Methods for large-scale genome-wide association studies

Georgios Kalantzis

St Edmund Hall

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Michaelmas 2022

Dedicated to my father,
whose departure serves as light on my quest through genetics.

Acknowledgements

I would like to thank my supervisor, Pier Palamara, for giving me the opportunity to work in the exciting field of statistical genetics. His guidance and inspiration throughout the last five years has been invaluable. During my studies I had the chance to collaborate with several people, including Matthew Page, Ali Pazokitoroudi, Sriram Sankararaman, Han Chen, and Sasha Gusev, to which all I am grateful. Similarly, I also thank Robbie Davies and Noah Zaitlen for agreeing to be my DPhil assessors.

Next, I would like to thank SABS, my graduate program; my studies were supported by the EPSRC and MRC grant EP/L016044/1. The computational aspects of this research were further supported by the Wellcome Trust Core Award Grant Number 203141/Z/16/Z and the NIHR Oxford BRMC. I am also thankful to Suellen Littleton, my college advisor, and to the Department of Statistics for providing the most pleasant working space and fantastic support during the lock-downs.

I also thank everyone else in the Palamara Lab for creating a delightful, stimulating, and frequently humorous atmosphere. In particular, I thank Arni, Jaro, and Juba for providing feedback on the thesis, and Hrushi, Arjun, and Brian for sharing or reviewing code. I am grateful to Luis, Miriam, and Sinan, my writing buddies, for helping me complete this thesis and making writing enjoyable. I would like to thank the rest of my Oxford friends, as well as my rowing buddies at the SEH Boat Club for making Oxford the most exciting place to live. I am grateful to my good non-Oxford friends who helped me reach at this stage through various ways.

Finally, I want to express my gratitude to my family for their unlimited love and encouragement, always helping me stay focused and be ambitious. This extends to Antonia, for whom there are no words to describe how lucky I am to have her by my side. Her patience and support helped me get through and be brave enough for any challenge.

Abstract

Genome-wide association studies (GWAS) have led to the identification of thousands of associations between genetic polymorphisms and complex traits or diseases, facilitating several downstream applications such as genetic risk prediction and drug target prioritisation. Biobanks containing extensive genetic and phenotypic data continue to grow, creating new opportunities for the study of complex traits, such as the analysis of rare genomic variation across multiple populations. These opportunities are coupled with computational challenges, creating the need for the development of novel methodology.

This thesis develops computational tools to facilitate large-scale association studies of rare and common variation. First, we develop methods to improve the analysis of ultra-rare variants, leveraging the sharing of identical-by-descent (IBD) genomic regions within large biobanks. We compare $\sim$ 400k genotyped UK Biobank (UKBB) samples with 50k exome-sequenced samples and devise a score that quantifies the extent to which a genotyped individual shares IBD segments with carriers of rare loss-of-function mutations. Our approach detects several associations and replicates 11/14 loci of a pilot exome sequencing study. Second, we develop a linear mixed model framework, `FMA`, that builds on previous techniques and is suitable for scalable and robust association testing. We benchmark `FMA` and several state-of-the-art approaches using synthetic and UKBB data, evaluating computational performance, statistical power, and robustness to known confounders, such as cryptic relatedness and population stratification. Finally, we integrate `FMA` with recently developed methods for genealogical analysis of complex traits, enabling it to perform scalable genealogy-based estimation of narrow-sense heritability and association.

Contents

Appendices

List of Figures

List of Tables

List of Abbreviations and Notation

ARG	Ancestral Recombination Graph
GRM	Genetic Relatedness Matrix
GWAS	Genome-Wide Association Study
LD	Linkage Disequilibrium
LMM	Linear Mixed Model
LOCO	Leave One Chromosome Out
MAF	Minor Allele Frequency
PCA	Principal Components Analysis
PCs	Top eigenvectors from PCA on the genotypes matrix, also referred to as top components of ancestry
SNP	Single Nucleotide Polymorphism
WES	Whole Exome Sequencing
WGS	Whole Genome Sequencing
$\mathbf{A}$	matrix of numbers
$\mathbf{b}$	vector
c	scalar
$\mathbf{I}_n$	identity matrix of size $n \times n$
$\mathbf{x}^\top\mathbf{y}$	dot product of two vectors

1 Introduction

The last decades have witnessed a dramatic decrease in the cost of DNA sequencing. During the Human Genome Project [1], which took place roughly 22 years ago, sequencing an individual's genetic code required large research teams and significant costs. Sequencing a human genome today is a fast and highly automated process requiring only a few hundred dollars [1]. In addition to whole genome sequencing (WGS), a wide range of technologies have enabled reading an informative subset of the genetic code, such as whole exome sequencing (WES) or array genotyping of single nucleotide polymorphisms (SNP) [2, 3]. These advances have enabled building rich data sets comprising genetic, medical, and behavioural information for thousands of people, such as the UK Biobank (UKBB) [4], PAGE [5], or ATLAS [6].

Statistical and computational methods can be applied to these data sets to extract knowledge about natural selection, the human dispersal across the globe, or recent demographic events [7–11]. Genome-wide association studies (GWAS) are a widely adopted experimental design used to systematically detect associations between genomic variants and heritable traits or diseases [12–18]. In the era of broad availability of SNP array data, GWAS have flourished and opened the way for several downstream biomedical applications, including risk prediction, understanding of

[1] https://ourworldindata.org/grapher/cost-of-sequencing-a-full-human-genome

disease aetiology, and faster drug development [19–23]. The field of GWAS is still growing with recent studies comprising millions of individuals [24, 25].

Most recent biobanks are rich sources of genotypic and phenotypic information which can assist in the detection of novel associations. However, a larger biobank will naturally contain more sample structure, such as cryptic relatedness or population structure, which are known confounders in association studies [26, 27]. To prevent that, studies often exclude related individuals or whole sub-populations at the cost of statistical power. Furthermore, methods that were efficient in earlier data sets might struggle with larger sizes or multiple phenotypes due to large asymptotic costs or inefficient implementations. As new large and diverse biobanks emerge [5, 6], association studies pose new statistical and computational challenges.

To date, most GWAS have detected associations with variants of high ($\geq 5\%$) or low (from 0.5% to 5%) frequencies, whereas association with rare variants ($< 0.5\%$) is still challenging [14, 20, 28–30]. This is primarily due to the small volumes of whole genome sequencing that is currently available compared to SNP array data. Statistical imputation of genotyped samples from a sequenced reference panel offers one route to analyse lower frequency variants [31, 32]. However, because low frequency variants tend to be population specific [33], genotype imputation is only accurate when the target sample is sufficiently close to the reference panel in terms of ancestry. In addition, association studies have predominantly focused on European samples, complicating complex trait analyses in under-represented populations [34, 35]. Addressing these issues will require the collection of larger and more diverse sequencing panels, as well as the development of novel computational methods.

There is currently a broad set of methods for association which can be used to test single variants or whole genomic regions, such as genes. Some of the most recent examples include `BOLT-LMM` [36], `SAIGE` [37], `fastGWA` [38], and `REGENIE` [39]. Methods that enable gene-based testing, often specialised for rare variation, include `SKAT` [40], `SAIGE-Gene` [41] and several others [42–45]. Depending on the specific association task being considered, each method presents advantages as well as disadvantages in terms of model specification, statistical power, robustness to confounding, or scalability. For example, BOLT-LMM, which usually achieves state-of-the-art statistical power on quantitative traits [36, 46], is less computationally efficient than REGENIE or fastGWA, and suffers from high type I errors when handling unbalanced case/control phenotypes [37, 39]. The development of scalable, powerful, and robust association methodology therefore remains an active area of research.

This thesis attempts to address some of the challenges raised above by developing strategies for scalable and robust association of rare and common variants.

1.1 Thesis overview

The next chapters are structured as follows. First, in Chapter 2, I provide an overview of the relevant background and related work for this thesis. Next, in Chapter 3, I describe how shared identity-by-descent segments can be utilised to implicitly impute rare likely-causal variants and detect association with complex traits. This analysis is part of published collaborative work [11] which I presented as a platform talk at the *American Society of Human Genetics* 2020 annual meeting, where I received a *Charles J. Epstein Trainee Award for Excellence in Human Genetics Research* (semifinalist).

Chapter 4 describes a framework for scalable mixed-model association of quantitative phenotypes, called FMA. I benchmark FMA as well as state-of-the-art GWAS methods through an extensive set of simulations using UKBB genotypes, empirically verifying key LMM properties [47]. I then apply each method to 20 real quantitative phenotypes using samples of varying size and up to 38.5 million imputed genotypes. In Chapter 5, I explore several modifications to the FMA algorithm that may be used to decrease the computational cost of model fitting for larger sample sizes.

Next, in Chapter 6, building on the recent work of Zhang et al. [48], I introduce efficient techniques for association and heritability estimation using ancestral recombination graphs, which may be used to facilitate complex trait analyses in data sets where the availability of sequencing data is limited. Finally, in Chapter 7, I draw conclusions and suggest directions for future research, and include a set of supplementary figures and tables in the appendix.

2
Background and related work

Contents

This chapter discusses background theory and is organised as a brief review of the most relevant work, while highlighting any key concepts that appear throughout the thesis. My goal is to mainly help the reader become familiar with the context of the chapters that follow, rather than describing all the theory thoroughly. The discussion will mostly focus on complex quantitative traits with a brief note on case/control phenotypes.

2.1 Fundamentals of GWAS

Complex heritable traits are often well modelled using a linear model as follows. Let N, M denote the number of samples and genetic variants respectively. Assume that $\mathbf{y}$ is the vector measuring the phenotype for all samples, $\mathbf{X}$ is the $N \times M$ genotypes matrix, and $\mathbf{e}$ is a vector representing noise (environmental contribution to phenotype), assumed to be normally distributed. The phenotype vector and the columns of $\mathbf{X}$ are assumed to be standardized. Then, under the additive model of genetic effects, the trait is assumed to be in a linear relationship with the genotype and the environment described by:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}, \tag{2.1}$$

where $\boldsymbol{\beta}$ is the vector of the M effects. However, in genetic studies $N \ll M$, thus we cannot solve the full regression problem and obtain an ordinary least squares estimator for $\boldsymbol{\beta}$. A simple solution to this issue is to perform univariate regression by only considering the i-th column of the genotype matrix $\mathbf{X}$ [4, 12, 15, 17, 49, 50], in which case we can estimate the effect size as

$$\hat{\beta}_i = \mathbf{x}_i^\top \mathbf{y} / (\mathbf{x}_i^\top \mathbf{x}_i). \tag{2.2}$$

The corresponding test statistic has a similar form and, therefore, a genome-wide analysis with linear regression has cost linear in data size, or $O(NM)$.

This approach however is susceptible to **population stratification**, a common confounder in association studies which creates considerable difficulties. In this context, population stratification refers to the phenomenon where allele frequencies are different between cases and controls – or otherwise stratified with the phenotype [51]. When the sample is structured so that differences in frequencies coincide with phenotypic differences, a simple regression model would find spurious signals of association. Many techniques to deal with this phenomenon have been proposed, but there is evidence that residual stratification can still confound genetic studies [10, 52, 53] and might differentially affect the analysis of rare and common variants [33].

A widely adopted solution to control for population stratification is to condition on **principal components** of ancestry [51, 54]. In particular, the top eigenvectors obtained from the PCA of the genotype matrix $\mathbf{X}$ are highly correlated with ancestry

[55], and resemble geographic clines [7, 21]. These vectors can be encoded in a matrix of fixed effects $\mathbf{W}$ to be used in the linear model:

$$\mathbf{y} = \beta_i \mathbf{x}_i + \mathbf{W}\boldsymbol{\gamma} + \mathbf{e}. \tag{2.3}$$

In practice, additional environmental variables – like age, sex, or smoking status – are often used as covariates and included in the matrix $\mathbf{W}$. Detecting an association then involves testing the null hypothesis $H_0 : \beta_i = 0$ for a candidate SNP i. To that end, we first remove any effects of the covariates by multiplying Eq. (2.3) with the projection matrix $\mathbf{P} = \mathbf{I}_N - \mathbf{W}(\mathbf{W}^\top \mathbf{W})^{-1}\mathbf{W}^\top$. This matrix has the property that $\mathbf{PW} = \mathbf{0}$, and the model thus takes the form $\tilde{\mathbf{y}} = \beta_i \tilde{\mathbf{x}}_i + \tilde{\mathbf{e}}$, where $\tilde{\mathbf{y}}$ and $\tilde{\mathbf{x}}_i$ are the residualised phenotypes and genotypes respectively. To keep the notation simple, however, I will simply write $\mathbf{y}$ or $\mathbf{x}_i$ and assume that the covariates, where necessary, have been regressed out beforehand and thus $\mathbf{y}$ or $\mathbf{x}_i$ are mean-centered.

An alternative approach for conditioning on covariates is to regress their effects out only from the phenotypes, and thus enable simpler calculations. This yields similar results in many scenarios, as genotypes are typically less affected by this transformation, but may lead to biased association statistics (often conservative). For instance, this is the default approach in the `GCTA` software suite [38, 56] and Jiang et al. [45] have empirically shown that the differences with exact conditioning are minimal.

2.2 Linear Mixed Models

Cryptic relatedness, which refers to the presence of an unknown subset of individuals who are close genetic relatives [57, 58], might also result in spurious associations. Cryptic relatedness is likely to be present in large biobanks which inevitably contain members of the same family. False positive associations due to relatedness can be avoided by detecting and excluding closely related individuals, at the cost of reduced statistical power. For instance, the UKBB contains roughly 410k white British individuals (as defined by Bycroft et al. [4]) among which thousands of pairs of related individuals have been detected, excluding which results in a 20% smaller sample [4].

A **linear mixed model** (LMM) provides an alterantive route to excluding related samples. In LMMs, the model 2.3 is modified to include a random variable that captures sample structure [26, 36, 38, 47, 59–63]. Such a model is often expressed as

$$\mathbf{y} = \beta_i \mathbf{x}_i + \mathbf{g} + \mathbf{e}, \tag{2.4}$$

where β_i and $\mathbf{e}$ are as before (Equation 2.3), and $\mathbf{g}$ is the genetic random effect. In particular, $\mathbf{g}$ is assumed to follow a multivariate normal distribution with mean zero and covariance $\sigma_g^2\mathbf{K}$ (i.e. $\mathcal{N}_N(\mathbf{0}, \sigma_g^2\mathbf{K})$), where σ_g^2 captures genetic variance and $\mathbf{K}$ denotes the $N \times N$ pedigree or **genetic relatedness matrix** (GRM); $K_{i,j}$ quantifies the genome shared between individuals i and j (often called kinship coefficient). Environmental effects are also assumed to be independent and identically distributed as $\mathbf{e} \sim \mathcal{N}_N(\mathbf{0}, \sigma_e^2\mathbf{I}_N)$, where σ_e^2 captures environmental variance. Assuming that $\mathbf{g}$ and $\mathbf{e}$ are uncorrelated, the phenotypic variance is described by $\mathbf{V} = \text{cov}(\mathbf{y}) = \sigma_g^2\mathbf{K} + \sigma_e^2\mathbf{I}_N$. Therefore, an LMM uses the covariance matrix $\mathbf{V}$ to account for population stratification and relatedness, without the need of excluding related samples or including principal components as fixed effects in the model [47, 64] (see empirical results in Chapter 4). Testing for H_0 with an LMM is achieved by looking at the magnitudes of the test statistic and effect size [65–67]

$$\chi^2_{\text{LMM}} = \frac{(\mathbf{x}_{\text{test}}^\top \mathbf{V}^{-1}\mathbf{y})^2}{\mathbf{x}_{\text{test}}^\top \mathbf{V}^{-1}\mathbf{x}_{\text{test}}}, \ \beta_{\text{test}} = \frac{\mathbf{x}_{\text{test}}^\top \mathbf{V}^{-1}\mathbf{y}}{\mathbf{x}_{\text{test}}^\top \mathbf{V}^{-1}\mathbf{x}_{\text{test}}}, \tag{2.5}$$

where χ^2_{LMM} is approximately distributed as χ^2 with one degree of freedom.

A second LMM formulation which more closely reflects the polygenic architecture of complex traits is often used in genetic studies. We assume an infinitesimal model [68], where all genomic variants have a (usually small) phenotypic contribution according to a Gaussian distribution. This model takes the form

$$\mathbf{y} = \beta_i \mathbf{x}_i + \mathbf{X}\mathbf{b} + \mathbf{e}. \tag{2.6}$$

where $\mathbf{b} \sim \mathcal{N}_M(\mathbf{0}, \frac{\sigma_g^2}{M}\mathbf{I}_M)$ and $\mathbf{e}$ is as before. The prior that we now place on the effect sizes enables to perform a genome-wide joint regression [36, 56]. In this case the variance-covariance is $\mathbf{V} = \frac{\sigma_g^2}{M}\mathbf{X}\mathbf{X}^\top + \sigma_e^2\mathbf{I}_N$. In practice, $\mathbf{K}$ for model 2.4 is estimated by $\frac{1}{M}\mathbf{X}\mathbf{X}^\top$ (which is known as the empirical kinship matrix), in which case the two formulations are equivalent [36, 56].

Based on the previous formulation and the mixed-model equations [67, 69], we can estimate the vector of phenotypic effects with the **best linear unbiased predictor** (BLUP):

$$\hat{\mathbf{b}} = \frac{\sigma_g^2}{M}\mathbf{X}^\top\mathbf{V}^{-1}\mathbf{y}. \tag{2.7}$$

Proof of Equation 2.7 For simplicity, let $\mathbf{K} = \frac{\sigma_g^2}{M}\mathbf{X}\mathbf{X}^\top$ and $\delta = \frac{\sigma_e^2}{\sigma_g^2}$, so that $\mathbf{V} = \mathbf{K} + \sigma_e^2\mathbf{I}_N$. Starting from the mixed-model equations [67] and assuming that $\mathbf{V}^{-1}$ exists,

$$\begin{aligned}
\hat{\mathbf{b}} &= (\mathbf{X}^\top\mathbf{I}_N\mathbf{X} + M\delta\mathbf{I}_M)^{-1}\mathbf{X}^\top\mathbf{I}_N\mathbf{y} && (2.8)\\
&= [\frac{1}{\delta M}\mathbf{I}_M - \frac{1}{\delta M}\mathbf{I}_M\mathbf{X}^\top(\mathbf{I}_N + \mathbf{X}\frac{1}{\delta M}\mathbf{I}_M\mathbf{X}^\top)^{-1}\mathbf{X}\frac{1}{\delta M}\mathbf{I}_M]\mathbf{X}^\top\mathbf{y} && (2.9)\\
&= \frac{1}{\delta M}\mathbf{X}^\top\mathbf{y} - \frac{1}{\delta M}\mathbf{X}^\top(\mathbf{I}_N + \frac{1}{\sigma_e^2}\mathbf{K})^{-1}\frac{1}{\sigma_e^2}\mathbf{K}\mathbf{y}\\
&= \frac{1}{\delta M}\mathbf{X}^\top\mathbf{y} - \frac{1}{\delta M}\mathbf{X}^\top\mathbf{V}^{-1}\mathbf{K}\mathbf{y}\\
&= \frac{1}{\delta M}\mathbf{X}^\top(\mathbf{y} - \mathbf{V}^{-1}\mathbf{K}\mathbf{y})\\
&= \frac{1}{\delta M}\mathbf{X}^\top[\mathbf{y} - \mathbf{V}^{-1}(\mathbf{V} - \sigma_e^2\mathbf{I}_N)\mathbf{y}] = \frac{1}{\delta M}\mathbf{X}^\top\sigma_e^2\mathbf{V}^{-1}\mathbf{y}\\
&= \frac{\sigma_g^2}{M}\mathbf{X}^\top\mathbf{V}^{-1}\mathbf{y},
\end{aligned}$$

as we get 2.9 from 2.8 using the Woodbury matrix identity. □

The BLUP estimates may be used to obtain a prediction of the phenotype using $\mathbf{X}\hat{\mathbf{b}} = \mathbf{K}\mathbf{V}^{-1}\mathbf{y}$. Based on this, the residual $\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}$ equals $\sigma_e^2\mathbf{V}^{-1}\mathbf{y}$, which is part of the numerator of the test statistics defined in Eq. 2.5. Therefore, by using the residualised phenotype, testing with a LMM conditions on polygenic effects. This removes the genome-wide effects of other variants, allowing to focus on the phenotypic contribution of the variant being tested, increasing signal-to-noise and resulting in higher statistical power [36, 47].

A key element for LMM-based association is the way the GRM is formed. Early implementations included every genotyped variant which was later shown to be problematic as including the focal SNP in the GRM led to reduced statistical power [47, 70]. Listgarden et al. [70] observed that this is due to the fact that the candidate marker is used both as a fixed and a random effect, a phenomenon which was called "proximal contamination" [70]. Furthermore, variants that are positioned closely along the genome are often highly correlated as the corresponding alleles are inherited together due to lack of recombination in the genealogical history of the sample [71]. Therefore, the genome consists of blocks of correlated variants which are said to be in **linkage disequilibrium** (LD). Based on that, a method that tackles proximal contamination, `Fast-LMM-Select` [70], works with region-based GRMs by excluding variants in close proximity with the tested ones from the GRM. This, however, can be computationally intensive. An improvement can be achieved with the **leave one chromosome out** (LOCO) scheme, whereby all variants from the same chromosome as the variant being tested are excluded from the formation of the GRM [47]. Thus, for human genetic studies, a LMM with LOCO is a combination of 22 models, one for each autosomal chromosome.

A further memory bottleneck might arise due to the covariance matrix $\mathbf{V}$ which can be prohibitive to form, let alone invert, for large sample sizes. Initial implementations required to explicitly compute $\mathbf{V}^{-1}$ [59, 60, 72] and are thus not scalable enough for modern biobanks, which comprise hundreds of thousands of individuals. Recent methodological advances have enabled the development of efficient approaches, such as `BOLT-LMM` [36, 46] which avoids the explicit formation of $\mathbf{V}$ and relies on the use of the conjugate gradient iteration [73], or `fastGWA` [38] which works with a sparse version of the kinship matrix. `REGENIE` [39] is a more recent approximation to model 2.6, which avoids working with the full covariance matrix and instead leverages Ridge regression. I discuss these methods more in Chapter 4.

In general, most LMM-based association algorithms involve two steps: first infer model parameters (hereafter “model fitting”), then use these to calculate test statistics. The first step typically requires a carefully selected set of variants (e.g. LD-pruned and within a given MAF range) whereas testing can be applied to any variant (e.g. hard-called or imputed genotypes [31]). The testing step tends to be similarly designed for most LMM algorithms and has an $\mathcal{O}(NM_{\text{test}})$ complexity, for N individuals and M_{test} variants. Thus, LMM algorithms mostly differ in the way they fit the model, with a CPU cost ranging from a few $\mathcal{O}(NM_{\text{model}})$ iterations to $\mathcal{O}(N^3M_{\text{model}})$, and various levels of memory usage depending on how genotypes are loaded.

In conclusion, LMMs offer increased power and improved control for population stratification and cryptic relatedness. The aforementioned methods constitute the state of the art. However, as I will discuss in greater detail in the remainder of my thesis, most methods rely on approximations that lead to different trade-offs between scalability, statistical robustness, and association power. The development of scalable methods for association is a multifaceted problem and continues to be an active research area.

2.2.1 Dichotomous phenotypes

A dichotomous, or case/control (c/c), phenotype can be encoded with a binary variable with 1 indicating a case (e.g. an individual affected by a disease) and 0 a control. The simple linear model (2.3) I described earlier may again be used under certain assumptions, but a more suitable approach is to instead use logistic regression [12, 17]. Additional approaches (e.g. Fisher’s exact test [15]) may be used in this setting, but an extensive review is beyond the scope of this thesis. The logistic linear model for individual j can be written as

$$\operatorname{logit}(\mu_j) = \beta_i X_{ji} + \mathbf{w}_j^\top \boldsymbol{\gamma}, \tag{2.10}$$

where $\operatorname{logit}(z) = \log(z) - \log(1-z)$ and $\mu_j = \Pr(y_j = 1 | X_{ji}, \mathbf{w}_j)$, the probability that individual j is a case given the covariates $\mathbf{w}_j$ (or the environment) and genotype X_{ji}.

A widely used approach for association is the generalised-LMM, which, similarly to the case of quantitative phenotypes, incorporates random effects to control for confounding and increase power. `SAIGE` [37] was the first such method which enabled large-scale analyses of binary traits, followed by `fastGWA-GLMM` [45]. Both mixed models take the form

$$\operatorname{logit}(\mu_j) = \beta_i X_{ji} + \mathbf{w}_j^\top \boldsymbol{\gamma} + g_j, \tag{2.11}$$

which is similar to Eq. 2.4 as the vector of random effects $\boldsymbol{g}$ is assumed to be distributed as $\mathcal{N}_N(\mathbf{0}, \sigma_g^2 \mathbf{K})$. `SAIGE` estimates the kinship matrix $\mathbf{K}$ using SNP data as done by most other LMM algorithms, while `fastGWA-GLMM` works with a sparse estimator.

In practice, phenotypes often present unbalanced (c/c $= 1:10$) or extremely-unbalanced (c/c $\leq 1:100$) ratios. Studying such traits is challenging as unbalanced case/control ratios violate the asymptotic assumptions required by most estimators used in association. This leads to substantially high rates of type I error, particularly for rare variants [37, 39, 45]. Zhou et al. [37] observed that as the c/c ratio drops, a standard GLMM finds increasingly more false associations with rare variants. A solution to this phenomenon involves a calibration of the test statistics. One such technique, used by `SAIGE` and `fastGWA-GLMM`, is the saddle-point approximation (SPA), which approximates the distribution of the test statistic using the entire cumulant generating function. `REGENIE`, on the other hand, works with Firth correction, suggesting that it provides higher accuracy than the use of SPA [39]. Further discussion on dichotomous phenotypes lies beyond the scope of this thesis.

2.3 Evaluating the results of association studies

Testing for association of variant i to a phenotype amounts to testing whether the null hypothesis that $\beta_i = 0$ may be rejected. The determination of an association depends on some level of significance and the number of tests performed (e.g. M_{test}). An established p-value cut-off for multiple-testing correction in GWAS of common variants is 5×10^{-8}, reflecting the effective number of variants usually tested and the correlation among them [13, 15, 16, 74].

I previously mentioned that closely positioned variants are usually highly correlated due to LD. As a result, a statistically significant association implies that the candidate SNP is either directly correlated with the phenotype, or is in high LD with a – possibly not genotyped – causal variant [13, 15]. In either case, an association signal typically involves many significantly associated variants at one locus. The findings of a GWAS are often visually summarized using the so-called **Manhattan plots**, which depict the $-\log_{10}(\cdot)$ of each p-value for any tested variants along a genomic region (see Fig. 3.3 for an example).

Finally, several strategies have been developed to indicate whether GWAS association statistics are inflated due to relatedness or population stratification. One traditional metric is the **genomic inflation factor** λ_{GC} [26, 57], defined as the observed median χ^2 statistic divided by the theoretical median of a χ^2 distribution with one degree of freedom. In general, $\lambda \approx 1$ indicates no stratification, whereas values larger than 1.05 suggest confounding [26]. The rationale behind this is that the genetic architecture of a phenotype is generally sparse and most of the genome should be non-causal. Thus, one way to account for confounding, usually termed **genomic control** [26, 33, 57], is to adjust the test statistics by diving with λ_{GC}. However, large sample sizes and high polygenicity can also cause inflation of the test statistics [75], therefore such an approach is likely to yield conservative estimates and should be used with caution.

A better approach to discriminate between confounding and polygenicity is provided by the **LD score regression** framework (`LDSC`) [76]. The LD score of a variant is defined as the sum of the squared correlation between one variant and all other variants within a region. For polygenic traits, variants with higher LD scores will tend to tag more associated variants and thus have elevated test statistics. In contrast, population stratification and cryptic relatedness inflate test statistics independently of variant LD scores. By regressing observed χ^2 statistics on variant LD scores, `LDSC` allows the deconvolution of association signal from confounders and can also estimate SNP-heritability. An appealing property of `LDSC` is that it only requires summary statistics and a reference LD panel to be applied. However, LD scores may be difficult to accurately estimate and such biases affect `LDSC`'s performance. `LDER` [77] addresses this issue by working with the eigenvalues of the LD matrix and offers higher accuracy when there the discovery and the reference panel exhibit differences.

2.4 Rare variant association

The aforementioned methods work well when the trait is affected by common variants with small or moderate effects and GWAS have successfully identified a growing number of loci associated with phenotypes [16, 20]. Complex traits are also affected by a large number of rare variants which are increasingly accessible thanks to the advent of large whole exome/genome sequenced cohorts. Some rare variants are expected to have moderate to high effects as they tend to be recent and thus have less time to be eliminated by natural selection. However, single-variant tests, as those described in the previous section, are underpowered for rare variants unless sample sizes are very large [16].

2.4.1 Burden and overdispersion tests

A more effective approach involves merging of multiple rare markers in one score per region which is then tested using linear models [42]. One such approach, referred to as burden test, proceeds by collapsing markers within a given region and has the added benefit of being computationally efficient [78]. More specifically, an indicator variable for the j-th sample is defined as

$$Z_j = \begin{cases} 1, & \text{if rare variants are present in the region} \\ 0 & \text{otherwise} \end{cases} \tag{2.12}$$

where the region of interest could be a particular gene. Besides a binary indicator, Z_j could be defined as the total number variants present in the region. Next, a vector containing all the Z_j's for a given region is used in place of $\mathbf{x}_i$ in a linear model as that of Eq. 2.3, and each region of interest can be independently tested.

The method described here gives all variants the same weight, but this approach may be generalised by placing weights on variants according to their MAF or functional role (as with the adaptive burden tests) [42]. Additionally, when sample sizes are too small, genes can be grouped together into biologically relevant gene sets (e.g. according to implicated pathways) and tested for enrichment of the underlying variants of each set [29, 50].

There is a known limitation of burden tests regarding the assumptions made for the effects. By merging many variants together we assume that they all have the same direction of effect and have comparable magnitudes. In practice, however, a region might contain alleles of opposite directions, or variants that have no effect at all (e.g. synonymous mutations). Violating these assumptions may lead to decreased statistical power.

A strategy that circumvents this issue is based on over-dispersion tests, such as the *sequence kernel association test* (SKAT) [40]. SKAT regresses the phenotype against a group of genetic variants, while including covariates, and allows different variants to have different directions and magnitudes of effects by relying on a mixed model formulation. More specifically, SKAT assumes the following model

$$\mathbf{y} = \mathbf{W}\boldsymbol{\gamma} + \mathbf{G}\boldsymbol{\beta} + \mathbf{e} \tag{2.13}$$

where $\mathbf{W}$ encodes a set of relevant covariates with fixed effects $\boldsymbol{\gamma}$ (as before), $\mathbf{G}$ is an $N \times p$ matrix containing the genotypes for the p variants within the candidate region with random effects $\boldsymbol{\beta}$, and $\mathbf{e}$ is the residual error term following a multivariate normal distribution with mean zero and covariance matrix $\sigma_e^2\mathbf{I}$.

