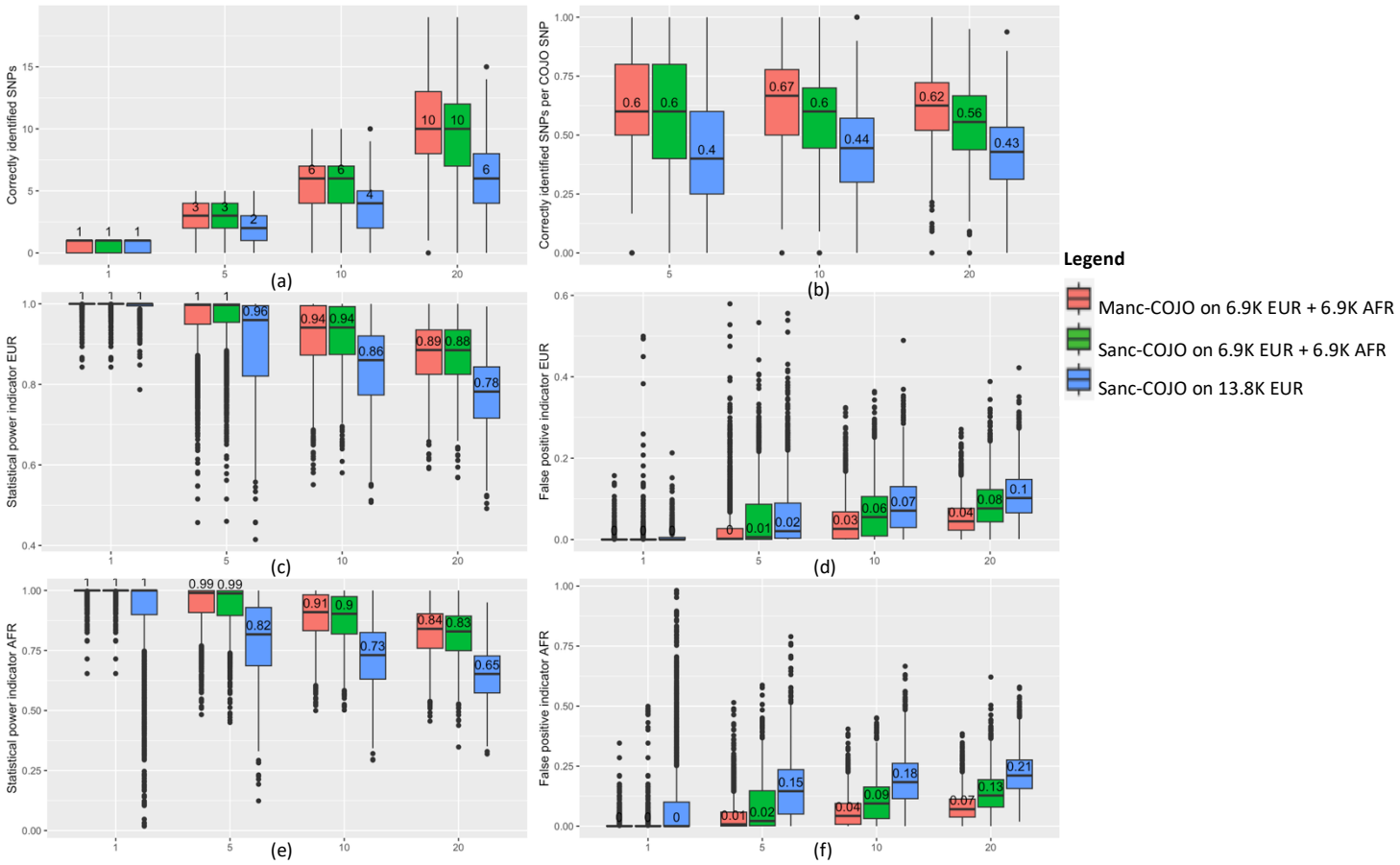
Supplementary Figure III.1. Partitioning of chromosome 22 for simulation. In our study, Chromosome 22 was partitioned separately for the African and European cohorts, resulting in 33 and 23 blocks, respectively. These blocks were partially overlapping, and a total of 52 overlapping sections were extracted (middle row of the figure). Within each of these 52 overlapping regions, blocks containing fewer than 1,000 SNPs after quality control were excluded (shown in grey in the middle row), leaving 38 blocks (shown in red) for use in the simulation.



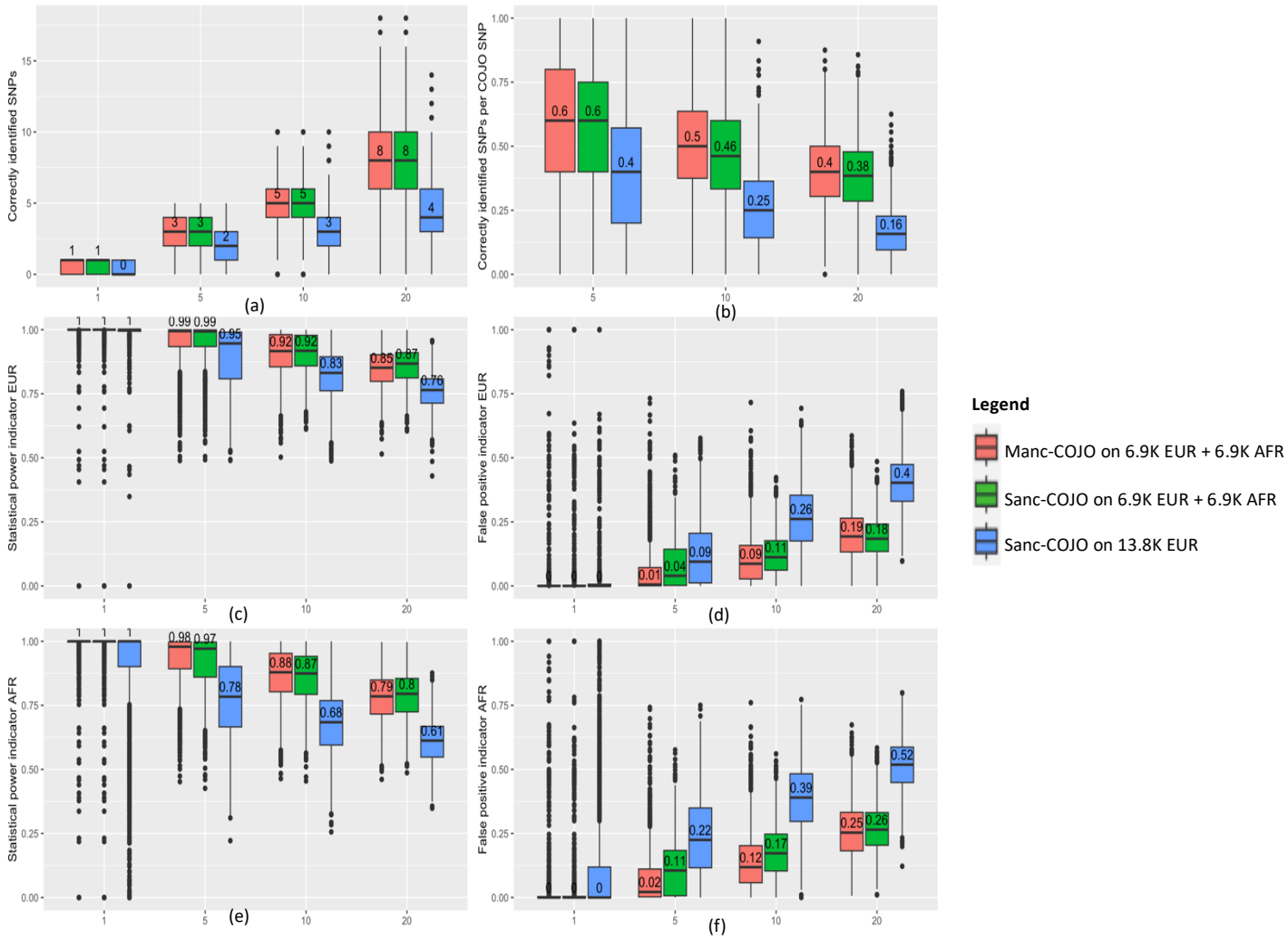
Supplementary Figure III.2. Simulation results within individual blocks are consistent with the pooled results. This figure summarises simulation results for each of the 38 simulated blocks. The plot in the top-left corresponds to Block 1, with Block 2 immediately to its right, and so on. The x-axis within each plot represents the number of simulated causal SNPs in each block—5, 10, and 20 for the first, second, and third columns, respectively. These results demonstrate that the simulation trends observed within individual blocks are consistent with the pooled analysis.



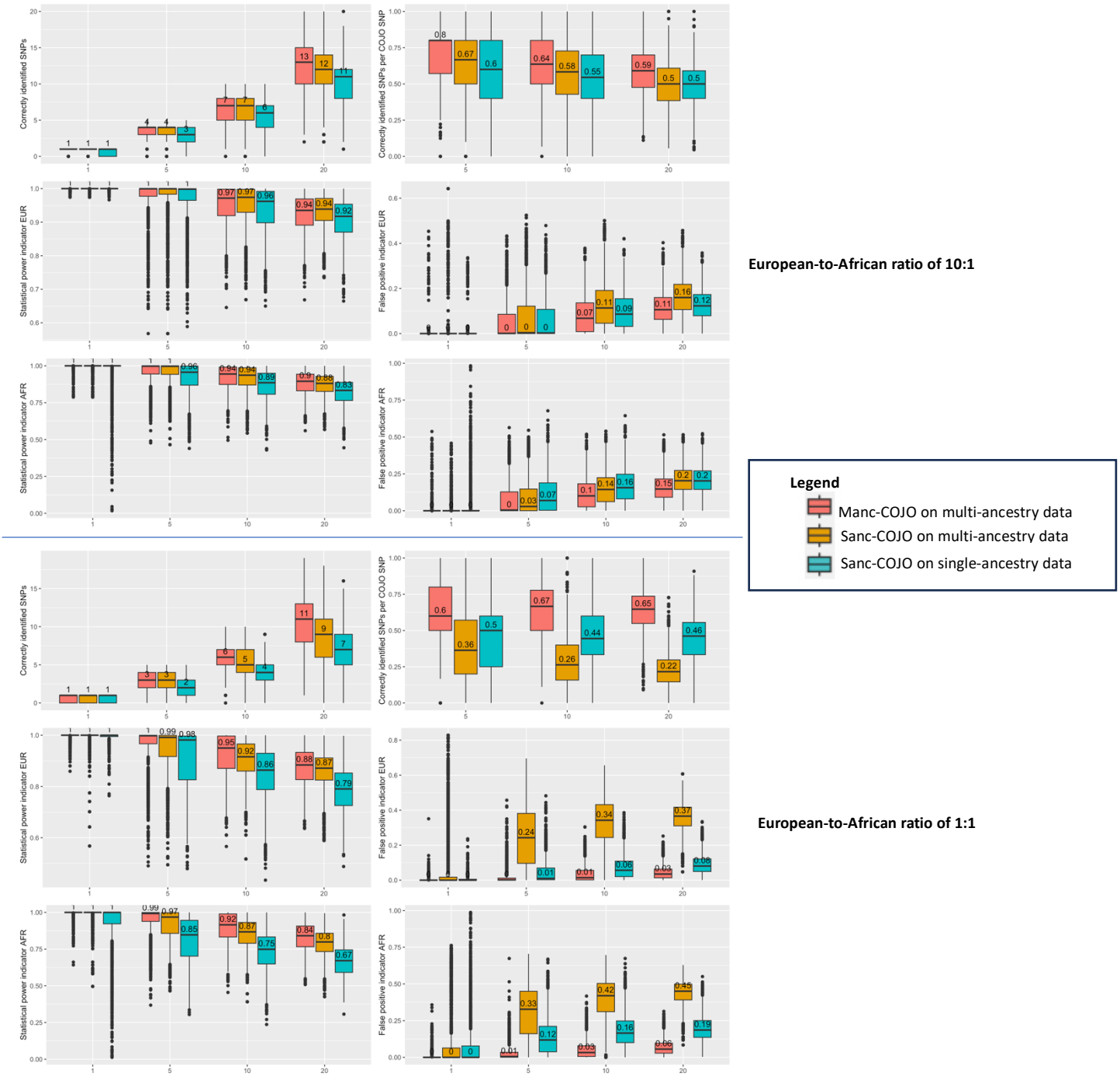
Supplementary Figure III.3. Manc-COJO demonstrates superior performance in simulations when causal SNPs are not genotyped. This figure summarise simulation results of 7,600 scenarios, and illustrates the performance of Manc-COJO in identifying causal SNPs, showing, on average, superior power and fewer false-positive discoveries across ancestries. We applied our algorithm to both a simulated European (EUR) and a simulated African (AFR) cohort, each with a sample size of 6,900. Our comparisons include Sanc-COJO applied to either 13,800 Europeans or a meta-analysis of 6,900 Europeans with 6,900 Africans. In all scenarios, in-sample LD was utilized. Only Hapmap3 SNPs (~1M) are included in the analysis. The X-axis represents the number of simulated causal SNPs. (a) We identified the COJO-SNP with the highest LD correlation (in European cohort) to each causal SNP and averaged these correlations across all causal SNPs. Higher value indicates better power in identifying causal variants. (b) For each COJO-SNP, we assessed the LD (in European cohort) with the closest true causal SNPs and averaged across COJO-SNP instances, providing an indicator of false-positive discoveries (i.e., instances where SNPs not in LD with causal SNPs were incorrectly identified will result in lower averaged LD-correlation). Higher value indicates higher false positive discoveries. (c)(d) Power and false positive indicators as in (a) and (b), except that African LD is used in calculation. Numbers in each plots are the median value of each scenario.



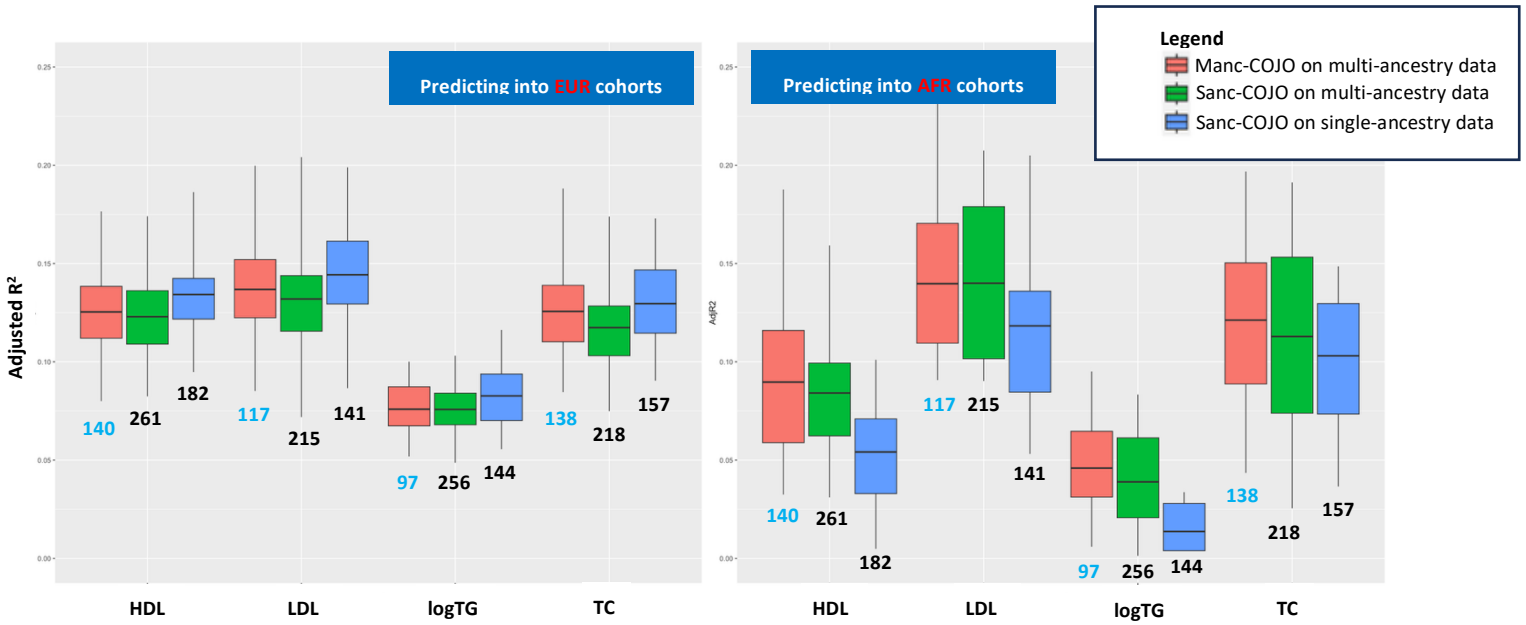
Supplementary Figure III.4. Manc-COJO demonstrates superior performance in simulations when effect sizes of causal SNPs are not identical across ancestries. This figure summarise simulation results of 7,600 scenarios, and illustrates the performance of Manc-COJO in identifying causal SNPs, showing, on average, superior power and fewer false-positive discoveries across ancestries. We applied our algorithm to both a simulated European (EUR) and a simulated African (AFR) cohort, each with a sample size of 6,900. Our comparisons include Sanc-COJO applied to either 13,800 Europeans or a meta-analysis of 6,900 Europeans with 6,900 Africans. In all cases, in-sample LD was utilized. The effect sizes of the causal SNPs varied across the simulated conditions, but the direction of association remained consistent. The X-axis represents the number of simulated causal SNPs. (a) The Y-axis indicates the number of causal SNPs correctly identified by each algorithm. (b) The Y-axis represents the proportion of identified SNPs that are causal. (c) As COJO algorithms may identify SNPs (termed COJO-SNPs) near the causal SNP rather than the causal SNP itself, we identified the COJO-SNP with the highest LD correlation (in the European cohort) to each causal SNP and averaged these correlations across all causal SNPs; higher values reflect greater power in detecting causal variants. (d) For each COJO-SNP, we calculated its LD (in the European cohort) with the nearest true causal SNP and averaged these values across all COJO-SNPs, providing an indicator of false-positive discoveries (i.e., lower averaged LD correlations indicate instances where SNPs not in LD with causal SNPs were incorrectly identified); higher values suggest more false-positive discoveries. (e)(f) Power and false-positive indicators as in (c) and (d), respectively, but calculated using African LD. Numbers in each plot represent the median value for each group. This figure illustrates that Manc-COJO, on average, identifies a greater number of causal SNPs, exhibits superior power, and produces fewer false-positive discoveries when sample sizes across ancestries are unequal.



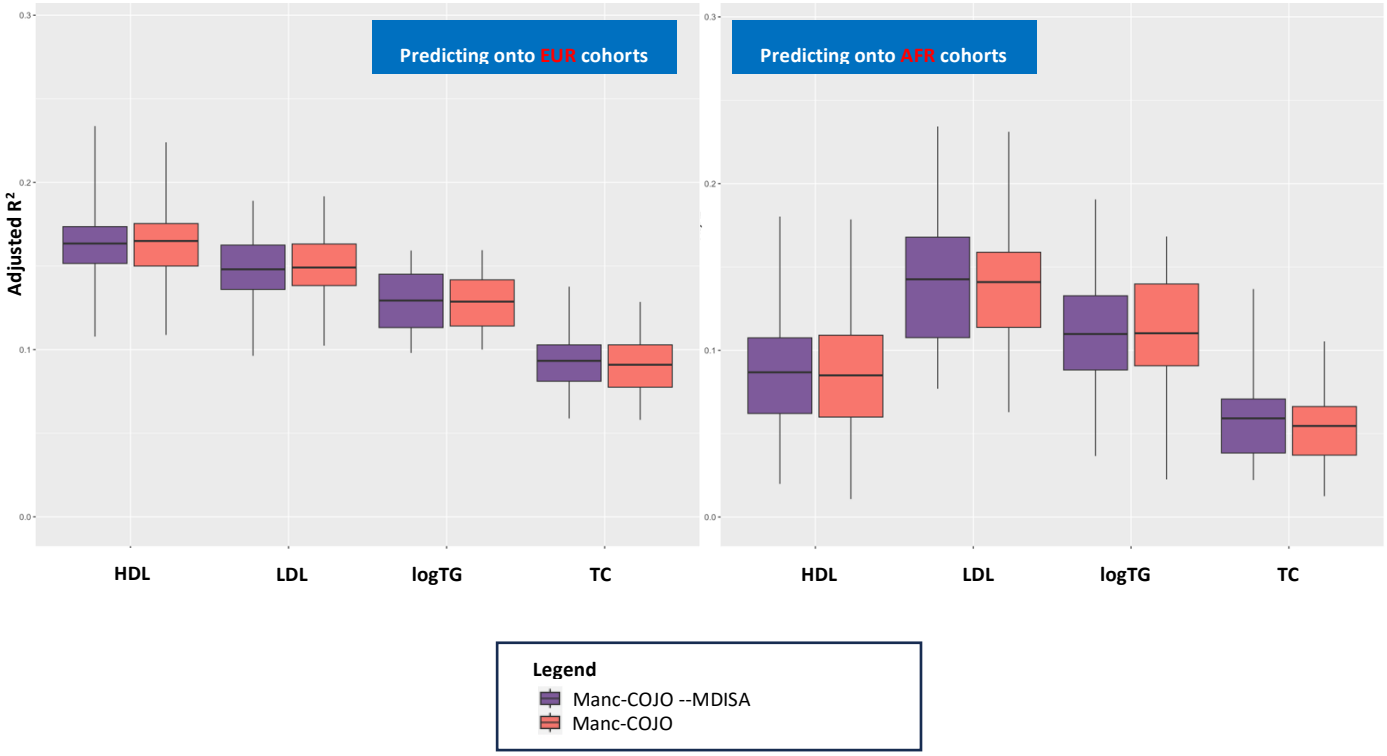
Supplementary Figure III.5. Manc-COJO is robust to imperfect LD reference in simulations. This figure summarise simulation results of 7,600 scenarios, and illustrates Manc-COJO is robust to imperfect LD reference. We applied our algorithm to both a simulated European (EUR) and a simulated African (AFR) cohort, each with a sample size of 6,900. Our comparisons include Sanc-COJO applied to either 13,800 Europeans or a meta-analysis of 6,900 Europeans with 6,900 Africans. An LD reference from the 1000 Genomes Project was applied in both European (N = 503) and African ancestry (N = 503). The X-axis represents the number of simulated causal SNPs. (a) The Y-axis denotes the number of causal SNPs correctly identified by each algorithm. (b) The Y-axis denotes the proportion of causal SNPs that are causal SNPs. (c) Instead of identifying the causal SNP, COJO algorithms may identify SNPs (termed COJO-SNP) in nearby locations. We identified the COJO-SNP with the highest LD correlation (in European cohort) to each causal SNP and averaged these correlations across all causal SNPs. Higher value indicates better power in identifying causal variants. (d) For each COJO-SNP, we assessed the LD (in European cohort) with the closest true causal SNPs and averaged across COJO-SNP instances, providing an indicator of false-positive discoveries (i.e., instances where SNPs not in LD with causal SNPs were incorrectly identified will result in lower averaged LD-correlation). Higher value indicates higher false positive discoveries. (e)(f) Power and false positive indicators as in (c) and (d), except that African LD is used in calculation. Numbers in each plot are the median value of each group.



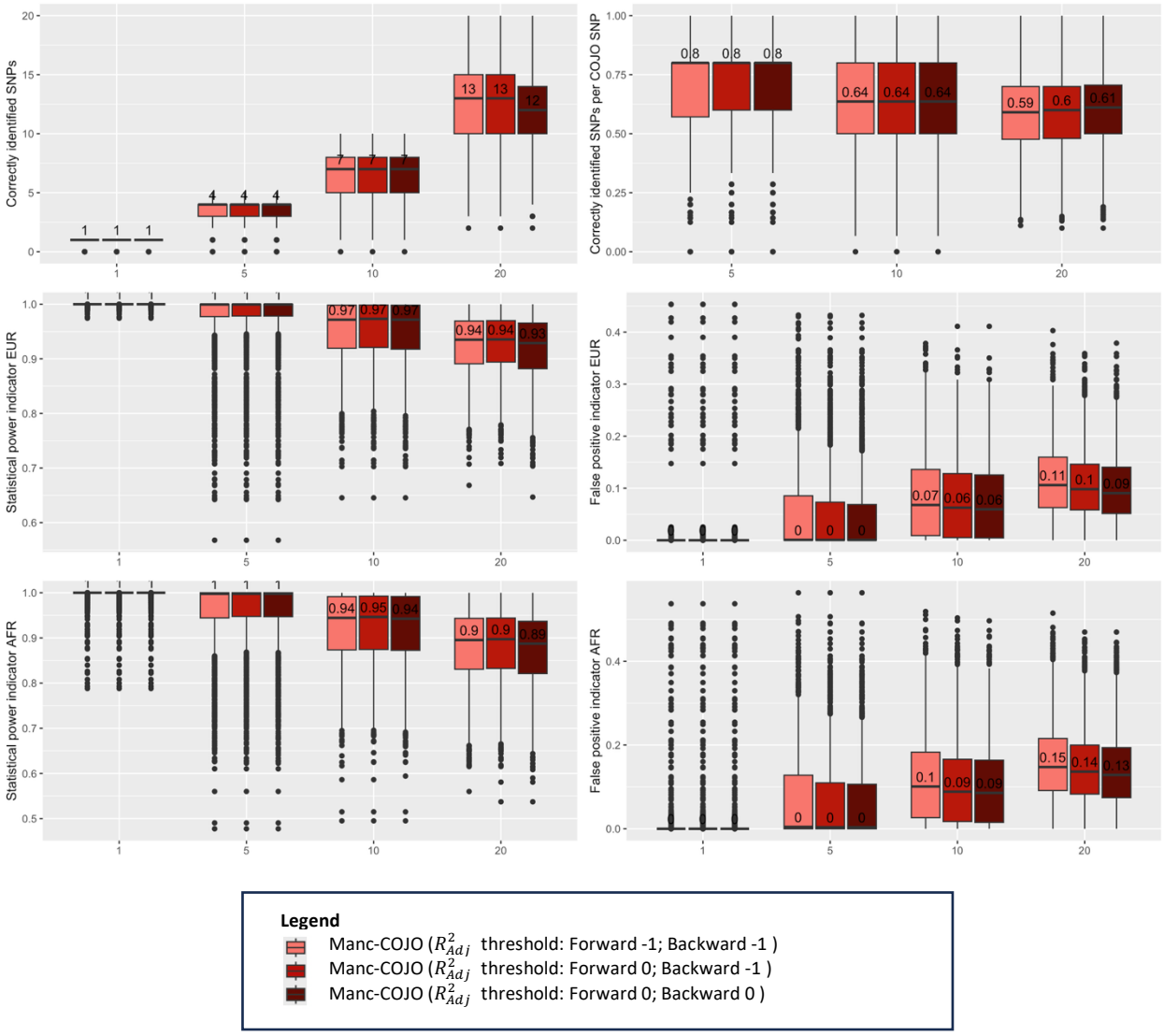
Supplementary Figure III.6. Effectiveness of European-only LD reference panels depends on ancestry composition of the study sample in simulations. This figure shows the performance of Sanc-COJO applied to multi-ancestry GWAS summary statistics (Sanc-COJO->Manc) using a European-only LD reference panel, under two different sample size scenarios: a European:African ratio of 10:1 (top panel) and 1:1 (bottom panel). Model performance is evaluated as in **Figure III.2**. When Europeans dominate the sample (10:1 ratio), using a European-only LD reference performs reasonably well, achieving higher power in both ancestries and lower false positive rates in African samples, compared to Sanc-COJO applied to European-only summary statistics. However, when the ancestry balance is equal (1:1 ratio), the use of a European-only LD reference becomes ineffective, leading to marked inflation in false-positive discoveries, particularly in African samples.



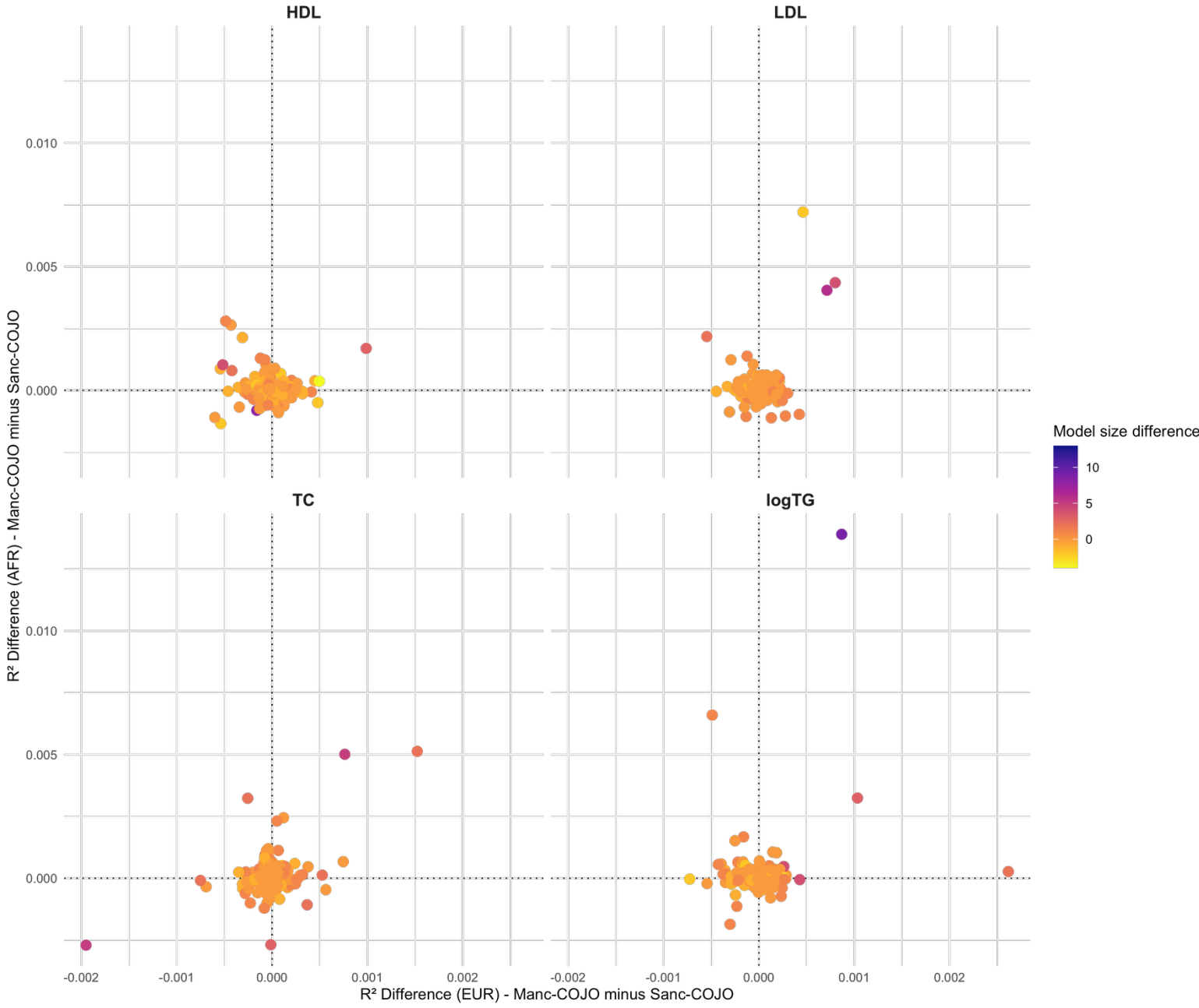
Supplementary Figure III.7. Manc-COJO demonstrates superior trans-ancestry portability compared to the single-ancestry Sanc-COJO model in down-sample analysis applied in real data. We conducted a down-sampling analysis to examine whether closer ancestry balance (approximately 1:1) would enhance detection of independent associations. This figure summarises the out-of-sample prediction accuracies of Manc-COJO on multi-ancestry data, Sanc-COJO on multi-ancestry data, and Sanc-COJO on single-ancestral data for HDL, LDL, logTG, and TC when predicting into independent European and African cohorts. Numbers in the figure are sizes of the model, and the model with the smallest number of SNPs is highlighted in blue for each trait. (i) Manc-COJO input: African-specific GWAS summary statistics from GLGC (N ~ 90,000), combined with new European-specific GWAS conducted in UK Biobank participants (N = 90,000) using pipelines consistent with the original GIANT and GLGC studies. The LD reference panels consist of 6,901 unrelated European individuals and 6,901 unrelated African individuals from the UK Biobank. (ii) Sanc-COJO on multi-ancestry data input: Meta-analysis of the European-specific and African-specific GWAS summary statistics described above. The LD reference panel iss a merged set of the unrelated European and African individuals from the UK Biobank (6,901 Europeans and 6,901 Africans). (iii) Sanc-COJO on single-ancestry data input: new European-specific GWAS conducted in UK Biobank participants (N = 200,000 for height and 180,000 for lipid traits) using pipelines consistent with the original GIANT and GLGC studies. The LD reference panel was a set 13,802 unrelated European and African individuals from the UK Biobank. To ensure fair comparisons across methods, we harmonized the GWAS summary statistics to have identical SNP sets and approximately equal sample sizes, and we kept identical LD reference panel sizes across methods. Overall, Manc-COJO models showed slightly lower out-of-sample prediction accuracy into European cohorts compared to single-ancestry Sanc-COJO models. However, Manc-COJO models were much smaller and consistently demonstrated much better trans-ancestry portability. Compared to the full-size analysis, the relative size difference between Manc-COJO and Sanc-COJO models decreased, suggesting that closer ancestry balance can enhance the detection of independent associations. Notably, Sanc-COJO models trained on multi-ancestry data continued to show sub-optimal out-of-sample performance across all scenarios.



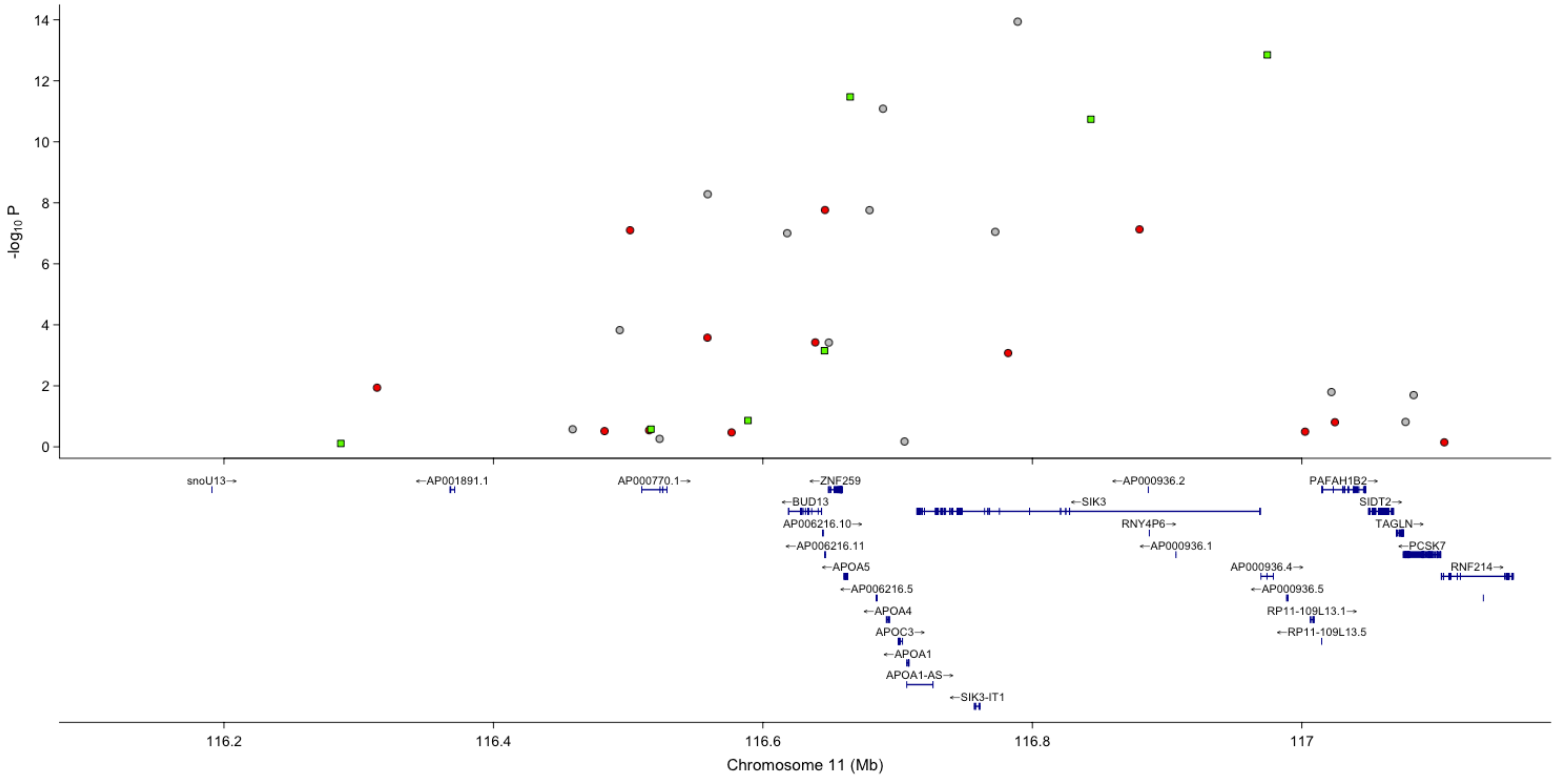
Supplementary Figure III.8. Manc-COJO models based on shared causal SNP assumptions explain most of the out-of-sample prediction R^2 in both European and African ancestries. This figure summarises the results from the Manc-COJO --MDISA algorithm, which allows for ancestry-specific causal variants. These results are compared to the standard Manc-COJO model, which assumes shared causal SNPs across ancestries. We evaluated the performance of Manc-COJO --MDISA using real data for HDL, LDL, logTG, and TC, applying the same settings as shown **in Figure III.6**. The minimal differences in out-of-sample prediction accuracy for both European (EUR) and African (AFR) ancestries suggest that the model based on shared causal SNP assumption can explain most of the out-of-sample prediction R^2 in both ancestries.



Supplementary Figure III.9. Impact of different adjusted R² threshold on Manc-COJO output in idealised situations. The simulation settings and layout of this figure are identical to those in Extended Figure 2. In this figure, we compare, under a well-powered scenario with a EUR:AFR ratio of 10:1, the impact of adding an additional optional variable selection algorithm based on the model's adjusted R². In general, applying a more stringent threshold reduces the false positive inclusion of non-causal SNPs in the model at the expense of reduced power to include causal SNPs.