SKAT tests $H_0 : \boldsymbol{\beta} = \mathbf{0}$ by assuming each β_j has mean zero and variance $q_j\tau$, where τ is a variance component and q_j is a pre-specified weight for variant j (e.g. a decreasing function of MAF) [40]. It can be easily shown that, under H_0, the variance-component score statistic is $Q = \hat{\boldsymbol{y}}^\top\mathbf{R}\hat{\boldsymbol{y}}$, where $\hat{\boldsymbol{y}}$ is the covariate-adjusted phenotype and $\mathbf{R} = \mathbf{G}\mathbf{Q}\mathbf{G}^\top$, with a diagonal matrix $\mathbf{Q}$ containing the variant weights. $\mathbf{R}$ can be seen as a kernel function $R(\cdot,\cdot)$ defined on genotypes, thus the name of the method. The use of suitable weights q_j and modelling $\boldsymbol{\beta}$ as a random effect offer SKAT increased statistical power compared to burden testing in cases where underlying assumptions are not met. Although it needs to be applied independently on every region, parameter estimation is efficient on an exome-wide scale provided that each gene contains a small number of variants.

Additionally, many of the aforementioned methods have been extended for region-based testing of dichotomous phenotypes, such as SAIGE-GENE [41], REGENIE[39], and fastGWA-GLMM [45], while being scalable for hundreds of thousands of individuals. As an example, Zhao et al. [44] recently optimised SKAT methods by adding the SPA feature and performed an exome-wide association study of rare variation (MAF$\leq$ 0.01) in UKBB. After analysing 791 diseases, and running a variety of tests including conditioning on nearby common variants, they detected 10 significant associations, providing further evidence that gene-based approaches may be more powerful, or complementary at least, to single-variant methods for rare variation, a theme that we will revisit in the next chapter.

2.4.2 Rare variant association using IBD sharing

Burden and overdispersion tests assume that the rare variants have been genotyped. A different family of approaches relies on the sharing of identical-by-descent (IBD) genomic segments between individuals to enable testing of rare variants that have not necessarily been typed. IBD is a fundamental measure of genetic relatedness and can be particularly suitable for rare variant association studies for the following reason. IBD segments are co-inherited by modern individuals from common ancestors that lived in the recent past, thus their presence suggests that individuals who are IBD also share all genomic variation within the affected region, including rare variants of recent origin [11, 79, 80]. IBD-based tests might also offer a gain in statistical power for rare variants as they automatically collapse all markers within each haplotype [81, 82].

Several algorithms for IBD detection exist relying on a mixture of scalable heuristics, such as string matching, and more accurate probabilistic approaches that model the underlying genealogical processes, such as Hidden Markov Models.

Heuristic strategies rely on the assumption that identical-by-state (IBS) regions are a good proxy for IBD regions. They are typically scalable and ideal for finding large IBD regions from recent ancestors (e.g. segments longer than 5 centimorgans)[79]. For shorter segments transmitted by more remote common ancestors, however, the assumed equivalence between IBS and IBD regions becomes less reliable. In this case, approaches modelling haplotypic or genealogical information tend to be more accurate, at the cost of additional computation. `FastSMC` [11] is a recent method for IBD detection which scales to large biobanks. It combines a fast heuristic search [81] for candidate IBD segments with accurate coalescent-based likelihood calculations [9], and enables estimating the age of the common ancestors who transmit IBD regions.

There are several ways in which genetic relatedness can be exploited for detecting associations. One approach suitable for dichotomous traits involves the comparison of IBD sharing rates among case/case pairs versus control/control pairs and a simple statistic to test for significant differences [82]. Another technique uses IBD segments to construct haplotype clusters [81]. Each cluster indicates a group of samples sharing specific rare alleles and the cluster membership can be tested for association with a phenotype using a LMM. Gusev et al. [81] followed this approach and detected significant associations in loci where SNPs failed to be associated, indicating insufficient genotyping to detect rare causal variation. Finally, IBD sharing could be used as a proxy to associate ultra-rare variants by comparing sequenced with genotyped samples in the same cohort; I present such an approach in Chapter 3.

2.4.3 Ancestral recombination graphs

The ancestral recombination graph (ARG) is a mathematical object that leverages relatedness to model the history of genetic recombination and coalescence events among a sample of genomes [83]. It is a directed acyclic graph in which nodes represent genomes in time and edges show the flow of genetic material among those genomes. An ARG can be constructed from genetic data, mainly from sequenced but also array genotypes, and enables the inference of several evolutionary parameters, such as population sizes and times to most recent common ancestors. Consequently, ARGs are increasingly used in population genetics to study evolution or demography [83–85], as well as improving association testing or polygenic predictio in comparison to SNP-based approaches [48, 86, 87].

Initial studies involving ARGs were limited to a handful of samples, due to the complexity of inferring an ARG from genomic data [83]. However, recent algorithmic improvements have enabled studies with tens or hundreds of thousands of samples [48, 84, 88]. The `ARG-Needle` algorithm, developed recently by Zhang et al. [48] has been used to infer the ARG from SNP array data of 337k UKBB individuals. The inferred ARG was then used to infer the presence of genetic variation, particularly rare and ultra-rare variants, that was not observed in the data used to construct the ARG, enabling to detect associations that would otherwise require sequencing data to be available. This analysis detected association with variants having frequency as low as 4×10^{-6}, some of which were missed by using imputation from an ancestry-matched reference panel, but were broadly validated based on UKBB exome sequencing data [48]. Efficient methods for ARG-inference, such as `ARG-Needle`, may therefore be used to complement genotype imputation in the analysis of complex traits for populations under-represented in ongoing sequencing studies [5, 6, 34].

Besides enabling the detection of association, the unobserved variants inferred through an ARG allow improving other linear-mixed-model analyses, by providing better estimates of genetic relatedness held within the GRM. As a result, an ARG-based approach has the potential to yield more accurate estimates of heritability, better prediction of phenotypic values, and improved power for detecting associations between genetic variants and traits, as demonstrated with simulations by Zhang et al. [48].

For association, in particular, there are two ways an ARG can yield higher power. First, the ARG clades can be used as a proxy to test for untyped genetic variation, after training the LMM with SNP data (as described above). Second, utilising a GRM obtained from a (sufficiently accurate) inferred-ARG to fit the LMM can improve power in contrast to a SNP-based GRM. The latter was alse demonstrated by Zhang et al. [48] using simulations, but their approach was based on explicitly forming and inverting large covariance matrices, posing several computational bottlenecks. I revisit all the aforementioned tasks in Chapter 6 where I explore how ARG-based complex-trait analyses can be improved computationally.

3

Leveraging IBD to detect association with rare variation

Contents

Genetic relatedness results in the presence of genomic regions that are shared identical-by-descent (IBD) between individuals, who are usually unaware of their distant genealogical relationship [79, 80]. This chapter describes how IBD sharing can be leveraged to characterize the contribution of rare genetic variation in disease aetiology, and offers an implicit way to test for rare variant association [81, 82]. To briefly summarise the intuition behind the proposed approach, consider a genotyped individual who shares IBD segments with a known carrier of an untyped mutation. This individual is also likely a carrier of that mutation, by inheriting it from the shared IBD ancestor, as shown in Fig. 3.1. When such a mutation is associated to disease, both individuals are at increased or decreased risk for disease and IBD sharing can thus be utilized to test for association with the phenotype.

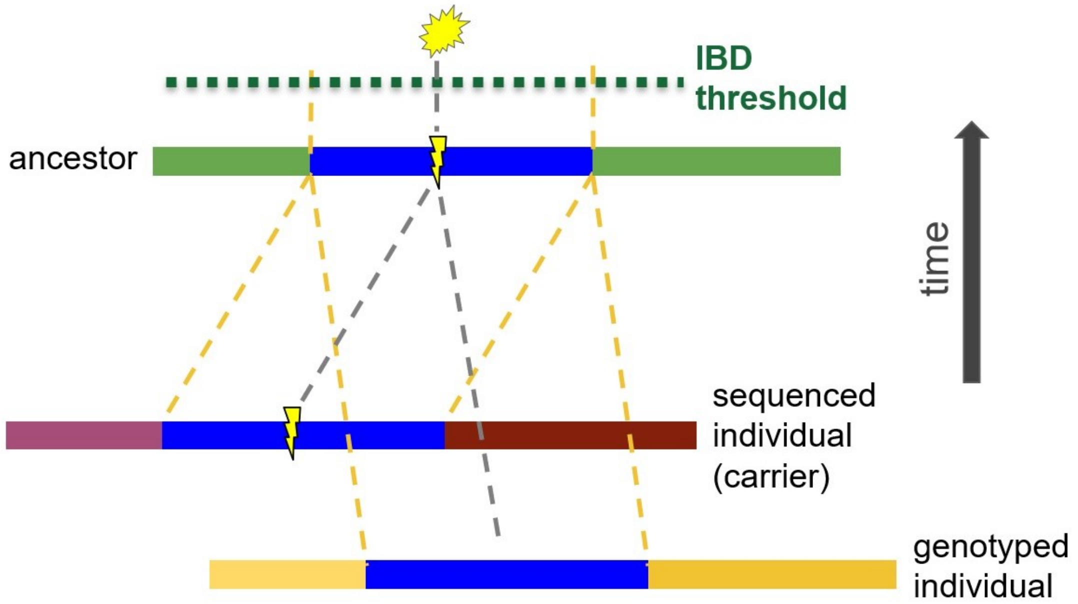


Figure 3.1: Explanation of IBD-based imputation. This figure depicts two samples, one of which is exome-sequenced, and a common ancestor who transmitted the blue segment. The mutation in focus, which we assume is observed only in the sequenced individual, occurred before the age of the common ancestor, thus the genotyped sample should also carry it. Using a high enough time threshold, as the one shown, would allow us to correctly impute this variant.

The following two sections are part of a collaboration currently published in [11]. The material presented here, which is slightly adapted for coherence with the rest of the thesis, is work I performed myself regarding the development and testing of the *LoF-segment burden*, validating the utility of accurately detecting IBD segments and extending a few initial findings. Sections 3.4 and 3.5 contain work I performed independently.

3.1 The LoF-Segment burden test for association

We applied `FastSMC` [11], an IBD detection method, to 487,409 phased samples from the UKBB and detected roughly 214 billion IBD-shared segments, revealing an intricate network of cryptic genetic relatedness among British individuals within the past two millennia. By focusing on the subset of 49,797 samples who had undergone whole-exome sequencing (WES) by a pilot study [89], we observed that sharing of rare alleles was correlated with the sharing of IBD segments, especially for ultra rare

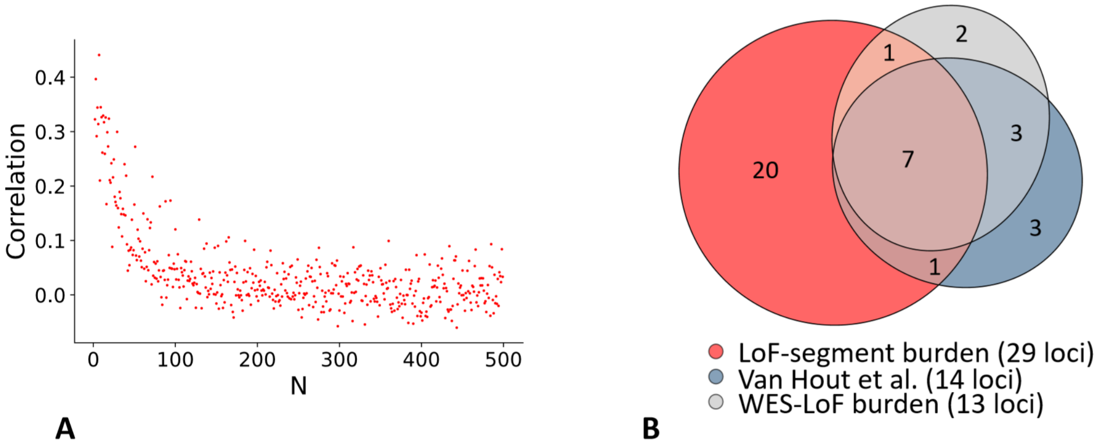


Figure 3.2: A. Correlation between IBD sharing (average number of IBD segments per pair of samples in the past 10 generations) and ultra-rare variants sharing (average number of F_N mutations per pair, for increasing numbers of carriers N). **B.** Venn diagram representing the sets of exome-wide significant associated loci for 7 blood-related traits by the IBD-based LoF-segment test we performed (red), the WES-based LoF burden test reported by Van Hout et al. [89] (petrol), and the WES-based burden test we performed (grey).

variants (e.g. $r = 0.3$ for MAF$\sim 0.003\%$, Fig. 3.2A). Based on this, we set out to detect association to rare and likely trait-associated variation in the non-sequenced cohort by quantifying the extent to which a genotyped sample shares IBD segments with exome-sequenced carriers of any loss-of-function (LoF) mutation at each gene. This set included variants annotated as stop-lost, start-lost, splice-acceptor, splice-donor, stop-gained, and frameshift, with MAF less than 0.01, following [89].

We used each IBD segment between exome-sequenced, LoF-carrying individuals and non-sequenced individuals as a surrogate for the latter carrying an untyped LoF mutation, which we then tested for association with a target phenotype. For a given gene and a given non-sequenced individual, we define a “LoF-segment” as any IBD segment shared with an exome-sequenced LoF mutation carrier, within the gene boundaries. We then compute a LoF-segment burden for each individual as the sum of probabilities (IBD quality scores) of all LoF-segments involving that individual, under the assumption that increased IBD probability and incidence corresponds to increased probability of sharing the LoF variant. Finally, this burden is tested for association with each phenotype (rank-based inverse normal transformed) using linear regression with covariates for age, sex, BMI, smoking status, and four principal components, similarly to Van Hout et al. [89].

Although this test captures uncertainty about the sharing of IBD segments through the use of IBD quality scores, it makes use of all LoF-segments, regardless of their age. As a result, it may be suboptimal in cases where the LoF arose after the most recent common ancestor, for which a LoF-segment is independent of the underlying LoF sharing, and thus should not contribute signal to the burden test. We thus augmented the LoF burden test by separately considering only LoF-segments older than a specified threshold. For each gene, we divided all LoF-segments into deciles based on the IBD quality score. For instance, segments with scores in the tenth decile (which corresponds to the interval [0.47,1]), strongly suggest the sharing of common ancestors that lived recently and have therefore transmitted extremely recent variation. Thus we constructed ten separate LoF-segment burdens, with increasingly more stringent quality score cut-offs (referred to as time transformations), and performed ten association tests for each gene, selecting the test that resulted in the lowest p-value after adjusting the significance thresholds by conservatively assuming independence for all tests. Because not all genes contained shared LoF-segments for testing, the total number of tested genes was reduced to 14,249. This resulted in a Bonferroni-corrected exome-wide significance threshold of $0.05/(10 \times 14{,}249)$ for our LoF-segment burden analysis.

Gene-based burden tests are meant to implicate specific genes with a known directional effect on the trait. The observed signal, however, may not always be driven by a causal variant and instead be due to tagging of causal variants in nearby genes. In this case, it is possible that the underlying rare causal variant is tagged by a common variant, which may have been detected in a previous GWAS. In particular, these common variants may provide better tagging of the underlying true causal variation than our LoF-segment burden score, and would thus remove or significantly reduce the association signal if included as covariates in the test. Based on this principle, for each gene and each trait, we selected up to three genotyped SNPs that were in proximity (±1 Mb from the gene), which were significantly

($p < 1 \times 10^{-8}$) associated by Loh et al. [46], and used them as covariates. We observed that this approach often improves the association signal (e.g. see Section 3.3), removing signals that were likely caused by tagging of common variants. We refer to analyses that include top associated SNPs as covariates as *SNP-adjusted*, for either the LoF-segment or WES-LoF burden test.

We validated our approaches, both *LoF-segment* and *not SNP-adjusted LoF-segment* burden tests, by testing for association between rare variation and 7 blood-related phenotypes analysed in the UKBB pilot exome study [89], and comparing to the results of that same study. Moreover, summary statistics for this analysis were not available, thus we performed an exome-wide burden testing. Specifically, we used the same testing framework we used in our LoF-segment burden analysis to test for association between phenotypes and burden of LoF variants within a gene in exome-sequenced individuals, adjusting for the same covariates and using the same rank-based inverse normal transformation of the phenotype. We refer to this analysis as WES-LoF burden analysis.

Both LoF-segment burden and WES-LoF burden analyses were restricted to unrelated individuals of White British ancestry, as defined in [4], and the LoF-segment burden analysis was further restricted to individuals for which exome sequencing data was not available. This resulted in 303,125 individuals for the two LoF-segment burden tests and 34,422 individuals for the WES-LoF. On top of these approaches, and to better account for sample structure or polygenicity (as explained in Section 2.2), we applied `BOLT-LMM` [46] on 446,050 European samples using $\sim$ 623k SNPs for model-fitting and testing for the LoF-segment burden scores. Finally, we note that the UK Biobank had released a statement regarding incorrectly mapped variants in the 50k WES “Functionally Equivalent” (FE) dataset, which we however believe did not introduce any significant biases in our analyses.

3.2 Results from the exome-wide association study

A comparison of our approaches with the pilot WES study [89] on 7 blood-related traits is summarised in Fig. 3.2B and Table 3.1. The LoF-segment burden replicated

11 out of the 14 previously reported associations at $p < 0.05/10 = 0.005$ (adjusted for testing of 10 transformations) and, strikingly, 8 of these associations were exome-wide significant in the non-sequenced cohort ($p < 0.05/(10 \times 14{,}249)$). Missing a few associations could be ascribed to the slightly different testing strategy we adopted, e.g. the use of a linear model, rather than a mixed model, and the exclusion of related samples. Indeed, testing the LoF-segment scores using `BOLT-LMM` increased the replication ratio to 13/14.

We next aimed at quantifying how effective IBD sharing (through LoF-segment burden testing) is at detecting associations, compared to testing directly based on exome sequencing data. We computed the phenotypic variance explained by the indirect IBD-based test and the direct exome-based test (after subtracting the effect of covariates from both), focusing on the 14 loci reported in [89]. The ratio of these variances was 19.64%, on average, corresponding to the decrease in effect-size (in units of variance) due to estimation error and inclusion of segments sharing the non-LoF haplotype. We note that, due to phase uncertainty, we expect the LoF-segment burden to explain at most 50% of the variance explained by direct sequencing. Assuming the ratio of variances corresponds to the squared correlation between the LoF-segment burden estimate and the true exome burden, the LoF-segment burden estimator has statistical power equivalent to a direct exome sequencing study of 19.64% × 303,125, or ∼60k samples [90] - effectively doubling the size of the exome study.

Table 3.2 reports the exome-wide set of associations detected by the LoF-segment burden. An indicative Manhattan plot is given in Fig. 3.3, plots for the rest of the analysed traits are given in the appendix (figures A.1 - A.6), and a QQ-plot verifying the calibration of the LoF-segment burden is shown in Fig. 3.4. Our proposed approach identified a total of 29 associated loci ($p < 0.05/(10 \times 14{,}249)$), 20 of which were not discovered in either of the exome sequencing studies, possibly due to lack of statistical power. These loci were obtained by clustering 111 genes according to their genomic location, which are all listed in Table 3.1. Without adjusting for common variation we would have identified 186 significant gene

	Gene	Trait	Van Hout et al.	WES LoF	LoF-segment	Bolt-LMM	$\mathbf{R^2_{prop}}$
1	*IL33*	Eosinophil count	3.30×10^{-10}	2.01×10^{-3}	8.64×10^{-15}	3.8×10^{-18}	72.26
2	*GP1BA*	Mean platelet volume	6.40×10^{-8}	8.84×10^{-8}	1.82×10^{-19}	2.4×10^{-30}	32.57
3	*TUBB1*	Platelet distribution width	2.50×10^{-23}	7.34×10^{-18}	7.38×10^{-12}	1.7×10^{-97}	07.25
4	*TUBB1*	Mean platelet volume	2.40×10^{-8}	3.01×10^{-7}	2.15×10^{-3}	1.4×10^{-25}	04.11
5	*TUBB1*	Platelet count	2.10×10^{-9}	7.45×10^{-7}	4.21×10^{-5}	2.4×10^{-17}	07.84
6	*HBB*	Red blood cell distribution width	5.80×10^{-8}	3.49×10^{-2}	2.25×10^{-3}	1.3×10^{-13}	23.99
7	*HBB*	Red blood cell count	1.70×10^{-9}	7.95×10^{-2}	2.68×10^{-2}	1.3×10^{-9}	18.23
8	*KLF1*	Red blood cell distribution width	1.50×10^{-13}	6.95×10^{-13}	3.49×10^{-34}	3.6×10^{-44}	32.99
9	*KLF1*	Mean corpuscular haemoglobin	1.70×10^{-16}	9.11×10^{-15}	6.79×10^{-21}	1.3×10^{-27}	16.76
10	*ASXL1*	Platelet distribution width	4.70×10^{-9}	1.44×10^{-6}	0.16×10^{0}	9.0×10^{-2}	00.98
11	*ASXL1*	Red blood cell distribution width	2.40×10^{-11}	8.23×10^{-4}	0.32×10^{0}	8.2×10^{-3}	01.03
12	*KALRN*	Mean platelet volume	2.70×10^{-23}	3.85×10^{-18}	3.79×10^{-12}	2.3×10^{-23}	07.33
13	*IQGAP2*	Mean platelet volume	1.10×10^{-19}	3.72×10^{-15}	4.40×10^{-34}	3.0×10^{-78}	27.43
14	*GMPR*	Mean corpuscular haemoglobin	1.10×10^{-8}	2.94×10^{-6}	7.60×10^{-11}	1.3×10^{-14}	22.18

Table 3.1: Power comparison with the UKBB pilot WES study. A comparison between the pilot WES-based burden test [89] (obtained using a linear mixed model), our WES-based burden (two-sided t-test; labelled as WES LoF), the LoF-segment (two-sided t-test), and the corresponding results using `Bolt-LMM`; the latter approach was not part of [11]. The Bonferroni-corrected exome-wide significance threshold for the first two approaches is 3.4×10^{-6}, after correcting for multiple testing with ~15k genes, and 3.51×10^{-7} for the LoF-segment burden, after adjusting for 14,249 genes and 10 time transformations. The last column estimates the proportion of the phenotypic variation (in %) of the sequenced samples that can be explained by the non-sequenced cohort; on average that is 19.64% for all the 14 reported associations, or 27.35% if focusing on the exome-wide significant signals.

associations spanning 33 genomic loci. This difference suggests that inclusion of significant common associations in rare variant burden tests may lead to improved interpretability and fewer false-positives due to tagging.

Our exome-wide significant associations can be partitioned in three groups: those reported by Van Hout et al. [89] (presented in Table 3.1), those with support from published GWAS, and previously unreported gene-trait associations (potentially novel). An example of the latter case is the association between platelet count and *MPL* ($p = 1.99 \times 10^{-7}$), which encodes the thrombopoietin receptor that acts as a primary regulator of megakaryopoiesis and platelet production. Associations with genes that have been previously implicated in genome-wide scans for common variants include associations between eosinophil count, *GFI1B* ($p = 1.92 \times 10^{-7}$) and *RPH3A* ($p = 7.63 \times 10^{-14}$) [23, 91], or platelet count, *IQGAP2* ($p = 6.52 \times 10^{-8}$) and *GP1BA* ($p = 1.43 \times 10^{-7}$) [23, 92]. We also identify genes that were previously associated with other blood-related phenotypes in different populations such as the association between platelet distribution width and *APOA5* ($p = 1.94 \times 10^{-8}$). This gene encodes proteins regulating the plasma triglyceride levels and linked common variants have been associated with platelet count in individuals of Japanese descent [93].

Additional associations include the one between red blood cell distribution width and *APOC3* ($p = 3.67 \times 10^{-11}$), which encodes a protein that interacts with proteins encoded by other genes (*APOA1, APOA4*) associated with the same trait. The association between *APOC3* and platelet count was also detected with our WES-LoF burden analysis ($p = 2.13 \times 10^{-7}$) and by previous studies based on common SNPs [23]. We also found links between *CHEK2* and mean corpuscular haemoglobin ($p = 1.43 \times 10^{-7}$) or mean platelet volume ($p = 1.93 \times 10^{-7}$). This gene plays an important role in tumor suppression and was found to be associated with other blood traits, such as platelet crit [23, 89] and red blood cell distribution width [91]. Overall, this analysis highlights the utility of applying `FastSMC` on a hybrid sequenced/genotyped cohort to identify novel, rare variant associations and/or characterize known signals in larger cohorts.

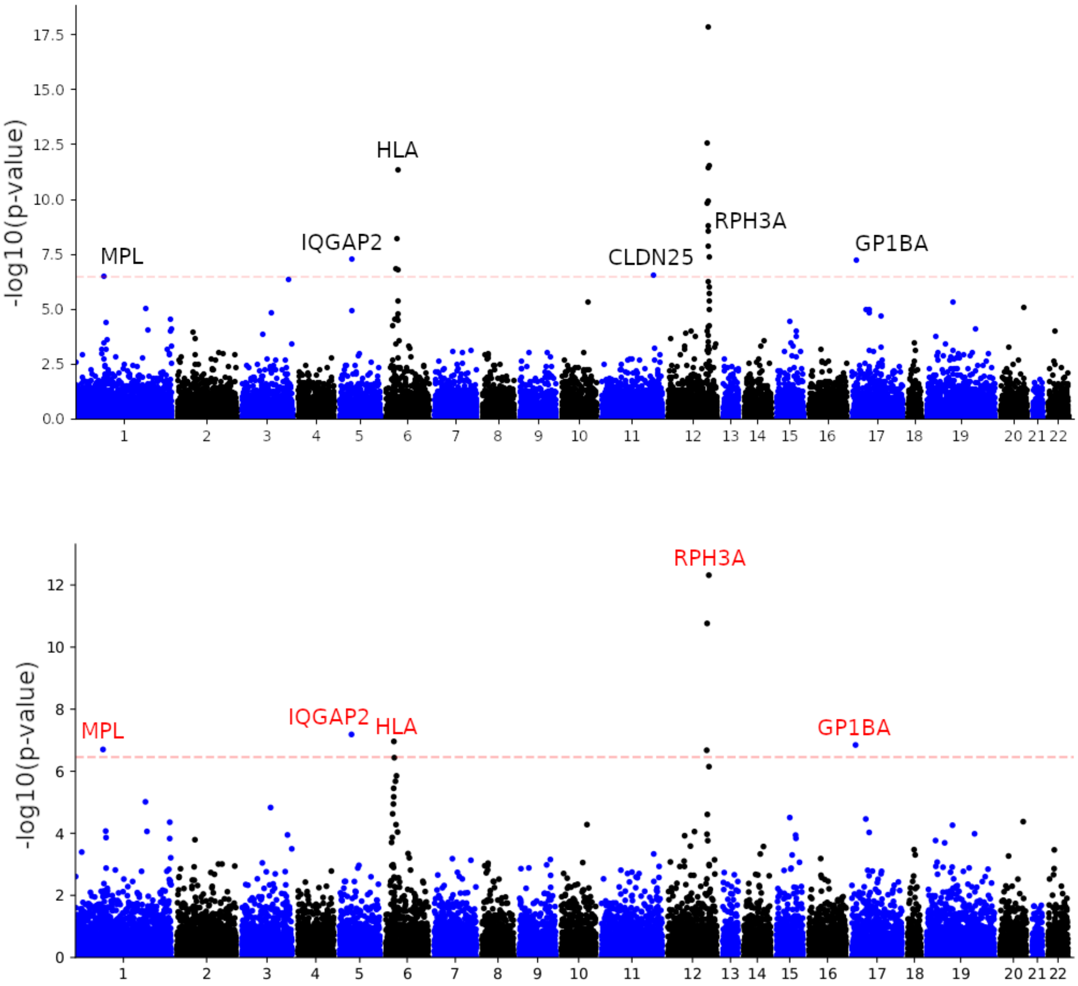


Figure 3.3: LoF-segment burden exome-wide Manhattan plots for platelet count with and without SNP-adjustment. Labelled genes are exome-wide significant after adjusting for multiple testing (t-test p $< 0.05/(14{,}249 \times 10) = 3.51 \times 10^{-7}$; dashed red line). We compare results before (**top**) and after (**bottom**) adjusting for common SNP associations [46]. Both LoF-segment burden analyses used 303,125 British individuals not included in the exome sequencing cohort. The cluster of genes in chromosome 12 labeled as *RPH3A* (the top association) contains *KCTD10, TCHP* and *RPH3A* and the signal with *CLDN25* was cleared after SNP-adjustment. Red labels in the lower plot indicate associations that were not detected in our WES-based LoF burden analysis or reported by Van Hout et al. [89].

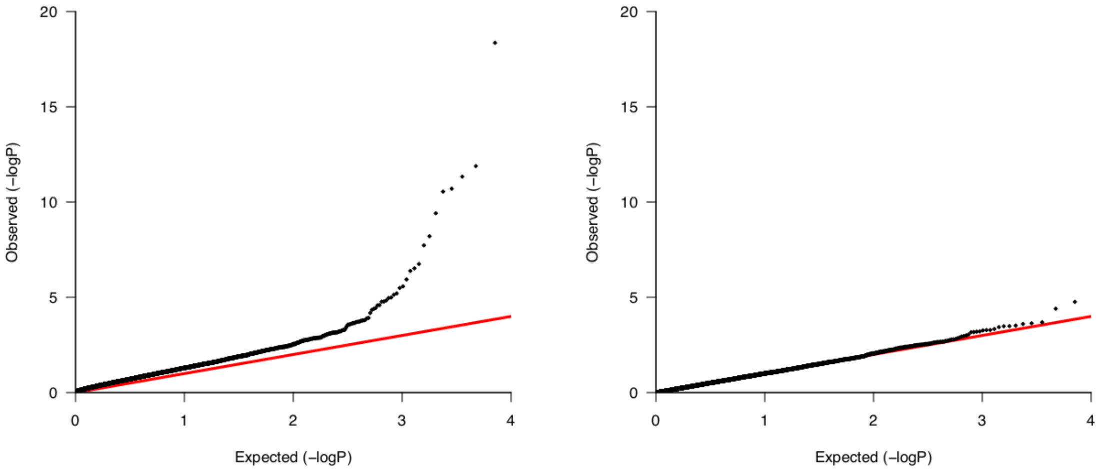


Figure 3.4: QQ-plots for the LoF-segment burden test. Quantile-quantile plots for the LoF-segment burden association on mean platelet volume (**left**) and on the same trait, but with randomly permuted phenotype values (**right**). For the latter case, and in contrast to the proper one, the observed values do not deviate from the expected ones, suggesting a well-calibrated test.