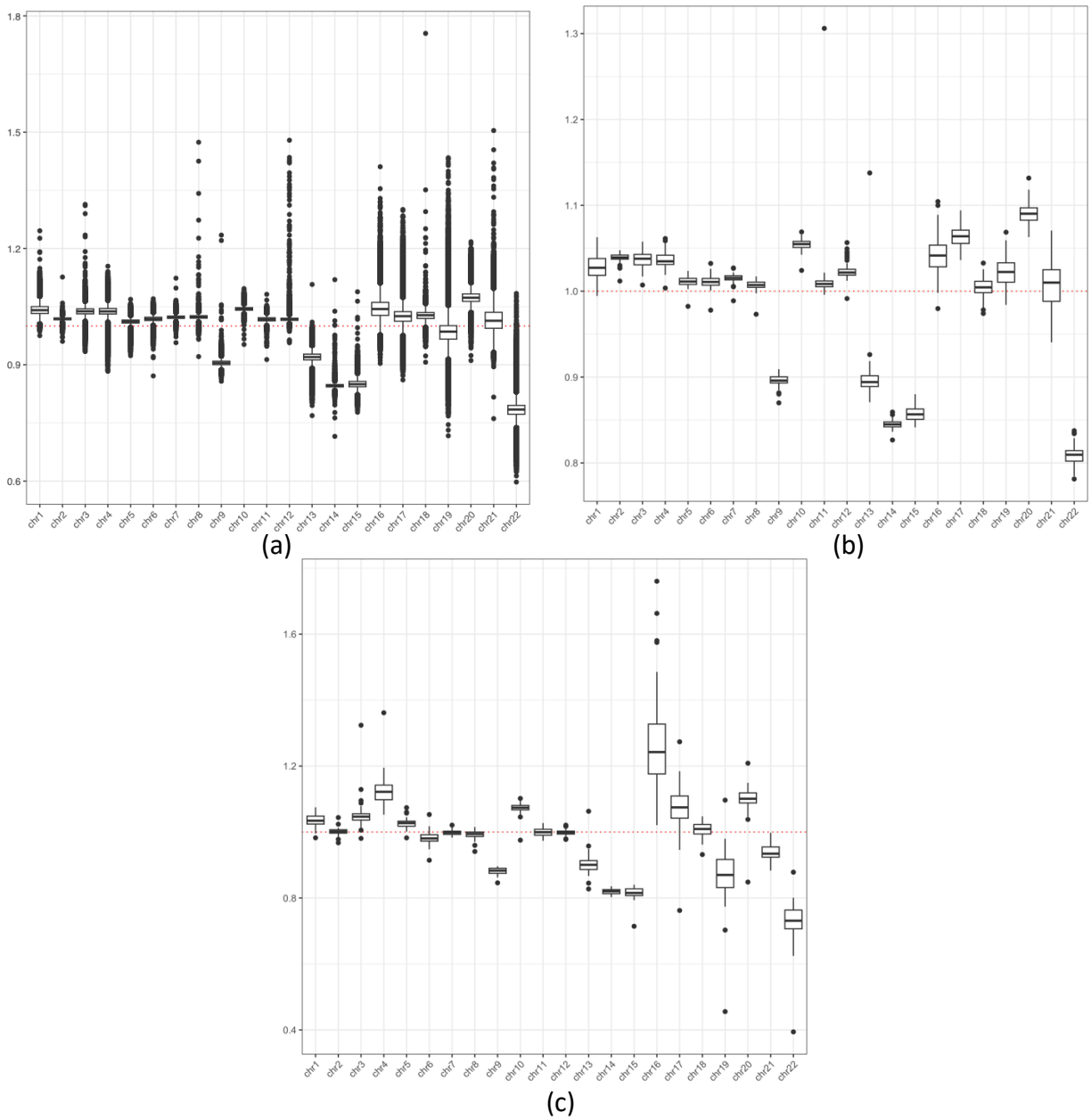


Supplementary Figure III.10. Comparison of Manc-COJO and single-ancestry Sanc-COJO model performance across LD-independent blocks. This figure compares the out-of-sample prediction accuracies of the Manc-COJO model (trained on GWAS summary statistics from 1 million Europeans and 0.1 million Africans) with the single-ancestry Sanc-COJO model (trained on 1.1 million Europeans) across LD-independent blocks for four lipid traits. Each point represents a single LD block, with the x-axis showing the difference in prediction accuracy in Europeans (Manc-COJO minus Sanc-COJO) and the y-axis showing the difference in Africans. Positive values indicate better performance of Manc-COJO; negative values indicate better performance of Sanc-COJO. The color of each point reflects the difference in model size for that block—purplish colors indicate more SNPs were selected by Manc-COJO than Sanc-COJO. Note that the x- and y-axes use different scales, reflecting the fact that prediction accuracies between the two models are generally more similar in European populations, while greater variability is observed in African populations. The greater number of points above the x-axis suggests that Manc-COJO more consistently improves prediction accuracy in African cohorts.

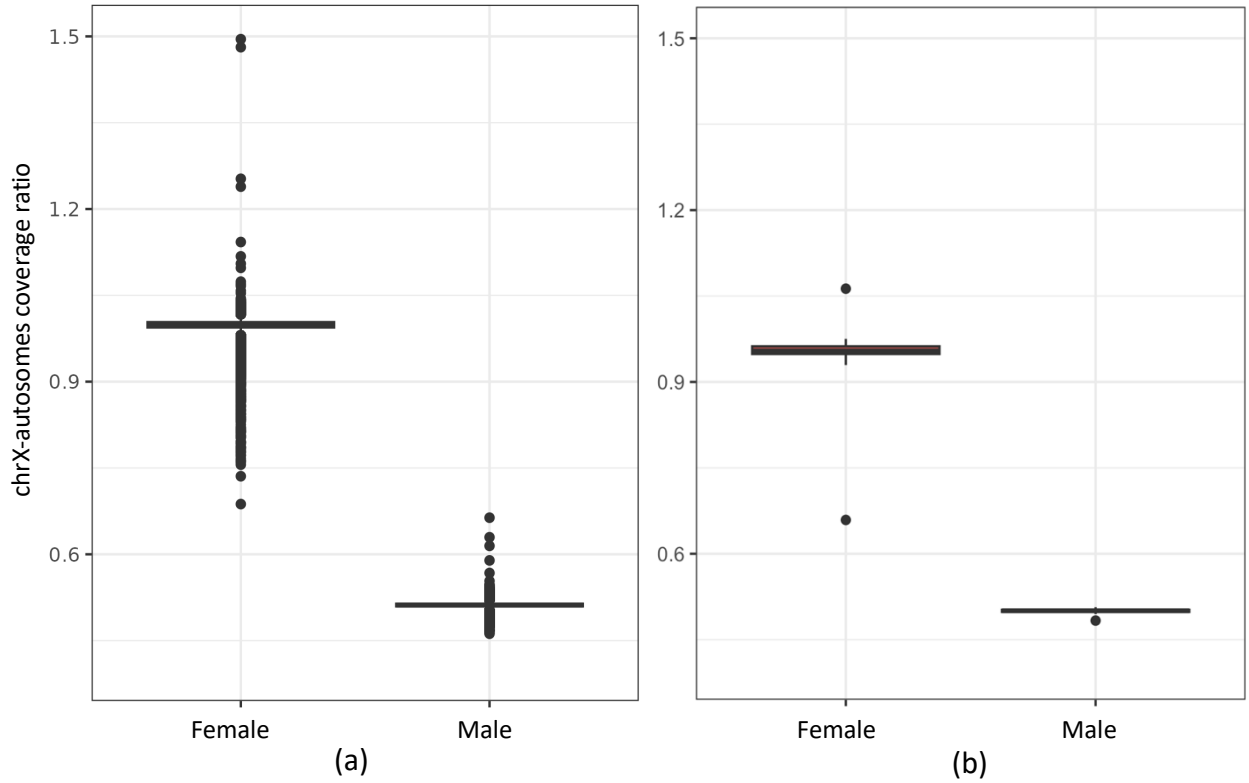


Supplementary Figure III.11. Comparison of Manc-COJO and Sanc-COJO outputs for logTG in a region on Chromosome 11. This figure illustrates the outputs of the Manc-COJO and single ancestry Sanc-COJO models for log-transformed triglycerides (logTG) within a region on chromosome 11 (base pairs 116,106,324–117,541,718; GRCh37 build). In this block, the Manc-COJO model identified 34 independent SNP associations, compared to 25 identified by the Sanc-COJO model. Notably, SNPs identified by the Manc-COJO model in this region explained 1.3% more phenotypic variance in the African cohort than those from the Sanc-COJO model. The x-axis denotes base-pair position (BP), and the y-axis shows $-\log_{10}(P)$ values from the logTG GWAS in the AFR cohort. Grey dots represent SNPs shared by both models. Red dots indicate SNPs uniquely identified by Manc-COJO, while green dots represent SNPs uniquely identified by Sanc-COJO.

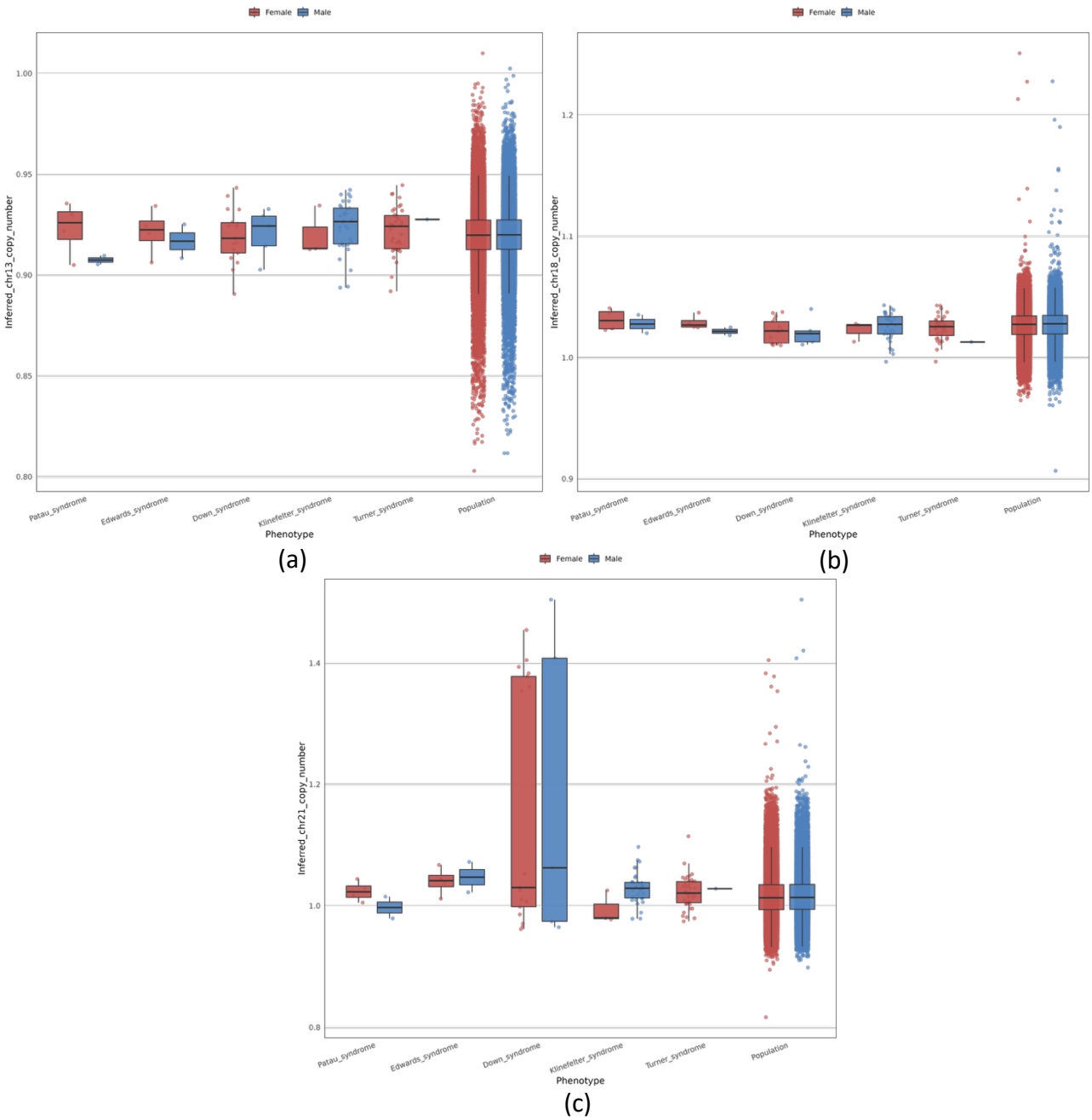
A3. Supplementary files for Chapter IV



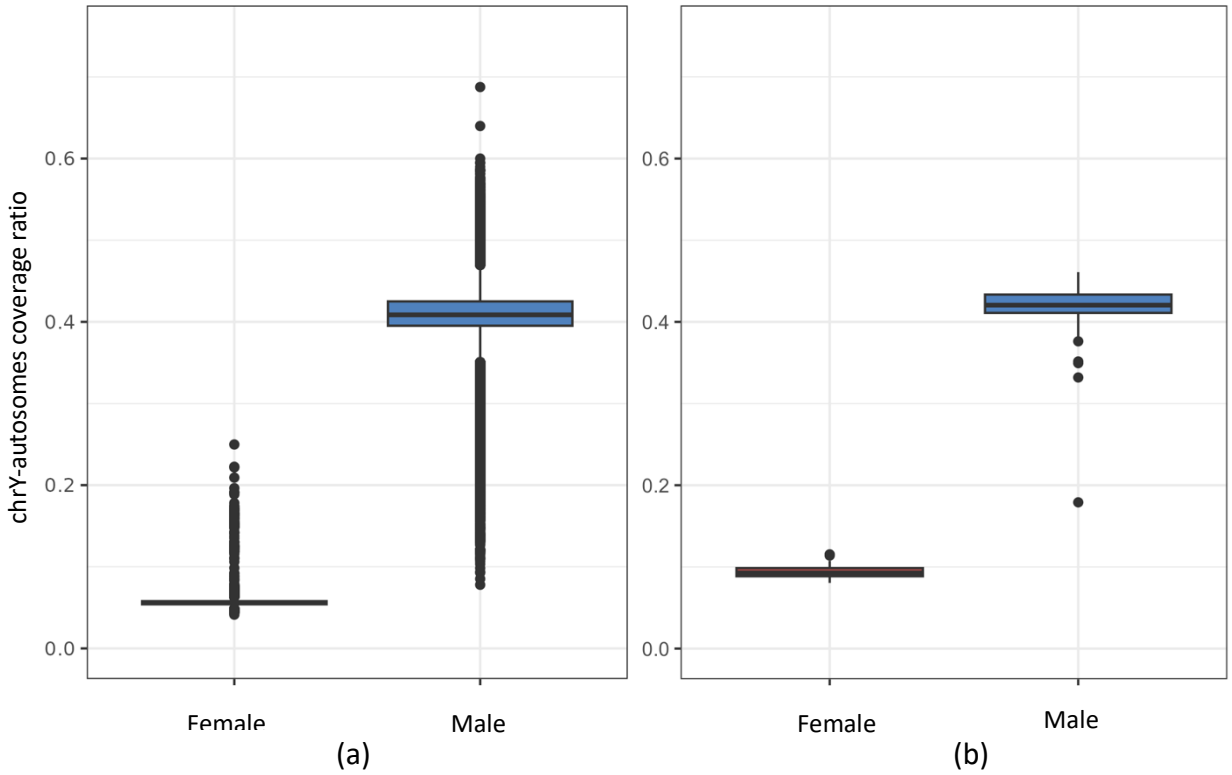
Supplementary Figure IV.1. Coverage biases also occur at the chromosomal level. We calculated the coverage ratio for each chromosome, using the chromosome of interest as the target region and all autosomes combined as the reference. (a) unrelUKB-EUR cohort sequenced on the Illumina NovaSeq 6000; (b) 1000G-GBR cohort sequenced on the Illumina NovaSeq 6000; (c) 1000G-LR cohort sequenced on the PacBio Sequel II platform. The Y-axis represents the coverage ratio between each chromosome and all autosomes, and red dotted line indicates the expected ratio of 1. Compared with regional biases, coverage biases across different sequencing platforms are more consistent in direction.



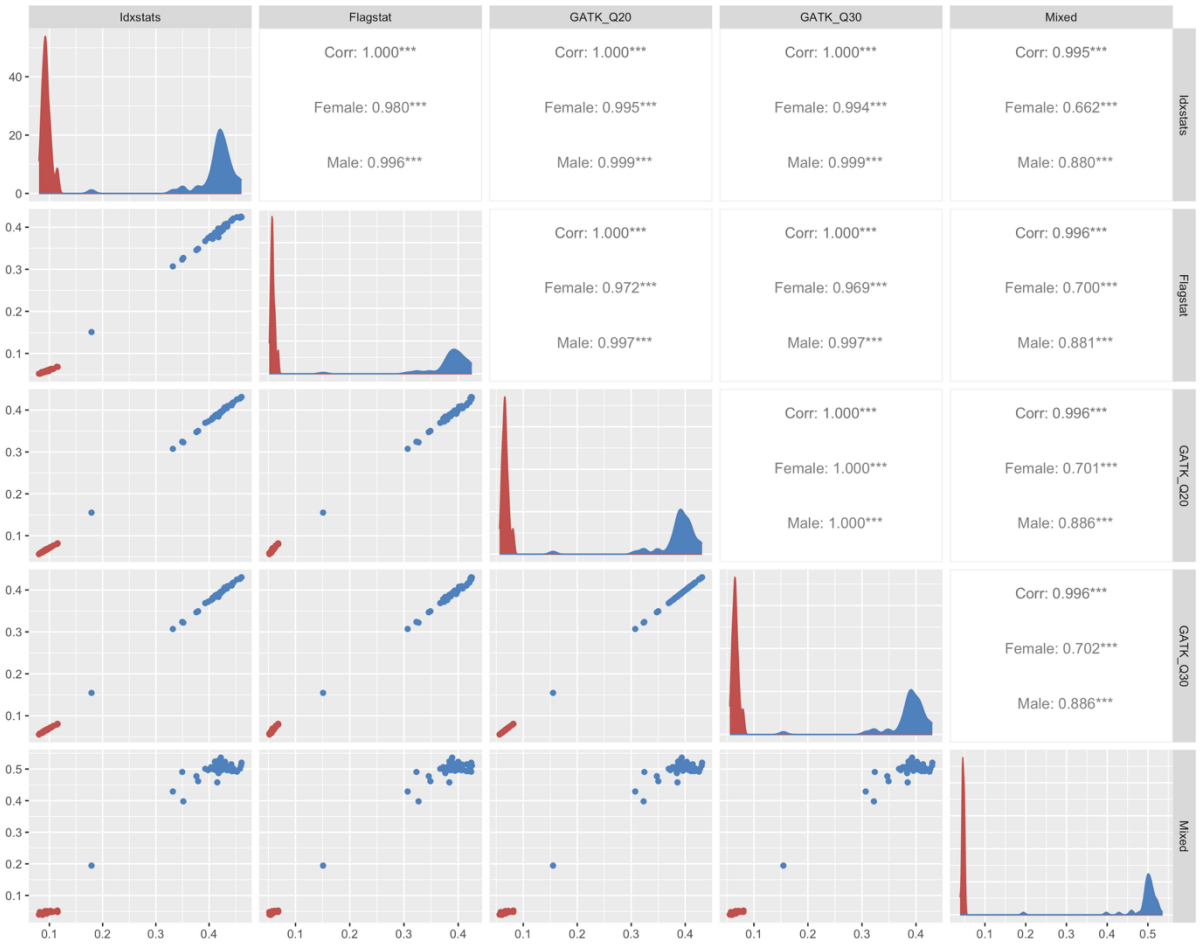
Supplementary Figure IV.2. Chromosome X–to–autosome ratio in UK Biobank and 1000 Genomes Project. Despite coverage biases at both regional and chromosomal levels, the ratio of coverage between targeted (chrX in this case) and reference sequences (autosome in this case) remains informative. A clear example is the comparison of sequencing coverage between sex chromosomes and autosomes (y-axis) across individuals of different sexes (x-axis; reported sex in UKB and biological sex in 1000G). In both (a) unrelUKB-EUR cohorts and (b) 1000G-GBR cohort, the observed mean chrX-to-autosome coverage ratios were generally consistent with theoretical expectations (1.0 in biological females, 0.5 in biological males). Approximately 0.2% of samples in UKB were classified as outliers, defined as those lying beyond 3 standard deviations from the mean within each sex. Some outliers may result from potential sample contamination, as discussed in the main text, while others likely reflect genuine biological variation, such as partial duplication or deletion of X chromosome, or mosaicism. These phenomena are not uncommon in large biobank datasets and have been reported in previous studies.



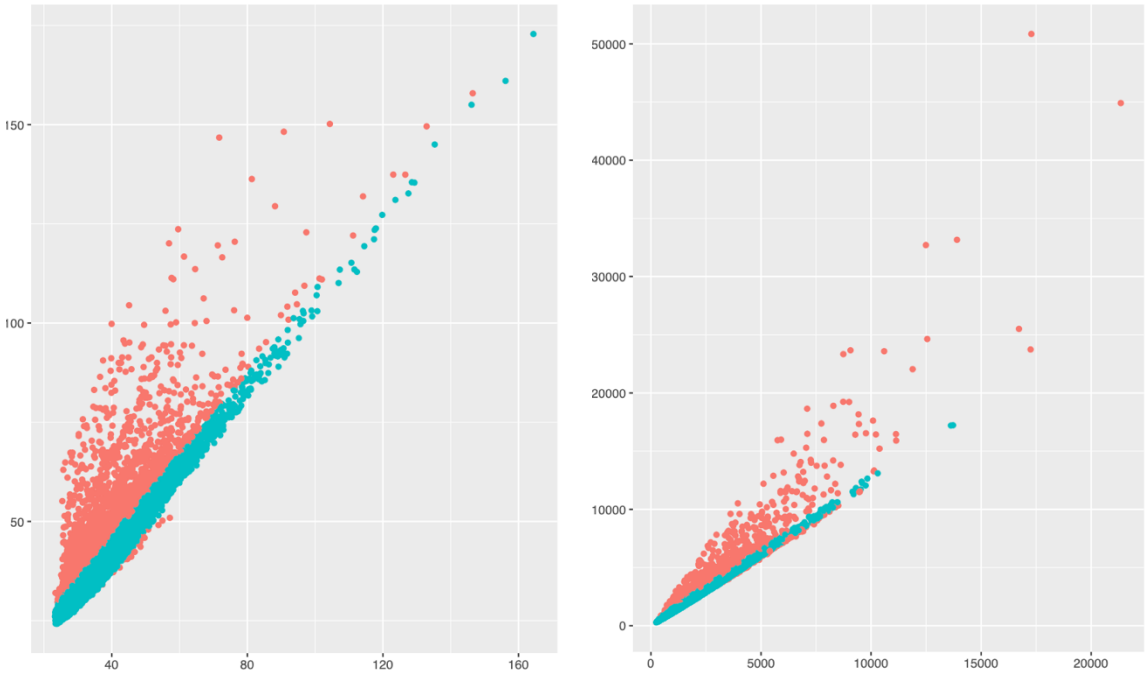
Supplementary Figure IV.3. Coverage ratio between chr13, 18, 21 and autosomes in the presence of whole-chromosome duplications. This figure examines whether the coverage of chromosomes 13, 18, and 21 is elevated in participants documented (based on ICD10 codes) with Patau syndrome (trisomy 13; N =6), Edwards syndrome (trisomy 18; N=6), and Down syndrome (trisomy 21; N=22), respectively, compared with the general population (unrelated UK Biobank participants who passed quality). Overall, the results conform to expectations. Individuals with whole-chromosome duplications associated with Patau or Edwards syndrome rarely survive beyond the first year of life, suggesting that the degree of duplication in UK Biobank participants must be minimal, or alternatively, that these records may reflect reporting inaccuracies. In contrast, individuals with trisomy 21 typically have a longer lifespan, which likely explains the observed the ability to detect the expected higher levels of chr21 coverage ratio among participants with Down syndrome.



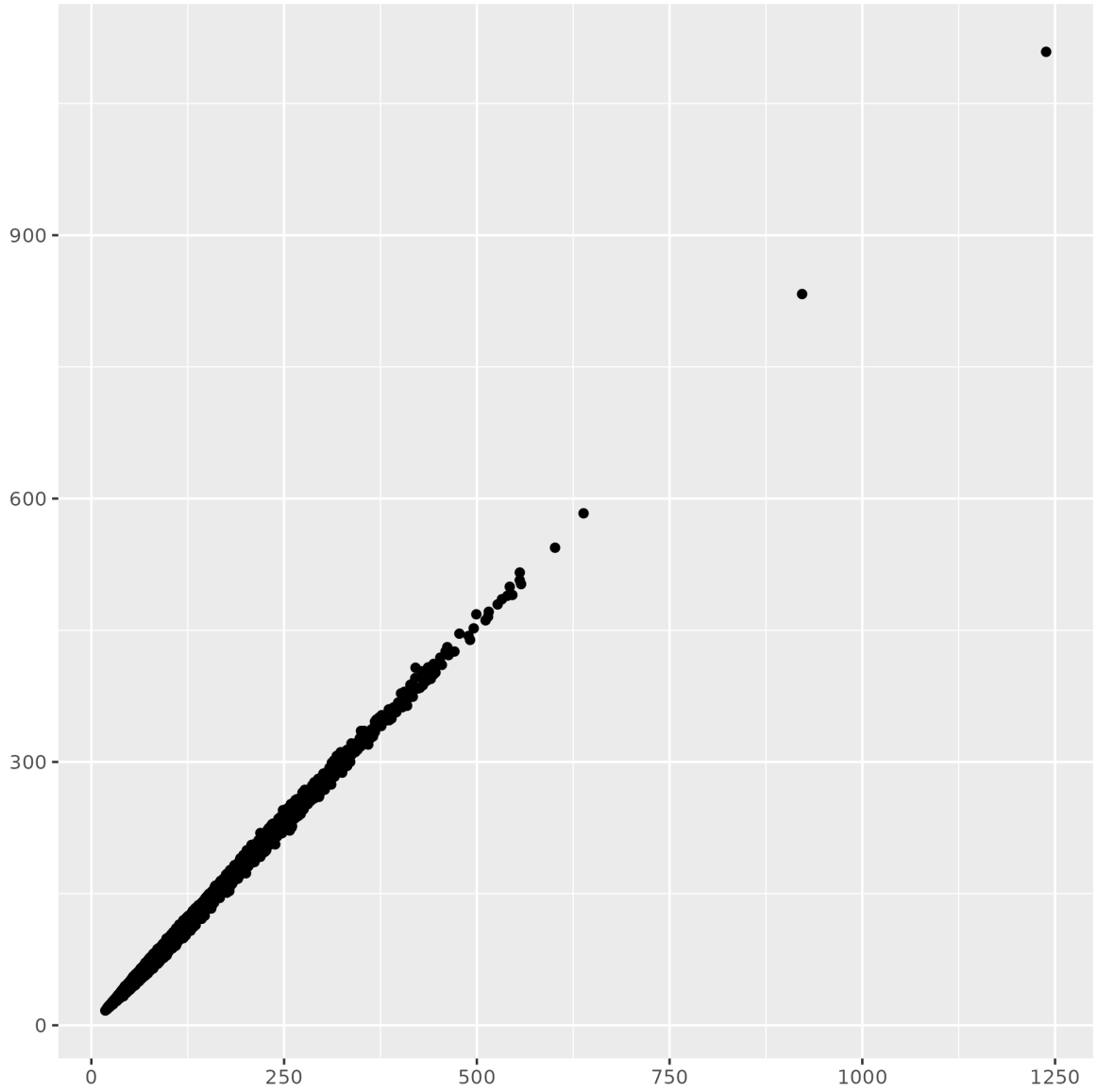
Supplementary Figure IV.4. Chromosome Y–to–autosome ratio in UK Biobank and 1000 Genomes Project. Although the overall ratios of sex chromosome to autosome coverage were broadly consistent with expectations, slight deviations from the idealised values were observed, particularly for chromosome Y. This plot shows the observed chrY-to-autosome coverage ratios in (a) unrelUKB-EUR cohorts and (b) 1000G-GBR cohort. The x-axis represents sex (reported sex in unrelUKB-EUR cohort and biological sex in 1000G-GBR cohort), and the y-axis shows the coverage ratio between chromosome Y and the autosomes. In both were generally consistent with theoretical expectations (1.0 in biological females, 0.5 in biological males). Deviations from the expected values (0 in biological females and 0.5 in biological males) likely reflect the influence of homologous regions between the sex chromosomes (e.g. pseudoautosomal regions, PARs) that lead to ambiguous read mapping, as well as technical factors such as contamination and alignment artefacts.



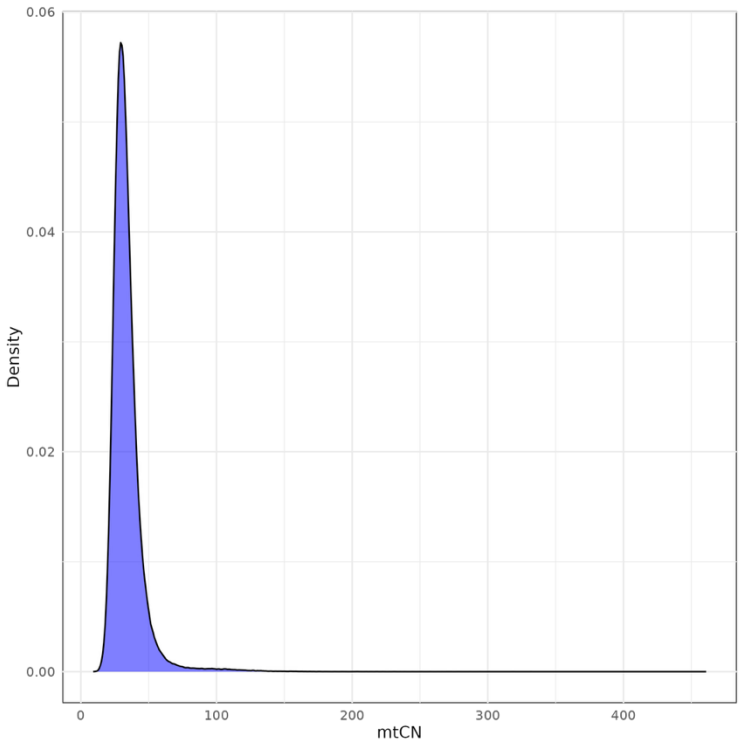
Supplementary Figure IV.5. Impact of different pipelines on the chrY-to-autosome ratio. In this figure, we applied the methodologies used in various pipelines previously described to calculate the chrY-to-autosome ratio. For chromosome Y, five approaches were applied, excluding the UCS-based method (as no UCS regions exist on chrY), consistent with Figure IV.3. Detailed descriptions of each approach are provided in the Methods section. Colours correspond to biological sex (red = female; blue = male). Row and column labels indicate each method: the lower triangle shows pairwise comparisons of inferred ratios between methods (each dot represents an individual from the 1000G-GBR cohort), the diagonal shows density plots of the inferred ratios, and the upper triangle presents correlation coefficient from linear regression (combined and stratified by sex). As observed for chromosome X, the results for the chromosome Y–to–autosome ratio were highly correlated across different approaches; however, when the target and reference regions became shorter (as in the mixed approach), estimation variability increased.



Supplementary Figure IV.6. Different pipelines yield similar results in most UK Biobank participants. In this figure, we compared autosomal coverage (left panel) and mitochondrial coverage (right panel) across all UK Biobank participants (N ≈ 490,000), calculated using the least stringent pipeline (Idxstat-based; x-axis) and the most stringent pipeline (UCS-based; see Methods). For the majority of individuals, results from the two pipelines were almost perfectly correlated. Outliers were defined as samples with regression residuals more than 3 standard deviations from the mean (highlighted in red). We identified 4,147 outliers for autosomal coverage and 1,118 outliers for mitochondrial coverage. Notably, 958 of the mitochondrial outliers were also classified as autosomal outliers.



Supplementary Figure IV.7. High correlation of mtCN estimates derived from different pipelines. This figure compares mitochondrial copy number (mtCN) estimates obtained using the Idxstat-based (x-axis) and UCS-based (y-axis) pipelines, after excluding outliers as described in the Methods section, across all European participants from the UK Biobank. The mtCN values derived from the two pipelines exhibit an almost perfect correlation, indicating that analyses such as GWAS or association testing with disease traits would yield highly consistent results regardless of the pipeline used. However, the absolute estimates differ between pipelines, which warrants caution in studies where the absolute mtCN values are critical. Owing to the absence of a definitive ground truth, it remains difficult to determine which pipeline provides more accurate estimates. Nevertheless, in our analyses, the relative mtCN values are of primary importance.



Supplementary Figure IV.8. Distribution of inferred mtCN in all UK Biobank EUR participants. In this figure, we made a density plot on all UK Biobank EUR participants that passed the quality control (N = 413,946). The mean mtCN in UK Biobank is estimated to be 33.9, which are consistent with previous observation.

Appendix

	ID	Ancestry	Genotyping Location		ID	Ancestry	Genotyping Location
1	HG00096	EUR	JAX	25	HG03371	AFR	JAX
2	HG00171	EUR	UW	26	HG03456	AFR	JAX
3	HG00268	EUR	JAX	27	HG03683	SAS	UW
4	HG00358	EUR	JAX	28	HG03732	SAS	JAX
5	HG00731	AMR	UW	29	HG03807	SAS	JAX
6	HG00733	AMR	UW	30	HG04036	SAS	UW
7	HG00864	EAS	UW	31	HG04217	SAS	JAX
8	HG01114	AMR	UW	32	NA12329	EUR	UW
9	HG01352	AMR	UW	33	NA18534	EAS	JAX
10	HG01457	AMR	JAX	34	NA18939	EAS	UW
11	HG01505	EUR	JAX	35	NA18989	EAS	JAX
12	HG01573	AMR	UW	36	NA19036	AFR	UW
13	HG01596	EAS	JAX	37	NA19238	AFR	UW
14	HG01890	AFR	JAX	38	NA19240	AFR	UW
15	HG02011	AFR	JAX	39	NA19317	AFR	JAX
16	HG02018	EAS	UW	40	NA19331	AFR	JAX
17	HG02106	AMR	JAX	41	NA19347	AFR	JAX
18	HG02492	SAS	JAX	42	NA19384	AFR	JAX
19	HG02554	AFR	UW	43	NA19650	AMR	JAX
20	HG02587	AFR	UW	44	NA19705	AFR	JAX
21	HG02666	AFR	JAX	45	NA19983	AFR	UW
22	HG03009	SAS	JAX	46	NA20509	EUR	JAX
23	HG03065	AFR	JAX	47	NA20847	SAS	UW
24	HG03248	AFR	JAX				

Supplementary Table IV.1. Participants included in the 1000G-LR cohort. This table summarises the 47 individuals from the 1000 Genomes Project who were genotyped using both short-read and long-read sequencing platforms and included in our analysis. UW=University of Washington; JAX=Jackson Lab

App

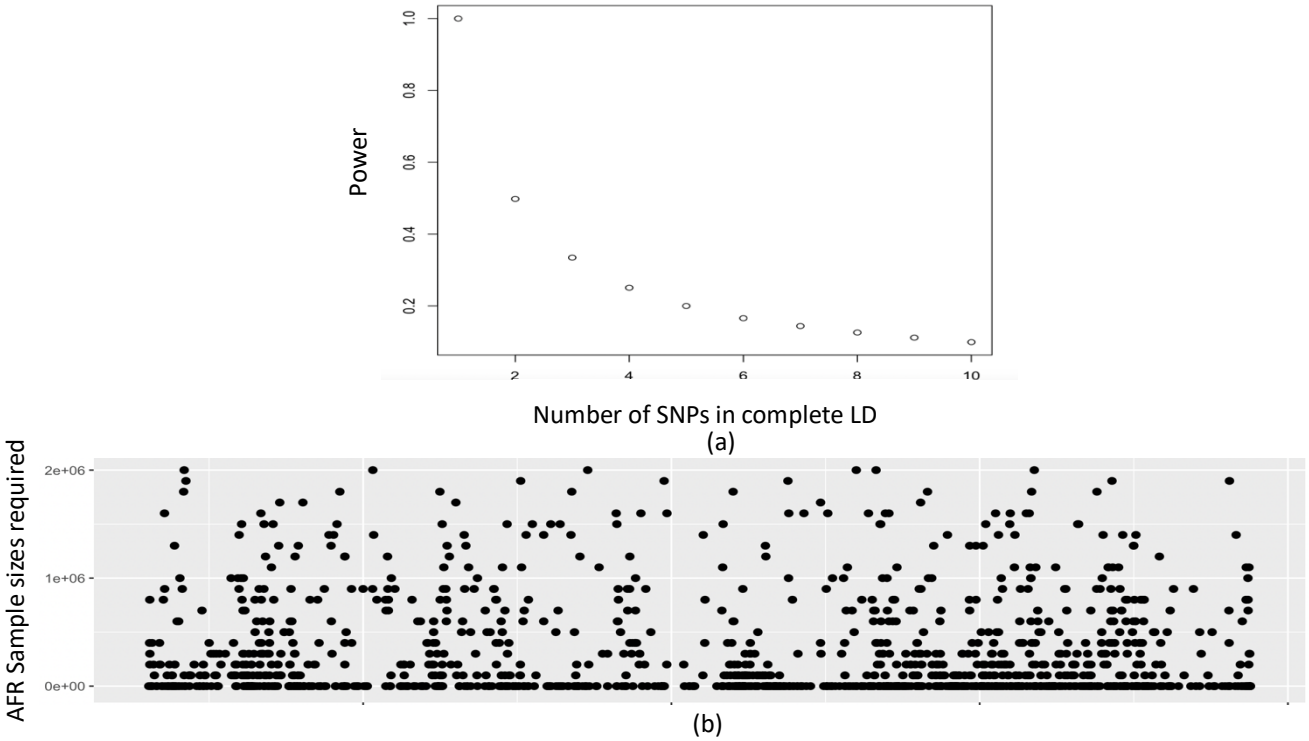
Definition	Model	Differences in Mean	SE	P	N_Case	N_Control
1_GPpsy	Raw_resid	0.01	0.00	1.62E-03	105927	198841
1_GPpsy	Basic_resid	-0.02	0.00	4.43E-10	105927	198841
1_GPpsy	Time_resid	-0.02	0.00	1.02E-10	105927	198841
1_GPpsy	All_resid	0.00	0.00	8.11E-01	105927	198841
2_Psypsy	Raw_resid	-0.02	0.01	3.42E-03	35666	198841
2_Psypsy	Basic_resid	-0.04	0.01	4.02E-10	35666	198841
2_Psypsy	Time_resid	-0.04	0.01	1.62E-10	35666	198841
2_Psypsy	All_resid	0.01	0.01	3.40E-01	35666	198841
3_DepAll	Raw_resid	-0.08	0.01	4.02E-25	19159	198841
3_DepAll	Basic_resid	-0.03	0.01	1.93E-04	19159	198841
3_DepAll	Time_resid	-0.03	0.01	5.07E-04	19159	198841
3_DepAll	All_resid	0.00	0.01	9.22E-01	19159	198841
4_SelfRepDep	Raw_resid	-0.00	0.01	6.20E-01	17704	198841
4_SelfRepDep	Basic_resid	-0.05	0.01	4.99E-09	17704	198841
4_SelfRepDep	Time_resid	-0.05	0.01	6.44E-11	17704	198841
4_SelfRepDep	All_resid	-0.01	0.01	4.36E-01	17704	198841
5_ICD10Dep	Raw_resid	-0.10	0.01	4.75E-20	8713	198841
5_ICD10Dep	Basic_resid	-0.13	0.01	2.18E-30	8713	198841
5_ICD10Dep	Time_resid	-0.13	0.01	4.42E-32	8713	198841
5_ICD10Dep	All_resid	-0.02	0.01	9.89E-02	8713	198841
6_LifetimeMDD	Raw_resid	0.07	0.01	2.56E-17	15102	198841
6_LifetimeMDD	Basic_resid	0.01	0.01	1.68E-01	15102	198841
6_LifetimeMDD	Time_resid	0.01	0.01	3.04E-01	15102	198841
6_LifetimeMDD	All_resid	0.00	0.01	6.78E-01	15102	198841
7_MDDRecur	Raw_resid	0.07	0.01	1.33E-11	9089	198841
7_MDDRecur	Basic_resid	0.01	0.01	5.74E-01	9089	198841
7_MDDRecur	Time_resid	0.00	0.01	6.87E-01	9089	198841
7_MDDRecur	All_resid	-0.00	0.01	9.49E-01	9089	198841
8_GpNoDep	Raw_resid	-0.10	0.01	2.98E-19	7741	198841
8_GpNoDep	Basic_resid	-0.03	0.01	2.79E-03	7741	198841
8_GpNoDep	Time_resid	-0.04	0.01	2.42E-03	7741	198841
8_GpNoDep	All_resid	-0.02	0.01	1.28E-01	7741	198841
9_MDDCurrent	Raw_resid	-0.13	0.05	1.26E-02	388	198841
9_MDDCurrent	Basic_resid	-0.17	0.05	6.00E-04	388	198841
9_MDDCurrent	Time_resid	-0.13	0.05	8.30E-03	388	198841
9_MDDCurrent	All_resid	-0.03	0.05	4.99E-01	388	198841