	Trait	Chr	Region (Mb)	Min. p-value	Candidate gene(s)
1	Eosinophil count	chr6	26.01:31.10	1.21×10^{-26}	HIST1H1A, HIST1H1C, HIST1H1T, HIST1H2BF, HIST1H3E, HIST1H4F, BTN3A2, BTN2A2, BTN3A3, BTN2A1, BTN1A1, ABT1, HIST1H2AG, HIST1H2AH, PRSS16, POM121L2, ZNF391, HIST1H2BM, PGBD1, HIST1H2AK, HIST1H2BO, OR2B2, OR2B6, ZNF165, ZSCAN16, ZKSCAN8, ZSCAN9, ZKSCAN4, NKAPL, **ZSCAN31**, ZSCAN12, ZSCAN23, GPX6, PSORS1C2
2	Eosinophil count	chr9	6.21:6.25	8.64×10^{-15}	**IL33** [23, 89]
3	Eosinophil count	chr9	135.82:135.86	1.92×10^{-7}	**GFI1B** [23]
4	Eosinophil count	chr12	113.01:113.41	7.63×10^{-14}	**RPH3A**, OAS3 [91]
5	Mean corpuscular haemoglobin	chr6	16.23:16.29	7.60×10^{-11}	**GMPR** [23, 89, 91]
6	Mean corpuscular haemoglobin	chr6	25.72:31.10	3.82×10^{-69}	HIST1H2BA, SLC17A2, HIST1H2AB, HFE [23], HIST1H4C, HIST1H2BD, **HIST1H4D**, HIST1H2BG, HIST1H2AE, HIST1H1D, BTN2A1, HIST1H2AG, HIST1H4I, ZNF184, PSORS1C2
7	Mean corpuscular haemoglobin	chr19	12.98:12.99	6.79×10^{-21}	**KLF1** [23, 89, 91]
8	Mean corpuscular haemoglobin	chr22	29.08:29.13	1.43×10^{-7}	**CHEK2**
9	Mean platelet thrombocyte volume	chr1	247.87:247.88	1.44×10^{-8}	**OR6F1**
10	Mean platelet thrombocyte volume	chr3	123.81:124.44	3.79×10^{-12}	**KALRN** [23, 89, 92]
11	Mean platelet thrombocyte volume	chr5	74.80:76.00	4.40×10^{-34}	POLK, **IQGAP2** [23, 89]
12	Mean platelet thrombocyte volume	chr6	26.18:27.92	1.46×10^{-8}	HIST1H4D, **POM121L2**, OR2B6
13	Mean platelet thrombocyte volume	chr12	122.51:124.49	6.29×10^{-10}	**MLXIP**, ZNF664
14	Mean platelet thrombocyte volume	chr16	90.03:90.03	2.61×10^{-7}	**CENPBD1**
15	Mean platelet thrombocyte volume	chr17	4.83:4.83	1.82×10^{-19}	**GP1BA** [23, 89]
16	Mean platelet thrombocyte volume	chr22	29.08:29.13	1.93×10^{-7}	**CHEK2**
17	Platelet count	chr1	43.80:43.82	1.99×10^{-7}	**MPL**
18	Platelet count	chr5	75.69:76.00	6.52×10^{-8}	**IQGAP2** [23]
19	Platelet count	chr6	26.59:26.60	1.1×10^{-7}	**ABT1** (within HLA region)
20	Platelet count	chr12	109.88:113.33	4.82×10^{-13}	KCTD10, TCHP, **RPH3A** [92]
21	Platelet count	chr17	4.83:4.83	1.43×10^{-7}	**GP1BA** [23, 92]
22	Platelet distr. width	chr11	116.66:116.66	1.94×10^{-8}	**APOA5**
23	Platelet distr. width	chr17	4.83:4.83	4.26×10^{-9}	**GP1BA** [23]
24	Platelet distr. width	chr20	57.59:57.60	7.38×10^{-12}	**TUBB1** [23, 89]
25	Red blood cell count	chr6	26.45:31.10	1.39×10^{-10}	BTN2A1, POM121L2, HIST1H2BM, HIST1H2BO, ZNF165, **ZSCAN9**, ZKSCAN4, PGBD1, ZSCAN31, GPX6, PSORS1C2
26	Red blood cell distr. width	chr6	26.01:28.48	3.03×10^{-15}	HIST1H1A, HIST1H3A, HIST1H1C, HIST1H1T, HIST1H4D, HIST1H4F, BTN3A2, HIST1H2AG, HIST1H2AH, **ZNF391**, HIST1H2BM, HIST1H2AM, ZNF165, ZSCAN16, NKAPL, PGBD1, ZSCAN31, ZSCAN12, GPX6
27	Red blood cell distr. width	chr9	135.82:135.86	8.1×10^{-8}	**GFI1B** [23, 91]
28	Red blood cell distr. width	chr11	116.69:116.70	3.67×10^{-11}	**APOC3**
29	Red blood cell distr. width	chr19	12.98:12.99	3.49×10^{-34}	**KLF1** [89]

Table 3.2: Associations detected by the LoF-segment burden. Exome-wide significant associations ($p < 0.05/(14{,}249 \times 10) = 3.51 \times 10^{-7}$) detected using the LoF-segment burden (SNP-adjusted). Associated genes are clustered in 29 loci. For each locus we report the set of associated genes, minimum p-value, and any previous studies identifying the same loci. The gene corresponding to the minimum p-value is highlighted in bold.

3.3 LoF-segment burden analysis of bone mineral density

In a collaboration with Dr. Matthew Page, I set out to apply this approach to validate and complement a recent analysis of heel bone mineral density (BMD; UKBB data-field 3148) performed at Union Chimique Belge (UCB). The analysis performed at UCB used the `SKAT` framework [40] to test for association between BMD and rare (MAF$\leq$ 1%) genotyped variants in the UKBB (unpublished; personal communication). I prepared the BMD phenotype conditioning on several variables, including the occurrence of any heel fractures, smoking and menopause status, and any known history of musculoskeletal or bone related diseases, as advised for increasing signal-to-noise. This resulted in a sample of size N = 223,915.

I first applied the LoF-segment approach, this time working with `BOLT-LMM` [36] to increase statistical power. I used most common SNPs (MAF$\geq$ 0.05, M = 352k) to fit the model and then tested the burden scores for association. This approach identified a few genes related to bone function[1] including *NME8* ($p = 1.1 \times 10^{-5}$, MAF $\sim$ 0.4%) and *PYY* ($p = 3.5 \times 10^{-5}$, MAF $\sim$ 0.1%).

As a next step, I considered a broader set of variants with those characterised as either “deleterious” by `SIFT` or “probably-damaging” by `PolyPhen` [94], with MAF not exceeding 1%. This yielded a set of 1,360,009 variants, which was roughly five times larger than what the LoF analysis considered (247,098 mutations), resulting in an average of 656 carriers per gene for 16,498 genes. This approach identified an association between BMD and *LRP5* ($p = 3.1 \times 10^{-6}$), which was also the strongest association in the `SKAT` analysis performed by UCB, and two additional genes. Autosomal dominant mutations in *LRP5* are known to cause the Van Buchem disease, a sclerosing bone dysplasia characterised by increased BMD [95]. Overall, this analysis demonstrates the effectiveness of IBD sharing for rare variant association especially for cohorts with limited sequencing.

[1] as found in the GWAS Catalogue, `https://www.ebi.ac.uk/gwas/`.

3.4 Utilising IBD to account for recent ancestry

Mathieson and McVean [33] demonstrated with simulations that rare variants may confound association studies when standard approaches used to control for common-variant stratification are adopted. This is because rare variants exhibit a different type of stratification, likely stronger, in comparison to common ones [33]. Recent studies using UKBB have shown that residual stratification has affected association studies of complex traits [10, 52, 53]. Moreover, a simulation study by Zaidi and Mathieson [96] showed that subtle stratification might bias polygenic risk prediction, and suggested that using IBD may lead to better control for stratification.

To explore that, I compared the correlations between 10 UKBB quantitative phenotypes and different types of principal components (PCs) of ancestry. To that end, I worked with the top eigenvectors either from the SNP-based GRM [4, 26, 51] (standard PCA), or the GRM formed by considering IBD segments shared within the last 10 generations [11] (IBD-based). As illustrated in Fig. 3.5, the gradient of significance for the PC-phenotype correlation was diverse and quite different from that of the corresponding eigenvalue. For the standard PCA for instance, PC-5 was the most significantly correlated covariate for every phenotype, whereas most of the top-15 had moderate correlation. For IBD-based PCA, the two most significantly-correlated vectors (18 and 47) had very distinctive patterns, which implies that using an arbitrary small number of top eigenvectors might only partially account for sample structure.

The IBD-based PCs could be used as covariates in association, in addition to standard PCA. This approach was utilised in the latest GWAS of reproductive success [97] to ensure that recent ancestry did not bias any findings. In particular, I passed the top-100 principal components obtained from the aforementioned IBD-based GRM to Gardner et al. [97] who compared this approach with standard PCA correction and observed no differences or biases. Because it captures fine-scale population structure [11, 98], IBD information can likely be used to more effectively control for population structure, although this will require further methodological development.

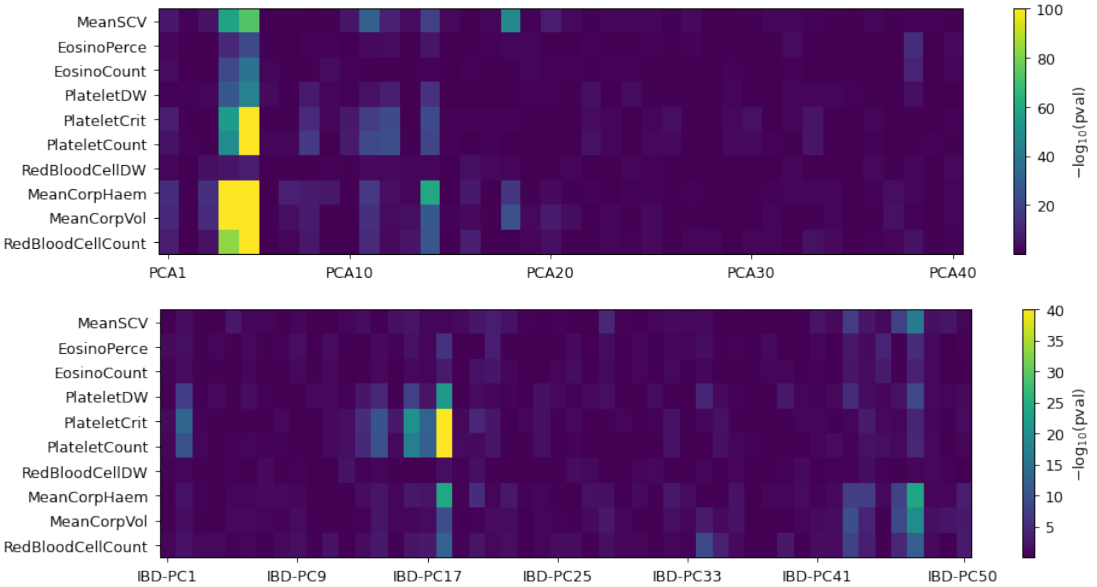


Figure 3.5: Correlation between complex traits and principal components of ancestry. This figure illustrates the significance of each phenotype-eigenvector correlation, for 10 quantitative UKBB phenotypes and either the top 40 eigenvectors of the SNP-based GRM [4] (**top**), or the top 50 eigenvectors of the IBD-based GRM [11] (**bottom**), using $N = 315$k White British. Each value is the $-\log(\cdot)$ of the p-value, capped to 100 to increase clarity.

3.5 Limitations of the LoF-Segment burden and next steps

In this chapter I presented results of using IBD sharing to detect association to ultra-rare variants leveraging a subset of sequenced individuals, as well as an exploratory analysis of using IBD-derived PCs to control for confounders. I will close this chapter with a few limitations of this approach, some of which motivated the work I performed next during my DPhil.

The LoF-segment burden does not rely on phasing information and the IBD segment shared with a carrier is equally likely to involve or not the haplotype that harbors the target LoF variant. In principle it may be possible to leverage phasing information to increase the accuracy of this approach, but phasing of rare sequenced variants is in general challenging [31, 42, 89, 99]. The LoF-segment burden approach may be seen as implicitly performing genotype imputation followed by burden testing. A recent study by Barton et al. [100] performed these steps explicitly and obtained high imputation accuracy down to MAF $\sim 5 \times 10^{-5}$ for most non-sequenced UKBB samples. This enabled a powerful association study of $N = 459{,}327$ for both common and rare variation effectively outperforming our LoF-segment and many other burden tests.

On the other hand, UKBB would gradually increase the availability of sequenced samples [101, 102], making within-cohort imputation less appealing. IBD-based imputation, in particular, depended on the set of variants to be imputed. Changing this set, e.g. from LoF to a broader class of likely pathogenic, requires new imputation and thus new testing analyses, which can be time-consuming. In contrast, performing explicit imputation [100] enables direct testing of any set of variants. Finally, our method was phenotype-specific, as we had to condition on different sets of common variants. Switching to `BOLT-LMM` [36] – the state of the art at the time – would improve things only marginally as it also supports only one phenotype per run. Testing for numerous traits would thus be laborious and this motivated me to work on a new LMM framework that will be more suitable for phenome-wide association studies; this work is presented in the next chapter.

4

The FMA framework for linear mixed model association

Contents

In this chapter I describe a new framework for linear mixed model (LMM) association testing, called Fast Mixed-model Association (FMA). I focus on empirically verifying the key LMM properties, which were discussed in Section 2.2, using simulations with genotyped variants. Then, I apply FMA to real phenotypes and imputed variants in Section 4.6 and use it to perform genealogy-based complex trait analyses in Chapter 6.

An LMM combines fixed effects, such as a set of covariates or a candidate

SNP being tested, with random effects, such as polygenic or environmental effects [36, 38, 56, 60, 70, 72]. Given N genotyped samples with phenotype vector $\mathbf{y}$, an LMM can be expressed as

$$\mathbf{y} = \beta_{\text{test}}\mathbf{x}_{\text{test}} + \mathbf{W}\boldsymbol{\gamma} + \boldsymbol{g} + \boldsymbol{e}, \tag{4.1}$$

where $\mathbf{x}_{\text{test}}$ is the N-dimensional variant we aim to test, β_{test} is its effect size (a scalar), $\mathbf{W}$ denotes the $N \times c$ matrix of covariates, $\boldsymbol{\gamma}$ represents the corresponding $c \times 1$ fixed effects (including an intercept term), $\boldsymbol{g}$ and $\boldsymbol{e}$ represent the N-dimensional genetic and environmental components, respectively. We assume that $\boldsymbol{g} \sim \mathcal{N}_N(\mathbf{0}, \sigma_g^2\mathbf{K})$ and $\boldsymbol{e} \sim \mathcal{N}_N(\mathbf{0}, \sigma_e^2\mathbf{I}_N)$, where σ_g^2 and σ_e^2 capture the genetic and environmental variance, typically normalised so that $\sigma_g^2 + \sigma_e^2 = 1$. $\mathbf{I}_N$ is an $N \times N$ identity matrix and $\mathbf{K}$ denotes the $N \times N$ genetic relatedness matrix (GRM). Assuming that $\boldsymbol{g}$ and $\boldsymbol{e}$ are uncorrelated, the variance-covariance is $\text{cov}(\mathbf{y}) = \mathbf{V} = \sigma_g^2\mathbf{K} + \sigma_e^2\mathbf{I}_N$. $\mathbf{K}$ is often estimated via $\frac{1}{M}\mathbf{X}\mathbf{X}^T$, where $\mathbf{X}$ is an $N \times M$ matrix of standardised genotypes. To keep the notation clear, I do not discriminate between the corresponding estimate $\hat{\mathbf{V}}$ and the theoretical term $\mathbf{V}$.

To detect association, we aim to test the null hypothesis $H_0 : \beta_{\text{test}} = 0$ for a candidate SNP. For that we typically compute a test statistic χ^2_{LMM} as the one defined in Eq. 2.5, which is approximately distributed as χ^2 with one degree of freedom and is closely related to the effect size β_{test} (also defined in Eq. 2.5). The inclusion of $\mathbf{V}^{-1}$ in the test statistics allows controlling for confounding due to relatedness and population stratification [26, 47, 59, 60, 62, 103, 104], and an increase in statistical power is obtained as a result of implicitly conditioning on genome-wide causal variants that may not be significantly associated [36, 47, 70] (see BLUP in section 2.2).

4.1 Overview of current approaches

LMMs for genome-wide association have been extensively optimized to improve statistical power and robustness while reducing computational costs [56, 59, 60, 70, 72, 103]. Svishcheva et al. [72] found that the fractions $\frac{\mathbf{x}_m^\top\mathbf{V}^{-1}\mathbf{x}_m}{\mathbf{x}_m^\top\mathbf{x}_m}$ are approximately constant for any genetic variant. Such a constant, referred to as *GRAMMAR-γ* or just γ, may be estimated using a small number of variants. Having the estimate $\hat{\gamma}$, the test statistics (defined in Eq. 2.5) take the form

$$\hat{\beta}_{\text{test}} = \frac{\mathbf{x}_{\text{test}}^\top \mathbf{V}^{-1}\mathbf{y}}{\mathbf{x}_{\text{test}}^\top \mathbf{x}_{\text{test}}}/\hat{\gamma},\ \chi^2_{\text{LMM}} = \frac{(\mathbf{x}_{\text{test}}^\top \mathbf{V}^{-1}\mathbf{y})^2}{\mathbf{x}_{\text{test}}^\top \mathbf{x}_{\text{test}}}/\hat{\gamma}, \tag{4.2}$$

This leads to substantial computational gains, as it allows avoiding to compute terms of the form $\mathbf{x}_{\text{test}}^\top \mathbf{V}^{-1}\mathbf{x}_{\text{test}}$ for each tested variant. Furthermore, Listgarten et al. [70] observed that fitting a candidate marker both as a fixed and random effect may result in reduced statistical power. This may be remedied by excluding the tested marker as well as variants which are in linkage disequilibrium (LD) with it when forming the GRM [70]. In the leave-one-chromosome-out (LOCO) approach [47], all markers on the same chromosome as the one tested are excluded from the GRM, thus computing association statistics as in Eq. 4.2 for all autosomal chromosomes requires 22 sets of $\mathbf{V}_{-c}^{-1}\mathbf{y}$ vectors, where each $\mathbf{V}_{-c}$ is the covariance matrix computed with all chromosomes excluding c.

More recently, several LMM association algorithms, such as `BOLT-LMM` [36, 46], `REGENIE` [39], and `fastGWA` [38, 45], introduced additional computational improvements. `BOLT-LMM` [36, 46] avoids explicit calculations involving the $N \times N$ GRM, by instead relying on numerical techniques such as the use of conjugate gradient iteration [73], which brings the overall computational cost down from $\mathcal{O}(N^2M)$ to $\mathcal{O}(N^{1.5}M)$. In addition, `BOLT-LMM` enables adopting a spike-and-slab prior, approximated as a mixture of two Gaussians, on the distribution of effect sizes. This approach, which I refer to as `BOLT:MoG`, provides improved modelling of the sparsity of effects, and offers an increase in association power. `BOLT-LMM` also allows computing association statistics using a simpler infinitesimal prior (`BOLT:Inf`), which does not require variational inference, thereby reducing computation at the cost of reduced gains in statistical power. The `fastGWA` algorithm [38, 45] is also based on the LMM described by Eq. 2.4, but instead works with a sparse version of the GRM and uses sparse matrix techniques to drastically improve computational performance. Computing such a sparse GRM requires $\mathcal{O}(N^2M)$ computation, but only needs to be performed once even when multiple traits are analysed in the same cohort. Finally, REGENIE [39] approximates the LMM framework by using a stacked ridge regression approach to estimate genetic effects. These approaches incorporate several of the previously described features: BOLT:Inf, BOLT:MoG, and fastGWA incorporate a *GRAMMAR-*γ calibration factor, while BOLT:Inf, BOLT:MoG, and REGENIE rely on the LOCO scheme. REGENIE is additionally optimized to process multiple traits in parallel.

4.2 Testing for association with FMA

In this section I describe FMA, a new framework for LMM-based association. This is similar to BOLT:Inf but makes improvements to the estimation of variance components and enables a parallel analysis of several quantitative traits. More in detail, FMA performs the following steps: (a) estimate the variance components σ_g^2, σ_e^2; (b) estimate the calibration factor γ and calculate the 22 sets of LOCO residuals $\mathbf{V}_{\text{-c}}^{-1}\boldsymbol{y}$; (c) calculate the test statistics based on Eq. 4.2. These steps are explained in the paragraphs and then summarised in Algorithm 1.

Fast variance-components estimation. FMA relies on a fast moment-based multiple variance component estimator, RHE-mc [105], which is based on the Haseman-Elston regression [106, 107]. This approach has a computational cost of $\mathcal{O}\Big(\frac{NMB}{\log_3(N)}\Big)$, where B is the number of random vectors (typically up to 50), which is drastically better than the $\mathcal{O}(N^{1.5}M)$ cost required by likelihood-based algorithms, such as BOLT-LMM. To facilitate parallel analysis of multiple traits, and in collaboration with Pazokitoroudi et al. [105], we developed an extension of RHE-mc that can handle numerous phenotypes with minimal computational overhead. As a further speed-up, we skip the calculation of standard errors, since association testing only requires to obtain point estimates for the variance components in order to build the covariance matrix.

Efficient conjugate gradient solver for multiple phenotypes. The next step for LMM-based association involves calculating the residuals $\mathbf{V}_{-c}^{-1}\mathbf{y}$ and any $\mathbf{V}_{-c}^{-1}\mathbf{x}_m$ products for the estimation of γ. Although these involve computing the inverse of $N \times N$ matrices, we may avoid the inversion by solving the system of linear equations $\mathbf{V}_{-c}\mathbf{u} = \mathbf{z}$ instead, using the conjugate gradient method [36, 73], where $\mathbf{z}$ is either a phenotype or a SNP. This procedure costs $\mathcal{O}(qNM)$, where q is the number of iterations needed for convergence, or $\mathcal{O}(N^{1.5}M)$ as q typically scales as $\sqrt{N}$ [36].

The inversion of $\mathbf{V}_{-c}$ is not the only computational bottleneck. Forming these matrices is also prohibitive for large samples as these require storing N^2 double precision values. An analysis of $N = 200\text{k}$, for instance, requires up to 300GB just for storing the covariance matrix. However, owing to the conjugate gradient method, we can solve the system of equations described above in a *matrix-free* way since the only operations involving $\mathbf{V}_{-c}$ are matrix-vector products involving the corresponding matrix $\mathbf{X}$ of standardised genotypes. The following scheme describes how we get a matrix-vector product using the GRM but without forming any intermediate matrices:

$$\begin{aligned}
\mathbf{X}\mathbf{X}^\top\mathbf{u} &= [(\mathbf{G} - \mathbf{1}_N\boldsymbol{\mu}^\top)\mathtt{diag}(\boldsymbol{\sigma})^{-1}][(\mathbf{G} - \mathbf{1}_N\boldsymbol{\mu}^\top)\mathtt{diag}(\boldsymbol{\sigma})^{-1}]^\top\mathbf{u} \\
&= (\mathbf{G} - \mathbf{1}_N\boldsymbol{\mu}^\top)\mathtt{diag}(\boldsymbol{\sigma})^{-2}(\mathbf{G}^\top\mathbf{u} - \mathbf{1}_N\boldsymbol{\mu}^\top\mathbf{u}) \\
&= \mathbf{G}\mathbf{D}\mathbf{G}^\top\mathbf{u} - \mathbf{1}_N\boldsymbol{\mu}^\top\mathbf{D}\mathbf{G}^\top\mathbf{u} - \mathbf{G}\mathbf{D}\boldsymbol{\mu}\mathbf{1}_N^\top\mathbf{u} + \mathbf{1}_N\boldsymbol{\mu}^\top\mathbf{D}\boldsymbol{\mu}\mathbf{1}_N^\top\mathbf{u}, \qquad (4.3)
\end{aligned}$$

$\mathbf{G}$ is the $N \times M$ matrix of raw allele counts (which is cheaper to store than $\mathbf{X}$), $\mathbf{1}_N$ is the $N \times 1$ vector of ones, $\boldsymbol{\mu}$ is the vector of variant means and $\boldsymbol{\sigma}$ contains the corresponding standard deviations. The key element of Eq. 4.3 is how to decompose the standardised version $\mathbf{X}$ of the genotypes into simple operations involving $\mathbf{G}$ and two matrix-vector products. The product $\mathbf{V}_{-c}\mathbf{u}$ for a LOCO covariance matrix can be formed by adding the corresponding $\mathbf{X}_t\mathbf{X}_t^\top\mathbf{u}$ terms for all $t \neq c$ chromosomes.

Previous LMM implementations leveraging conjugate gradients supported multiple right hand sides (RHS) (for LOCO residuals or SNPs for estimating γ) [36, 37] only for one phenotype. Dealing with multiple phenotypes requires different covariance matrices, since the variance components change, but the main computational bottleneck is the formation of the GRM, which is shared across all traits. With this in mind, I developed a conjugate-gradient solver which forms multiple covariance matrices on the fly, at a small memory cost which is linear to the number of traits.

More in detail, assume we analyse N samples, P phenotypes, and 22 chromosomes. Using 1 SNP per chromosome, per trait, to estimate γ, we need to solve $22 \cdot P \cdot 2$ systems of equations of size N, since each LOCO-based covariance matrix corresponds to P phenotypes and P variants. We can concatenate all the RHS vectors and form a matrix $\mathbf{U}$, then, for each chromosome t, obtain the $N \times 44P$ array $\mathbf{X}_t\mathbf{X}_t^\top\mathbf{U}$ using the expression 4.3, and subsequently update each column according to the corresponding variance component. Assuming that the genotypes are partitioned in chunks of size $M_b \times N$ (each being sufficiently small to span one chromosome), we process a chunk $\mathbf{X}_b$ at each step and calculate the product $\mathbf{X}_b^\top\mathbf{X}_b\mathbf{U}$.

Implementation in Python To increase scalability with respect to both number of variants and sample size, `FMA` streams genotypes from disk in contrast to strategies that load the whole genotype matrix or GRM in memory. This allows `FMA` to potentially scale to several millions of individuals. Streaming operations are performed using the `HDF5`[1] format, which leads to significant speed-ups compared to other file formats. Any utilisation of files in the `plink` format in Python is performed using `PySnpTools`[2].

Repeatedly reading data from disk creates significant I/O overheads which are offset in `FMA` through additional optimization, such as code vectorisation and multi-threading. Moreover, `FMA` may process several chunks of genotypes in parallel. This improves speed at the cost of higher memory footprint, but the amount of memory required grows linearly with the number of parallel processes (see Fig. A.7 for more details) making it preferable to run a few parallel processes (e.g. 4-6), especially for large data sets.

[1] https://docs.h5py.org/en/stable/

[2] https://github.com/fastlmm/PySnpTools

Algorithm 1

Genome-wide association with FMA

```
Input: genotypes; phenotypes
Preprocessing: SNP QC; Phenotype QC; LD scores; HDF5 file;
if condition to covariates then
    regress covariates from phenotypes;
end
estimate $h^2$ with multi-trait RHEmc;
select markers for $\gamma$ estimation;
initialise phenotype-specific variables for the CG iteration;
while max-tol $\geq 5 \cdot 10^{-4}$ do
    for each chunk $\mathbf{G}$ of genotypes do
        read $\mathbf{G}$ from disk;
        multiply with $\mathbf{G}$ and $\mathbf{G}^\top$ and standardise (Eq. 4.3);
        make any LOCO adjustments;
        for each phenotype do
            calculate CG residuals, directions, and new tol;
        end
    end
end
save $\mathbf{Y}_{\texttt{FMA-LOCO}}$ residuals to disk;
collect all $\mathbf{V}^{-1}\mathbf{x}$ to estimate $\gamma$ and save to disk;
if test array genotypes then
    for each chunk $\mathbf{G}$ of genotypes do
        read $\mathbf{G}$ from disk;
        calculate $\mathbf{Y}_{\texttt{FMA-LOCO}}$ and mean-center;
        for each phenotype do
            calculate test statistics according to Eq. 4.2;
        end
    end
end
if test imputed genotypes then
    run PLINK using $\mathbf{Y}_{\texttt{FMA-LOCO}}$ ;
    for each phenotype do
        convert PLINK's $t$-test to $\chi^2$;
        calibrate according to $\gamma$;
    end
end
Output: Test statistics; LOCO residuals; $\gamma$ coefficients
```

4.3 Association using multiple genetic components

The standard mixed model (2.4) can be generalized to allow for the presence of multiple variance components, each with distinct assumptions for the distribution of effect sizes. Consider a partition of the genome in K components $\mathbf{g}_1, \mathbf{g}_2, ..., \mathbf{g}_K$, of sizes $M_1, M_2, ..., M_K$ respectively:

$$\mathbf{y} = \beta_{\text{test}}\mathbf{x}_{\text{test}} + \mathbf{W}\boldsymbol{\gamma} + \sum_{k=1}^{K} \mathbf{g}_k + \mathbf{e}. \tag{4.4}$$

We assume that variants from component k follow a Gaussian distribution with mean zero and variance σ_k^2/M_k, so that $\mathbf{g}_k \sim \mathcal{N}_N(0, \sigma_k^2\mathbf{X}_k\mathbf{X}_k^\top/M_k)$, where $\mathbf{X}_k$ is the $N \times M_k$ matrix of standardized genotypes for component k, $M_1 + ... + M_K = M$, $\sigma_1^2 + ... + \sigma_K^2 + \sigma_e^2 = 1$, and $\mathbf{e} \sim \mathcal{N}_N(0, \sigma_e^2\mathbf{I})$ as before. By relying on the `RHE-mc` algorithm [37], `FMA` enables fitting models that include multiple variance components with cost $\mathcal{O}\Big(K^2(K+NB) + \frac{NMB}{\max(\log_3(N), \log_3(M))}\Big)$; note that because $N \gg K$ this cost is significantly better than the $\mathcal{O}(KN^{1.5}M)$ required to perform maximum-likelihood estimation as in `BOLT-LMM` [108] (for estimating h^2_{snp} only).

A genetic component can be any class of variants with a specific characteristic, e.g. low/high allele frequency or biological function, which are assumed to be drawn from the same underlying distribution. Here I focus on partitions of the genome according to MAF and LD scores. The single component model, which is referred to as "infinitesimal", is a special case and can be obtained by the trivial partition where $k = 1$, $M_1 = M$, and $\sigma_1^2 = \sigma_g^2$. Models with multiple genetic components have been extensively used in heritability estimation, offering less biased estimates [105, 109–111], as well as in SNP-based complex trait prediction, providing increased accuracy [112, 113].