Supplementary Table IV.2. Summary of two-sample t-test to compare mtCN between MDD cases and controls. This table summarises the mean differences and corresponding *p-value*s from two-sample *t*-tests comparing cases and controls, based on the scaled residuals derived from models incorporating different sets of covariates. MDD cases and controls were defined according to the criteria outlined in the Method section.

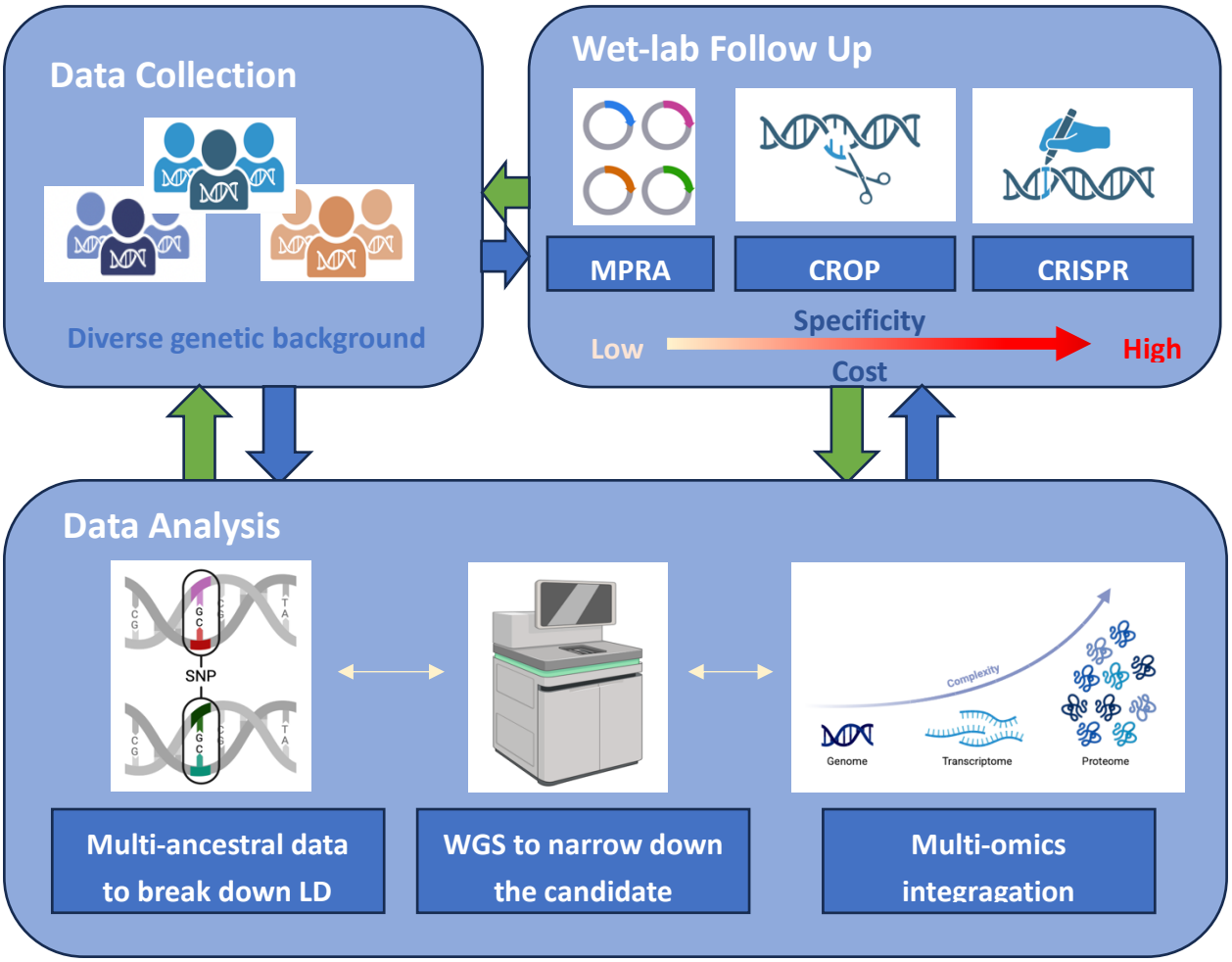
Definition	Residual	b	se	p	N_Case	N_Control
1_BipolarI	Raw_resid	-0.15	0.05	2.37E-03	416	198841
1_BipolarI	Basic_resid	-0.09	0.05	7.17E-02	416	198841
1_BipolarI	Time_resid	-0.09	0.05	8.18E-02	416	198841
1_BipolarI	All_resid	0.02	0.05	6.71E-01	416	198841
2_BipolarII	Raw_resid	-0.02	0.06	7.40E-01	241	198841
2_BipolarII	Basic_resid	0.05	0.06	4.05E-01	241	198841
2_BipolarII	Time_resid	0.05	0.06	4.73E-01	241	198841
2_BipolarII	All_resid	0.10	0.06	1.13E-01	241	198841
3_BIP_ICD10	Raw_resid	-0.13	0.02	3.59E-07	1623	198841
3_BIP_ICD11	Basic_resid	-0.15	0.02	2.43E-09	1623	198841
3_BIP_ICD12	Time_resid	-0.15	0.02	1.76E-09	1623	198841
3_BIP_ICD13	All_resid	0.01	0.02	5.85E-01	1623	198841
4_BIP_Combined	Raw_resid	-0.13	0.02	1.08E-08	2117	198841
4_BIP_Combined	Basic_resid	-0.13	0.02	7.48E-09	2117	198841
4_BIP_Combined	Time_resid	-0.13	0.02	5.05E-09	2117	198841
4_BIP_Combined	All_resid	0.02	0.02	3.63E-01	2117	198841
5_BIP_Current	Raw_resid	-0.11	0.27	6.81E-01	14	198841
5_BIP_Current	Basic_resid	-0.21	0.27	4.22E-01	14	198841
5_BIP_Current	Time_resid	-0.17	0.27	5.32E-01	14	198841
5_BIP_Current	All_resid	0.06	0.27	8.29E-01	14	198841

Supplementary Table IV.3. Summary of two-sample t-test to compare mtCN between BIP cases and controls. This table summarises the mean differences and corresponding *p-value*s from two-sample *t*-tests comparing cases and controls, based on the scaled residuals derived from models incorporating different sets of covariates. BIP cases and controls were defined according to the criteria outlined in the Method section.

A4. Supplementary files for Chapter V



Supplementary Figure V.1 Sample sizes required to completely break-down LD. Most genetic data to date are centred on European populations, making it imperative to determine the minimum sample size required from African (and other) ancestries to break down LD and identify independent associations. (a) In a simulation scenario with a single causal variant shared across two ancestries, there are *M-1* variants in perfect LD proximal to this causal variant. The probability of identifying the causal variant is approximately 1/*M*. (b) As a realistic example, we analysed real genetic data from an LD-independent block (spanning base pair positions 25,662,928 to 26,469,735 hg37 on chromosome 22) overlapping in both African and European populations. LD correlation outside this block is assumed to be 0. We set the true causal effect size at 0.1 and fixed the European population sample size at 1,000,000 (as inputs for equation III.19). For each of the 3,719 SNPs within this block, we designated one SNP as the causal variant in turn. We then calculated the minimum sample size required in the African population to achieve a power greater than 0.8 for tagging the true causal variant (Y-axis). For a substantial proportion of SNPs (68% in this block), completely resolving LD with a realistic sample size is impossible when they are the true causal variants as too many SNPs are in perfect LD with them in both ancestries. However, for the remaining 32% (shown in the plot, where each dot represents an SNP and the X-axis denotes base pair position on chromosome 22), integrating multi-ancestral data makes it feasible to fully break down LD.



Supplementary Figure V.2 Proposed framework for future genetic research. This figure presents an idealised pipeline for future genetic research that is both cost-effective and prioritises health equity across diverse ancestries. Resources should be directed towards underrepresented ancestry groups during data collection. This approach not only enhances confidence in detecting ancestry-specific associations but also aids in breaking down LD, thereby reducing the cost of fine-mapping variants in wet-lab settings. In addition to the traditional unidirectional flow of analysis (blue arrows), future analyses should adopt a bidirectional approach (green arrows). Functional variants identified through wet-lab approaches can be reintegrated into dry-lab analyses to facilitate the identification of independent associations and guide subsequent data collection efforts. Conversely, outcomes from statistical genetic analyses can inform future data collection priorities. Within statistical genetic analysis, multi-ancestry methods, such as Manc-COJO, can disentangle LD and refine the list of

R

References

Refenrences

1. Yengo, L., Vedantam, S., Marouli, E., Sidorenko, J., Bartell, E., Sakaue, S., Graff, M., Eliasen, A.U., Jiang, Y., and Raghavan, S. (2022). A saturated map of common genetic variants associated with human height. Nature *610*, 704-712.
2. Cerezo, M., Sollis, E., Ji, Y., Lewis, E., Abid, A., Bircan, K.O., Hall, P., Hayhurst, J., John, S., and Mosaku, A. (2025). The NHGRI-EBI GWAS Catalog: standards for reusability, sustainability and diversity. Nucleic acids research *53*, D998-D1005.
3. Visscher, P.M., Yengo, L., Cox, N.J., and Wray, N.R. (2021). Discovery and implications of polygenicity of common diseases. Science *373*, 1468-1473. 10.1126/science.abi8206.
4. Relling, M.V., and Evans, W.E. (2015). Pharmacogenomics in the clinic. Nature *526*, 343-350.
5. Hueber, W., Patel, D.D., Dryja, T., Wright, A.M., Koroleva, I., Bruin, G., Antoni, C., Draelos, Z., Gold, M.H., and Group, P.S. (2010). Effects of AIN457, a fully human antibody to interleukin-17A, on psoriasis, rheumatoid arthritis, and uveitis. Science translational medicine *2*, 52ra72-52ra72.
6. Visscher, P.M., Wray, N.R., Zhang, Q., Sklar, P., McCarthy, M.I., Brown, M.A., and Yang, J. (2017). 10 years of GWAS discovery: biology, function, and translation. The American Journal of Human Genetics *101*, 5-22.
7. Kamiza, A.B., Toure, S.M., Vujkovic, M., Machipisa, T., Soremekun, O.S., Kintu, C., Corpas, M., Pirie, F., Young, E., and Gill, D. (2022). Transferability of genetic risk scores in African populations. Nature Medicine *28*, 1163-1166.
8. Lewis, C.M., and Vassos, E. (2020). Polygenic risk scores: from research tools to clinical instruments. Genome medicine *12*, 1-11.
9. Uffelmann, E., Huang, Q.Q., Munung, N.S., De Vries, J., Okada, Y., Martin, A.R., Martin, H.C., Lappalainen, T., and Posthuma, D. (2021). Genome-wide association studies. Nature Reviews Methods Primers *1*, 1-21.
10. Pritchard, J.K., and Przeworski, M. (2001). Linkage disequilibrium in humans: models and data. The American Journal of Human Genetics *69*, 1-14.
11. Cai, N., Revez, J.A., Adams, M.J., Andlauer, T.F.M., Breen, G., Byrne, E.M., Clarke, T.K., Forstner, A.J., Grabe, H.J., Hamilton, S.P., et al. (2020). Minimal phenotyping yields genome-wide association signals of low specificity for major depression. Nat Genet *52*, 437-447.
12. Sullivan, P. (2009). A framework for interpreting genome-wide association studies of psychiatric disorders. Molecular psychiatry.
13. Siva, N. (2008). 1000 Genomes project. Nature biotechnology *26*, 256-257.
14. McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A.R., Teumer, A., Kang, H.M., Fuchsberger, C., Danecek, P., Sharp, K., et al. (2016). A reference panel of 64,976 haplotypes for genotype imputation. Nat Genet *48*, 1279-1283. 10.1038/ng.3643.

15. Taliun, D., Harris, D.N., Kessler, M.D., Carlson, J., Szpiech, Z.A., Torres, R., Taliun, S.A.G., Corvelo, A., Gogarten, S.M., and Kang, H.M. (2021). Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. Nature *590*, 290-299.
16. Lam, M., Awasthi, S., Watson, H.J., Goldstein, J., Panagiotaropoulou, G., Trubetskoy, V., Karlsson, R., Frei, O., Fan, C.-C., and De Witte, W. (2020). RICOPILI: rapid imputation for COnsortias PIpeLIne. Bioinformatics *36*, 930-933.
17. Yang, J., Bakshi, A., Zhu, Z., Hemani, G., Vinkhuyzen, A.A., Lee, S.H., Robinson, M.R., Perry, J.R., Nolte, I.M., and van Vliet-Ostaptchouk, J.V. (2015). Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. Nature genetics *47*, 1114-1120.
18. Abdellaoui, A., Yengo, L., Verweij, K.J., and Visscher, P.M. (2023). 15 years of GWAS discovery: Realizing the promise. The American Journal of Human Genetics.
19. Porta, M. (2014). A dictionary of epidemiology (Oxford university press).
20. Marees, A.T., de Kluiver, H., Stringer, S., Vorspan, F., Curis, E., Marie-Claire, C., and Derks, E.M. (2018). A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. International journal of methods in psychiatric research *27*, e1608.
21. Anderson, C.A., Pettersson, F.H., Clarke, G.M., Cardon, L.R., Morris, A.P., and Zondervan, K.T. (2010). Data quality control in genetic case-control association studies. Nature protocols *5*, 1564-1573.
22. Investigators, A.o.U.R.P. (2019). The “All of Us” research program. New England Journal of Medicine *381*, 668-676.
23. Cook, M.B., Sanderson, S.C., Deanfield, J.E., Reddington, F., Roddam, A., Hunter, D.J., and Ali, R. (2025). Our Future Health: a unique global resource for discovery and translational research. Nature Medicine *31*, 728-730.
24. Scholtens, S., Smidt, N., Swertz, M.A., Bakker, S.J., Dotinga, A., Vonk, J.M., Van Dijk, F., van Zon, S.K., Wijmenga, C., and Wolffenbuttel, B.H. (2015). Cohort Profile: LifeLines, a three-generation cohort study and biobank. International journal of epidemiology *44*, 1172-1180.
25. Chen, Z., Chen, J., Collins, R., Guo, Y., Peto, R., Wu, F., and Li, L. (2011). China Kadoorie Biobank of 0.5 million people: survey methods, baseline characteristics and long-term follow-up. International journal of epidemiology *40*, 1652-1666.
26. Davis, L. (2019). Psychiatric Genomics, Phenomics, and Ethics Research In A 270,000-Person Biobank (BioVU). European Neuropsychopharmacology *29*, S739-S740.
27. MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., Junkins, H., McMahon, A., Milano, A., and Morales, J. (2017). The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic acids research *45*, D896-D901.
28. Canela-Xandri, O., Rawlik, K., and Tenesa, A. (2018). An atlas of genetic associations in UK Biobank. Nature genetics *50*, 1593-1599.

29. Kurki, M.I., Karjalainen, J., Palta, P., Sipilä, T.P., Kristiansson, K., Donner, K., Reeve, M.P., Laivuori, H., Aavikko, M., and Kaunisto, M.A. (2022). FinnGen: Unique genetic insights from combining isolated population and national health register data. medRxiv.
30. Tryka, K.A., Hao, L., Sturcke, A., Jin, Y., Wang, Z.Y., Ziyabari, L., Lee, M., Popova, N., Sharopova, N., and Kimura, M. (2014). NCBI's Database of Genotypes and Phenotypes: dbGaP. Nucleic acids research *42*, D975-D979.
31. Nagai, A., Hirata, M., Kamatani, Y., Muto, K., Matsuda, K., Kiyohara, Y., Ninomiya, T., Tamakoshi, A., Yamagata, Z., and Mushiroda, T. (2017). Overview of the BioBank Japan Project: study design and profile. Journal of epidemiology *27*, S2-S8.
32. Jiang, L., Zheng, Z., Qi, T., Kemper, K.E., Wray, N.R., Visscher, P.M., and Yang, J. (2019). A resource-efficient tool for mixed model association analysis of large-scale data. Nat Genet *51*, 1749-1755. 10.1038/s41588-019-0530-8.
33. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J., and Sham, P.C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet *81*, 559-575. 10.1086/519795.
34. Loh, P.R., Tucker, G., Bulik-Sullivan, B.K., Vilhjalmsson, B.J., Finucane, H.K., Salem, R.M., Chasman, D.I., Ridker, P.M., Neale, B.M., Berger, B., et al. (2015). Efficient Bayesian mixed-model analysis increases association power in large cohorts. Nat Genet *47*, 284-290. 10.1038/ng.3190.
35. Zhang, Y., and Pan, W. (2015). Principal component regression and linear mixed model in association analysis of structured samples: competitors or complements? Genet Epidemiol *39*, 149-155. 10.1002/gepi.21879.
36. Yang, J., Lee, S.H., Goddard, M.E., and Visscher, P.M. (2011). GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet *88*, 76-82.
37. Chen, D., Liu, C., and Xie, J. (2016). Multi-locus Test and Correction for Confounding Effects in Genome-Wide Association Studies. Int J Biostat *12*. 10.1515/ijb-2015-0091.
38. Evans, D.M., Spencer, C.C., Pointon, J.J., Su, Z., Harvey, D., Kochan, G., Oppermann, U., Dilthey, A., Pirinen, M., and Stone, M.A. (2011). Interaction between ERAP1 and HLA-B27 in ankylosing spondylitis implicates peptide handling in the mechanism for HLA-B27 in disease susceptibility. Nature genetics *43*, 761-767.
39. Pirinen, M., Donnelly, P., and Spencer, C.C. (2012). Including known covariates can reduce power to detect genetic effects in case-control studies. Nature genetics *44*, 848-851.
40. Schisterman, E.F., Cole, S.R., and Platt, R.W. (2009). Overadjustment bias and unnecessary adjustment in epidemiologic studies. Epidemiology *20*, 488-495.
41. Pe'er, I., Yelensky, R., Altshuler, D., and Daly, M.J. (2008). Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. Genetic

Epidemiology: The Official Publication of the International Genetic Epidemiology Society *32*, 381-385.