Current approaches for LOCO-based association are limited to the use of a single variance component, whereas `FMA` allows LMM association with multiple genetic components. This is achieved with covariance matrices formed as follows:

$$\mathbf{V}_{-c} = \sigma_e^2\mathbf{I}_N + \sum_{k=1}^{K} \frac{\sigma_k^2}{M_{k\ominus c}} \sum_{t\neq c} \mathbf{X}_{k\otimes t}\mathbf{X}_{k\otimes t}^\top, \tag{4.5}$$

where each SNP belongs to a unique component and one chromosome and $k \otimes t$ denotes the intersection of component k with chromosome t. As a result, to form the covariance matrix $\mathbf{V}_{-c}$, component k consists of all M_k variants besides those in chromosome c; $M_{k \ominus c}$ denotes the cardinality of this set. In this work I consider `FMA:C1`, `FMA:C8`, and `FMA:C16` corresponding to one component, a combination of 2 MAF with 4 LD classes, and a combination of 4 MAF with 4 LD classes respectively. For `C8` I combined 2 MAF ranges, $(0.01, 0.05]$ and $(0.05, 0.5]$, with LD-score quartiles, and for `C16` I combined 4 MAF ranges, $(0, 0.028]$, $(0.028, 0.067]$, $(0.067, 0.212]$, and $(0.212, 0.5]$, with LD-score quartiles [105].

4.4 Setup for evaluation with synthetic phenotypes

I performed extensive simulations to evaluate the computational and statistical properties of `BOLT-LMM` (using both `MoG` and `Inf` algorithms) [36, 46] `REGENIE` [39], `fastGWA` [38], `FMA`, and linear regression as implemented in PLINK [114] (`LinReg`). I used $M = 387{,}700$ SNP array variants from chromosomes 1 to 10 of UKBB and extracted four sets of $N = 50{,}000$ individuals, selected based on ancestry and relatedness described by Bycroft et al. [4]. The first set, referred to as “UWB”, included 50k randomly selected unrelated White British (self-reported [4]). A second set, “RWB”, included $N = 25$k samples from the UWB set and $N = 25$k individuals sharing close familial relationships, with an average pairwise kinship coefficient of 0.11 [4]. The third, “EUR”, comprised $N = 25$k samples from the first set and $N = 25$k European individuals who did not self-report as British [4]. Finally, I analysed a fourth set, “CSR”, involving continental structure and relatedness, consisting of $\sim$ 34k British, 9k Europeans, and 7k individuals of African or East Asian ancestries, where 9k pairs of individuals had an average kinship coefficient of 0.10.

I generated sets of 50 synthetic phenotypes with narrow-sense heritability $h^2 = 25\%$. I selected either 4,000 or 19,000 causal variants from the odd chromosomes, corresponding to $\sim$1% or $\sim$5% of all variants, while variants on even chromosomes had no contribution to the phenotype [36, 38, 39]. I simulated population stratification for the EUR and CSR samples by introducing correlation between the top-10 principal components of ancestry (computed by Bycroft et al. [4]) and the phenotype. In particular, the phenotypes were synthesized as

$$\mathbf{y} = \sum_{j=1}^{m} \beta_j \mathbf{g}_j + \sum_{t=1}^{10} \gamma_j \mathbf{s}_j + \mathbf{e} \tag{4.6}$$

for m casual variants, where $\mathbf{g}_j$ is SNP-j, $\mathbf{s}_t$ represents the top-t principal component, and $\mathbf{e}$ represents environmental effects. For the effect sizes, I followed the setup of Pazokitoroudi et al. [105] to obtain realistic genetic architectures by adjusting according to MAF and LD-score, so that $\beta_j \sim \mathcal{N}(0, \frac{h^2 w_j}{m}(2f_j(1-f_j))^{-0.75})$ [105, 115] where w_j and f_j are respectively the LD-score (average LD within a 200kb window) and allele frequency of variant j (in-sample). Finally, the weights γ_t followed an exponential decay, adjusted so that the total contribution of stratification was 5% of the phenotypic variance (only for the EUR and CSR samples).

I measured efficacy in controlling for population stratification by running different methods both with and without conditioning on the top-5 principal components of ancestry. Components 6 to 10 were not included as fixed effects, with the goal of measuring the robustness to residual stratification that is not captured by the top PCs (approximately 1.5% of phenotypic variance). When running `FMA`, I used 50 random vectors to estimate heritability with `RHE-mc`, using either a single variance component (`FMA:C1`), or multiple components based on MAF and LD as in [105] (defined in the previous section). I ran `REGENIE` using the `-lowmem` flag to keep memory requirements low while storing intermediate files to disk, and `-bsize 1000`, as recommended [39]. Because `BOLT:MoG`, `BOLT:Inf`, and `fastGWA` are not optimized for parallel trait analysis, I ran each sequentially for all traits. All methods were ran on an Intel Skylake 2.6 GHz CPU architecture, providing access to 5 CPU cores and up to 75 GB of RAM.

4.5 Results from simulations

4.5.1 Statistical power to detect association

I first assessed the statistical performance, measured in terms of percentage increase in average χ^2 values compared to LinReg, by applying each method to a sample of 50k unrelated British, and the two types of genetic architecture. The results are summarised in Fig. 4.1, where I show the performance gain over LinReg, and numerical values, including mean test statistics and statistical power, are reported in Tables B.1 and B.2. BOLT:MoG clearly outperformed all the other methods, with the largest gain observed in the scenario involving 1% polygenicity, demonstrating the advantage of non-infinitesimal modelling under sparse genetic architectures [36, 38, 46]. FMA:C8, FMA:C16, BOLT:Inf, and REGENIE all had similar performance when the genetic architecture was sparse (power $12.4-12.5\%$), but REGENIE showed slightly less power when polygenicity was set to 5% (Fig. 4.1). fastGWA achieved the least power and in fact behaved similarly to linear regression. This could be explained by the lack of relatedness in the sample which leads to the sparse GRM being almost a diagonal matrix.

This scenario only includes unrelated samples of homogeneous ancestry, therefore we expect to observe controlled type I error rates. Indeed, FMA:C8, FMA:C16, BOLT:Inf, REGENIE and LinReg all consistently yielded controlled error rates (Tables B.1,B.2). For similar reasons, running REGENIE, fastGWA, or LinReg with covariates did not make any noticeable differences. Finally, BOLT:MoG had significantly elevated error rates (0.051, p-value$< 1 \times 10^{-4}$) for the 1% polygenicity case, but the inflation was minor and may be due to the overestimation in h^2 (see Fig. 4.3 and Section 4.5.3).

4.5.2 Robustness to relatedness and population stratification

Family structure and cryptic relatedness can inflate test statistics [36, 59, 60], so next, I benchmarked association algorithms in scenarios that include the presence of related individuals. I considered 50k British samples this time including 12,500 pairs of related individuals and measured type I error rates by computing the

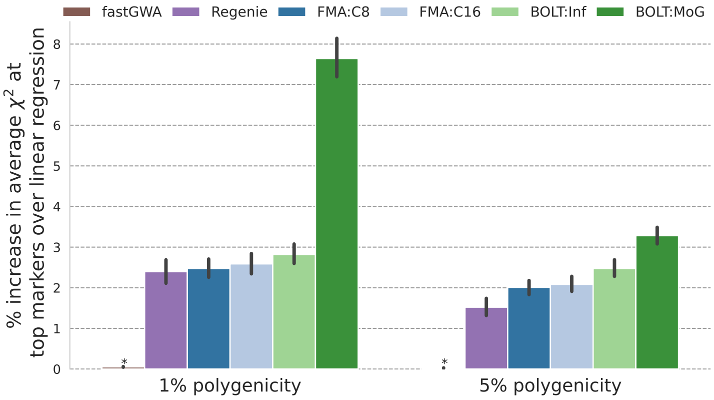
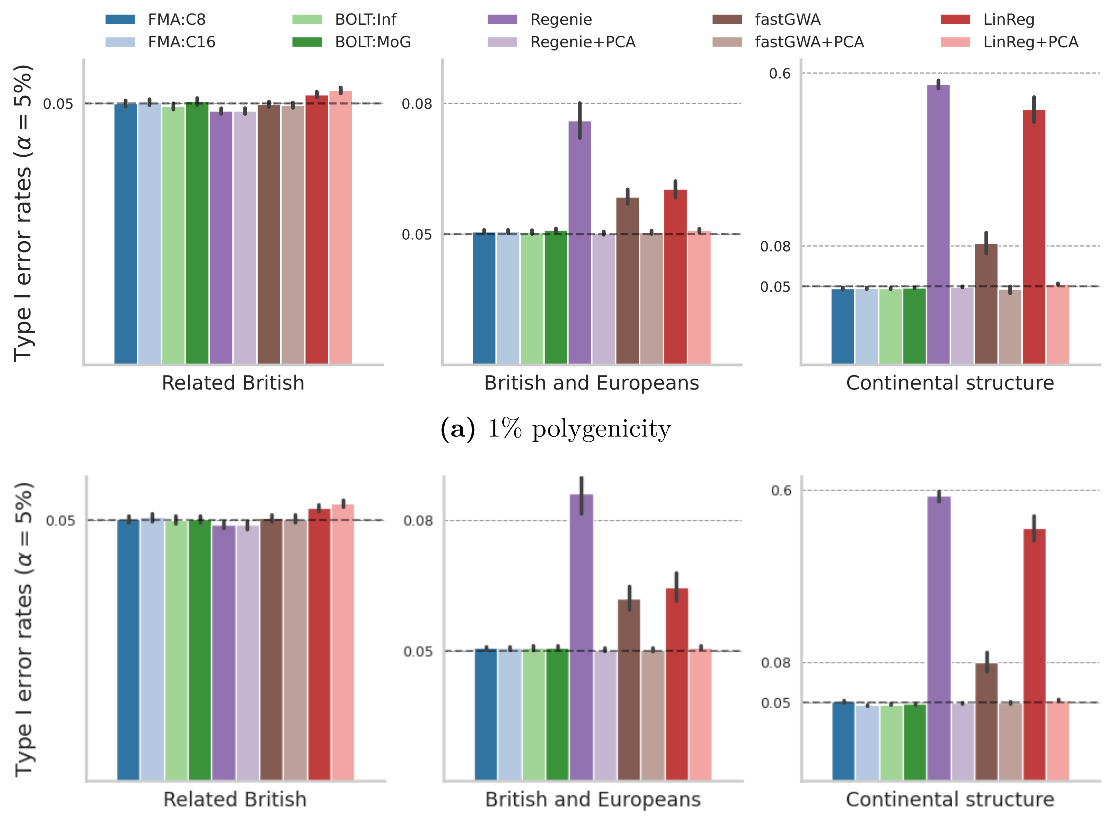


Figure 4.1: Benchmarking of statistical power in 50k UWB samples and synthetic phenotypes. I measure power as the increase (i.e. ratio) over Linear Regression in average test statistic at the causal markers with the highest effects (roughly 234 and 895), for each of the 50 replicates. The results for the unrelated British samples (UWB) and the two types of architecture I investigated (1% or 5% polygenicity; $h^2 = 0.25$) are plotted here. Similar trends were observed by looking at the average test statistic at top inferred variants or the total number of significant loci.

fraction of null variants identified as causal using 5% as threshold. To determine if an error rate was significantly larger than 5%, I used 4.5×10^{-3} as a threshold accounting for multiple testing and the 11 methods in the benchmarking. As Figure 4.2 (left column) and Tables B.3,B.4 show, every LMM approach was adequately calibrated, whereas `LinReg` was always significantly inflated (type I errors ranged in $0.051 - 0.052$). This inflation was not fixed after adding covariates, an observation which agrees with previous studies demonstrating that principal components of ancestry are not sufficient to control for relatedness [36, 38, 39, 47, 51].

I next sought to investigate which association methods are robust to population stratification, using the EUR sample consisting of 25k British and 25k non-British Europeans. I ran `FMA:C8`, `FMA:C16`, `BOLT:Inf`, and `BOLT:MoG` without any covariates, and `REGENIE`, `fastGWA`, and `LinReg` both with and without covariates. The results are summarised in Figure 4.2 (middle column) and Tables B.5 and B.6. I found that proper covariance-based methods, e.g. `FMA:C8` and `BOLT:Inf`, retained controlled type I error rates in both cases of polygenicity. On the other hand, approximate mixed-model approaches, i.e. `REGENIE` and `fastGWA`, required the use of covariates in order to achieve the same calibration. In line with previous



(a) 1% polygenicity

(b) 5% polygenicity

Figure 4.2: Type I error performance on synthetic phenotypes. Type I error rates are calculated as the proportion of falsely determined causal out of all ∼ 196k non-causal variants within even chromosomes, at a nominal level of 5%. Showing are results for the sets of related British (RWB), British and Europeans (EUR), and a combination of European, African and South-East Asian samples including relatives (CSR), for the case of **(a)** 1% and **(b)** 5% polygenicity. `FMA`, `BOLT:Inf`, and `BOLT:MoG` were invoked without any covariates, whereas `REGENIE`, `fastGWA`, and linear regression were invoked either without covariates, or after conditioning on the top-5 principal components of ancestry. See Tables B.3-B.8 for more details.

studies comparing LMMs and PCA when correcting for population stratification [26, 47, 51, 64], `LinReg` with covariates on the EUR cases was better than without but still inflated, indicating confounding from residual stratification (EUR). This was expected since I used the top 10 components to simulate stratification and only up to 5 to account for that.

As a final experiment, I assessed the aforementioned methods in the CSR sample comprising both population stratification and relatedness. The results are illustrated in Fig. 4.2 (right column) with more details following in tables B.7

and B.8. As before, FMA, BOLT:Inf, and BOLT:MoG all yielded controlled error rates, in contrast to fastGWA and REGENIE which required the use of covariates and were inflated otherwise. LinReg was significantly inflated even after conditioning on principal components, as expected given the presence of related individuals in the sample. Moreover, REGENIE without covariates was severely confounded with error rates reaching 56%. This may be caused by overfitting of the Ridge predictors to the underlying differences in LD patterns [39]. Overall, these results highlight the importance of using both an accurate GRM and the LOCO feature in LMM-based association.

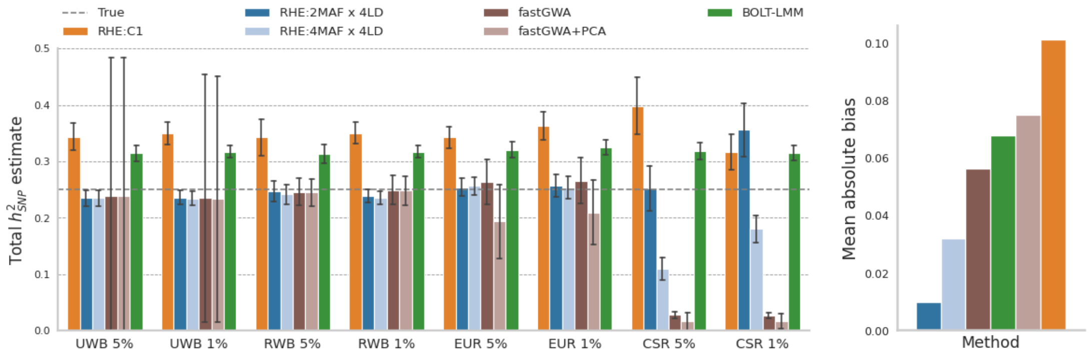
4.5.3 Biased σ_g^2 estimation and implication in association

I investigated the efficacy of several approaches for estimating narrow-sense heritability, equivalently σ_g^2, which is an intermediate step for several LMM algorithms for association, given the samples of increasing structure described above. As presented previously, I considered two multiple variance-components (VC) approaches: 2 MAF classes coupled with 4 LD classes (RHE:2MAF x 4LD, or C8) and 4 MAF coupled with 4 LD (RHE:4MAF x 4LD, or C16). I compared these with single-component methods, namely BOLT-LMM and fastGWA, both of which employ restricted maximum likelihood (REML) algorithms, and also assessed RHE-mc using one component containing all the available variants excluding the *HLA* region (RHE:C1, as in [105]).

Figure 4.3 shows the estimates for all 8 scenarios. RHE:2MAF x 4LD and RHE:4MAF x 4LD produced similar estimates and were unbiased in most cases, while keeping small standard errors, in contrast to single component approaches which showed systematic biases. Overall, and considering the mean absolute bias, RHE:C8 was the least biased method, followed by RHE:C16. fastGWA yielded unbiased estimates for the cases of RWB and EUR, but was unstable in the UWB and CSR cases as it either had high variance (estimates ranging (0.0,0.7) while the true value was 0.25) or repeatedly under-estimated heritability (estimates ≤ 0.04). Moreover, BOLT-LMM and RHE:C1 over-estimated heritability in every scenario, with the latter yielding an average bias of roughly 0.10. Interestingly, BOLT-LMM yielded a consistent overestimation, with estimates in the range $[0.275 - 0.359]$. This could be a result of mis-specifying the contribution of common and rare variants by assuming the same distribution for all variants, and agrees with a similar study by Jiang et al. [38] (see their Ext. Fig. 4) showing that `BOLT-LMM` yields a similar bias irrespectively of environmental effects. To further investigate the differences between single and multiple component approaches, I used `BOLT-REML` [3] [108] and found that using the `2MAF x 4LD` annotation indeed yields less biased estimates than using a single component (Fig. A.10). This observation is consistent with previous studies demonstrating the benefits of using multiple VC for avoiding model-misspecification in h^2 estimation [105, 110, 111, 115–117], e.g. when the genetic architecture depends on MAF or LD as is the case here.

The aforementioned biases might have implications in statistical power for association. For instance, `FMA:C1`, which was based on heavily over-estimated σ_g^2s, yielded elevated type I error rates in most cases (Tables B.1-B.6). `BOLT:MoG`, which was based on modestly over-estimated σ_g^2s, also had inflated error rates for UWB or EUR with low polygenicity. This is in accordance with a separate experiment I performed to assess `FMA`'s robustness to biased VC estimates, illustrated in Fig. 4.4. This showed that test-statistics inflation is proportional to the bias in σ_g^2 estimation, with the inflation being less than 1% for as long as bias was up to 0.10.

[3] this software is part of `BOLT-LMM` but specialised for narrow-sense h^2 estimation



(a) h^2_{SNP} estimation in each of the 8 synthetic cases

(b) Total bias

Figure 4.3: h^2_{SNP} estimation in UKBB samples and synthetic phenotypes. (a) Estimates for SNP-based heritability by six methods, averaging for the 50 replicates of each case, with error bars showing the standard deviation of the 50 estimates. I analysed sets of 50k samples that were unrelated British (UWB), British including pairs of related individuals (RWB), British and Europeans (EUR), and a combination of European, African and South-East Asian samples including relatives (CSR). I used chromosomes 1-10 ($M = 387{,}700$) and either 1% or 5% polygenicity, corresponding to roughly 4,000 or 19,000 variants. **(b)** Mean bias for each of the six methods estimating h^2_{SNP}, calculated as the mean absolute error between the true value (0.25) and the mean estimate (as in the left panel).

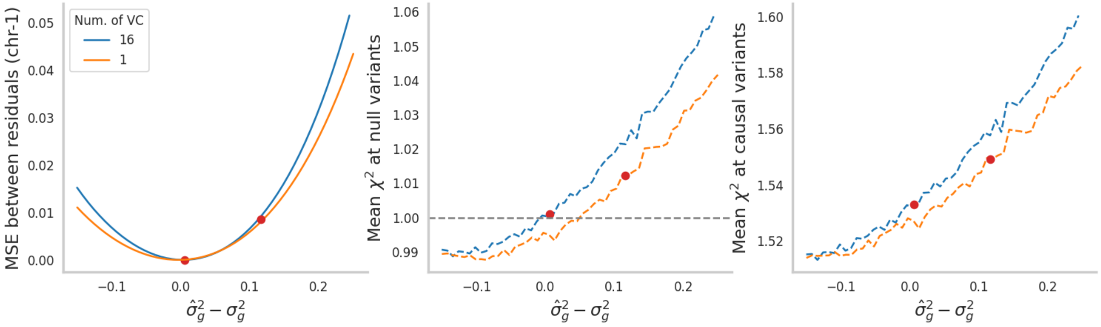


Figure 4.4: Robustness of FMA to biased σ^2_g estimates. Left: I compare the LOCO residuals obtained using perturbed variance-component (VC) estimates to those corresponding to the original case, where the true σ^2_g was 0.25, and plot the mean of the squared differences. Next, I calculate test statistics for each case and measure inflation as the average χ^2 at 196k non-causal variants (**centre**) and at 19k causal variants (**right**). Red markers indicate the estimates obtained by RHE-mc.

4.6 Analysis of 20 real phenotypes

To demonstrate the applicability of `FMA` to real studies, I focused on a set of 20 quantitative phenotypes with high phenotyping rates and various levels of heritability, a summary of which is given in Table 4.1. I explored subsets of UKBB genotypes ranging from 50,000 individuals (consisting only of unrelated British) to 446,050 (consisting of all Europeans); the smaller sets were mainly used for a computational comparison. For all methods requiring model fitting, I used a fixed set of 623,128 SNPs, consisting mostly of those variants with frequency higher than 1% (possibly lower on the smaller subsets), stored in the bed/bim/fam format [49]. Testing was performed both on that set of array data (for all subsets), and on the imputed genotypes provided by UKBB (version 3, based on the HRC+UK10K reference panel [4, 32]) using 30 and 38.5 million variants for $N = 50$k and $N = 446$k, respectively.

I ran each method while conditioning on sex, age, age^2 and 20 principal components of ancestry as covariates, using 12 computational cores and up to 150GB of memory, using any recommended options. In particular, I ran `REGENIE` using `-bsize 1000` and `-lowmem`, as recommended for model fitting, and then `-bsize 400, -ref-first, -minMAC 5` and `-minINFO 0.50` for testing. `FMA` used 6 parallel processes and one variant per chromosome for estimating the calibration factors. Regarding the imputed genotypes, `BOLT:MoG` and `REGENIE` where applied to the original data in the `bgen` format [4]. `PLINK` v2 [114] was restricted to the sample of $N = 328$k unrelated British (as a baseline), invoked with `-glm`, `-mach-r2-filter 0.50`, `-mac 5`, after converting the data to the `pgen/psam/pvar` format. This format was also used for `fastGWA`, which was invoked with similar arguments.

To obtain `FMA` test statistics on imputed genotypes, after model fitting, I first ran `PLINK` using the corresponding LOCO residuals $\mathbf{V}_{-c}^{-1}\mathbf{y}$ as phenotypes (one set per chromosome, per trait), and then used Python to convert `PLINK`'s t-test values to χ^2 and adjust according to each calibration factor (as in Eq. 4.2). This approach is similar to `fastGWA` which, by default, avoids the explicit conditioning on covariates to reduce the runtime of step-2 in association [38, 45].

Phenotype	UKBB code	% missing	$\hat{h}^2_{\text{snp}}$
Body mass index	21001	0.30	0.2992
Diastolic blood pressure	4079	5.98	0.1454
Eosinophil count	30150	0.17	0.1977
Eosinophil percentage	30210	0.17	0.2008
Forced vital capacity	3062	8.15	0.1649
Glycated haemoglobin	30750	4.31	0.1419
HDL cholesterol	30760	11.86	0.2175
Mean corp haemoglobin	30050	0.00	0.2457
Mean corp volume	30040	0.00	0.2928
Mean platelet volume	30100	0.00	0.4344
Mean sphered cell volume	30270	1.63	0.2447
Monocyte count	30130	0.17	0.1562
Platelet count	30080	0.00	0.3421
Platelet crit	30090	0.05	0.2972
Platelet distr width	30110	0.05	0.2694
Red blood cell count	30010	0.00	0.2741
Red blood cell distr width	30070	0.00	0.1508
Systolic blood pressure	4080	5.99	0.1500
Total cholesterol	30690	3.97	0.1470
White blood cell count	30000	0.00	0.1885

Table 4.1: List of 20 real phenotypes analysed. I selected traits based on phenotypic rates and popularity among other studies and mention the UKBB field codes for reference. The third column reports the fraction of missing values for the $N = 446{,}050$ sample, which is the total number of individuals included in the analysis. I also give the estimates of total heritability obtained by `RHE-mc` [105] using the annotation with 2MAF × 4LD components; the variance-components estimates are shown in Fig. A.11.

4.6.1 Results based on genotyped variants

Figure 4.5 gives a summary of each method's performance measured by the gain in χ^2 statistics over `LinReg`, which can be seen as a surrogate for statistical power [36, 46], computed using the 623k genotyped variants I used for model fitting. These results follow very similar trends to those observed previously with simulations. In particular, `BOLT:MoG` achieved the highest mean χ^2, followed by `FMA:C8`, `BOLT:Inf`, and `REGENIE`. `fastGWA` behaves similarly to `LinReg` when $N = 50$k, $N = 164$k, or $N = 328$k, as these samples consist of unrelated individuals resulting in a sparse GRM which is effectively diagonal. `LinReg` is expected to be confounded by relatedness in the $N = 446$k sample, in contrast to the LMM approaches, therefore

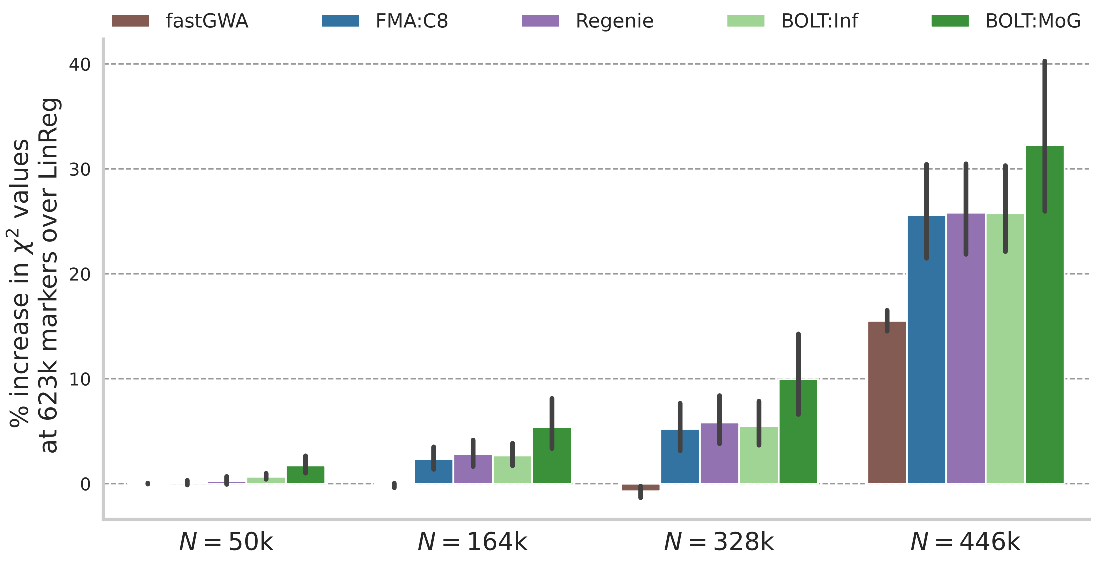


Figure 4.5: Percentage of increase in χ^2 values over linear regression for 20 real quantitative phenotypes. I compare the increase in χ^2 values over `LinReg` across $M = 623{,}128$ genotyped variants – the same as those used for model fitting – and bars correspond to standard deviations for the 20 phenotypes. The first three samples consist of unrelated British individuals, and the fourth one contains all Europeans (UKBB). I report pairwise comparisons in Table B.9.

for this case I compared with the findings of the $N = 328$k sample, following Loh et al. [46]. For most methods, the gain over `LinReg` due to the phenotype residualisation implicitly performed by LMMs was proportional to sample size, in accordance to previous studies [36, 47, 75]. Similar results are observed when considering the most significant associations, as shown in Fig. A.12.

I tested which methods achieved significantly higher performance than others by comparing the average test statistic at top variants with paired t-tests (Table B.9). `BOLT:MoG`, `BOLT:Inf`, `FMA:C8`, and `REGENIE` obtained significantly higher χ^2 statistics than both `fastGWA` and `LinReg` ($p < 0.05/15$). No other comparisons yielded significant differences, after adjusting for multiple testing of 15 pairs. For instance, `REGENIE` had a marginally higher average χ^2 than `FMA:C8` (238.36 vs 235.37; $p = 0.036$) but the p-value did not pass the adjusted threshold for significance.

4.6.2 Results based on imputed variants

I next applied `FMA:C8`, `REGENIE`, `fastGWA`, and `LinReg` to all 20 phenotypes and 38.5 million imputed variants, using $N = 446$k samples of European ancestry for all methods except `LinReg`, which was restricted to $N = 328$k unrelated British samples to avoid confounding due to relatedness [46, 51]. I also ran `BOLT:MoG` but only for a computational assessment. A summary of this experiment is given in Table 4.2, including average test statistics, number of associations, and LDscore regression [76] intercepts.

To assess if any method is inflated, due to relatedness or residual stratification, I compared the LDscore regression attenuation ratios (AR) with those by `LinReg` on the set of unrelated British, following Loh et al. [46]. In particular, I ran LDscore regression [76] using LD scores calculated from the 1000 Genomes reference panel (Phase 3) [3], the "baselineLD" model [109] (which stratifies variants according to 96 annotations), and regression weights from Hap Map 3 [2], resulting in about 1 million variants. I observed no significant differences (Figure 4.7), implying that all mixed-model approaches are similarly calibrated to `LinReg` restricted to the smaller sample. For instance, the mean AR for `FMA:C8` was 0.0901 (0.0053), the one for `LinReg` was 0.0893 (0.0072), and the difference was not significantly different from 0. This was also the case for `REGENIE` and `fastGWA`, having an average AR of 0.0876 (0.0053) and 0.0884 (0.0061), respectively.

Next, I assessed the rate at which associated variants replicate in an independent cohort using summary statistics obtained from Biobank Japan (BBJ) [118]. I looked at the number of genome-wide significant variants ($p < 5e \times 10^{-8}$) and the number of loci (considering windows of 100,000 base pairs) which had nominal significance in BBJ ($p < 0.05$). Among the 13 phenotypes for which summary statistics were available (of the 20 analysed), `FMA:C8`, `REGENIE`, and `fastGWA` replicated 19,224, 19,101, and 16,999 variants respectively (the average ratios were 40.8%, 40.8%, and 42.2% respectively; see Table 4.2). In terms of independent loci, `FMA:C8` replicated a total of 4,810, whereas `REGENIE` and `fastGWA` had a smaller total of 4,755 and

4,723 respectively. Although this difference is not significant, it shows that FMA can have performance better, or at least comparable, to the state of the art.