42. Consortium, I.H. (2005). A haplotype map of the human genome. Nature *437*, 1299.
43. Wu, Y., Zheng, Z., Visscher, P.M., and Yang, J. (2017). Quantifying the mapping precision of genome-wide association studies using whole-genome sequencing data. Genome biology *18*, 86.
44. Xiao, R., and Boehnke, M. (2011). Quantifying and correcting for the winner's curse in quantitative-trait association studies. Genet Epidemiol *35*, 133-138. 10.1002/gepi.20551.
45. Willer, C.J., Li, Y., and Abecasis, G.R. (2010). METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics *26*, 2190-2191.
46. Shahar, D.J. (2017). Minimizing the variance of a weighted average. Open Journal of Statistics *7*, 216-224.
47. Lin, D.-Y., and Sullivan, P.F. (2009). Meta-analysis of genome-wide association studies with overlapping subjects. The American Journal of Human Genetics *85*, 862-872.
48. Turley, P., Walters, R.K., Maghzian, O., Okbay, A., Lee, J.J., Fontana, M.A., Nguyen-Viet, T.A., Wedow, R., Zacher, M., and Furlotte, N.A. (2018). Multi-trait analysis of genome-wide association summary statistics using MTAG. Nature genetics *50*, 229-237.
49. Schaid, D.J., Chen, W., and Larson, N.B. (2018). From genome-wide associations to candidate causal variants by statistical fine-mapping. Nat Rev Genet *19*, 491-504. 10.1038/s41576-018-0016-z.
50. Burgess, D.J. (2022). Fine-mapping causal variants—why finding 'the one'can be futile. Nature Reviews Genetics *23*, 261-261.
51. Barrett, J.C., Fry, B., Maller, J., and Daly, M.J. (2005). Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics *21*, 263-265. 10.1093/bioinformatics/bth457.
52. Pruim, R.J., Welch, R.P., Sanna, S., Teslovich, T.M., Chines, P.S., Gliedt, T.P., Boehnke, M., Abecasis, G.R., and Willer, C.J. (2010). LocusZoom: regional visualization of genome-wide association scan results. Bioinformatics *26*, 2336-2337. 10.1093/bioinformatics/btq419.
53. Yang, J., Ferreira, T., Morris, A.P., Medland, S.E., Genetic Investigation of, A.T.C., Replication, D.I.G., Meta-analysis, C., Madden, P.A., Heath, A.C., Martin, N.G., et al. (2012). Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. Nat Genet *44*, 369-375, S361-363. 10.1038/ng.2213.
54. Ning, Z., Lee, Y., Joshi, P.K., Wilson, J.F., Pawitan, Y., and Shen, X. (2017). A Selection Operator for Summary Association Statistics Reveals Allelic Heterogeneity of Complex Traits. Am J Hum Genet *101*, 903-912. 10.1016/j.ajhg.2017.09.027.
55. Chen, W., Larrabee, B.R., Ovsyannikova, I.G., Kennedy, R.B., Haralambieva, I.H., Poland, G.A., and Schaid, D.J. (2015). Fine mapping causal variants with an approximate Bayesian method using marginal test statistics. Genetics *200*, 719-736.

56. Hormozdiari, F., Kostem, E., Kang, E.Y., Pasaniuc, B., and Eskin, E. (2014). Identifying causal variants at loci with multiple signals of association. Genetics *198*, 497-508.
57. Wang, G., Sarkar, A., Carbonetto, P., and Stephens, M. (2020). A simple new approach to variable selection in regression, with application to genetic fine-mapping. BioRxiv, 501114.
58. Hutchinson, A., Watson, H., and Wallace, C. (2020). Improving the coverage of credible sets in Bayesian genetic fine-mapping. PLoS Computational Biology *16*, e1007829.
59. Abell, N.S., DeGorter, M.K., Gloudemans, M.J., Greenwald, E., Smith, K.S., He, Z., and Montgomery, S.B. (2022). Multiple causal variants underlie genetic associations in humans. Science *375*, 1247-1254. 10.1126/science.abj5117.
60. Marigorta, U.M., and Navarro, A. (2013). High trans-ethnic replicability of GWAS results implies common causal variants. PLoS Genet *9*, e1003566. 10.1371/journal.pgen.1003566.
61. Igartua, C., Myers, R.A., Mathias, R.A., Pino-Yanes, M., Eng, C., Graves, P.E., Levin, A.M., Del-Rio-Navarro, B.E., Jackson, D.J., Livne, O.E., et al. (2015). Ethnic-specific associations of rare and low-frequency DNA sequence variants with asthma. Nat Commun *6*, 5965. 10.1038/ncomms6965.
62. Wang, Y., Guo, J., Ni, G., Yang, J., Visscher, P.M., and Yengo, L. (2020). Theoretical and empirical quantification of the accuracy of polygenic scores in ancestry divergent populations. Nature communications *11*, 1-9.
63. Asimit, J.L., Hatzikotoulas, K., McCarthy, M., Morris, A.P., and Zeggini, E. (2016). Trans-ethnic study design approaches for fine-mapping. Eur J Hum Genet *24*, 1330-1336. 10.1038/ejhg.2016.1.
64. Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res *38*, e164. 10.1093/nar/gkq603.
65. Kircher, M., Witten, D.M., Jain, P., O'Roak, B.J., Cooper, G.M., and Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet *46*, 310-315. 10.1038/ng.2892.
66. Ioannidis, N.M., Rothstein, J.H., Pejaver, V., Middha, S., McDonnell, S.K., Baheti, S., Musolf, A., Li, Q., Holzinger, E., Karyadi, D., et al. (2016). REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. Am J Hum Genet *99*, 877-885. 10.1016/j.ajhg.2016.08.016.
67. Nasser, J., Bergman, D.T., Fulco, C.P., Guckelberger, P., Doughty, B.R., Patwardhan, T.A., Jones, T.R., Nguyen, T.H., Ulirsch, J.C., Lekschas, F., et al. (2021). Genome-wide enhancer maps link risk variants to disease genes. Nature *593*, 238-243. 10.1038/s41586-021-03446-x.
68. Consortium, E.P. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57-74. 10.1038/nature11247.

69. Roadmap Epigenomics, C., Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., et al. (2015). Integrative analysis of 111 reference human epigenomes. Nature *518*, 317-330. 10.1038/nature14248.
70. Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., and Young, N. (2013). The genotype-tissue expression (GTEx) project. Nature genetics *45*, 580-585.
71. Consortium, G. (2020). Cell type-specific genetic regulation of gene expression across human tissues. Science (New York, NY) *369*, eaaz8528.
72. Wainberg, M., Sinnott-Armstrong, N., Mancuso, N., Barbeira, A.N., Knowles, D.A., Golan, D., Ermel, R., Ruusalepp, A., Quertermous, T., Hao, K., et al. (2019). Opportunities and challenges for transcriptome-wide association studies. Nat Genet *51*, 592-599. 10.1038/s41588-019-0385-z.
73. Suhre, K., McCarthy, M.I., and Schwenk, J.M. (2021). Genetics meets proteomics: perspectives for large population-based studies. Nature Reviews Genetics *22*, 19-37.
74. Yang, J., Bakshi, A., Zhu, Z., Hemani, G., Vinkhuyzen, A.A., Lee, S.H., Robinson, M.R., Perry, J.R., Nolte, I.M., van Vliet-Ostaptchouk, J.V., et al. (2015). Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. Nat Genet *47*, 1114-1120.
75. Yang, J., Benyamin, B., McEvoy, B.P., Gordon, S., Henders, A.K., Nyholt, D.R., Madden, P.A., Heath, A.C., Martin, N.G., Montgomery, G.W., et al. (2010). Common SNPs explain a large proportion of the heritability for human height. Nat Genet *42*, 565-569.
76. Bulik-Sullivan, B.K., Loh, P.R., Finucane, H.K., Ripke, S., Yang, J., Schizophrenia Working Group of the Psychiatric Genomics, C., Patterson, N., Daly, M.J., Price, A.L., and Neale, B.M. (2015). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat Genet *47*, 291-295.
77. Ni, G., Moser, G., Schizophrenia Working Group of the Psychiatric Genomics, C., Wray, N.R., and Lee, S.H. (2018). Estimation of Genetic Correlation via Linkage Disequilibrium Score Regression and Genomic Restricted Maximum Likelihood. Am J Hum Genet *102*, 1185-1194. 10.1016/j.ajhg.2018.03.021.
78. Lee, S.H., Wray, N.R., Goddard, M.E., and Visscher, P.M. (2011). Estimating missing heritability for disease from genome-wide association studies. The American Journal of Human Genetics *88*, 294-305.
79. Wainschtein, P., Jain, D., Zheng, Z., Group, T.O.A.W., Consortium, N.T.-O.f.P.M., Cupples, L.A., Shadyab, A.H., McKnight, B., Shoemaker, B.M., Mitchell, B.D., et al. (2022). Assessing the contribution of rare variants to complex trait heritability from whole-genome sequence data. Nat Genet *54*, 263-273. 10.1038/s41588-021-00997-7.

80. Wainschtein, P., Jain, D., Zheng, Z., Cupples, L.A., Shadyab, A.H., McKnight, B., Shoemaker, B.M., and Mitchell, B.D. (2022). Assessing the contribution of rare variants to complex trait heritability from whole-genome sequence data. Nature genetics *54*, 263-273.
81. Wainschtein, P., Zhang, Y., Schwartzentruber, J., Kassam, I., Sidorenko, J., Fiziev, P.P., Wang, H., McRae, J., Border, R., and Zaitlen, N. (2025). Estimation and mapping of the missing heritability of human phenotypes. Nature, 1-9.
82. Zuk, O., Hechter, E., Sunyaev, S.R., and Lander, E.S. (2012). The mystery of missing heritability: Genetic interactions create phantom heritability. Proc Natl Acad Sci U S A *109*, 1193-1198. 10.1073/pnas.1119675109.
83. van Rheenen, W., Peyrot, W.J., Schork, A.J., Lee, S.H., and Wray, N.R. (2019). Genetic correlations of polygenic disease traits: from theory to practice. Nat Rev Genet *20*, 567-581. 10.1038/s41576-019-0137-z.
84. Lee, S.H., Yang, J., Goddard, M.E., Visscher, P.M., and Wray, N.R. (2012). Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. Bioinformatics *28*, 2540-2542.
85. Evans, L.M., Tahmasbi, R., Vrieze, S.I., Abecasis, G.R., Das, S., Gazal, S., Bjelland, D.W., De Candia, T.R., Goddard, M.E., and Neale, B.M. (2018). Comparison of methods that use whole genome data to estimate the heritability and genetic architecture of complex traits. Nat. Genet. *50*, 737-745.
86. Ni, G., Moser, G., Ripke, S., Neale, B.M., Corvin, A., Walters, J.T., Farh, K.-H., Holmans, P.A., Lee, P., and Bulik-Sullivan, B. (2018). Estimation of genetic correlation via linkage disequilibrium score regression and genomic restricted maximum likelihood. The American Journal of Human Genetics *102*, 1185-1194.
87. Shi, H., Mancuso, N., Spendlove, S., and Pasaniuc, B. (2017). Local Genetic Correlation Gives Insights into the Shared Genetic Architecture of Complex Traits. Am J Hum Genet *101*, 737-751. 10.1016/j.ajhg.2017.09.022.
88. Werme, J., van der Sluis, S., Posthuma, D., and de Leeuw, C.A. (2022). An integrated framework for local genetic correlation analysis. Nat Genet *54*, 274-282. 10.1038/s41588-022-01017-y.
89. Consortium, I.S., Purcell, S.M., Wray, N.R., Stone, J.L., Visscher, P.M., O'Donovan, M.C., Sullivan, P.F., and Sklar, P. (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature *460*, 748-752. 10.1038/nature08185.
90. Erbe, M., Hayes, B., Matukumalli, L., Goswami, S., Bowman, P., Reich, C., Mason, B., and Goddard, M. (2012). Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. Journal of dairy science *95*, 4114-4129.

91. Moser, G., Lee, S.H., Hayes, B.J., Goddard, M.E., Wray, N.R., and Visscher, P.M. (2015). Simultaneous discovery, estimation and prediction analysis of complex traits using a Bayesian mixture model. PLoS genetics *11*, e1004969.

92. Yengo, L., Sidorenko, J., Kemper, K.E., Zheng, Z., Wood, A.R., Weedon, M.N., Frayling, T.M., Hirschhorn, J., Yang, J., and Visscher, P.M. (2018). Meta-analysis of genome-wide association studies for height and body mass index in∼ 700000 individuals of European ancestry. Human molecular genetics *27*, 3641-3649.

93. Lloyd-Jones, L.R., Zeng, J., Sidorenko, J., Yengo, L., Moser, G., Kemper, K.E., Wang, H., Zheng, Z., Magi, R., Esko, T., et al. (2019). Improved polygenic prediction by Bayesian multiple regression on summary statistics. Nat Commun *10*, 5086.

94. MacLeod, I.M., Bowman, P.J., Vander Jagt, C.J., Haile-Mariam, M., Kemper, K.E., Chamberlain, A.J., Schrooten, C., Hayes, B.J., and Goddard, M.E. (2016). Exploiting biological priors and sequence variants enhances QTL discovery and genomic prediction of complex traits. BMC Genomics *17*, 144. 10.1186/s12864-016-2443-6.

95. Zheng, Z., Liu, S., Sidorenko, J., Wang, Y., Lin, T., Yengo, L., Turley, P., Ani, A., Wang, R., and Nolte, I.M. (2024). Leveraging functional genomic annotations and genome coverage to improve polygenic prediction of complex traits within and between ancestries. Nature Genetics *56*, 767-777.

96. Prive, F., Arbel, J., and Vilhjalmsson, B.J. (2020). LDpred2: better, faster, stronger. Bioinformatics. 10.1093/bioinformatics/btaa1029.

97. Lee, S.H., Goddard, M.E., Wray, N.R., and Visscher, P.M. (2012). A better coefficient of determination for genetic profile analysis. Genet Epidemiol *36*, 214-224.

98. Wray, N.R., Yang, J., Goddard, M.E., and Visscher, P.M. (2010). The genetic interpretation of area under the ROC curve in genomic profiling. PLoS Genet *6*, e1000864. 10.1371/journal.pgen.1000864.

99. Kerr, K.F., Wang, Z., Janes, H., McClelland, R.L., Psaty, B.M., and Pepe, M.S. (2014). Net reclassification indices for evaluating risk prediction instruments: a critical review. Epidemiology *25*, 114-121. 10.1097/EDE.0000000000000018.

100. Zheng, J., Baird, D., Borges, M.C., Bowden, J., Hemani, G., Haycock, P., Evans, D.M., and Smith, G.D. (2017). Recent Developments in Mendelian Randomization Studies. Curr Epidemiol Rep *4*, 330-345. 10.1007/s40471-017-0128-6.

101. Hemani, G., Zheng, J., Elsworth, B., Wade, K.H., Haberland, V., Baird, D., Laurin, C., Burgess, S., Bowden, J., and Langdon, R. (2018). The MR-Base platform supports systematic causal inference across the human phenome. elife *7*.

102. Tan, M.H. (1980). HDL-cholesterol: the negative risk factor for coronary heart disease. Ann Acad Med Singap *9*, 491-495.

103. Emdin, C.A., Khera, A.V., and Kathiresan, S. (2017). Mendelian randomization. Jama *318*, 1925-1926.

104. Davies, N.M., Holmes, M.V., and Davey Smith, G. (2018). Reading Mendelian randomisation studies: a guide, glossary, and checklist for clinicians. BMJ *362*, k601. 10.1136/bmj.k601.
105. Schwartz, G.G., Olsson, A.G., Abt, M., Ballantyne, C.M., Barter, P.J., Brumm, J., Chaitman, B.R., Holme, I.M., Kallend, D., Leiter, L.A., et al. (2012). Effects of dalcetrapib in patients with a recent acute coronary syndrome. N Engl J Med *367*, 2089-2099. 10.1056/NEJMoa1206797.
106. Berry, D.J., Vimaleswaran, K.S., Whittaker, J.C., Hingorani, A.D., and Hyppönen, E. (2012). Evaluation of genetic markers as instruments for Mendelian randomization studies on vitamin D. PloS one *7*, e37465.
107. Bowden, J., Davey Smith, G., Haycock, P.C., and Burgess, S. (2016). Consistent estimation in Mendelian randomization with some invalid instruments using a weighted median estimator. Genetic epidemiology *40*, 304-314.
108. Bowden, J., Davey Smith, G., and Burgess, S. (2015). Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. Int J Epidemiol *44*, 512-525. 10.1093/ije/dyv080.
109. Zhu, Z., Zhang, F., Hu, H., Bakshi, A., Robinson, M.R., Powell, J.E., Montgomery, G.W., Goddard, M.E., Wray, N.R., and Visscher, P.M. (2016). Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. Nature genetics *48*, 481-487.
110. Verbanck, M., Chen, C.-Y., Neale, B., and Do, R. (2018). Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. Nature genetics *50*, 693-698.
111. Storm, C.S., Kia, D.A., Almramhi, M., and Wood, N.W. (2020). Using Mendelian randomization to understand and develop treatments for neurodegenerative disease. Brain communications *2*, fcaa031.
112. Schmidt, A.F., Finan, C., Gordillo-Marañón, M., Asselbergs, F.W., Freitag, D.F., Patel, R.S., Tyl, B., Chopade, S., Faraway, R., and Zwierzyna, M. (2020). Genetic drug target validation using Mendelian randomisation. Nature communications *11*, 1-12.
113. Wishart, D.S., Knox, C., Guo, A.C., Shrivastava, S., Hassanali, M., Stothard, P., Chang, Z., and Woolsey, J. (2006). DrugBank: a comprehensive resource for in silico drug discovery and exploration. Nucleic acids research *34*, D668-D672.
114. Gaulton, A., Hersey, A., Nowotka, M., Bento, A.P., Chambers, J., Mendez, D., Mutowo, P., Atkinson, F., Bellis, L.J., and Cibrián-Uhalte, E. (2017). The ChEMBL database in 2017. Nucleic acids research *45*, D945-D954.
115. Feuk, L., Carson, A.R., and Scherer, S.W. (2006). Structural variation in the human genome. Nature Reviews Genetics *7*, 85-97.
116. Hood, L., and Rowen, L. (2013). The human genome project: big science transforms biology and medicine. Genome medicine *5*, 79.
117. Consortium, U.B.W.-G.S. (2025). Whole-genome sequencing of 490,640 UK Biobank participants. Nature.