Phenotype	Method	Mean χ^2	LDSC intercept	Atten ratio	Loci repl/ed	SNPs repl/ed ratio
BMI	fastGWA	3.13	1.120	0.057	642	44.7%
BMI	FMA:C8	3.25	1.115	0.051	612	44.4%
BMI	Regenie	3.29	1.102	0.045	570	45.5%
Diastolic blood pressure	fastGWA	2.15	1.057	0.049	228	33.4%
Diastolic blood pressure	FMA:C8	2.16	1.079	0.068	242	33.8%
Diastolic blood pressure	Regenie	2.21	1.068	0.056	247	34.2%
Eosinophil count	fastGWA	2.32	1.088	0.067	298	29.3%
Eosinophil count	FMA:C8	2.36	1.104	0.077	323	28.7%
Eosinophil count	Regenie	2.43	1.108	0.076	320	28.6%
Eosinophil percentage	fastGWA	2.36	1.091	0.066	-	-
Eosinophil percentage	FMA:C8	2.41	1.099	0.070	-	-
Eosinophil percentage	Regenie	2.49	1.105	0.070	-	-
Forced vital capacity	fastGWA	2.52	1.119	0.078	-	-
Forced vital capacity	FMA:C8	2.40	1.116	0.083	-	-
Forced vital capacity	Regenie	2.57	1.115	0.073	-	-
Glycated haemoglobin	fastGWA	2.05	1.127	0.122	230	32.3%
Glycated haemoglobin	FMA:C8	2.06	1.137	0.130	232	31.7%
Glycated haemoglobin	Regenie	2.11	1.134	0.121	245	31.1%
HDL cholesterol	fastGWA	2.57	1.139	0.088	334	36.6%
HDL cholesterol	FMA:C8	2.69	1.124	0.073	294	34.2%
HDL cholesterol	Regenie	2.89	1.164	0.087	321	32.9%
Mean corp haemoglobin	fastGWA	2.39	1.177	0.128	370	53.9%
Mean corp haemoglobin	FMA:C8	2.51	1.198	0.131	391	51.6%
Mean corp haemoglobin	Regenie	2.59	1.188	0.119	381	51.0%
Mean corp volume	fastGWA	2.70	1.237	0.139	492	53.3%
Mean corp volume	FMA:C8	2.93	1.255	0.132	467	49.7%
Mean corp volume	Regenie	3.01	1.251	0.125	458	49.7%
Mean platelet volume	fastGWA	3.24	1.236	0.105	-	-
Mean platelet volume	FMA:C8	3.83	1.276	0.098	-	-
Mean platelet volume	Regenie	3.96	1.294	0.099	-	-
Mean sphered cell volume	fastGWA	2.52	1.210	0.138	-	-
Mean sphered cell volume	FMA:C8	2.68	1.205	0.122	-	-
Mean sphered cell volume	Regenie	2.75	1.217	0.124	-	-
Monocyte count	fastGWA	1.94	1.086	0.091	306	44.9%
Monocyte count	FMA:C8	1.95	1.098	0.103	304	44.1%
Monocyte count	Regenie	2.01	1.108	0.107	298	44.3%
Platelet count	fastGWA	3.11	1.196	0.093	587	45.6%
Platelet count	FMA:C8	3.46	1.238	0.097	607	42.1%
Platelet count	Regenie	3.53	1.227	0.090	600	42.2%
Platelet crit	fastGWA	2.78	1.204	0.115	-	-
Platelet crit	FMA:C8	3.00	1.208	0.104	-	-
Platelet crit	Regenie	3.07	1.225	0.109	-	-
(continued next)						

		Mean	LDSC	Atten	Loci	SNPs
Phenotype	Method	χ^2	intercept	ratio	repl/ed	repl/ed ratio
Platelet distr width	fastGWA	2.72	1.123	0.071	-	-
Platelet distr width	FMA:C8	2.88	1.143	0.076	-	-
Platelet distr width	Regenie	3.04	1.140	0.069	-	-
RB cell count	fastGWA	2.82	1.170	0.093	450	44.6%
RB cell count	FMA:C8	2.98	1.174	0.088	506	42.2%
RB cell count	Regenie	3.08	1.188	0.090	512	42.3%
RB cell distr width	fastGWA	2.08	1.050	0.046		-
RB cell distr width	FMA:C8	2.13	1.061	0.054	-	-
RB cell distr width	Regenie	2.17	1.061	0.052	-	-
Systolic blood pressure	fastGWA	2.19	1.091	0.076	245	41.8%
Systolic blood pressure	FMA:C8	2.19	1.091	0.077	260	41.8%
Systolic blood pressure	Regenie	2.25	1.094	0.075	253	41.2%
Total cholesterol	fastGWA	1.93	1.069	0.074	180	45.8%
Total cholesterol	FMA:C8	1.97	1.080	0.082	193	45.1%
Total cholesterol	Regenie	2.03	1.093	0.091	193	44.5%
WB cell count	fastGWA	2.20	1.086	0.072	361	42.4%
WB cell count	FMA:C8	2.22	1.105	0.086	379	41.6%
WB cell count	Regenie	2.26	1.097	0.077	357	42.3%
Averages	fastGWA	2.49	1.134	0.088	363	42.2%
	FMA:C8	2.60	1.145	0.090	370	40.8%
	Regenie	2.69	1.149	0.088	366	40.8%

Table 4.2: GWAS summary of 20 phenotypes and $N = 446$k. This is a summary of the results by applying `FMA:C8`, `REGENIE`, and `fastGWA` to $N = 446{,}050$ UKBB samples of European ancestry, using $M = 623{,}128$ genotyped variants for model-fitting, and testing on 38.5 million markers (22 autosomes). I report the mean χ^2 statistics, the LD score regression intercept, and the corresponding attenuation ratio, after selecting about 800k SNPs for which LD scores were available, using the "baselineLD" annotation [36, 76]. The last two columns show the fractions for the number of loci (after clumping) and the ratio of single variants that were replicated in Biobank Japan, for the 13/20 phenotypes for which summary statistics were freely available [118]. The corresponding UKBB codes are given in Table 4.1.

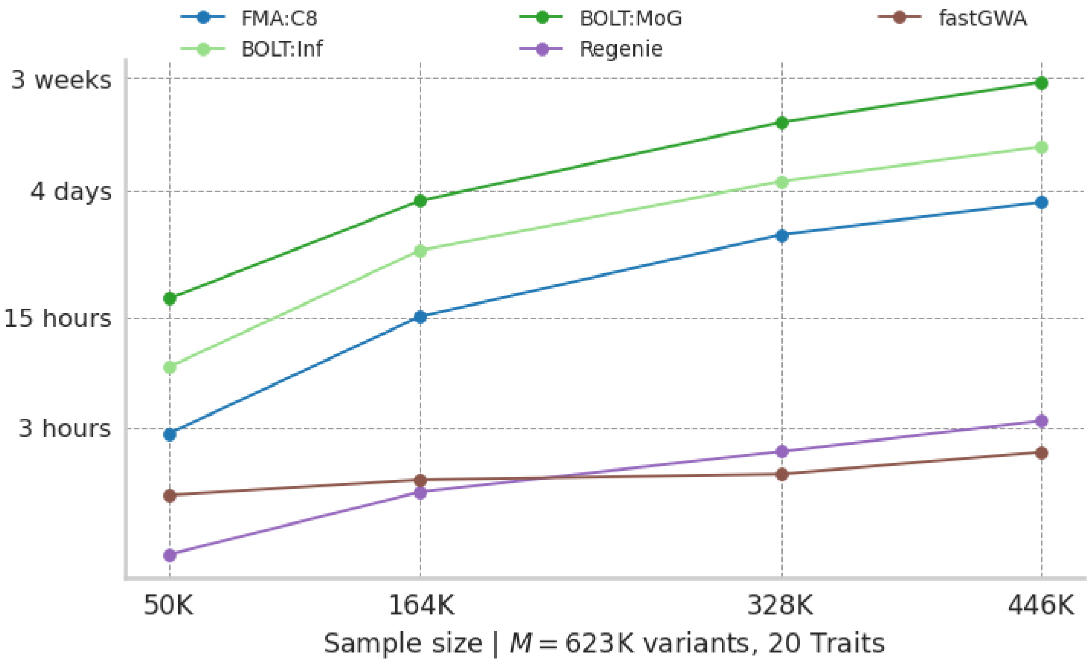


Figure 4.6: Time complexity for model fitting in GWAS. I illustrate the duration for the model-fitting step of each method (without testing), considering 20 real quantitative phenotypes, $M = 623{,}128$ variants, and samples of increasing size. I used 12 CPU cores and up to 150GB of RAM. For `BOLT:Inf`, `BOLT:MoG`, and `fastGWA`, which process each trait separately, I considered the total running time of the 20 jobs (one for each phenotype); the cost for building the GRM for `fastGWA` is not included.

4.7 Computational benchmarking

Figure 4.6 summarises the running times for model fitting (step-1 of association) of the 20 real phenotypes. A similar comparison for the 50 synthetic phenotypes is reported in Fig. A.8. `REGENIE` achieves the fastest training for $N \leq 164$k, but requires more time than `fastGWA` for larger samples. Note, however, that I do not account for the cost of preparing `fastGWA`'s sparse GRM, which is a computationally intense pre-processing step (i.e. 74 hours for $N = 446$k). `FMA:C8` comes next achieving a significant speed-up over `BOLT:Inf`, e.g. 81 vs 183 hours for $N = 446$k, owing to the support of multiple phenotypes which compensates for the time lost in repeatedly reading data from disk. `BOLT:MoG` is the slowest method with orders of magnitude higher times, e.g. almost 3 weeks for $N = 446$k. These patterns resemble the underlying implementation of each algorithm. `REGENIE` performs a single pass over the genotypes (and then another on the intermediate

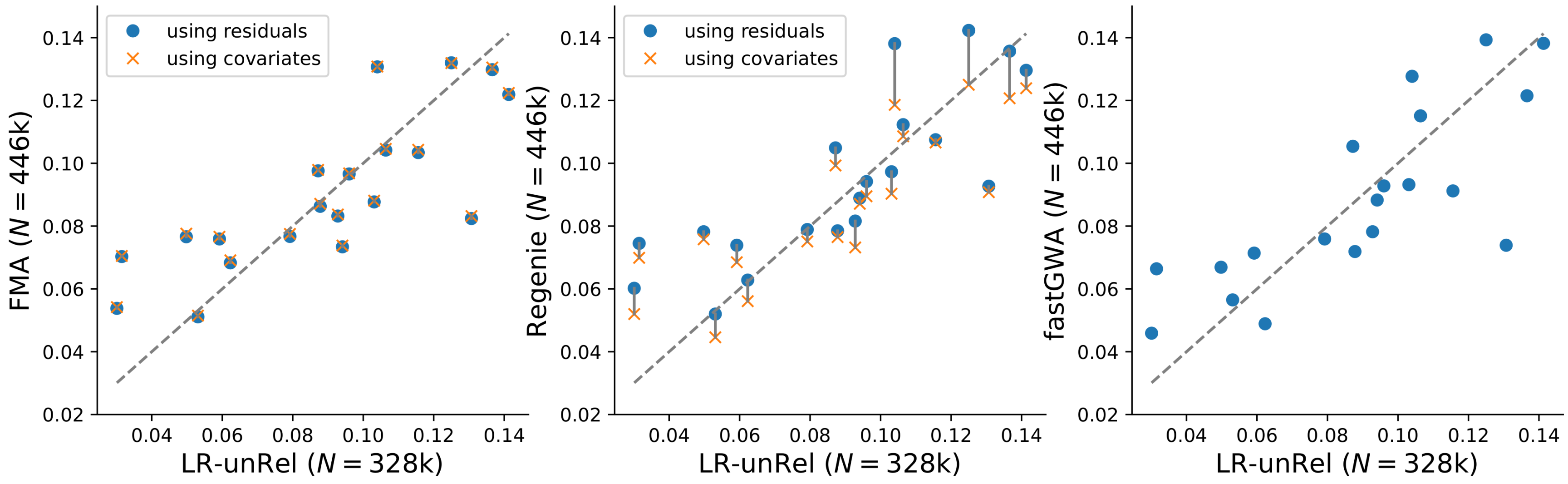


Figure 4.7: LDScore regression attenuation ratios for 20 real quantitative phenotypes. I applied LDScore regression [76] on ~ 1 million HapMap 3 variants with MAF$\geq 1\%$, and variants for which LD-score estimates were provided (based on 1000G European samples), using the “baselineLD” annotation [109], similarly to Loh et al. [36]. I compare each method with `LinReg` run on 328k unrelated British samples. For `FMA REGENIE`, attenuation ratios were calculated using two approaches to condition on covariates: directly including covariates while testing, or using phenotypes that were pre-residualized by regressing out covariates before testing. Note that `fastGWA` works with residualised phenotypes by default [38].

Sample	Method	Step-1 (h)	Step-2 (h)	Total (h)	RAM (GB)
50k	`FMA:C8`	2.6	5.0	7.6	12
	`BOLT:MoG`	19.7	55.9	75.6	9.1
	`REGENIE`	0.5	7.4	7.9	2.5
	`fastGWA`	1.5	2.1	3.6	5.3
	`LinReg`	-	5.9	5.9	15
446k	`FMA:C8`	81	29	110	85
	`BOLT:MoG`	445	742	1187	69
	`REGENIE`	5	154	159	13
	`fastGWA`	4	43	47	14
	`LinReg`$_{N=328k}$	-	134	134	15

Table 4.3: Computational benchmarking for large-scale GWAS. I compare the computational resources needed for the analysis of 20 real quantitative phenotypes (Table 4.1) using either N = 50,000 or N = 446,050 UKBB samples, and testing for 30m or 38.5m imputed variants respectively. I used M = 623,128 genotyped variants for model fitting (step-1) and each method was given 12 CPU cores and sufficient memory or disk space. I note that for `FMA:C8` I used `PLINK` v2 ([114]) to test for imputed genotypes in the `pgen` format (Supplementary Notes); times for $N = 446$k are rounded to the nearest integer for clarity; `fastGWA` required 74 hours to prepare the sparse GRM. Times for $N = 446$k are rounded to the nearest integer for clarity.

results), whereas covariance-based approaches – `fastGWA`, `FMA:C8`, `BOLT:Inf`, and `BOLT:MoG` – employ iterative algorithms and the number of iterations is proportional to the sample size. This effectively results in a complexity of $\mathcal{O}(N^{1.5}M)$ and the corresponding differences can be explained by the intermediate steps, such as the use of LOCO, how/if the GRM is formed, etc.

The next step in association, calculating test statistics, can be more demanding than step-1 when millions of variants are considered. This is emphasized in Table 4.3 where I present each method's total cost for large-scale studies. For $N = 446$k, `fastGWA` was the fastest method overall requiring a total of 47 hours. Interestingly, `FMA:C8` was the one achieving the shortest step-2 with 29 hours, which can be explained by two factors. First, `FMA:C8` avoids the explicit condition on covariates during testing, as is the case for `REGENIE` and `fastGWA`, and instead applies `PLINK` on the LOCO residuals. Second, `FMA:C8` makes a single pass over the imputed genotypes, instead of reading each file once for each phenotype, as is the case for `fastGWA`. `BOLT:MoG` could, in theory, complete step-2 in a similar duration, but I followed the recommended guidelines suggesting the explicit use of principal components as covariates for reduced running time [46]. The same ranking was observed in the $N = 50$k sample.

To ensure fairness, I ran `REGENIE` similarly to `FMA:C8` or `fastGWA`, without explicitly conditioning on covariates. As expected, this was faster requiring about 61 hours in total. However, the corresponding attenuation ratios were significantly higher with an average of 0.0942 (0.0058) (t-test $p = 0.0011$; Fig. 4.7) and the total number of loci replicated in BBJ dropped to 4,741. Furthermore, I compared this approach to `fastGWA` and `FMA:C8` and found that their attenuation ratios were significantly lower (t-test $p = 0.0038$ or 0.0013 respectively; Fig. A.13). These observations in combination with my simulation study imply that `REGENIE` requires explicit conditioning on covariates during the computation of association statistics, which increases the computational requirements. `FMA:C8`, in contrast, remains calibrated when association statistics are computed using a previously residualised phenotype, resulting in a significant speed-up during testing.

In terms of memory, `FMA` had the largest requirements, e.g. 85 GB of RAM for a full-UKBB analysis of 20 traits, or 12 GB for the smaller sample, similar to `BOLT:MoG`'s cost for one phenotype (69 and 9 respectively). `fastGWA` and `REGENIE` were more efficient, with `REGENIE` requiring the least amount of memory, owing to the feature of writing intermediate files to disk. I note that `REGENIE`'s disk usage depends both on the number of traits and number of model-SNPs used, whereas `FMA` only stores the $N \times M$ genotypes in an uncompressed format and memory usage depends on N and the number of phenotypes (see Fig. A.7).

Translation to actual cost. The aforementioned observations yield a similar comparison when actual running costs are considered. To demonstrate that, I will focus on the $N = 446$k sample which shows the largest variability. Computing such costs depends on the architecture of each system and how jobs are prioritised, but here I will only present rough estimates based on two scenarios; a hypothetical cluster with infinite resources and a fixed cost per CPU, and a more realistic setup whereby an analysis takes place on the UKBB Research Analysis Platform (RAP)[4].

First, assuming that utilising a single CPU core costs £0.10 per hour, the cost in money is proportional to each method's running time. In this scenario, the most affordable method would be fastGWA requiring $43 \cdot 12 \cdot 0.10 = 51.6$£, followed by FMA:C8 (£132), REGENIE (£190.8), and BOLT:MoG (£1,440). In practice, however, there are several limitations for memory and higher costs might occur when more RAM is needed. To that end, assume that the analysis needs to take place inside the UKBB-RAP, which offers machine-specific rates. In this scenario, BOLT:MoG and FMA:C8 should be assigned to machines with large amounts of RAM, such as m5.4xlarge or i3.4xlarge, whereas REGENIE or fastGWA have a low memory footprint and can thus use machines with lower costs, such as c5d.2xlarge. REGENIE, however, makes heavy disk utilisation, thus a more suitable machine would be one c5.2xlarge (which employs an SSD instead), with a slightly larger cost. Combining these assignments with the computational requirements described earlier, the corresponding total costs will be £535, £69, £48, and £11 for BOLT:MoG, FMA:C8, REGENIE, and fastGWA respectively, as summarised in Table 4.4.

I note that these are only rough estimates and, given the wide range of machines available and the specifications for each approach, better optimisation may be achievable. Furthermore, this comparison builds on the observations made earlier regarding the implicit use of covariates within FMA, and the corresponding overhead for REGENIE, whereby covariates are accounted for explicitly. Overall, the above comparisons are in line with the observations made throughout this chapter regarding the trade-offs between efficiency and statistical power or robustness.

[4] based on the online rate card, December 2022, provided by *DNAnexus*

Method	Instance Type	Rate (£/hour)	Jobs and time	Total (£)
`FMA`	i3.4xlarge	0.6256	1 × 110	68.8
`BOLT:MoG`	m5.4xlarge	0.5632	20 × 47.5	534.8
`REGENIE`	c5.2xlarge	0.2520	1 × 211.5	53.3
`fastGWA`	c5d.2xlarge	0.1984	20 × 2.8	11.1

Table 4.4: Estimated costs in GBP (£) for a large-scale GWAS for each method, using machines available in UKBB-RAP. Times per machine are as in Table 4.3, adjusted to account for the number of available cores. To clarify the adjustment, c5.2xlarge comprises 8 cores and would thus require roughly 33% more compute time than using 12 cores (159 hours), assuming linear gains; similar adjustments were followed whenever necessary.

5
Approximations to FMA for higher scalability

Contents

In this section, I discuss how `FMA` can incorporate approximations to the original model for higher scalability during step-1 and how these translate to statistical power. These model extensions are motivated by the difference in the asymptotic cost for model fitting in comparison to more efficient methods, as depicted in Fig. 4.6. In general, `FMA` has an $\mathcal{O}(N^{1.5}M)$ complexity, owing to the conjugate gradients iteration, the number of which increases with sample size. Having that in mind, I explored a few ways – namely `FMA:Pruned`, `FMA:Cluster`, and `FMA:AprLoco` – to reduce the computational burden without changing the core algorithm. In the following, I mostly refer to an estimator for the covariance matrix $\mathbf{V}$, so I do not discriminate from the estimand to keep the notation clear.

5.1 Variance-component pruning

One type of hyper-parameter in LMM-based association is the selection of genetic markers to be used in model fitting, e.g. for building the GRM.This may be addressed in different ways, with criteria involving causal plausibility based on linear regression [70], and MAF or LD thresholds [36, 37, 39, 105]. Therefore, the first model approximation I explored involves utilizing multiple variance components (VC) for heritability estimation, to detect sets of variants that have negligible contribution to the phenotype. Such sets can therefore be omitted during the conjugate gradient iteration to reduce the running time. More in detail, the `FMA:Pruned` approach is obtained by first applying `RHE-mc` to an annotation consisting of K equally sized and contiguous bins (e.g. $K = 90$ for simulated or $K = 130$ for real phenotypes), and then calculating the residuals with the preserved bins. This way, the I/O cost is reduced and the GRMs are sparser increasing the conjugate-gradient's convergence rate.

As a sanity check, I applied `FMA:Pruned` to synthetic phenotypes described in Section 4.4. In these particular simulations, only odd chromosomes contain causal variants, thus the set of markers with reduced contribution to heritability is clearly defined. As expected, `RHE-mc` correctly estimated the null contribution of even chromosomes, and `FMA:Pruned` performed well in this setting, leading to a ×2 speed-up (1.8 hours against 3.9 for `FMA:C8`, on average). As shown in Tables B.1-B.6, `FMA:Pruned` had similar performance to other approaches when polygenicity was 5%, but yielded inflated test statistics for the case of 1%. This approach, however, was not applicable to the CSR samples because the VC estimates had high errors, making pruning infeasible.

I followed a similar procedure for real phenotypes, applying `FMA:Pruned` to the set of $N = 328$k samples (described in Section 4.6). In this case, the estimates of $\hat{h}^2_{\text{SNP}}$ varied substantially, so the inclusion threshold I used was 5×10^{-4}. This resulted in keeping 512,539 of the 623,128 total variants, and the corresponding running time was decreased by roughly 40% (51 vs 30 hours). The speed-up can be explained by the smaller I/O cost (18% fewer variants) in combination with the higher convergence rate for the conjugate gradients (45 vs 50 iterations). Interestingly, FMA:Pruned had statistical performance similar to that of the full FMA model, as illustrated in Figure 5.1. Considering the 623k genotyped variants and the 20 phenotypes I tested, FMA:Pruned had a slightly larger average test statistic (2.25 vs 2.22) and the two approaches yielded the same average number of significant markers (4203.3 vs 4202.7).

To conclude, VC-pruning decreased the overall computational cost, with negligible differences in performance compared to FMA:C8, but the total running time was still one order of magnitude larger than that of fastGWA or REGENIE (about ×15 and ×13 slower respectively), when applied to the $N = 328$k sample. Although I did not explore this further, alternative pruning strategies, for instance based on functional annotations, could be considered in future work.

5.2 Distributed calculations using IBD clusters

Since FMA's convergence rate is inversely proportional to the sample size, I explored splitting the $N = 446$k samples into smaller sets and processing each in a distributed way, an approach which I will refer to as FMA:Cluster. This reduces the memory footprint and increases the convergence rate as each GRM has a smaller size. Such a technique would be particularly suitable in cases involving large and significantly structured samples, where the GRM is expected to be block diagonal, and where clusters of relatives may be efficiently used to estimate variance components. This principle is also leveraged by fastGWA [38, 45].

Intuitively, close relatives are highly informative in the calculation of variance components [119]. In FMA:Cluster, the samples are partitioned into subsets, which are analyzed separately. Subsets are chosen so that groups of related individuals are likely to belong to the same group, reducing computational costs, while preserving signal.

To be more precise, assume the existence of k such clusters. The GRM of each subset would be the same as the resulting block of the complete GRM restricted to the appropriate set of individuals. As a result, the same would hold for the

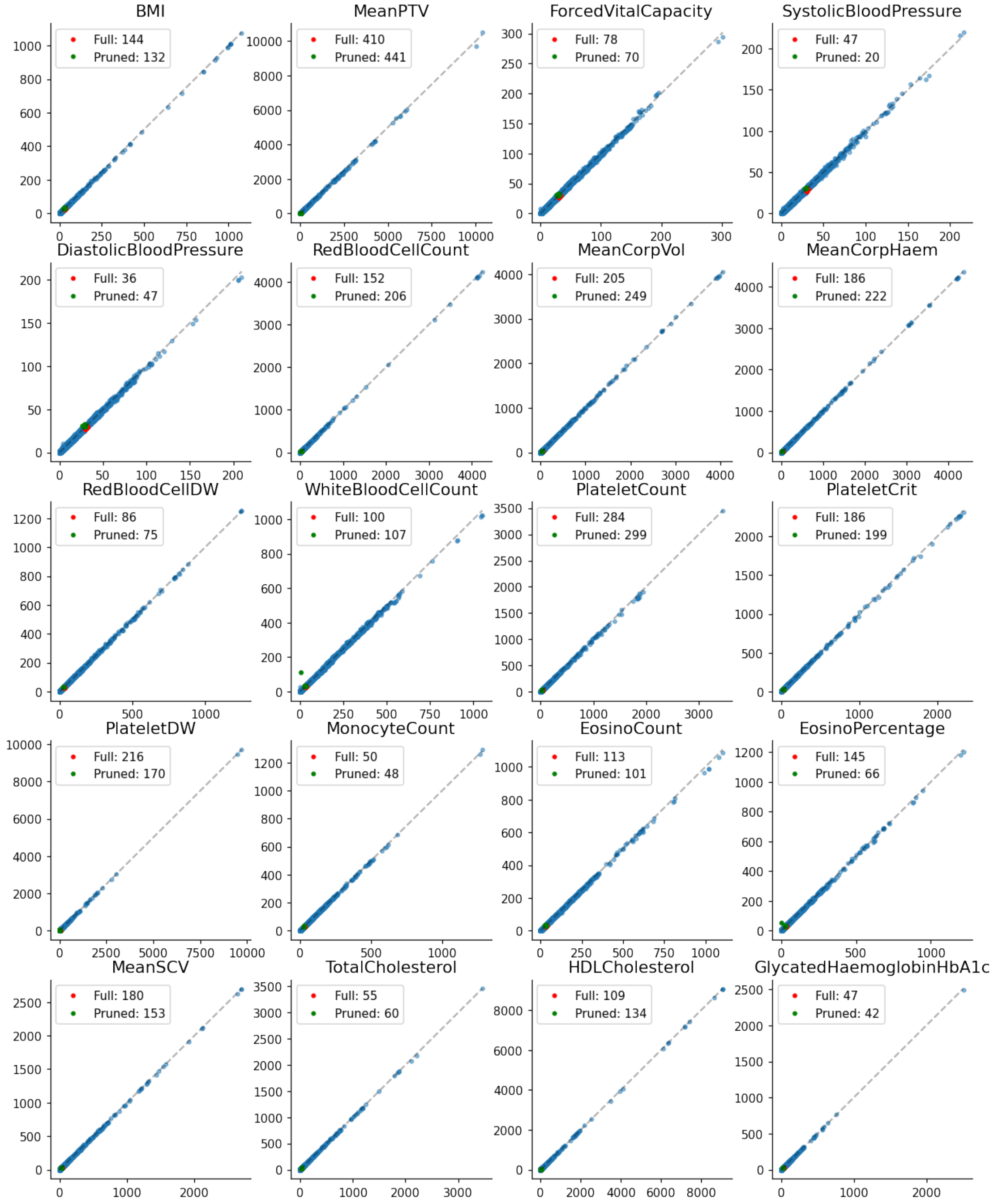


Figure 5.1: QQ-plots comparing `FMA:Pruned` with `FMA:C8` for $N = 328$k. The former was obtained by discarding genetic components with low contribution to h^2_{SNP}. Results on 20 real phenotypes, considering $M = 623$k genotyped variants with MAF> 0.01. Each QQ-plot compares the χ^2 statistic for each variant, highlighting the number of genome-wide associations ($p < 5 \times 10^{-8}$) detected by only one method. In total, the two approaches achieved similar numbers of uniquely-discovered associations (2841 vs 2829), indicating high concordance.

corresponding sample covariance $\mathbf{V}_* = \sigma_g^2\mathbf{K}_* + \sigma_e^2\mathbf{I}_*$, if σ_g^2 is the same across clusters, therefore $\mathbf{V} = \texttt{diag}(\mathbf{V}_1, ..., \mathbf{V}_k)$. Recall that the inverse of a block-diagonal matrix is a diagonal matrix containing the inverse of each block. Combining all these, and assuming that $\mathbf{y}_s$ is the vector of phenotypes for cluster s,

$$\mathbf{V}^{-1}\mathbf{y} = [\mathbf{V}_1^{-1}\mathbf{y}_1, ..., \mathbf{V}_k^{-1}\mathbf{y}_k]. \tag{5.1}$$

`FMA:C8` and `FMA:Cluster` calculate the left and right hand side, respectively, of this equation, thus the two approaches will be equivalent if $\mathbf{V}$ is indeed diagonal (but `FMA:Cluster` is more efficient). In practice, however, the non-diagonal blocks of $\mathbf{V}$ often have non-zero values, in which case `FMA:Cluster` will be based on a matrix $\mathbf{V}_{\text{sp}}$ obtained after sparsifying $\mathbf{V}$, assuming $\mathbf{V}_{\text{sp}}^{-1}\mathbf{y} \simeq \mathbf{V}^{-1}\mathbf{y}$. `FMA:Cluster` is thus conceptually similar to `fastGWA` [38], which is also based on a matrix like $\mathbf{V}_{\text{sp}}$.

One way to create such a partition is by utilising IBD sharing. Similarly to the analysis of Chapter 3, and based on Nait Saada et al. [11], I created a GRM for $N = 433$k UKBB individuals based on 213 billion shared IBD segments detected by `FastSMC`. Then I applied an algorithm for agglomerative hierarchical clustering[1] which yielded 4 clusters of genetically related individuals with sizes ranging in 54k−68k, and another set containing the remaining 197k (out of the original $N = 446$k set used throughout this chapter). For the `FMA:Cluster` approach I worked as follows: I applied `FMA` to each of the five subsets, using the same set of VC estimates, to obtain $\mathbf{V}_s^{-1}\mathbf{y}_s$ $(s = 1, \ldots, 5)$ which I then merged in order to calculate test statistics for the whole sample.

Computationally, `FMA:Cluster` was drastically better than `FMA:C8`, without considering the costs for IBD inference which may be seen as a preprocessing step, as in `fastGWA`. Assuming 12 cores and up to 85 GB of RAM – which is what `FMA:C8` required for the analysis of 20 phenotypes (Section 4.6) – I allocated 6 cores for the analysis of the $N = 197$k sample, and 6 cores for the (sequential) analysis of the smaller samples. The running time was bounded by the processing of the first batch which took 13 hours, which translates to a ×6 speed-up. Moreover, the first

[1] `https://github.com/khabbazian/sparseAHC/`

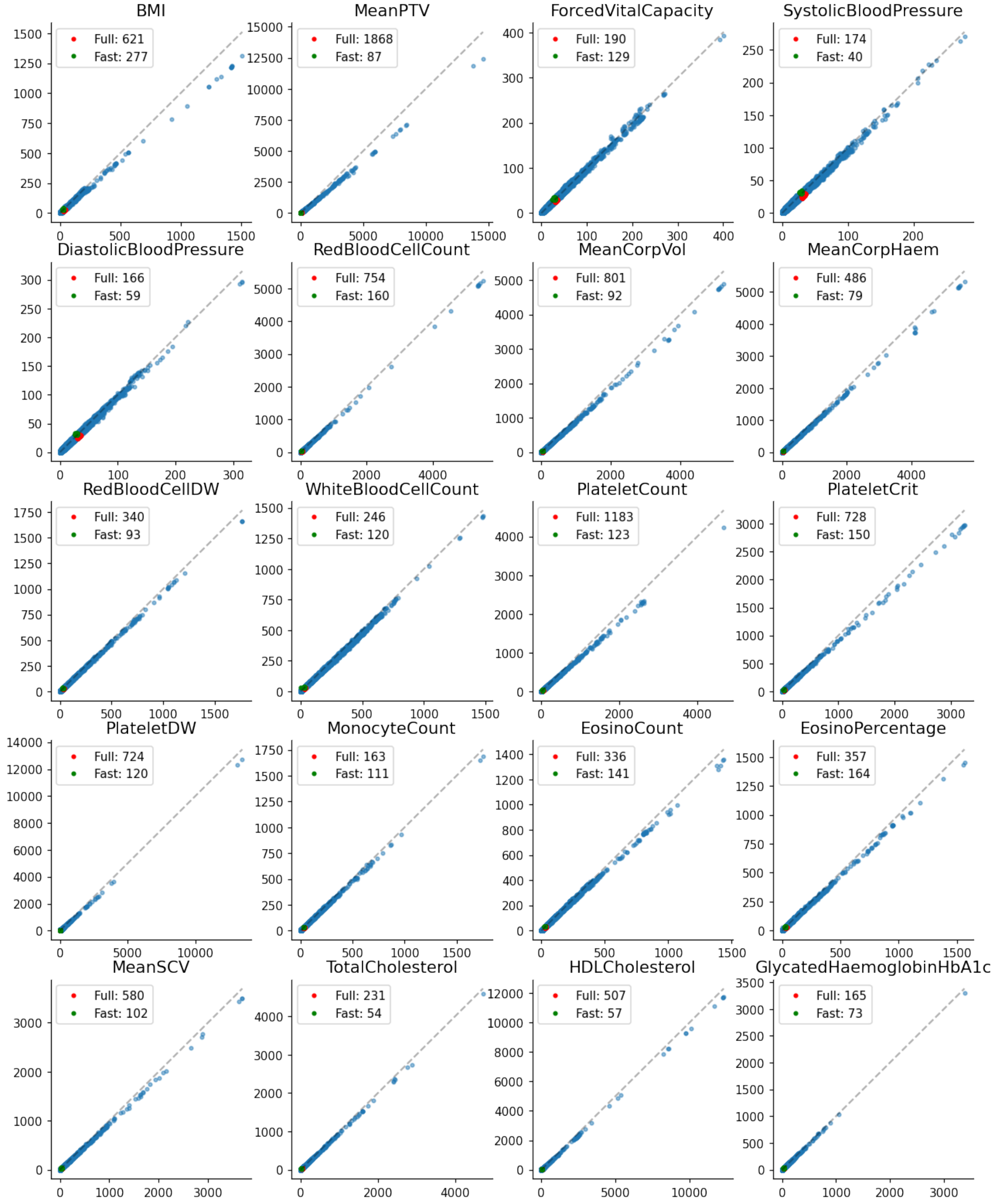


Figure 5.2: QQ-plots comparing `FMA:Cluster` with `FMA:C8` for $N = 446$k. The former was obtained by partitioning to sample to 5 clusters based on IBD, and discarding 3/16 components with low contribution to h^2_{SNP}. Results on 20 real phenotypes, considering $M = 623$k genotyped variants with MAF> 0.01. Each QQ-plot compares the χ^2 statistic between `FMA:C8` and `FMA:Cluster` for each variant, highlighting the number of genome-wide associations ($p <$ 5e-8) detected by only one method. In total, `FMA:Cluster` had roughly ×4 fewer uniquely discovered associations (2231 vs 10620), indicating a significant decrease in statistical power.

batch required about 44 GB of memory, whereas the second one took up to 14, thus the total memory used was $\sim 30\%$ less than what FMA:C8 needed (Table 4.3).

These computational gains, however, had a trade-off in statistical performance, as shown in Fig. 5.2. Considering the 623k variants tested, the average test statistic for FMA:Cluster was 5% lower than that of FMA:C8 (2.53 vs 2.66), and a similar decrease was observed in terms of significant associations (5,292 vs 5,711) which was statistically significant (paired t-test $p = 1.2 \times 10^{-4}$). As far as other methods are concerned, FMA:Cluster had less power than REGENIE (5,292 vs 5,741; $p = 6.8 \times 10^{-5}$), but was still more powered than fastGWA (5,292 vs 5,034; $p = 2.7 \times 10^{-4}$). This is in accordance with the assumption that FMA:Cluster's power is driven by exploiting broad clusters of relatedness, similarly to how fastGWA trades power for scalability by using a sparse GRM. Using a more accurate estimate of kinship for fastGWA (currently using the empirical GRM from genotyped variants, as recommended [38]), or by considering more distant relatives, could yield performance closer to that of FMA:Cluster, but this was not part of my investigation.

5.3 Approximate LOCO

As previously shown, leaving one chromosome out improves power, but increases the computational burden of calculating the residuals by roughly ×22, if 22 autosomal chromosomes are considered. REGENIE avoids that overhead by predicting the genome-wide effects once and then masking those accordingly to build 22 predictions. I investigated a similar strategy within FMA, where I calculated one set of residuals $\mathbf{V}^{-1}\mathbf{y}$ and adjusted it appropriately to more rapidly obtain 22 LOCO ones; I refer to this approach as FMA:AprLoco.

In general, the sample covariance matrix can be written as

$$\mathbf{V} = \frac{\sigma_g^2}{M}\mathbf{X}\mathbf{X}^\top + \sigma_e^2\mathbf{I}_N = \frac{\sigma_g^2}{M}\sum_{c=1}^{22}\mathbf{X}_c\mathbf{X}_c^\top + \sigma_e^2\mathbf{I}_N, \tag{5.2}$$

which is obtained by writing the (standardised) genotypes matrix $\mathbf{X}$ as a vector $[\mathbf{X}_1, ..., \mathbf{X}_{22}]$ of 22 matrices, one for each autosome. When we leave one chromosome out this relation becomes

$$\mathbf{V}_{-c} = \frac{\sigma^2_{g,-c}}{M_{-c}} \sum_{t \neq c} \mathbf{X}_t \mathbf{X}_t^\top + \sigma^2_{e,-c} \mathbf{I}_N, \tag{5.3}$$

where $M_{-c}, \sigma^2_{g,-c}, \sigma^2_{e,-c}$ are the number of variants left, genetic component, and environmental component respectively of all-but-c chromosomes. Due to polygenicity, we expect the fraction of heritability explained by the held out region to be proportional to the size of the region itself, so $\sigma^2_{g,-c}/M_{-c} = \sigma^2_g/M$ and $\sigma^2_{e,-c} \approx \sigma^2_e$. Therefore, the LOCO residuals can be written as $\mathbf{V}^{-1}_{-c}\mathbf{y} = (\mathbf{V} - \frac{\sigma^2_g}{M}\mathbf{X}_c\mathbf{X}_c^\top)^{-1}\mathbf{y}$.

Eq. 2.7 in Section 2.2 describes how to obtain the BLUP $\hat{\mathbf{b}}$ from the residual $\mathbf{V}^{-1}\mathbf{y}$. Based on that, we may get an approximate-LOCO BLUP as $\hat{\mathbf{b}}_{-c} = \frac{\sigma^2_g}{M}\mathbf{X}^\top_{-c}\mathbf{V}^{-1}\mathbf{y}$, and the corresponding residual will be $\mathbf{y} - \mathbf{K}_{-c}\mathbf{V}^{-1}\mathbf{y}$, by setting $\mathbf{K}_{-c} = \frac{\sigma^2_g}{M}\sum_{t \neq c}\mathbf{X}_t\mathbf{X}_t^\top$. This is used to obtain a set of 22 residuals by only calculating $\mathbf{V}^{-1}\mathbf{y}$ (the genome-wide term) and performing a few more operations, which is substantially more efficient than calculating the 22 sets of $\mathbf{V}^{-1}_{-c}\mathbf{y}$. However, this is an approximation to the true-LOCO residual, with a deviation given by

$$\sigma^2_e\mathbf{V}^{-1}_{-c}\mathbf{y} - (\mathbf{y} - \mathbf{K}_{-c}\mathbf{V}^{-1}\mathbf{y}) = -\mathbf{K}_{-c}\mathbf{V}^{-1}_{-c}\mathbf{y} + \mathbf{K}_{-c}\mathbf{V}^{-1}\mathbf{y} = \mathbf{K}_{-c}(\mathbf{V}^{-1} - \mathbf{V}^{-1}_{-c})\mathbf{y}. \tag{5.4}$$

In simulations, the mean squared error between the approximate LOCO and the original residuals remained fairly small. However, the corresponding test statistics were inflated, suggesting that further work is required for `FMA:AprLoco` to be used as a viable method. By comparing to the test statistics obtained by linear regression, I noticed that the bias was proportional to the LD score of each variant. To account for that, I ran `FMA:AprLoco` with covariates, estimated the bias within quartiles of the LD distribution, and re-scaled the test statistics accordingly. Note that working with GRMs requires a γ-type of calibration, but I omitted this part on `FMA:AprLoco` to keep the approach simple.

I applied `FMA:AprLoco` to both simulated and real phenotypes. As expected, the speed-up was significant in most cases, requiring 49 minutes for $N = 50$k and 50 traits, or 9.7 hours for $N = 446$k and 20 traits (×5.5 and ×8 faster than `FMA:C8`, respectively). In simulations (as described in Section 4.4), `FMA:AprLoco` achieved average χ^2 values and power similar to `BOLT:MoG`, while keeping type I errors less than 5% (tables B.1-B.8). The increase in power, however, is likely due to inflation of larger χ^2 values, rather than improved performance. In addition, `FMA:AprLoco` showed moderate statistical power when applied to real phenotypes (Section 4.6), having an average χ^2 value of 2.24, and a total of 3,881 significant variants, which was significantly lower than `FMA:C8` (2.66 and 5,711 respectively). Although `FMA:AprLoco` does not require any preprocessing, in contrast to `FMA:Pruned` or `FMA:Cluster`, step-2 can be slow as we need to process each variant twice. Therefore, this approach would have limited applicability to large scale studies with imputed genotypes.

6
Improving the scalability of ARG-based complex trait analyses

Contents

The ancestral recombination graph (ARG) was briefly introduced in Section 2.4.2. In this chapter, building on Zhang et al.'s work [48] and on the tools developed in the previous chapters, I develop methodology to increase the scalability of ARG-based complex trait analyses. More in detail, I use `FMA` to increase the computational efficiency of estimating narrow-sense heritability and detecting association using genealogical information. As with the rest of the thesis, I focus on quantitative phenotypes.

6.1 Utilising `FMA` for genealogy-wide association

Zhang et al. [48] developed a framework to detect association between a phenotype and the edges of the ARG of a set of samples, inferred by `ARG-Needle`. This approach has the potential to reveal association to variation that has not been observed in the data utilized to infer the ARG, such as indels, structural variants, or other rare variants that have not been genotyped or imputed. The authors refer to analyses that test for association of ARG branches as “genealogy-wide association”, or as “ARG-MLMA” to indicate the use of a linear mixed model to test ARG branches. As previously shown, mixed model association provides several advantages over the use of a simple linear model. These advantages, which include better control for relatedness and stratification, as well as increased association power due to the conditioning on polygenic effects, also apply to the case of genealogical association.

The ARG-MLMA approach introduced by Zhang et al. [48] relied on the `BOLT:MoG` algorithm to compute LMM association statistics. More in detail, because `BOLT-LMM` does not export LOCO-residualised phenotypes, their approach involved running `BOLT:MoG` 22 times, each time excluding a different chromosome, to obtain BLUP effect size estimates [36]. These BLUP estimates were then used in conjunction with `PLINK` to obtain LOCO phenotype predictions, which were then used to obtain LOCO residualised phenotypes. `BOLT:MoG` was invoked another time, using all chromosomes, to estimate the genome-wide calibration factor $\hat{\gamma}$. The set of residualised phenotypes and the calibration factors were finally passed to `ARG-Needle` to test for association between clades of a previously inferred ARG and the phenotype. Overall, this approach required a complex scripting pipeline and significant computational time.

The `FMA` framework, developed in Chapter 4, can be utilised to simplify ARG-MLMA, in place of `BOLT-LMM`. I followed a procedure similar to that of Section 4.6, where I tested for imputed genotypes, to compute the LOCO residuals $\mathbf{V}_{-c}^{-1}\mathbf{y}$ and estimate the calibration factor γ. I used this approach for an analysis involving standing height, focusing on the set of $N = 337$k unrelated British and $M = 623$k common variants (MAF$\geq 1\%$) used in previous experiments. `FMA:C8` required 20.5 hours to calculate the LOCO residuals and the calibration factor (inlcuding the time for h^2 estimation with `RHEmc`), which was roughly ×11 faster than using the pipeline based on `BOLT:Inf` and `PLINK`, as in [48]. Because `FMA` is optimised to parallelly process multiple phenotypes, the computational advantages over BOLT:Inf would have been even greater if more phenotypes were considered.

Figure 6.1 shows an indicative Manhattan plot of this analysis. ARG-MLMA using the LOCO residuals obtained by FMA:C8 is compared to standard association testing based on genotyped SNPs or variants imputed using the HRC+UK10K reference panel [4, 32, 48]. In this example, ARG-MLMA was applied to testing the edges of an ARG inferred using only SNP array data. As shown in the locus-plot for the 60.7Mb region of Chromosome 8, ARG-MLMA enabled detecting signal that would otherwise require imputation from a sequenced reference panel, as well as association peaks that would not be observed using imputed data. I note that the original study utilised the non-infinitesimal version of BOLT-LMM (BOLT:MoG), which better captures sparsity and often achieves more power at the cost of higher running times (see Chapter 4).

6.2 ARG-based GRMs for heritability estimation

Zhang et al. [48] have shown that an accurately inferred ARG may also be used to obtain unbiased estimates of narrow-sense heritability. This approach, however, relied on the explicit calculation of an ARG-based genomic relationship matrix (ARG-GRM; for additional details see [48], Supplementary Note 2), which was used within the GCTA framework to obtain heritability estimates. Computing the ARG-GRM, however, has $\mathcal{O}(N^2M)$ time and $\mathcal{O}(N^2)$ space complexity; using GCTA to estimate heritability further increases computational costs due to the need to invert a covariance matrix within the REML algorithm [38, 56, 120].

In this section, I introduce a randomised method-of-moments (MoM) approach for scalable ARG-based heritability estimation. This approach, which I refer to as ARG-RHE, is closely related to the strategy used by RHE-mc [105], leveraged in the first step of FMA. ARG-RHE, however, enables working directly with the ARG rather than relying on a provided set of genotyped, imputed, or sequenced variants. Using simulations, I show that this approach is orders of magnitude more scalable

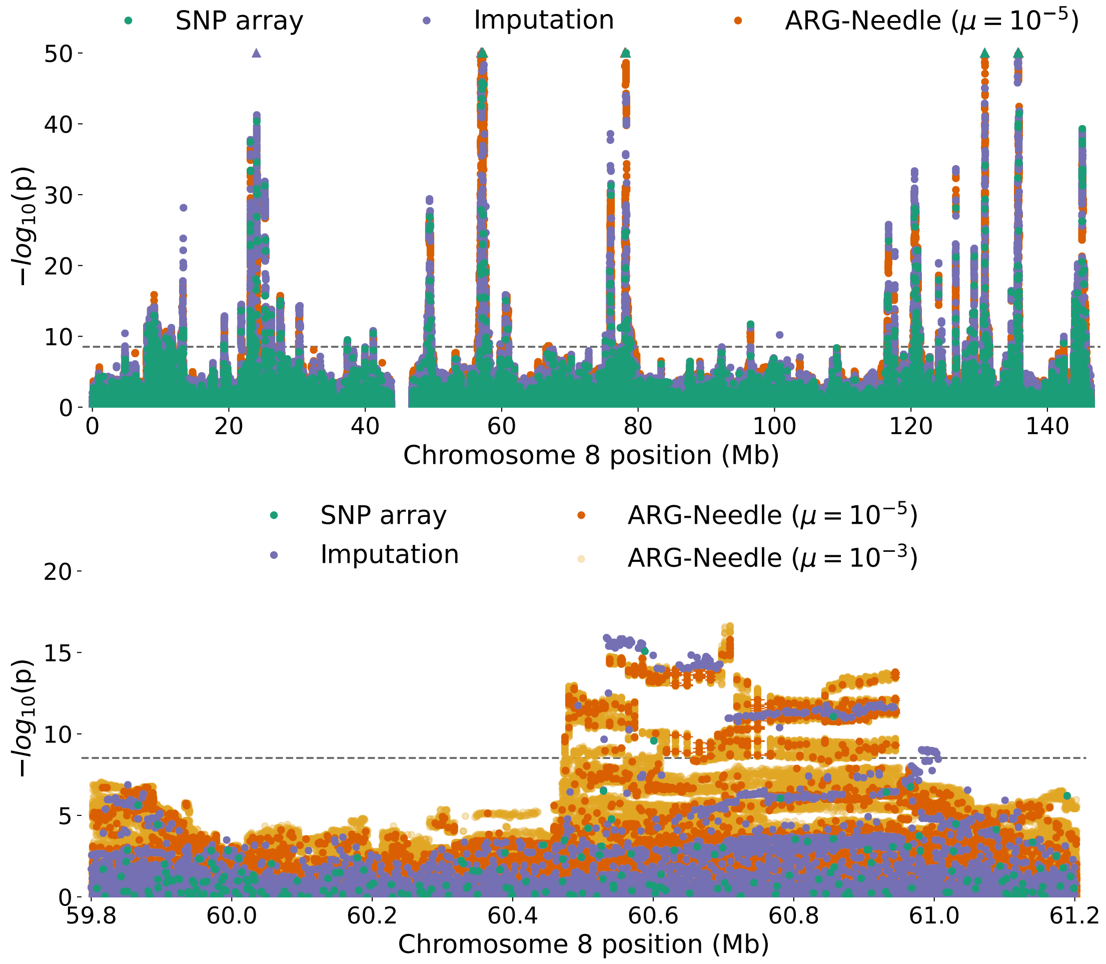


Figure 6.1: Genealogy-wide association of height in UKBB. Manhattan plots for LMM-based association between ARG branches and height, using 337k unrelated British. P-values for `ARG-Needle` were obtained using `FMA`, whereas for SNP array and imputed genotypes (HRC+UK10K v3 [4, 32]) were obtained with `BOLT:MoG`, provided by Zhang et al. [48]. Dotted lines correspond to $p = 3 \times 10^{-9}$, which was a permutation-based threshold for significance [48], and triangles indicate associations with $p < 10^{-50}$.

than the REML-based approach used previously [48], enabling to potentially scale such analyses to hundreds of thousands of individuals.

More in detail, `ARG-RHE` implements an approach that is related to the one used by Zhang et al. [48] to obtain Monte Carlo estimates of an ARG-GRM (referred to as "Monte Carlo ARG-GRMs" [48]). To that end, the ARG is traversed to generate mutations on its edges in order to form a genotype matrix $\mathbf{X}$. To obtain an ARG-GRM, these mutations are used on the fly to compute $\mathbf{X}\mathbf{X}^\top$, without the need of writing $\mathbf{X}$ to disk. In `ARG-RHE`, we are instead interested in performing matrix multiplication operations of the kind $\mathbf{XU}$, or $\mathbf{X}^\top\mathbf{U}$ (as in schema 4.3), for a given matrix $\mathbf{U}$ of appropriate dimensionality, without explicitly forming the genotype matrix $\mathbf{X}$. Additionally, the MoM estimator requires products with the trace of the GRM. To achieve this in a matrix-free way, I utilised the Hutchinson's trace estimator using random vectors [121], similarly to previous methods [105, 107].

In `ARG-RHE`, standard errors are obtained with a block jackknife estimator, following Pazokitoroudi et al. [105], by sequentially processing contiguous parts of the ARG. Assuming J independent blocks, these are combined in all-but-one samples of observations, and thus J estimates of heritability are obtained. Assuming that $\hat{\theta}_{(j)}, j = 1,...,J$, are the J jackknife estimates, the variance of the `ARG-RHE` estimator is calculated as the variance of those J estimates, or $(J-1)/J\sum_{j=1}^{J}(\hat{\theta}_{(j)} - \overline{\theta})^2$, where $\overline{\theta}$ is the mean of the J estimates [122].

Note that, as explained in [48], the mutations that are generated on the ARG to obtain $\mathbf{X}$ are not necessarily the same variants used to infer the ARG. However, Zhang et al. [48] showed with simulations that both the true ARG and an ARG inferred from SNP array data can provide ARG-GRMs which yield improved heritability estimates compared to those obtained directly from using SNP array data (see Figure 3b in [48]).

To test the approach described above, and to extend the work of Zhang et al. [48], I used the `msprime` coalescent simulator [85] to synthesize an ARG, with sample sizes ranging from 2,500 to 100,000 diploid individuals, using chromosomes of length $L = 10$Mb, recombination rate $1e \times 10^{-8}$, and a European demographic model [123]. I generated sequencing data for the simulated ARGs using a mutation rate $1e \times 10^{-8}$, which resulted in datasets comprising 146k to 204k genetic variants (for each sample size). Next, to emulate the genotyping of array variants, I filtered each set of mutations according to frequency so that the resulting distribution matches the one of UKBB genotypes, as done in [48], obtaining roughly 2,260 variants per sample.

I used these simulations to compare the speed and accuracy of two approaches for ARG-based heritability estimation. First, the `ARG-RHE` strategy described above; second, the approach described in [48], where an ARG-GRM (or its Monte Carlo estimate) is first computed and then provided to GCTA to obtain a maximum likelihood estimate, which I call ARG-GCTA. To this end, I generated new sets of mutations from the simulated ARGs, using a ×2 larger mutation rate, for either approach.

In addition to ARG-GCTA and ARG-RHE, I assessed the accuracy of two more approaches for h^2 estimation by applying RHE [105] to either sequencing or array genotypes. Note that all of these approaches implicitly rely on the following definition for a GRM:

$$\mathbf{K}_\alpha(i,j) = \frac{1}{M} \sum_{k=1}^{M} \frac{(x_{ki} - 2p_k)(x_{kj} - 2p_k)}{[2p_k(1-p_k)]^{-\alpha}}, \tag{6.1}$$

for the corresponding set of available variants, where x_{ki} is the allele count (0,1, or 2) of individual i at variant k, p_k is the frequency of variant k, and α is the negative selection parameter[115], for which I set to $\alpha = -1$ in this analysis. I created phenotypes by sampling effect sizes from a standard normal distribution, varying the levels of polygenicity (5% or 100%) and heritability (0.25 or 0.50), while using standardised genotypes (implying $\alpha = -1$). All methods were invoked using one genetic component containing all variants, and applied to 20 repeats of each scenario (sample size, polygenicity, and true heritability). For the computational benchmarking, I compared ARG-RHE to ARG-GCTA, which I had to independently run on each replicate, after building the GRM once per sample.

The results of this experiment are summarised in Figure 6.2, where I focus on the scenario with $h^2 = 0.50$ and 5% polygenicity; similar trends were observed in other cases (Fig. A.15). Numerical results and standard errors are given in Table B.10. First, both ARG-RHE and ARG-GCTA yielded estimates that were similar to those obtained by using a sequencing-based GRM, but with a small downward bias that may be explained by the use of a relatively low mutation rate to obtain Monte Carlo estimates [48]. Using the GRM constructed from array genotypes resulted in a systematic under-estimation, reflecting known limitations of using SNP array data to estimate narrow-sense heritability [111, 117].

The computational cost for the two ARG-based approaches is shown in the lower part of Fig. 6.2. ARG-RHE was dramatically faster than ARG-GCTA, requiring one order of magnitude more time for larger samples. For instance, ARG-GCTA needed roughly 19 hours for 25k diploid samples, which was ×77 more than ARG-RHE's cost (15 minutes), considering the total time for 20 repeats. This difference is driven by two factors, namely the $\mathcal{O}(N^2M)$ cost for building the GRM and the lack of parallelization for multiple phenotypes during model-fitting, in contrast to ARG-RHE which implicitly works with the GRM and can handle multiple phenotypes.

6.3 ARG-based GRMs for association

As previously described, LMM-based association studies use the GRM during model-fitting to account for structure and improve power by residualising the phenotype from polygenic effects [36–38, 47, 59, 62, 63]. The use of ARG-GRMs in this setting was shown to potentially improve association power further, compared to the use of genotyped SNPs, as the ARG enables capturing the contribution of rare or untyped variants leading to better LOCO residualised phenotypes (see Figure 3c in [48]). Zhang et al. [48], however, required to explicitly compute an ARG-GRM, which was then provided as input to GCTA to perform LMM association. As in the case of heritability estimation, these REML-based operations scale poorly for large sample sizes.

I developed an alternative approach, which I refer to as ARG-FMA, which uses a strategy similar to that used in the previous section for estimating heritability, where the ARG is used instead of the genotype matrix $\mathbf{X}$. For association, which includes ARG-RHE as an intermediate step, the aim is to use the ARG to perform the conjugate gradient iteration to obtain LOCO-residualised phenotypes, as described in Section 4.2. This is again achieved by sampling mutations on the ARG and multiplying the corresponding vectors into user-provided vectors and matrices to perform the computations described in Eq. 4.3. These operations enable computing LOCO-residuals without forming the GRM, which, as shown in previous chapters, leads to sub-quadratic asymptotic costs.

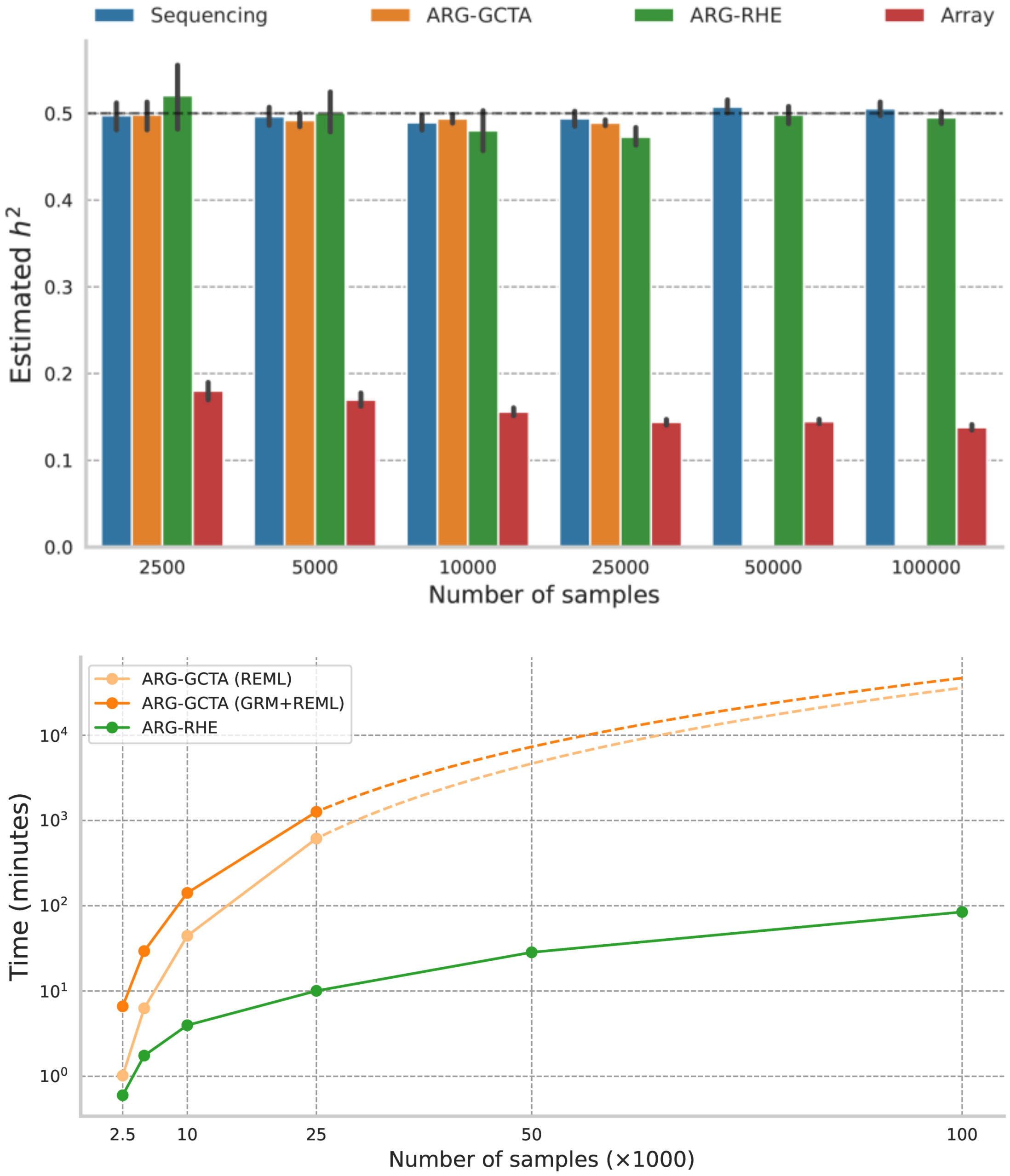


Figure 6.2: Estimation of heritability using ARG-based GRMs. Comparison of estimates (upper plot) and running times (lower) for estimating narrow-sense heritability using ARG-based GRMs, considering 1 synthetic chromosome with $L = 10$Mb and phenotypes generated with $h^2 = 0.50$ and 5% polygenicity. In the top plot, the dashed line represents the true heritability and bars represent standard errors across 20 independent simulations; numerical results including standard errors of the estimator are given in Table B.10. ARG-GCTA was not applied to samples with $N \geq 50{,}000$ because such an analysis would take more than 2 days (based on a polynomial extrapolation).

The simulation setup used to test this strategy is similar to the previously described one, this time using multiple small chromosomes instead of a large one, in order to apply the LOCO scheme. In particular, I used `msprime` to simulate ARGs spanning 10 independent chromosomes, each of length $L = 1$Mb, using 1e-8 recombination rate, for samples that ranged from 2,500 to 100,000 diploid individuals. To create phenotypes, I assumed a fully infinitesimal architecture (polygenicity = 100%) and drew effect sizes from a standard normal distribution, while considering different values for the negative selection parameter, namely $\alpha \in \{0, -0.5, -1\}$, and $h^2 = 0.50$. I benchmarked different LMM approaches using GRMs constructed from sequenced variants (`LMM-Seq`), array genotypes (`LMM-Array`), or mutations sampled from the ARG (`ARG-GCTA` or `ARG-FMA`) with mutation rate 1.65×10^{-7}. I measured power to detect association with the set of simulated array variants by comparing the mean χ^2 value of each approach to that obtained by linear regression.