118. Biobank, M.B.A.L.W.S.J.W.V.A.M.J.G.C.M.S., and 15, A.o.U.R.D.P.T.C.S.H.h.o.o.-.-.-W.X.h.o.o.-.-.-R.E.A. (2024). Genomic data in the all of us research program. Nature *627*, 340-346.
119. Logsdon, G.A., Vollger, M.R., and Eichler, E.E. (2020). Long-read human genome sequencing and its applications. Nature Reviews Genetics *21*, 597-614.
120. Chaisson, M.J., Sanders, A.D., Zhao, X., Malhotra, A., Porubsky, D., Rausch, T., Gardner, E.J., Rodriguez, O.L., Guo, L., and Collins, R.L. (2019). Multi-platform discovery of haplotype-resolved structural variation in human genomes. Nature communications *10*, 1784.
121. Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A.V., Mikheenko, A., Vollger, M.R., Altemose, N., Uralsky, L., and Gershman, A. (2022). The complete sequence of a human genome. Science *376*, 44-53.
122. Betschart, R.O., Thiéry, A., Aguilera-Garcia, D., Zoche, M., Moch, H., Twerenbold, R., Zeller, T., Blankenberg, S., and Ziegler, A. (2022). Comparison of calling pipelines for whole genome sequencing: an empirical study demonstrating the importance of mapping and alignment. Scientific Reports *12*, 21502.
123. Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv preprint arXiv:1303.3997.
124. Langmead, B., Wilks, C., Antonescu, V., and Charles, R. (2019). Scaling read aligners to hundreds of threads on general-purpose processors. Bioinformatics *35*, 421-432.
125. Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics *34*, 3094-3100.
126. Van der Auwera, G.A., and O'Connor, B.D. (2020). Genomics in the cloud: using Docker, GATK, and WDL in Terra (O'Reilly Media).
127. Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., and Davies, R.M. (2021). Twelve years of SAMtools and BCFtools. Gigascience *10*, giab008.
128. Behera, S., Catreux, S., Rossi, M., Truong, S., Huang, Z., Ruehle, M., Visvanath, A., Parnaby, G., Roddey, C., and Onuchic, V. (2025). Comprehensive genome analysis and variant detection at scale using DRAGEN. Nature Biotechnology *43*, 1177-1191.
129. Poplin, R., Chang, P.-C., Alexander, D., Schwartz, S., Colthurst, T., Ku, A., Newburger, D., Dijamco, J., Nguyen, N., and Afshar, P.T. (2018). A universal SNP and small-indel variant caller using deep neural networks. Nature biotechnology *36*, 983-987.
130. Babadi, M., Fu, J.M., Lee, S.K., Smirnov, A.N., Gauthier, L.D., Walker, M., Benjamin, D.I., Zhao, X., Karczewski, K.J., and Wong, I. (2023). GATK-gCNV enables the discovery of rare copy number variants from exome sequencing data. Nature genetics *55*, 1589-1597.
131. Gardner, E.J., Lam, V.K., Harris, D.N., Chuang, N.T., Scott, E.C., Pittard, W.S., Mills, R.E., Devine, S.E., and Consortium, G.P. (2017). The Mobile Element Locator Tool (MELT): population-scale mobile element discovery and biology. Genome research *27*, 1916-1929.

132. Chen, X., Schulz-Trieglaff, O., Shaw, R., Barnes, B., Schlesinger, F., Källberg, M., Cox, A.J., Kruglyak, S., and Saunders, C.T. (2016). Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. Bioinformatics *32*, 1220-1222.

133. Hujoel, M.L., Handsaker, R.E., Sherman, M.A., Kamitaki, N., Barton, A.R., Mukamel, R.E., Terao, C., McCarroll, S.A., and Loh, P.-R. (2024). Protein-altering variants at copy number-variable regions influence diverse human phenotypes. Nature Genetics *56*, 569-578.

134. Gupta, R., Kanai, M., Durham, T.J., Tsuo, K., McCoy, J.G., Kotrys, A.V., Zhou, W., Chinnery, P.F., Karczewski, K.J., and Calvo, S.E. (2023). Nuclear genetic control of mtDNA copy number and heteroplasmy in humans. Nature *620*, 839-848.

135. Wang, T., Antonacci-Fulton, L., Howe, K., Lawson, H.A., Lucas, J.K., Phillippy, A.M., Popejoy, A.B., Asri, M., Carson, C., and Chaisson, M.J. (2022). The Human Pangenome Project: a global resource to map genomic diversity. Nature *604*, 437-446.

136. Gong, B., Li, D., Łabaj, P.P., Pan, B., Novoradovskaya, N., Thierry-Mieg, D., Thierry-Mieg, J., Chen, G., Bergstrom Lucas, A., and LoCoco, J.S. (2024). Targeted DNA-seq and RNA-seq of Reference Samples with Short-read and Long-read Sequencing. Scientific Data *11*, 892.

137. Visscher, P.M., Wray, N.R., Zhang, Q., Sklar, P., McCarthy, M.I., Brown, M.A., and Yang, J. (2017). 10 years of GWAS discovery: biology, function, and translation. Am. J. Hum. Genet. *101*, 5-22.

138. Yang, J., Bakshi, A., Zhu, Z., Hemani, G., Vinkhuyzen, A.A., Lee, S.H., Robinson, M.R., Perry, J.R., Nolte, I.M., van Vliet-Ostaptchouk, J.V., et al. (2015). Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. Nat. Genet. *47*, 1114-1120.

139. Visscher, P.M., Yang, J., and Goddard, M.E. (2010). A commentary on 'common SNPs explain a large proportion of the heritability for human height'by Yang et al.(2010). Twin Res. Hum. Genet. *13*, 517-524.

140. Yang, J., Benyamin, B., McEvoy, B.P., Gordon, S., Henders, A.K., Nyholt, D.R., Madden, P.A., Heath, A.C., Martin, N.G., Montgomery, G.W., et al. (2010). Common SNPs explain a large proportion of the heritability for human height. Nat. Genet. *42*, 565-569.

141. Yang, J., Lee, S.H., Goddard, M.E., and Visscher, P.M. (2011). GCTA: a tool for genome-wide complex trait analysis. Am. J. Hum. Genet. *88*, 76-82.

142. Bulik-Sullivan, B.K., Loh, P.R., Finucane, H.K., Ripke, S., Yang, J., Schizophrenia Working Group of the Psychiatric Genomics, C., Patterson, N., Daly, M.J., Price, A.L., and Neale, B.M. (2015). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat. Genet. *47*, 291-295.

143. Lee, S.H., Ripke, S., Neale, B.M., Faraone, S.V., Purcell, S.M., Perlis, R.H., Mowry, B.J., Thapar, A., Goddard, M.E., Witte, J.S., et al. (2013). Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. Nat. Genet. *45*, 984-994.

144. Wray, N.R., Ripke, S., Mattheisen, M., Trzaskowski, M., Byrne, E.M., Abdellaoui, A., Adams, M.J., Agerbo, E., Air, T.M., and Andlauer, T.M. (2018). Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. Nat. Genet. *50*, 668-681.

145. Lee, J.J., Wedow, R., Okbay, A., Kong, E., Maghzian, O., Zacher, M., Nguyen-Viet, T.A., Bowers, P., Sidorenko, J., Karlsson Linner, R., et al. (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.

146. Escott-Price, V., and Hardy, J. (2022). Genome-wide association studies for Alzheimer's disease: bigger is not always better. Brain commun. *4*, fcac125.

147. Okbay, A., Wu, Y., Wang, N., Jayashankar, H., Bennett, M., Nehzati, S.M., Sidorenko, J., Kweon, H., Goldman, G., Gjorgjieva, T., et al. (2022). Polygenic prediction of educational attainment within and between families from genome-wide association analyses in 3 million individuals. Nat. Genet. *54*, 437-449.

148. de Vlaming, R., Okbay, A., Rietveld, C.A., Johannesson, M., Magnusson, P.K., Uitterlinden, A.G., van Rooij, F.J., Hofman, A., Groenen, P.J., and Thurik, A.R. (2017). Meta-GWAS Accuracy and Power (MetaGAP) calculator shows that hiding heritability is partially due to imperfect genetic correlations across studies. PLoS Genet. *13*, e1006495.

149. Dudbridge, F. (2013). Power and predictive accuracy of polygenic risk scores. PLoS Genet. *9*, e1003348.

150. Daetwyler, H.D., Villanueva, B., and Woolliams, J.A. (2008). Accuracy of predicting the genetic risk of disease using a genome-wide approach. PLoS ONE *3*, e3395.

151. Goddard, M.E., Hayes, B.J., and Meuwissen, T.H. (2011). Using the genomic relationship matrix to predict the accuracy of genomic selection. J. Anim. Breed. Genet. *128*, 409-421.

152. Hill, W., and Robertson, A. (1968). Linkage disequilibrium in finite populations. Theor. Appl. Genet. *38*, 226-231.

153. Fisher, R.A. (1919). XV.—The correlation between relatives on the supposition of Mendelian inheritance. Earth Environ. Sci. Trans. R. Soc. Edinb. *52*, 399-433.

154. Boyle, E.A., Li, Y.I., and Pritchard, J.K. (2017). An Expanded View of Complex Traits: From Polygenic to Omnigenic. Cell *169*, 1177-1186.

155. Lee, S.H., Wray, N.R., Goddard, M.E., and Visscher, P.M. (2011). Estimating missing heritability for disease from genome-wide association studies. Am. J. Hum. Genet. *88*, 294-305.

156. Visscher, P.M., Hemani, G., Vinkhuyzen, A.A., Chen, G.-B., Lee, S.H., Wray, N.R., Goddard, M.E., and Yang, J. (2014). Statistical power to detect genetic (co) variance of complex traits using SNP data in unrelated samples. PLoS Genet. *10*, e1004269.

157. Wray, N.R., Yang, J., Hayes, B.J., Price, A.L., Goddard, M.E., and Visscher, P.M. (2013). Pitfalls of predicting complex traits from SNPs. Nat. Rev. Genet. *14*, 507-515.

158. Howard, D.M., Adams, M.J., Clarke, T.-K., Hafferty, J.D., Gibson, J., Shirali, M., Coleman, J.R., Hagenaars, S.P., Ward, J., and Wigmore, E.M. (2019). Genome-wide meta-analysis of depression identifies 102 independent variants and highlights the importance of the prefrontal brain regions. Nat. Neurosci. *22*, 343-352.

159. Ni, G., Moser, G., Ripke, S., Neale, B.M., Corvin, A., Walters, J.T., Farh, K.-H., Holmans, P.A., Lee, P., and Bulik-Sullivan, B. (2018). Estimation of genetic correlation via linkage disequilibrium score regression and genomic restricted maximum likelihood. Am. J. Hum. Genet. *102*, 1185-1194.

160. Ni, G., Zeng, J., Revez, J.A., Wang, Y., Zheng, Z., Ge, T., Restuadi, R., Kiewa, J., Nyholt, D.R., and Coleman, J.R. (2021). A comparison of ten polygenic score methods for psychiatric disorders applied across multiple cohorts. Biol. Psychiatry *90*, 611-620.

161. Lloyd-Jones, L.R., Zeng, J., Sidorenko, J., Yengo, L., Moser, G., Kemper, K.E., Wang, H., Zheng, Z., Magi, R., Esko, T., et al. (2019). Improved polygenic prediction by Bayesian multiple regression on summary statistics. Nat. Commun. *10*, 5086.

162. Borenstein, M., Hedges, L.V., Higgins, J.P., and Rothstein, H.R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. Res. Synth. Methods *1*, 97-111.

163. Cai, N., Revez, J.A., Adams, M.J., Andlauer, T.F.M., Breen, G., Byrne, E.M., Clarke, T.K., Forstner, A.J., Grabe, H.J., Hamilton, S.P., et al. (2020). Minimal phenotyping yields genome-wide association signals of low specificity for major depression. Nat. Genet. *52*, 437-447.

164. Sullivan, P.F., Agrawal, A., Bulik, C.M., Andreassen, O.A., Børglum, A.D., Breen, G., Cichon, S., Edenberg, H.J., Faraone, S.V., and Gelernter, J. (2018). Psychiatric genomics: an update and an agenda. Am. J. Psychiatry *175*, 15-27.

165. Lloyd-Jones, L.R., Zeng, J., Sidorenko, J., Yengo, L., Moser, G., Kemper, K.E., Wang, H., Zheng, Z., Magi, R., Esko, T., et al. (2019). Improved polygenic prediction by Bayesian multiple regression on summary statistics. Nat. Commun. *10*, 5086.

166. Turley, P., Walters, R.K., Maghzian, O., Okbay, A., Lee, J.J., Fontana, M.A., Nguyen-Viet, T.A., Wedow, R., Zacher, M., and Furlotte, N.A. (2018). Multi-trait analysis of genome-wide association summary statistics using MTAG. Nat. Genet. *50*, 229-237.

167. Abdellaoui, A., Yengo, L., Verweij, K.J., and Visscher, P.M. (2023). 15 years of GWAS discovery: realizing the promise. The American Journal of Human Genetics *110*, 179-194.

168. Galarneau, G., Palmer, C.D., Sankaran, V.G., Orkin, S.H., Hirschhorn, J.N., and Lettre, G. (2010). Fine-mapping at three loci known to affect fetal hemoglobin levels explains additional genetic variation. Nature genetics *42*, 1049-1051.

169. Trynka, G., Hunt, K.A., Bockett, N.A., Romanos, J., Mistry, V., Szperl, A., Bakker, S.F., Bardella, M.T., Bhaw-Rosun, L., and Castillejo, G. (2011). Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. Nature genetics *43*, 1193-1201.

170. Yang, J., Ferreira, T., Morris, A.P., Medland, S.E., Consortium, G.I.o.A.T., Replication, D.G., Consortium, M.-a., Madden, P.A., Heath, A.C., Martin, N.G., and Montgomery, G.W. (2012). Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. Nature genetics *44*, 369-375.

171. Suzuki, K., Hatzikotoulas, K., Southam, L., Taylor, H.J., Yin, X., Lorenz, K.M., Mandla, R., Huerta-Chagoya, A., Melloni, G.E., and Kanoni, S. (2024). Genetic drivers of heterogeneity in type 2 diabetes pathophysiology. Nature *627*, 347-357.

172. Adams, M.J., Streit, F., Meng, X., Awasthi, S., Adey, B.N., Choi, K.W., Chundru, V.K., Coleman, J.R., Ferwerda, B., and Foo, J.C. (2025). Trans-ancestry genome-wide study of depression identifies 697 associations implicating cell types and pharmacotherapies. Cell *188*, 640-652. e649.

173. Martin, A.R., Kanai, M., Kamatani, Y., Okada, Y., Neale, B.M., and Daly, M.J. (2019). Clinical use of current polygenic risk scores may exacerbate health disparities. Nature genetics *51*, 584-591.

174. Purdue, M.P., Dutta, D., Machiela, M.J., Gorman, B.R., Winter, T., Okuhara, D., Cleland, S., Ferreiro-Iglesias, A., Scheet, P., and Liu, A. (2024). Multi-ancestry genome-wide association study of kidney cancer identifies 63 susceptibility regions. Nature genetics *56*, 809-818.

175. Peterson, R.E., Kuchenbaecker, K., Walters, R.K., Chen, C.-Y., Popejoy, A.B., Periyasamy, S., Lam, M., Iyegbe, C., Strawbridge, R.J., and Brick, L. (2019). Genome-wide association studies in ancestrally diverse populations: opportunities, methods, pitfalls, and recommendations. Cell *179*, 589-603.

176. Meng, X., Navoly, G., Giannakopoulou, O., Levey, D.F., Koller, D., Pathak, G.A., Koen, N., Lin, K., Adams, M.J., and Rentería, M.E. (2024). Multi-ancestry genome-wide association study of major depression aids locus discovery, fine mapping, gene prioritization and causal inference. Nature genetics *56*, 222-233.

177. Fu, J., Zhang, Q., Wang, J., Wang, M., Zhang, B., Zhu, W., Qiu, S., Geng, Z., Cui, G., and Yu, Y. (2024). Cross-ancestry genome-wide association studies of brain imaging phenotypes. Nature Genetics *56*, 1110-1120.

178. Thibord, F., Klarin, D., Brody, J.A., Chen, M.-H., Levin, M.G., Chasman, D.I., Goode, E.L., Hveem, K., Teder-Laving, M., and Martinez-Perez, A. (2022). Cross-ancestry investigation of venous thromboembolism genomic predictors. Circulation *146*, 1225-1242.

179. Byun, J., Han, Y., Li, Y., Xia, J., Long, E., Choi, J., Xiao, X., Zhu, M., Zhou, W., and Sun, R. (2022). Cross-ancestry genome-wide meta-analysis of 61,047 cases and 947,237 controls identifies new susceptibility loci contributing to lung cancer. Nature genetics *54*, 1167-1177.

180. Zhou, H., Kember, R.L., Deak, J.D., Xu, H., Toikumo, S., Yuan, K., Lind, P.A., Farajzadeh, L., Wang, L., and Hatoum, A.S. (2023). Multi-ancestry study of the genetics of problematic alcohol use in over 1 million individuals. Nature Medicine *29*, 3184-3192.

181. Budu-Aggrey, A., Kilanowski, A., Sobczyk, M.K., Team, a.R., Shringarpure, S.S., Mitchell, R., Reis, K., Reigo, A., Team, E.B.R., and Mägi, R. (2023). European and multi-ancestry genome-wide association meta-analysis of atopic dermatitis highlights importance of systemic immune regulation. Nature communications *14*, 6172.
182. Cui, R., Elzur, R.A., Kanai, M., Ulirsch, J.C., Weissbrod, O., Daly, M.J., Neale, B.M., Fan, Z., and Finucane, H.K. (2024). Improving fine-mapping by modeling infinitesimal effects. Nature genetics *56*, 162-169.
183. Teumer, A., Li, Y., Ghasemi, S., Prins, B.P., Wuttke, M., Hermle, T., Giri, A., Sieber, K.B., Qiu, C., and Kirsten, H. (2019). Genome-wide association meta-analyses and fine-mapping elucidate pathways influencing albuminuria. Nature communications *10*, 4130.
184. Marigorta, U.M., and Navarro, A. (2013). High trans-ethnic replicability of GWAS results implies common causal variants. PLoS genetics *9*, e1003566.
185. Hou, K., Ding, Y., Xu, Z., Wu, Y., Bhattacharya, A., Mester, R., Belbin, G.M., Buyske, S., Conti, D.V., and Darst, B.F. (2023). Causal effects on complex traits are similar for common variants across segments of different continental ancestries within admixed individuals. Nature genetics *55*, 549-558.
186. Hu, S., Ferreira, L.A., Shi, S., Hellenthal, G., Marchini, J., Lawson, D.J., and Myers, S.R. (2025). Fine-scale population structure and widespread conservation of genetic effect sizes between human groups across traits. Nature Genetics, 1-11.
187. Shi, H., Burch, K.S., Johnson, R., Freund, M.K., Kichaev, G., Mancuso, N., Manuel, A.M., Dong, N., and Pasaniuc, B. (2020). Localizing components of shared transethnic genetic architecture of complex traits from GWAS summary data. The American Journal of Human Genetics *106*, 805-817.
188. Galinsky, K.J., Reshef, Y.A., Finucane, H.K., Loh, P.R., Zaitlen, N., Patterson, N.J., Brown, B.C., and Price, A.L. (2019). Estimating cross-population genetic correlations of causal effect sizes. Genetic epidemiology *43*, 180-188.
189. Koyama, S., Liu, X., Koike, Y., Hikino, K., Koido, M., Li, W., Akaki, K., Tomizuka, K., Ito, S., and Otomo, N. (2024). Population-specific putative causal variants shape quantitative traits. Nature Genetics *56*, 2027-2035.
190. Hill, W.G., and Robertson, A. (1966). The effect of linkage on limits to artificial selection. Genetics Research *8*, 269-294.
191. Roze, D., and Barton, N.H. (2006). The Hill–Robertson effect and the evolution of recombination. Genetics *173*, 1793-1811.
192. Crismani, W., and Mercier, R. (2012). What limits meiotic crossovers? Cell Cycle *11*, 3527-3528.
193. Bergstrom, C.T., and Dugatkin, L.A. (2016). Evolution (WW Norton & Company).
194. Alkan, C., Sajjadian, S., and Eichler, E.E. (2011). Limitations of next-generation genome sequence assembly. Nature methods *8*, 61-65.

195. Bonett, D.G. (2008). Meta-analytic interval estimation for bivariate correlations. Psychological methods *13*, 173.
196. Li, A., Liu, S., Bakshi, A., Jiang, L., Chen, W., Zheng, Z., Sullivan, P.F., Visscher, P.M., Wray, N.R., and Yang, J. (2023). mBAT-combo: a more powerful test to detect gene-trait associations from GWAS data. The American Journal of Human Genetics *110*, 30-43.
197. Wang, Y., Guo, J., Ni, G., Yang, J., Visscher, P.M., and Yengo, L. (2020). Theoretical and empirical quantification of the accuracy of polygenic scores in ancestry divergent populations. Nature communications *11*, 3865.
198. Schaid, D.J., Chen, W., and Larson, N.B. (2018). From genome-wide associations to candidate causal variants by statistical fine-mapping. Nature Reviews Genetics *19*, 491-504.
199. van de Bunt, M., Cortes, A., Consortium, I., Brown, M.A., Morris, A.P., and McCarthy, M.I. (2015). Evaluating the performance of fine-mapping strategies at common variant GWAS loci. PLoS genetics *11*, e1005535.
200. Berisa, T., and Pickrell, J.K. (2015). Approximately independent linkage disequilibrium blocks in human populations. Bioinformatics *32*, 283.
201. Graham, S.E., Clarke, S.L., Wu, K.-H.H., Kanoni, S., Zajac, G.J., Ramdas, S., Surakka, I., Ntalla, I., Vedantam, S., and Winkler, T.W. (2021). The power of genetic diversity in genome-wide association studies of lipids. Nature *600*, 675-679.
202. Gao, B., and Zhou, X. (2024). MESuSiE enables scalable and powerful multi-ancestry fine-mapping of causal variants in genome-wide association studies. Nature genetics *56*, 170-179.
203. Miretti, M.M., Walsh, E.C., Ke, X., Delgado, M., Griffiths, M., Hunt, S., Morrison, J., Whittaker, P., Lander, E.S., and Cardon, L.R. (2005). A high-resolution linkage-disequilibrium map of the human major histocompatibility complex and first generation of tag single-nucleotide polymorphisms. The American Journal of Human Genetics *76*, 634-646.
204. De Bakker, P.I., McVean, G., Sabeti, P.C., Miretti, M.M., Green, T., Marchini, J., Ke, X., Monsuur, A.J., Whittaker, P., and Delgado, M. (2006). A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC. Nature genetics *38*, 1166-1172.
205. Dilthey, A., Cox, C., Iqbal, Z., Nelson, M.R., and McVean, G. (2015). Improved genome inference in the MHC using a population reference graph. Nature genetics *47*, 682-688.
206. Palmer, D.S., Zhou, W., Abbott, L., Wigdor, E.M., Baya, N., Churchhouse, C., Seed, C., Poterba, T., King, D., and Kanai, M. (2023). Analysis of genetic dominance in the UK Biobank. Science *379*, 1341-1348.
207. Phillips, P.C. (2008). Epistasis—the essential role of gene interactions in the structure and evolution of genetic systems. Nature Reviews Genetics *9*, 855-867.
208. Polderman, T.J., Benyamin, B., De Leeuw, C.A., Sullivan, P.F., Van Bochoven, A., Visscher, P.M., and Posthuma, D. (2015). Meta-analysis of the heritability of human traits based on fifty years of twin studies. Nature genetics *47*, 702-709.

209. Hill, W.G., Goddard, M.E., and Visscher, P.M. (2008). Data and theory point to mainly additive genetic variance for complex traits. PLoS genetics *4*, e1000008.
210. Consortium, G.P. (2015). A global reference for human genetic variation. Nature *526*, 68.
211. Anderson, S., Bankier, A.T., Barrell, B.G., de Bruijn, M.H., Coulson, A.R., Drouin, J., Eperon, I.C., Nierlich, D.P., Roe, B.A., and Sanger, F. (1981). Sequence and organization of the human mitochondrial genome. Nature *290*, 457-465.
212. Liu, Y.J., Sulc, J., and Auwerx, J. (2025). Mitochondrial genetics, signalling and stress responses. Nature Cell Biology, 1-15.
213. Andreazza, A.C., Barros, L.F., Behnke, A., Ben-Shachar, D., Berretta, S., Chouinard, V.-A., Do, K., Edwin Thanarajah, S., Ehrenreich, H., and Falkai, P. (2025). Brain and body energy metabolism and potential for treatment of psychiatric disorders. Nature Mental Health, 1-9.
214. Chung, J.K., Lee, S.Y., Park, M., Joo, E.-J., and Kim, S.A. (2019). Investigation of mitochondrial DNA copy number in patients with major depressive disorder. Psychiatry research *282*, 112616.
215. Cai, N., Li, Y., Chang, S., Liang, J., Lin, C., Zhang, X., Liang, L., Hu, J., Chan, W., and Kendler, K.S. (2015). Genetic control over mtDNA and its relationship to major depressive disorder. Current Biology *25*, 3170-3177.
216. Meloni, A., Paribello, P., Pinna, M., Contu, M., Ardau, R., Chillotti, C., Congiu, D., Gennarelli, M., Minelli, A., and Buson, L. (2025). Mitochondrial DNA copy number is significantly increased in bipolar disorder patients and is correlated with long-term lithium treatment. European Neuropsychopharmacology *91*, 37-44.
217. Wang, D., Li, Z., Liu, W., Zhou, J., Ma, X., Tang, J., and Chen, X. (2018). Differential mitochondrial DNA copy number in three mood states of bipolar disorder. BMC psychiatry *18*, 149.
218. Chung, J.K., Ahn, Y.M., Kim, S.A., and Joo, E.-J. (2022). Differences in mitochondrial DNA copy number between patients with bipolar I and II disorders. Journal of Psychiatric Research *145*, 325-333.
219. He, Y., Tang, J., Li, Z., Li, H., Liao, Y., Tang, Y., Tan, L., Chen, J., Xia, K., and Chen, X. (2014). Leukocyte mitochondrial DNA copy number in blood is not associated with major depressive disorder in young adults. PloS one *9*, e96869.
220. Lukmanji, A., Williams, J.V., Bulloch, A.G., Bhattarai, A., and Patten, S.B. (2019). Seasonal variation in symptoms of depression: A Canadian population based study. Journal of Affective Disorders *255*, 142-149.
221. Melrose, S. (2015). Seasonal affective disorder: an overview of assessment and treatment approaches. Depression research and treatment *2015*, 178564.
222. Hasler, B.P., Allen, J.J., Sbarra, D.A., Bootzin, R.R., and Bernert, R.A. (2010). Morningness–eveningness and depression: Preliminary evidence for the role of the behavioral activation system and positive affect. Psychiatry research *176*, 166-173.

223. Shafiee, M., Tayefi, M., Hassanian, S.M., Ghaneifar, Z., Parizadeh, M.R., Avan, A., Rahmani, F., Khorasanchi, Z., Azarpajouh, M.R., and Safarian, H. (2017). Depression and anxiety symptoms are associated with white blood cell count and red cell distribution width: a sex-stratified analysis in a population-based study. Psychoneuroendocrinology *84*, 101-108.

224. Foley, É.M., Parkinson, J.T., Mitchell, R.E., Turner, L., and Khandaker, G.M. (2023). Peripheral blood cellular immunophenotype in depression: a systematic review and meta-analysis. Molecular psychiatry *28*, 1004-1019.

225. Longchamps, R.J., Castellani, C.A., Yang, S.Y., Newcomb, C.E., Sumpter, J.A., Lane, J., Grove, M.L., Guallar, E., Pankratz, N., and Taylor, K.D. (2020). Evaluation of mitochondrial DNA copy number estimation techniques. PloS one *15*, e0228166.

226. Battle, S.L., Puiu, D., Group, T.m.W., Verlouw, J., Broer, L., Boerwinkle, E., Taylor, K.D., Rotter, J.I., Rich, S.S., and Grove, M.L. (2022). A bioinformatics pipeline for estimating mitochondrial DNA copy number and heteroplasmy levels from whole genome sequencing data. NAR Genomics and Bioinformatics *4*, lqac034.

227. Wang, L., Han, J., Fearnley, L.G., Milton, M., Rafehi, H., Reid, J., Gerring, Z.F., Masaldan, S., Lang, T., and Speed, T.P. (2024). Peripheral immune cell abundance differences link blood mitochondrial DNA copy number and Parkinson's disease. npj Parkinson's Disease *10*, 219.

228. Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W.J., Mattick, J.S., and Haussler, D. (2004). Ultraconserved elements in the human genome. Science *304*, 1321-1325.

229. Snetkova, V., Pennacchio, L.A., Visel, A., and Dickel, D.E. (2022). Perfect and imperfect views of ultraconserved sequences. Nature Reviews Genetics *23*, 182-194.

230. Byrska-Bishop, M., Evani, U.S., Zhao, X., Basile, A.O., Abel, H.J., Regier, A.A., Corvelo, A., Clarke, W.E., Musunuri, R., and Nagulapalli, K. (2022). High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. Cell *185*, 3426-3440. e3419.

231. Schloissnig, S., Pani, S., Ebler, J., Hain, C., Tsapalou, V., Söylev, A., Hüther, P., Ashraf, H., Prodanov, T., and Asparuhova, M. (2025). Structural variation in 1,019 diverse humans based on long-read sequencing. Nature, 1-11.

232. Consortium, U.B.W.-G.S. (2025). Whole-genome sequencing of 490,640 UK Biobank participants. Nature *645*, 692.

233. Lander, E.S., and Waterman, M.S. (1988). Genomic mapping by fingerprinting random clones: a mathematical analysis. Genomics *2*, 231-239.

234. Kuhn, R.M., Haussler, D., and Kent, W.J. (2013). The UCSC genome browser and associated tools. Briefings in bioinformatics *14*, 144-161.

235. Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841-842.

236. Karimzadeh, M., Ernst, C., Kundaje, A., and Hoffman, M.M. (2018). Umap and Bismap: quantifying genome and methylome mappability. Nucleic acids research *46*, e120-e120.

237. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Subgroup, G.P.D.P. (2009). The sequence alignment/map format and SAMtools. bioinformatics *25*, 2078-2079.

238. Yang, J., Lee, S.H., Goddard, M.E., and Visscher, P.M. (2011). GCTA: a tool for genome-wide complex trait analysis. The American Journal of Human Genetics *88*, 76-82.

239. Cai, N., Revez, J.A., Adams, M.J., Andlauer, T.F., Breen, G., Byrne, E.M., Clarke, T.-K., Forstner, A.J., Grabe, H.J., and Hamilton, S.P. (2020). Minimal phenotyping yields genome-wide association signals of low specificity for major depression. Nature genetics *52*, 437-447.

240. Bakshi, A., Zhu, Z., Vinkhuyzen, A.A., Hill, W.D., McRae, A.F., Visscher, P.M., and Yang, J. (2016). Fast set-based association analysis using summary data from GWAS identifies novel gene loci for human complex traits. Scientific reports *6*, 32894.

241. Wang, Q., Li, M., Wu, T., Zhan, L., Li, L., Chen, M., Xie, W., Xie, Z., Hu, E., and Xu, S. (2022). Exploring epigenomic datasets by ChIPseeker. Current protocols *2*, e585.

242. Yu, G., and He, Q.-Y. (2016). ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization. Molecular BioSystems *12*, 477-479.

243. Bulik-Sullivan, B.K., Loh, P.-R., Finucane, H.K., Ripke, S., Yang, J., Patterson, N., Daly, M.J., Price, A.L., and Neale, B.M. (2015). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nature genetics *47*, 291-295.

244. Li, A., Lin, T., Walker, A., Tan, X., Zhao, R., Yao, S., Sullivan, P.F., Hjerling-Leffler, J., Wray, N.R., and Zeng, J. (2025). Benchmarking methods integrating GWAS and single-cell transcriptomic data for mapping trait-cell type associations. medRxiv, 2025.2005. 2024.25328275.

245. Kamath, T., Abdulraouf, A., Burris, S., Langlieb, J., Gazestani, V., Nadaf, N.M., Balderrama, K., Vanderburg, C., and Macosko, E.Z. (2022). Single-cell genomic profiling of human dopamine neurons identifies a population that selectively degenerates in Parkinson's disease. Nature neuroscience *25*, 588-595.

246. Jardine, L., Webb, S., Goh, I., Quiroga Londoño, M., Reynolds, G., Mather, M., Olabi, B., Stephenson, E., Botting, R.A., and Horsfall, D. (2021). Blood and immune development in human fetal bone marrow and Down syndrome. Nature *598*, 327-331.

247. Aviv, R., Teichmann, S.A., Lander, E.S., Ido, A., Christophe, B., Ewan, B., Bernd, B., Campbell, P., Piero, C., and Menna, C. (2017). The human cell atlas. Elife *6*.

248. Kim, H.J., Wang, K., Chen, C., Lin, Y., Tam, P.P., Lin, D.M., Yang, J.Y., and Yang, P. (2021). Uncovering cell identity through differential stability with Cepo. Nature computational science *1*, 784-790.

249. Zhang, M.J., Hou, K., Dey, K.K., Sakaue, S., Jagadeesh, K.A., Weinand, K., Taychameekiatchai, A., Rao, P., Pisco, A.O., and Zou, J. (2022). Polygenic enrichment distinguishes disease associations of individual cells in single-cell RNA-seq data. Nature genetics *54*, 1572-1580.

250. De Leeuw, C.A., Mooij, J.M., Heskes, T., and Posthuma, D. (2015). MAGMA: generalized gene-set analysis of GWAS data. PLoS computational biology *11*, e1004219.

251. Benjamini, Y., and Speed, T.P. (2012). Summarizing and correcting the GC content bias in high-throughput sequencing. Nucleic acids research *40*, e72-e72.
252. Koller, A., Fazzini, F., Lamina, C., Rantner, B., Kollerits, B., Stadler, M., Klein-Weigel, P., Fraedrich, G., and Kronenberg, F. (2020). Mitochondrial DNA copy number is associated with all-cause mortality and cardiovascular events in patients with peripheral arterial disease. Journal of internal medicine *287*, 569-579.
253. Wang, H., He, Y., Sun, Z., Ren, S., Liu, M., Wang, G., and Yang, J. (2022). Microglia in depression: an overview of microglia in the pathogenesis and treatment of depression. Journal of neuroinflammation *19*, 132.
254. Wang, Y.-H., Fu, C.-L., Chen, L.-B., Zhang, C.-Y., Chen, J.-S., Zhang, Q.-M., Liang, Y., Yang, R.-L., Li, Y., and Zhang, Y.-N. (2025). Microglia contribute to bipolar depression through Serinc2-dependent phospholipid synthesis. Proceedings of the National Academy of Sciences *122*, e2500116122.
255. Antonarakis, S.E. (2022). Short arms of human acrocentric chromosomes and the completion of the human genome sequence. Genome Research *32*, 599-607.
256. Starke, H., Seidel, J., Henn, W., Reichardt, S., Volleth, M., Stumm, M., Behrend, C., Sandig, K.R., Kelbova, C., and Senger, G. (2002). Homologous sequences at human chromosome 9 bands p12 and q13-21.1 are involved in different patterns of pericentric rearrangements. European Journal of Human Genetics *10*, 790-800.
257. Rath, S.P., Gupta, R., Todres, E., Wang, H., Jourdain, A.A., Ardlie, K.G., Calvo, S.E., and Mootha, V.K. (2024). Mitochondrial genome copy number variation across tissues in mice and humans. Proceedings of the National Academy of Sciences *121*, e2402291121.
258. Klein, R.J., Zeiss, C., Chew, E.Y., Tsai, J.-Y., Sackler, R.S., Haynes, C., Henning, A.K., SanGiovanni, J.P., Mane, S.M., and Mayne, S.T. (2005). Complement factor H polymorphism in age-related macular degeneration. Science *308*, 385-389.
259. Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., and O'Connell, J. (2018). The UK Biobank resource with deep phenotyping and genomic data. Nature *562*, 203-209.
260. Li, Y., Xiao, J., Ming, J., Zeng, Y., and Cai, M. (2025). Funmap: integrating high-dimensional functional annotations to improve fine-mapping. Bioinformatics *41*, btaf017.
261. Rummel, C.K., Gagliardi, M., Ahmad, R., Herholt, A., Jimenez-Barron, L., Murek, V., Weigert, L., Hausruckinger, A., Maidl, S., and Hauger, B. (2023). Massively parallel functional dissection of schizophrenia-associated noncoding genetic variants. Cell *186*, 5165-5182. e5133.
262. Jaiswal, S., and Ebert, B.L. (2019). Clonal hematopoiesis in human aging and disease. Science *366*, eaan4673.
263. Chapman, C.D., Benedict, C., and Schiöth, H.B. (2018). Experimenter gender and replicability in science. Science advances *4*, e1701427.

264. Silver, D., Huang, A., Maddison, C.J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., and Lanctot, M. (2016). Mastering the game of Go with deep neural networks and tree search. nature *529*, 484-489.

265. Ahituv, N., Zhu, Y., Visel, A., Holt, A., Afzal, V., Pennacchio, L.A., and Rubin, E.M. (2007). Deletion of ultraconserved elements yields viable mice. PLoS biology *5*, e234.

266. Dumanski, J.P., Rasi, C., Lönn, M., Davies, H., Ingelsson, M., Giedraitis, V., Lannfelt, L., Magnusson, P.K., Lindgren, C.M., and Morris, A.P. (2015). Smoking is associated with mosaic loss of chromosome Y. Science *347*, 81-83.

267. Avsec, Ž., Latysheva, N., Cheng, J., Novati, G., Taylor, K.R., Ward, T., Bycroft, C., Nicolaisen, L., Arvaniti, E., and Pan, J. (2025). AlphaGenome: advancing regulatory variant effect prediction with a unified DNA sequence model. bioRxiv, 2025.2006. 2025.661532.