Figure 6.3 (upper part) summarises the results for $\alpha = -0.5$ and various sample sizes. Sequence-based and ARG-based GRMs achieved a considerably larger boost than the array-based GRM, replicating the original study [48]. As expected from the LMM properties, all mixed-model approaches yielded a power advantage over linear regression, which was proportional to sample size. All approaches achieved the largest gains in power (e.g. around 40%) for $\alpha = 0$, when effect sizes have low variance, and had lower gains ($5 - 8\%$) for $\alpha = -1$, when causal variants are harder to detect (Fig. 6.5), in accordance to previous studies [48, 115].

The computational performance of `ARG-GCTA` and `ARG-FMA` is illustrated in the lower part of Fig. 6.3. `ARG-FMA` is slower for $N \leq 5{,}000$, due to large constant costs, but becomes significantly faster than `ARG-GCTA` as sample size increases, reflecting the improved asymptotic behaviour. For instance, `ARG-FMA` required 52.5 hours to process $N = 50$k, which would be $\times 21$ less than what `ARG-GCTA` is extrapolated to need. This can be ascribed to the lower asymptotic cost of `ARG-FMA`, which requires $\mathcal{O}(N^{1.5}M)$ to perform the conjugate gradient iteration, compared to the $\mathcal{O}(N^2M)$ cost to build each GRM and an additional $\mathcal{O}(N^{2.5})$ cost to perform LMM association within `GCTA` [56] (due to matrix inversion). Finally, as illustrated in

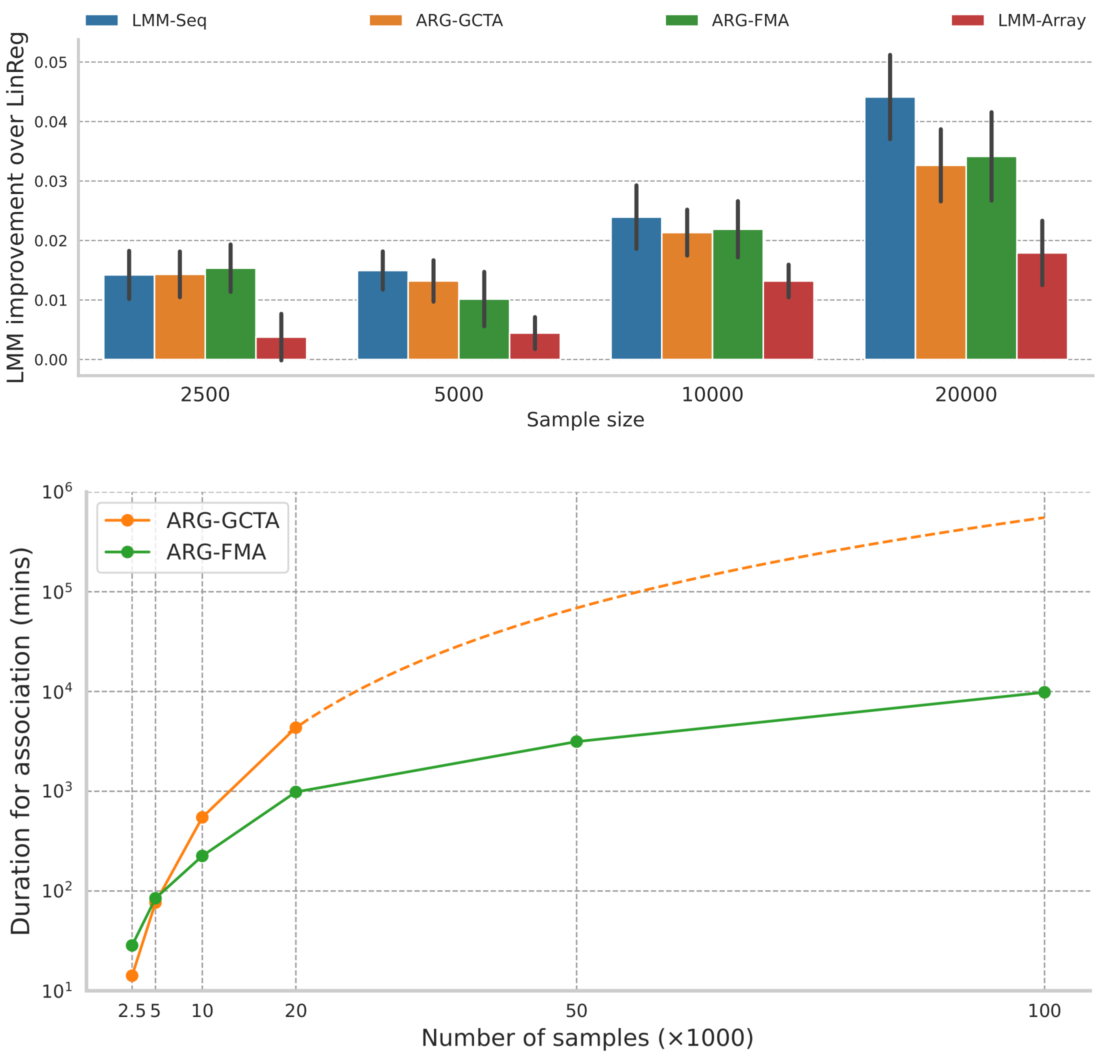


Figure 6.3: Mixed-model association using ARG-based GRMs. Comparison of statistical power (upper) and running times (lower) for mixed-model association using 10 synthetic chromosomes of length $L = 1$Mb and phenotypes with $h^2 = 0.50$ and $\alpha = -0.5$, considering 20 replicates. Power is measured as the relative improvement of mean χ^2 statistic in comparison to linear regression with array genotypes. `ARG-FMA`'s cost includes `ARG-RHE`running time (which is independent of α). `ARG-GCTA` did not complete for $N \geq 50{,}000$ because of excessive compute time (about 10 days).

Fig. 6.4, `ARG-FMA`'s memory footprint is significantly lower than that of `ARG-GCTA`, owing to the matrix-free operations enabling GRM-based analyses without the $\mathcal{O}(N^2)$ memory cost. This also holds for `ARG-RHE` which is similarly implemented.

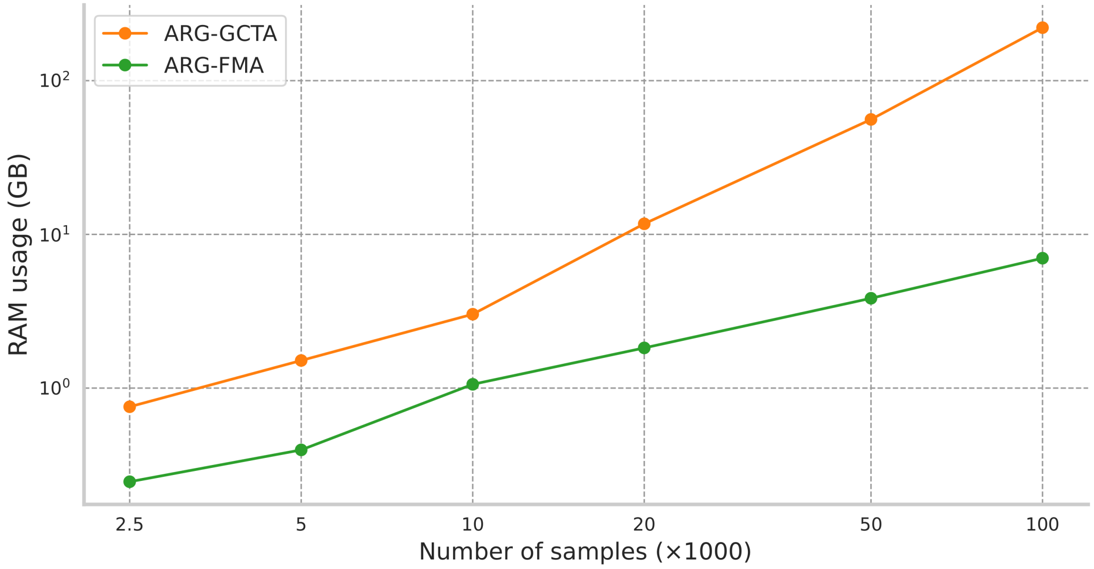


Figure 6.4: RAM usage for ARG-based LMM association. The memory footprint was measured using a Python profiler for `ARG-FMA` and the `top` tool for `ARG-GCTA`. These costs correspond to 10 traits for `ARG-FMA`, but one for `ARG-GCTA`.

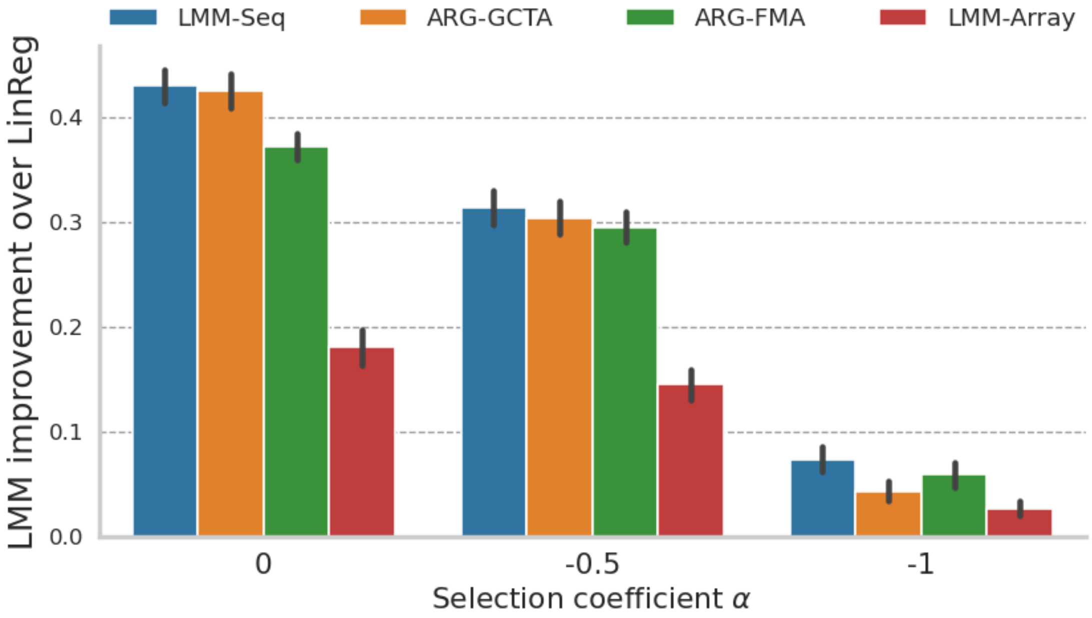


Figure 6.5: Mixed-model association with different selection coefficients. Comparison of statistical power measured as the relative improvement of mean χ^2 statistic in comparison to linear regression with array genotypes. This analysis involved a sample of N = 2,500 individuals, using 5 synthetic chromosomes of length L = 0.2Mb. Bars illustrate standard deviations for the 20 phenotypes simulated.

6.4 Remarks and limitations

I conclude with a few remarks and direct limitations of the methods presented in this chapter; additional thoughts for future research are discussed in Chapter 7. First, although `FMA` is a multi-threaded software, the benchmarking was limited to one computational core per task as `ARG-Needle` does not currently support multi-threading. Given this limitation, my simulations for association were based on short genomes focusing on demonstrating the asymptotic advantage of `ARG-FMA` (using the conjugate gradient iteration) over `ARG-GCTA` (working with pre-calculated GRMs). There is potential to improve the implementation, with one quick avenue involving parallel processing of different chunks, as is the case for LOCO association. Second, my analysis was based on `GCTA` which is the only implementation of an exact mixed-model allowing for pre-calculated GRMs. Other software were not applicable to this context; for instance `fastGWA` [38] works with a sparse GRM which would not have the beneficial properties of an accurate GRM (as discussed in the previous chapters), and `REGENIE` [39] requires genotyped variants for model-fitting thus extra work would be needed to prepare the appropriate input. Third, estimation of narrow-sense heritability is currently only implemented for the case of $\alpha = -1$ due to a limitation of the method-of-moments estimator, which requires standardised genotypes with unit variance. We plan to relax this requirement, and also support for multiple variance components (as in RHE-mc [105]), which reduce the relevance of which value of α is assumed by enabling non-parametric estimation of h^2.

7 Discussion

In this thesis, I explored methods for large-scale genome-wide association studies (GWAS) and introduced `FMA`, a flexible linear mixed model for association testing. The key contributions of this work are the following:

- In Chapter 3, I leveraged identity-by-descent information to implicitly perform within-cohort imputation and detect associations between rare loss-of-function variation (down to a MAF $\sim 5 \times 10^{-4}$) and 7 phenotypes in UKBB [4, 11].

- Chapter 4 described `FMA`, an efficient algorithm for mixed-model association which builds on `BOLT:Inf` [36]. I performed extensive benchmarking of state-of-the-art methods using both simulations and real phenotypes, assessing statistical power, robustness to confounding, and efficiency.

- In Chapter 5, I explored three modifications to the `FMA` algorithm that may further improve the method's scalability for association. These include the selection of subsets of informative variants for model fitting, the use of inferred relatedness clusters, and the use of an approximate leave-one-chromosome-out scheme.

- Chapter 6 described methodology to integrate FMA with the ARG-Needle library [48] to improve the scalability of complex trait analyses based on the ancestral recombination graph (ARG)

FMA is fully developed in Python and will be released as an open-source package, facilitating future development. The rest of this chapter provides a few general conclusions, highlights the limitations of this work, and suggests future directions.

The power-scalability trade-off for LMMs

The analyses of chapters 4 and 5 highlight the trade-off that exists between speed (measured in CPU hours) and statistical power (measured using true positive rates or mean χ^2 values). In particular, BOLT:MoG consistently achieved higher power than all the other methods, but was significantly slower for studies with large samples or multiple phenotypes. On the other hand, REGENIE and fastGWA, which were the most efficient methods, demonstrated a statistical performance that depended on sample structure and genetic architecture. REGENIE required the use of covariates during testing to properly control for sample structure, and fastGWA was well calibrated but had decreased power. FMA:C8 was more efficient than BOLT:Inf, requiring $\sim$ 8 hours to perform a GWAS for 20 phenotypes and 50k samples on a conventional machine (Table 4.3), but was slower than REGENIE and fastGWA. It had the same statistical power as BOLT:Inf and REGENIE, and more power than fastGWA but, like all other methods, less power than BOLT:MoG.

These trade-offs between statistical power, robustness, and computational efficiency also emerged in the analyses described in Chapter 5. The three LMM approximations I explored resulted in promising gains in efficiency which, however, translated into a loss in power. Perhaps not surprisingly, larger speed-ups yielded lower accuracy. One conclusion that was evident in several parts of this thesis is that the key LMM features — a high resolution GRM and the LOCO scheme — are necessary to maximise power and robustness to confounding.

FMA:C8 showed an advantage in statistical power over REGENIE in synthetic phenotypes, achieving higher test statistics at causal variants (Fig. 4.1 and Tables B.1-B.8), which however was not observed in the analysis of real phenotypes (Fig. 4.5, Table B.9). This difference could be attributed to the length of the genomes analysed, as I used all 22 chromosomes for the analysis of real phenotypes, but only 10 for simulations due to computational constrains. To validate that, I considered an additional synthetic scenario based on 22 chromosomes (Fig. A.9) and found that REGENIE attained power similar to that of FMA:C8 or BOLT:Inf. This indicates that REGENIE's decreased power in these experiments may be linked to the use of an approximate LOCO scheme.

Apart fron associations studies with common variants, these observations might have implications to rare variant assocation, as recently illustrated by a Jurgens et al. [124]. In brief, the authors used a LMM built with a sparse GRM, similar to fastGWA, to control for sample structure while testing for rare variants. They demonstrated that adding a SNP-based predictor as a covariate increased statistical power, as that enabled conditioning to common genetic effects. This was the case when using a newer version of SAIGE-Gene [125], which is also based on a sparse GRM. In contrast, the gains were not significant when using REGENIE which inherently accounts for polygenic effects, similar to how FMA, or BOLT:Inf work. Overall, given the broad avaialbility of methods for association studies, researchers need to adopt the one that is optimised for a specific goal, while considering efficiency, robustness to confounding, or statistical power.

Limitations and future steps

Next I discuss a few limitations and potential next steps, in particular those related to the development of FMA. First, the current implementation is based on implicitly conditioning on covariates by regressing them out from the phenotypes as the first step of model fitting. This is a common technique to reduce the computational burden [38, 56], but might result in conservative test statistics; an intermediate approach is to use the exact covariate adjustment only when

$\chi^2 \geq 4$, indicating signal [45]. My initial observations implied that the resulting loss in power is marginal, but future work involves a more careful implementation leading to unbiased test statistics.

Another key difference between FMA and other LMM packages is the way genotypes are handled. I followed a streaming approach by repeatedly reading chunks of genomes from an uncompressed file. This enabled analyses with relatively low memory usage, proportional to the number of phenotypes (Fig. A.7), but at an increased cost due to the I/O overhead of repeatedly reading data within FMA's iterative algorithm. Therefore, a faster implementation can be obtained by reading the genotypes once and keeping them in memory, while using a custom data structure to minimise memory usage (e.g. 1 bit per genotype, as in BOLT-LMM [36]). Although FMA relies on highly optimized Python libraries, reimplementing some of its core functionality in a compiled language, such as C++, will likely lead to computational gains.

I note that I did not consider phenotype pre-processing steps, such as the use of a rank-based inverse normal transformation (INT), which may be leveraged to increase statistical power [38, 89, 126]. This is because my main objective was to benchmark different algorithms, rather than detect novel associations. FMA can be easily extended to incorporate automated filtering steps, such as INT.

FMA can fit models with multiple variance components, by relying on RHE-mc, which yield better estimates of heritability (Fig. 4.3). This provides increased robustness because test-statistics may become uncalibrated for significantly misestimated σ_g^2 coefficients (Fig. 4.4), which are applied to scale covariance matrices. This study was limited to MAF/LD-based components and further work using functional annotations could lead to additional gains over traditional mixed models. For example, a recent study using a Bayesian multi-component approach reported a gain over BOLT:MoG [113].

A natural extension of the proposed framework involves the analysis of binary phenotypes for large-scale association studies. Traditional mixed-model approaches might be severely inflated at rare variants when the case/control ratio is low, and recent techniques rely on test statistic calibration to avoid that, such as the saddle-point approximation [37, 45], or Firth correction [39]. My study focused on quantitative phenotypes, but future work could involve implementing any of the aforementioned tools for binary phenotypes.

In Section 2.2 I described how LMM-based association is related to both heritability estimation and polygenic prediction through the relation $\sigma_e^2 \hat{\mathbf{V}}^{-1}\mathbf{y} = \mathbf{y} - \mathbf{X}\hat{\mathbf{b}}$ (see Eq. 2.7). Therefore, association requires estimating σ_g^2, σ_e^2 (to estimate $\mathbf{V}$), and calculating the BLUP $\mathbf{X}\hat{\mathbf{b}}$. `FMA` can be modified to calculate $\mathbf{X}\hat{\mathbf{b}}$ from the residual $\hat{\mathbf{V}}^{-1}\mathbf{y}$ (as described in Section 5.3) and therefore perform in-sample prediction. As a next step, it may be possible to extend `FMA` to provide out-of-sample predictions, similarly to the `cv-BLUP` approach [69], which requires (implicitly) accessing the trace of $\mathbf{K}\mathbf{V}^{-1}$.

Finally, Chapter 6 demonstrated the utility of combining recently developed techniques for the analysis of complex traits and for inference of genealogical relationships, building on the work of Zhang et al. [48]. Although this work was limited to simulations involving single variance components, a natural next step is the application of such methods to inferred genealogies in real data using multiple components. To that end, there is work in progress to incorporate `RHE-mc` [105] within the ARG-Needle library [48]. Besides the potential increase in power, utilising ARG-based GRMs for association may better account for population stratification [33, 48, 63, 96, 98, 127]. The development of these improved methodologies for ARG-based studies of complex traits may facilitate analyses in under-represented populations, for which an ancestrally-matched sequenced cohort for imputation may not be available [6, 17, 34, 35].

Appendices

A
Supplementary figures

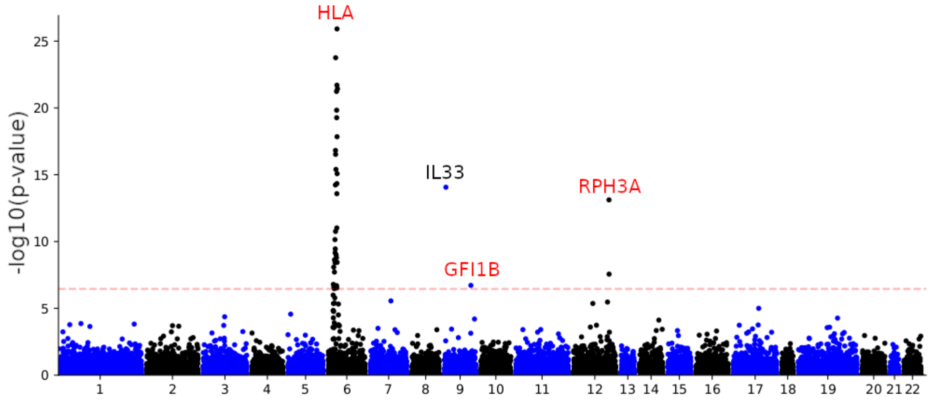


Figure A.1: LoF-segment burden exome-wide Manhattan plot for eosinophill count. Labelled genes are exome-wide significant (after adjusting for multiple testing, t-test p-value $< 0.05/(14{,}249 \times 10) = 3.51 \times 10^{-7}$; dashed red line). The LoF-segment burden analysis (with SNP adjustment) used 303,125 UK Biobank samples not included in the exome sequencing cohort. We identified one locus previously reported by Van Hout et al. [89] (black label), and additional loci on chromosomes 6,9 and 12 (labels in red).

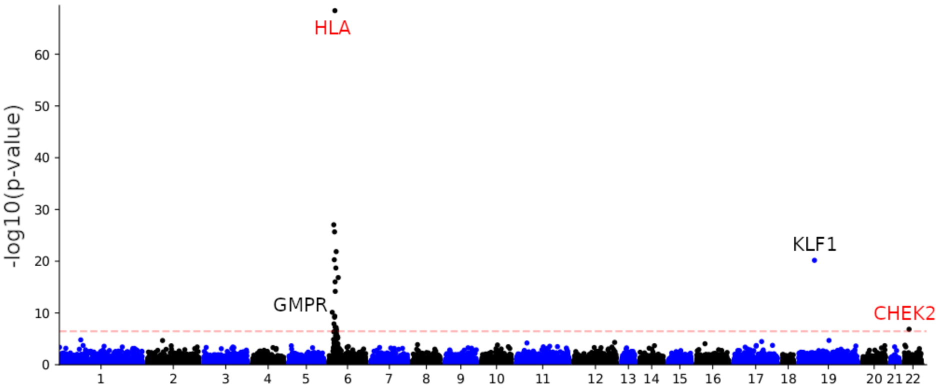


Figure A.2: LoF-segment burden exome-wide Manhattan plot for mean corpuscular haemoglobin. Labelled genes are exome-wide significant (after adjusting for multiple testing, t-test p-value $< 0.05/(14{,}249 \times 10) = 3.51 \times 10^{-7}$; dashed red line). The LoF-segment burden analysis (with SNP adjustment) used 303,125 UK Biobank samples not included in the exome sequencing cohort. We identified two loci previously reported by Van Hout et al. [89], *KLF1* and *GMPR* (gene labels in black), and two novel associations at HLA and *CHEK2* (labels in red).

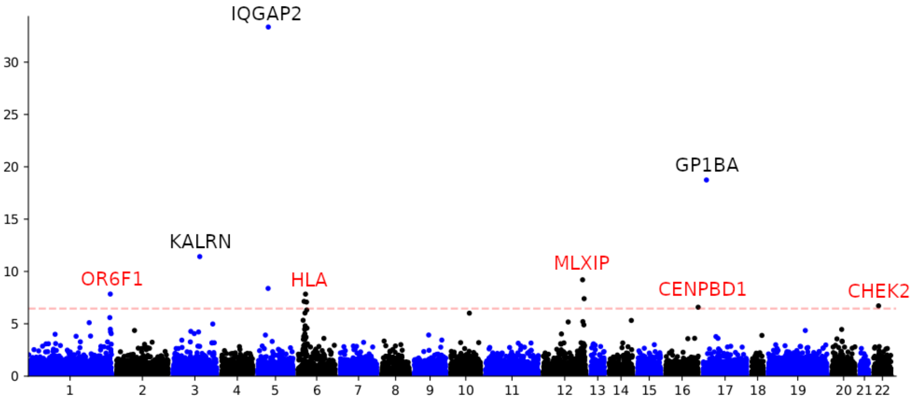


Figure A.3: LoF-segment burden exome-wide Manhattan plot for the mean platelet (thrombocyte) volume. Labelled genes are in exome-wide significance (after adjusting for multiple testing, t-test p-value $< 0.05/(14{,}249 \times 10) = 3.51 \times 10^{-7}$; dashed red line); the y-axis shows $-\log(\text{p-value})$. The test statistic was computed using 303,125 samples within the UK Biobank cohort. The Path-IBD method detected three previously-reported associated loci, *KALRN, GP1BA* and *IQGAP2* (gene labels in black), and additional hits at chromosomes 1,6,12,16 and 22 (labels in red).

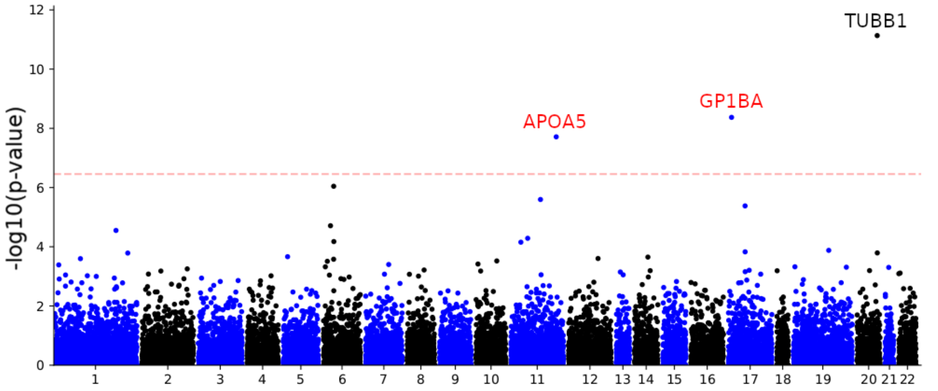


Figure A.4: LoF-segment burden exome-wide Manhattan plot for platelet distr width. Labelled genes are exome-wide significant (after adjusting for multiple testing, t-test p-value $< 0.05/(14{,}249 \times 10) = 3.51 \times 10^{-7}$; dashed red line). The LoF-segment burden analysis (with SNP adjustment) used 303,125 UK Biobank samples not included in the exome sequencing cohort. We identified one locus previously reported by Van Hout et al. [89], *TUBB1* (black label), and two additional genes, *APOA5* and *GP1BA* (labels in red).

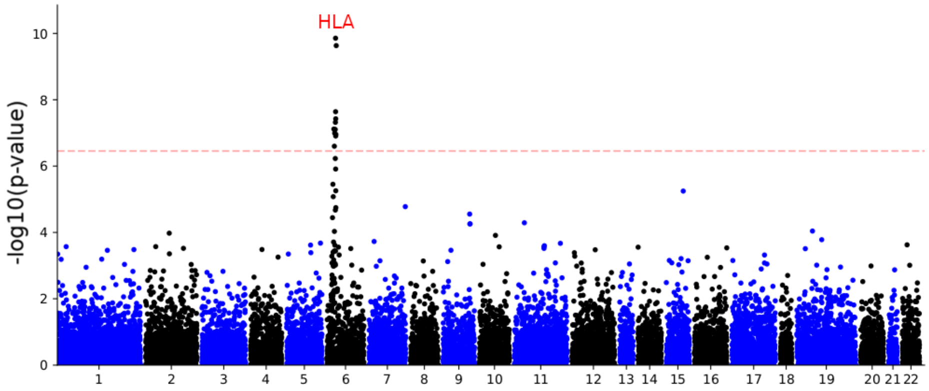


Figure A.5: LoF-segment burden exome-wide Manhattan plot for red blood cell count. Labelled genes are exome-wide significant (after adjusting for multiple testing, t-test p-value $< 0.05/(14{,}249 \times 10) = 3.51 \times 10^{-7}$; dashed red line). The LoF-segment burden analysis (with SNP adjustment) used 303,125 UK Biobank samples not included in the exome sequencing cohort. We detected one novel association at the HLA locus which was not detected by either of the WES-based LoF burden tests.

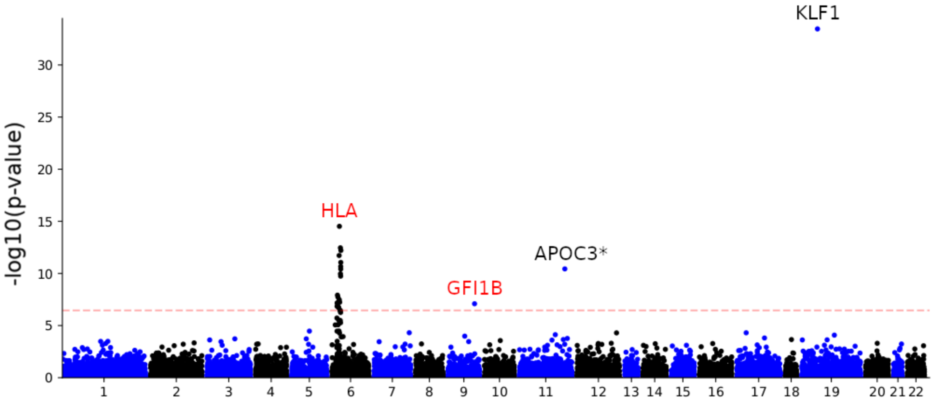


Figure A.6: LoF-segment burden exome-wide Manhattan plot for red blood cell distribution width. Labelled genes are exome-wide significant (after adjusting for multiple testing, t-test p-value $< 0.05/(14{,}249 \times 10) = 3.51 \times 10^{-7}$; dashed red line). The LoF-segment burden analysis (with SNP adjustment) used 303,125 UK Biobank samples not included in the exome sequencing cohort. We identified two previously-reported loci (black labels), *KLF1* (detected by Van Hout et al. [89]) and *APOC3* (detected by our WES-based LoF burden analysis), and two additional (labels in red).

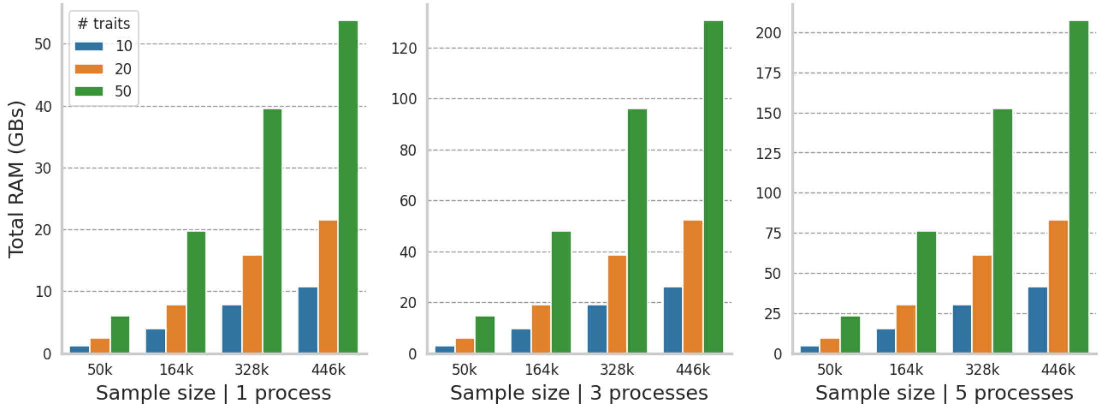


Figure A.7: Memory usage for `FMA`. This behaviour is mainly driven by the matrices required for the conjugate gradients iteration, most of which have dimension $N \times [22N_T(1+N_c)]$, for N samples, N_T traits, and N_c markers for estimating the calibration factor, assuming 22 chromosomes. The number of such matrices is $2+5P$, for P parallel processes, as each process requires a few such matrices to be kept in memory, and each contains floats requiring 8 bytes per element. For example, a full-UKBB analysis of $N = 446$k and 10 phenotypes will cost $\sim$ 40GB.

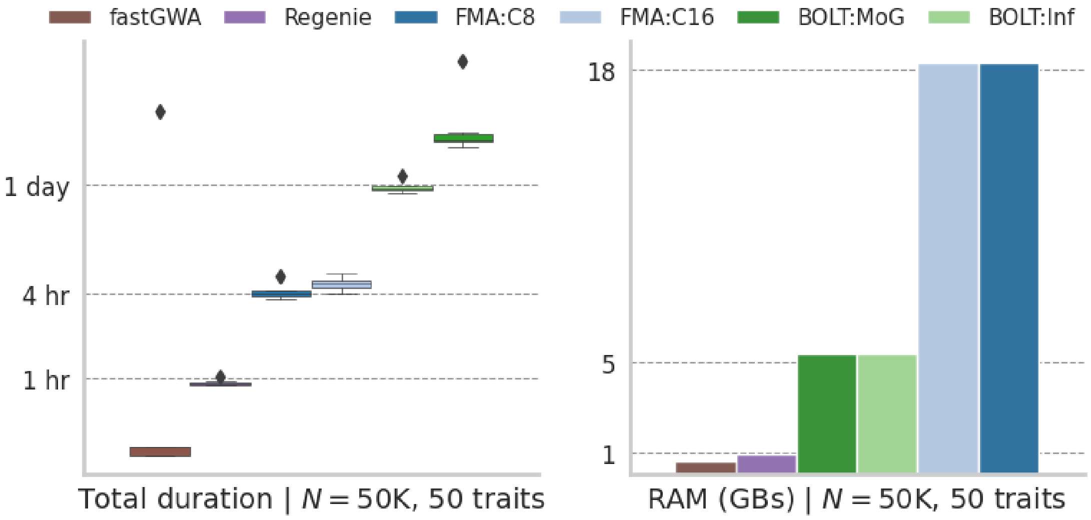


Figure A.8: Total computational costs for $N = 50$k samples and 50 synthetic phenotypes. Left: Total running times for model fitting and testing for each of the six methods, averaged for the 8 cases of synthetic phenotypes, using $M = 387{,}700$ variants. Diamonds correspond to outliers. `fastGWA` had a large variance because the algorithm did not converge for the CSR samples. **Right:** The corresponding memory usage, considering all the steps required for association (e.g. h^2 estimation, model-fitting etc).

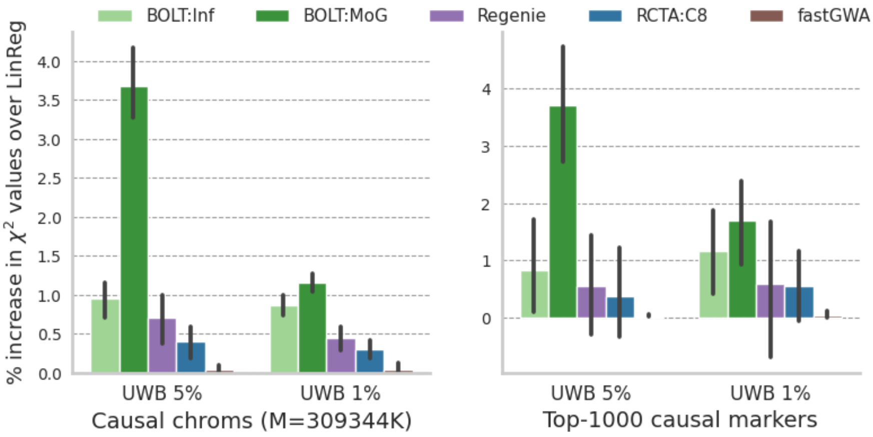


Figure A.9: Comparison of power for $N = 50$k unrelated British samples (UWB) using 22 chromosomes. Similarly to Fig. 4.1, we considered two types of genetic architecture (1% or 5% polygenicity; $h^2 = 0.25$), but limited this experiment to 10 replicates only due to computational constrains.

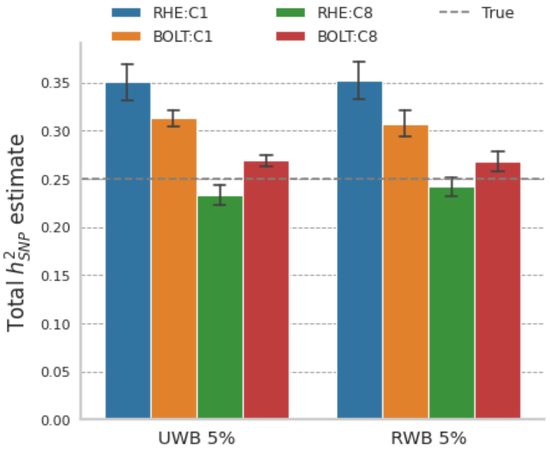


Figure A.10: Comparison between single and multiple components approaches for h^2_{SNP} estimation. Showing are results for the sets of unrelated (UWB) and related British (RWB), for the case of 5% polygenicity, where the true heritability is set at 0.25. I report estimates for sets of 10 replicates per case due to computational constraints, as `BOLT-REML` would require 10+ days for the complete set of 50 phenotypes. C1 stands for a single component, whereas C8 corresponds to the `2MAF x 4LD` annotation.

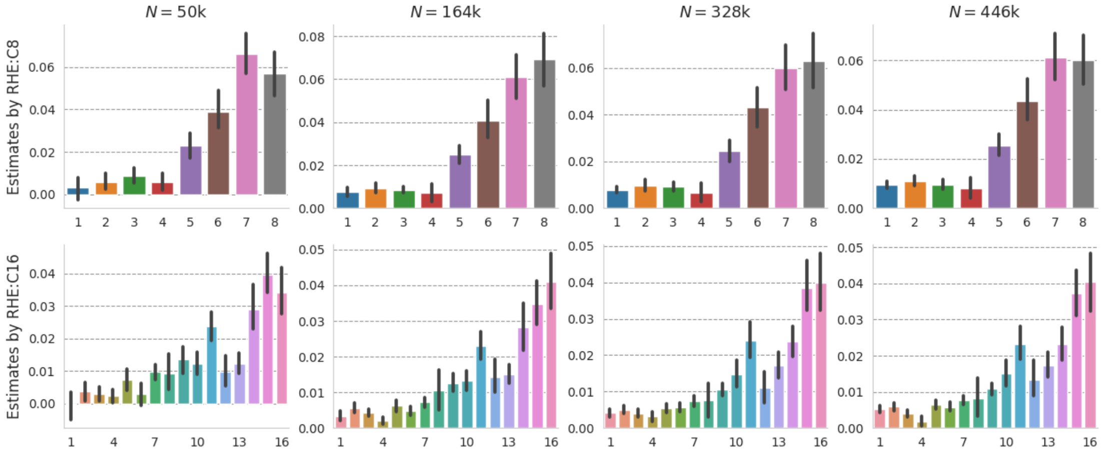


Figure A.11: Variance component estimates averaged across the 20 real phenotypes. Per component estimates of heritability using `RHE:mc` either with 8 (2MAF × 4LD), or 16 (4MAF × 4LD) components, applied to $N = 50$k, $N = 164$k, $N = 328$k, and $N = 446$k individuals respectively, and $M = 623{,}128$ genotyped variants.

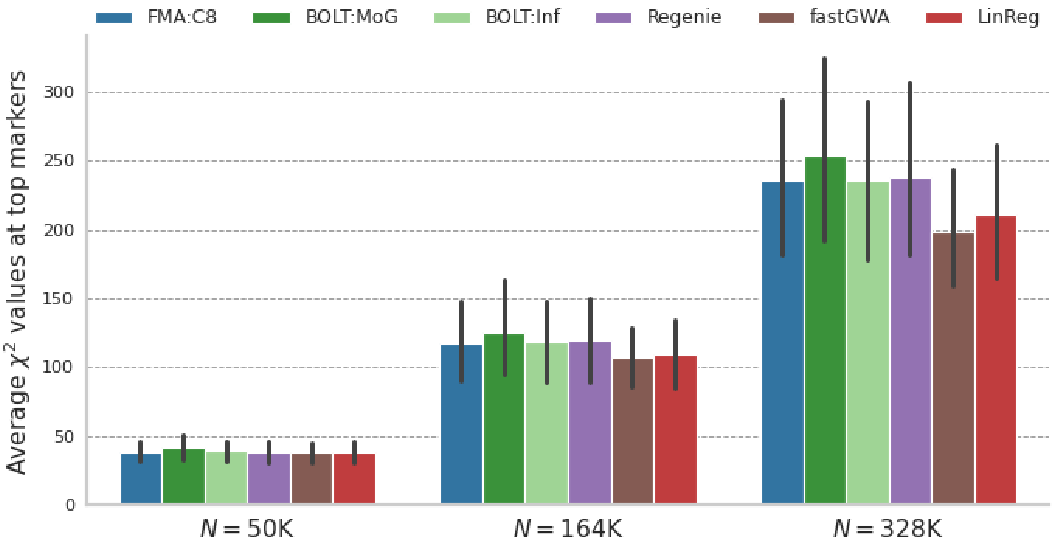


Figure A.12: Average χ^2 at top genotyped variants for 20 real quantitative phenotypes. I compare the average χ^2 value across "top" variants, defined as the intersection of `BOLT:MoG`'s, `fastGWA`'s, and `REGENIE`'s 1000 most significant associations, resulting in 378, 1859, and 4071 markers for $N = 50$k, $N = 164$k, and $N = 328$k, on average, respectively. Bars correspond to standard deviations for the 20 phenotypes. All three samples consist of unrelated British (UKBB).

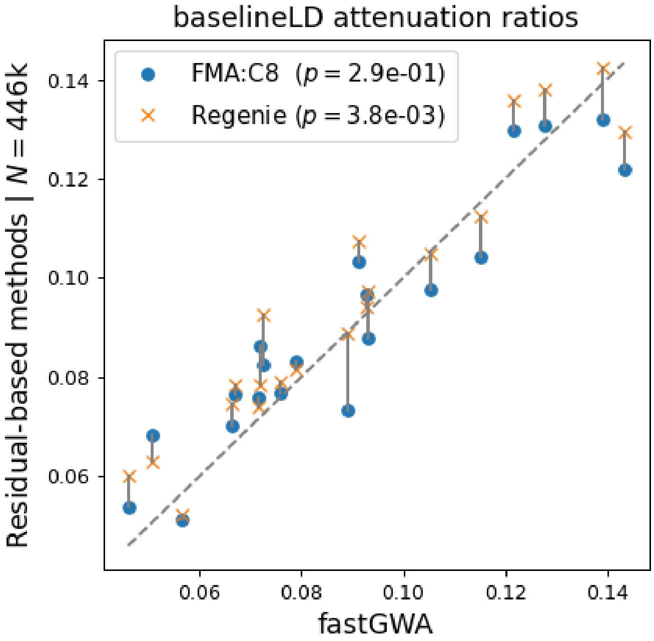


Figure A.13: Comparison of attenuation ratios for residual-based approaches. Complementing Fig. 4.7, I compare the LD score regression attenuation ratios between `fastGWA` and `FMA:C8` or `REGENIE`, all by using residualised phenotypes instead of explicitly conditioning on covariates. `REGENIE`'s attenuation ratios were significantly larger than any other approach.

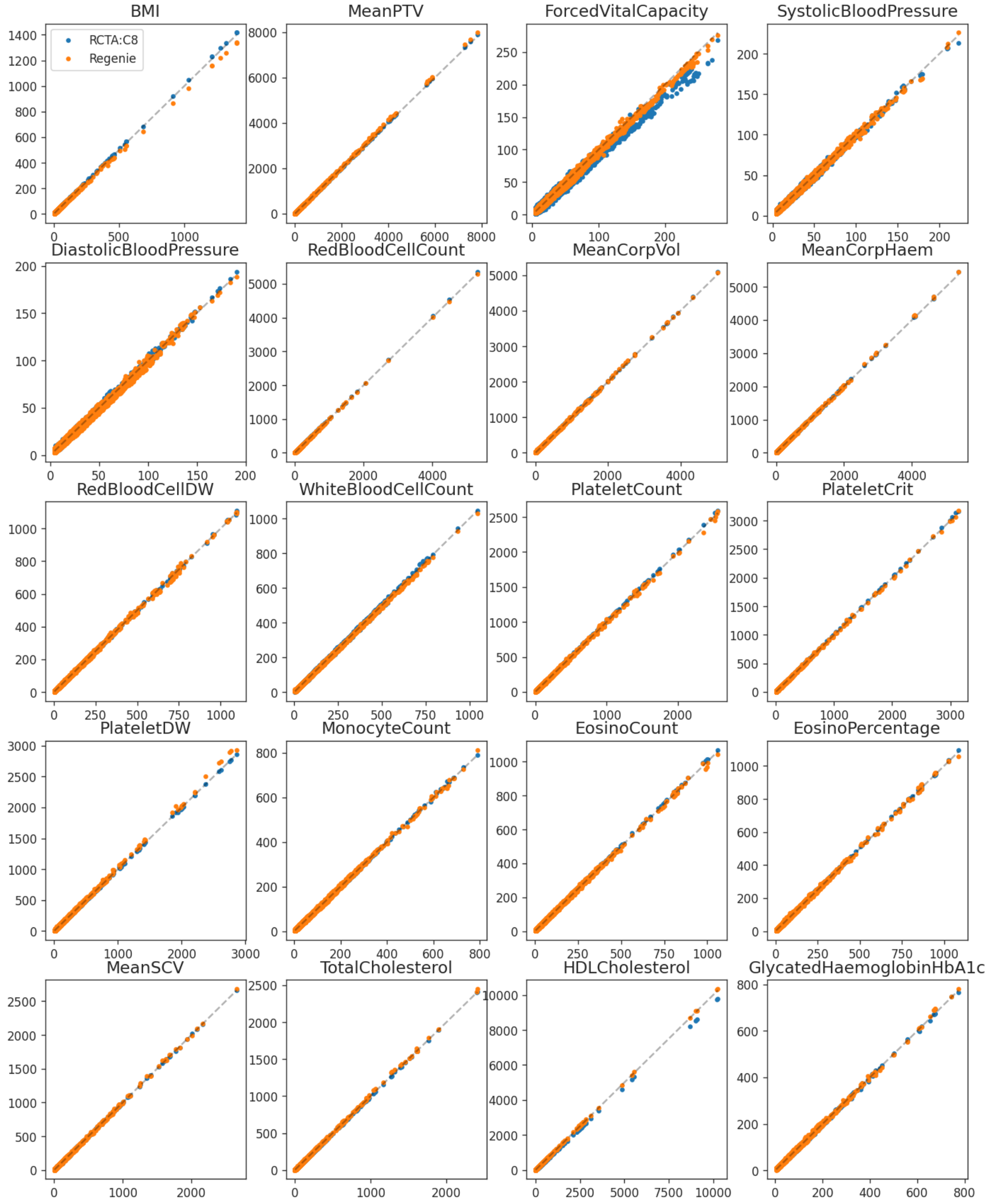
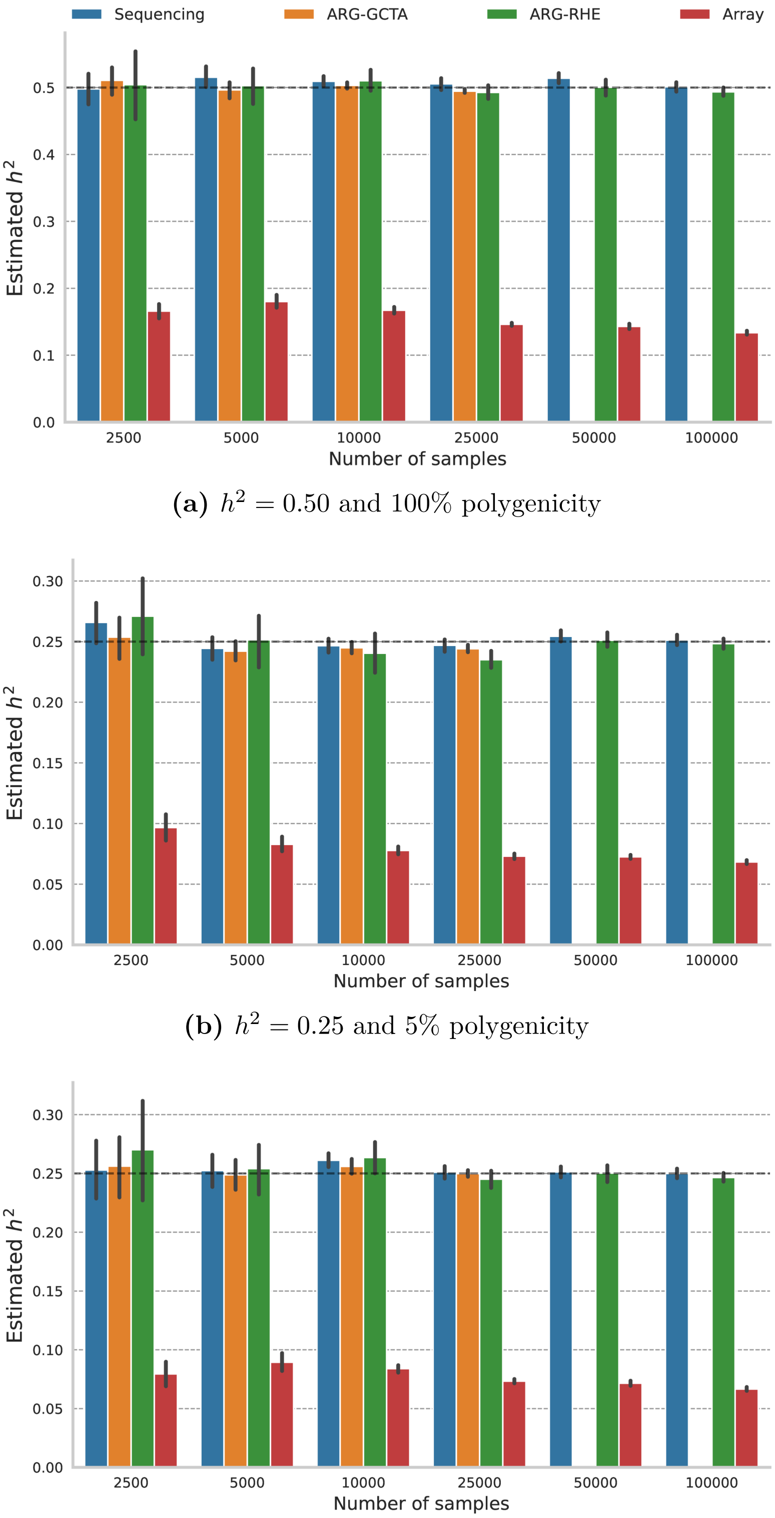


Figure A.14: QQ-plots comparing 3 different methods on 20 UKBB quantitative phenotypes and genotyped variants. X-axis corresponds to `BOLT:Inf` and y-axis is either `FMA:C8` or `REGENIE`, where each method was applied to the $N = 446\text{K}$ sample of British and Europeans. Shown are variants from the largest decile of the distribution, i.e. roughly 62,300 variants, determined by `BOLT:Inf`.



(a) $h^2 = 0.50$ and 100% polygenicity

(b) $h^2 = 0.25$ and 5% polygenicity

(c) $h^2 = 0.25$ and 100% polygenicity

Figure A.15: Estimation of heritability using ARG-based GRMs. Estimation of narrow-sense heritability using ARG-based GRMs, considering different levels of polygenicity and heritability, similarly to Figure 6.2. `ARG-GCTA` was not applied to samples with $N \geq 50{,}000$ because of time constraints.

B
Supplementary tables

	Mean χ^2 at all causal	Mean χ^2 at non-causal	Power	Type I error (se; pval)
FMA:C1	1.99 (0.013)	1.011 (0.002)	0.125 (6.55e-04)	**0.051** (1.95e-04; 0.0e+00)
FMA:C8	1.97 (0.012)	1.000 (0.002)	0.124 (6.11e-04)	0.050 (1.85e-04; 8.8e-01)
FMA:C16	1.97 (0.012)	1.001 (0.002)	0.124 (5.85e-04)	0.050 (1.86e-04; 9.3e-01)
FMA:AprLoco	1.99 (0.015)	0.973 (0.002)	0.127 (8.65e-04)	0.048 (2.60e-04; 0.0e+00)
FMA:Pruned	2.04 (0.013)	1.029 (0.002)	0.129 (6.39e-04)	**0.053** (2.18e-04; 0.0e+00)
BOLT:Inf	1.98 (0.012)	1.006 (0.002)	0.125 (6.47e-04)	0.050 (1.84e-04; 4.5e-01)
BOLT:MoG	2.05 (0.013)	1.011 (0.002)	0.130 (7.14e-04)	**0.051** (1.86e-04; 0.0e+00)
Regenie	1.97 (0.012)	1.001 (0.001)	0.124 (5.57e-04)	0.050 (1.78e-04; 6.2e-01)
Regenie+PCA	1.97 (0.012)	1.000 (0.001)	0.124 (5.59e-04)	0.050 (1.78e-04; 7.0e-01)
fastGWA	1.94 (0.012)	1.002 (0.001)	0.123 (6.51e-04)	0.050 (1.82e-04; 2.3e-01)
fastGWA+PCA	1.94 (0.012)	1.002 (0.001)	0.123 (6.55e-04)	0.050 (1.84e-04; 4.1e-01)
Lin Reg	1.94 (0.012)	1.002 (0.001)	0.124 (6.45e-04)	0.050 (1.82e-04; 9.6e-02)
LinReg+PCA	1.94 (0.012)	1.002 (0.001)	0.123 (6.56e-04)	0.050 (1.84e-04; 3.5e-01)

Table B.1: Results for the UWB sample with 1% **polygenicity.** I report the average test statistic at 3,900 causal variants (sampled uniformly from the odd chromosomes), and at 196,509 variants from the even chromosomes which are treated as null. Power is calculated as the proportion of detected out of all causal variants, and type I error as the proportion of falsely determined causal, out of all non-causal. Parentheses report the standard errors of the mean, besides the last column which also reports the p-value of the corresponding z-test. For Type I error I used 4.5e-3 for nominal significance, adjusting for the 11 methods assessed in Chapter 4 and bold fonts indicate significant inflation based on that. `FMA:AprLoco` and `FMA:Pruned` are described in Chapter 5.

	Mean χ^2 at all causal	Mean χ^2 at non-causal	Power	Type I error (se; pval)
FMA:C1	1.57 (0.005)	1.013 (0.002)	0.107 (4.23e-04)	**0.051** (2.19e-04; 0.0e+00)
FMA:C8	1.55 (0.004)	1.001 (0.001)	0.106 (4.30e-04)	0.050 (1.93e-04; 5.5e-01)
FMA:C16	1.55 (0.004)	1.002 (0.001)	0.106 (4.29e-04)	0.050 (1.89e-04; 1.9e-01)
FMA:AprLoco	1.60 (0.005)	0.978 (0.002)	0.109 (4.21e-04)	0.048 (2.68e-04; 0.0e+00)
FMA:Pruned	1.55 (0.004)	0.995 (0.001)	0.106 (4.08e-04)	0.049 (1.84e-04; 1.2e-03)
BOLT:Inf	1.56 (0.004)	1.008 (0.001)	0.107 (4.02e-04)	0.050 (1.89e-04; 3.7e-02)
BOLT:MoG	1.57 (0.004)	1.009 (0.001)	0.108 (4.26e-04)	0.050 (1.74e-04; 1.0e-02)
Regenie	1.55 (0.004)	1.001 (0.001)	0.106 (3.74e-04)	0.050 (1.81e-04; 3.1e-01)
Regenie+PCA	1.55 (0.004)	1.001 (0.001)	0.106 (3.77e-04)	0.050 (1.80e-04; 3.9e-01)
fastGWA	1.53 (0.004)	1.003 (0.001)	0.104 (4.06e-04)	0.050 (1.89e-04; 1.2e-01)
fastGWA+PCA	1.53 (0.004)	1.002 (0.001)	0.104 (4.16e-04)	0.050 (1.80e-04; 3.1e-01)
Lin Reg	1.53 (0.004)	1.003 (0.001)	0.105 (4.11e-04)	0.050 (1.88e-04; 2.6e-01)
LinReg+PCA	1.53 (0.004)	1.003 (0.001)	0.104 (4.14e-04)	0.050 (1.80e-04; 2.5e-01)

Table B.2: Results for the UWB sample with 5% **polygenicity.** Description as in Table B.1, with the difference that this case involves 18,803 causal variants.

	Mean χ^2	Mean χ^2		
	at all causal	at non-causal	Power	Type I error (se; pval)
FMA:C1	2.00 (0.019)	1.005 (0.002)	0.126 (8.08e-04)	**0.051** (1.90e-04; 3.1e-03)
FMA:C8	1.99 (0.018)	1.000 (0.001)	0.125 (7.60e-04)	0.050 (1.74e-04; 8.9e-01)
FMA:C16	1.99 (0.018)	1.001 (0.002)	0.125 (7.40e-04)	0.050 (1.69e-04; 4.3e-01)
FMA:AprLoco	1.99 (0.015)	0.973 (0.002)	0.127 (8.65e-04)	0.048 (2.60e-04; 0.0e+0)
FMA:Pruned	2.05 (0.020)	1.024 (0.002)	0.130 (8.60e-04)	**0.053** (2.54e-04; 0.0e+00)
BOLT:Inf	1.99 (0.018)	1.004 (0.001)	0.126 (7.83e-04)	0.050 (1.59e-04; 9.5e-01)
BOLT:MoG	2.06 (0.020)	1.006 (0.002)	0.130 (7.67e-04)	0.050 (1.94e-04; 2.3e-01)
Regenie	1.98 (0.018)	0.992 (0.001)	0.125 (7.36e-04)	0.049 (1.58e-04; 0.0e+00)
Regenie+PCA	1.98 (0.018)	0.992 (0.001)	0.125 (7.41e-04)	0.049 (1.59e-04; 0.0e+00)
fastGWA	1.95 (0.017)	0.999 (0.001)	0.122 (7.38e-04)	0.050 (1.59e-04; 4.5e-01)
fastGWA+PCA	1.95 (0.017)	0.998 (0.001)	0.122 (7.41e-04)	0.050 (1.55e-04; 1.4e-01)
Lin Reg	1.98 (0.017)	1.013 (0.001)	0.125 (7.84e-04)	**0.051** (1.68e-04; 0.0e+00)
LinReg+PCA	1.98 (0.017)	1.013 (0.001)	0.124 (7.80e-04)	**0.051** (1.68e-04; 0.0e+00)

Table B.3: Results for the RWB sample with 1% **polygenicity.** Description as in Table B.1, with the difference that this case involves relatedness.

	Mean χ^2	Mean χ^2		
	at all causal	at non-causal	Power	Type I error (se; pval)
FMA:C1	1.59 (0.005)	1.008 (0.002)	0.107 (4.15e-04)	**0.051** (2.26e-04; 0.0e+00)
FMA:C8	1.58 (0.005)	1.001 (0.002)	0.106 (3.90e-04)	0.050 (2.13e-04; 5.9e-01)
FMA:C16	1.58 (0.005)	1.003 (0.002)	0.106 (3.65e-04)	0.050 (2.23e-04; 9.8e-02)
FMA:AprLoco	1.60 (0.005)	0.978 (0.002)	0.109 (4.21e-04)	0.048 (2.68e-04; 0.0e+00)
FMA:Pruned	1.59 (0.005)	0.994 (0.002)	0.107 (4.18e-04)	0.049 (2.10e-04; 2.0e-04)
BOLT:Inf	1.59 (0.005)	1.005 (0.002)	0.108 (3.91e-04)	0.050 (2.13e-04; 8.1e-01)
BOLT:MoG	1.59 (0.005)	1.005 (0.002)	0.108 (4.02e-04)	0.050 (2.14e-04; 6.4e-01)
Regenie	1.58 (0.005)	0.995 (0.002)	0.106 (3.82e-04)	0.049 (1.97e-04; 4.7e-03)
Regenie+PCA	1.57 (0.005)	0.995 (0.002)	0.106 (3.77e-04)	0.049 (1.96e-04; 3.9e-03)
fastGWA	1.56 (0.005)	1.003 (0.002)	0.105 (3.83e-04)	0.050 (2.06e-04; 2.0e-01)
fastGWA+PCA	1.56 (0.005)	1.002 (0.002)	0.105 (3.79e-04)	0.050 (2.08e-04; 4.2e-01)
Lin Reg	1.58 (0.005)	1.017 (0.002)	0.108 (3.80e-04)	**0.051** (2.17e-04; 0.0e+00)
LinReg+PCA	1.58 (0.005)	1.017 (0.002)	0.107 (3.82e-04)	**0.052** (2.17e-04; 0.0e+00)

Table B.4: Results for the RWB sample with 5% **polygenicity.** Description as in Table B.1, with the difference that this case involves 18,921 causal variants.

	Mean χ^2	Mean χ^2		
	at all causal	at non-causal	Power	Type I error (se; pval)
FMA:C1	1.87 (0.009)	1.011 (0.002)	0.122 (7.04e-04)	**0.051** (2.07e-04; 0.0e+00)
FMA:C8	1.86 (0.008)	1.004 (0.002)	0.121 (7.00e-04)	0.051 (2.07e-04; 8.2e-03)
FMA:C16	1.86 (0.008)	1.005 (0.002)	0.121 (6.37e-04)	**0.051** (1.96e-04; 1.9e-03)
FMA:AprLoco	1.85 (0.009)	0.973 (0.002)	0.121 (7.92e-04)	0.047 (2.05e-04; 0.0e+00)
FMA:Pruned	1.92 (0.009)	1.026 (0.002)	0.126 (6.97e-04)	**0.053** (2.19e-04; 0.0e+00)
BOLT:Inf	1.87 (0.009)	1.008 (0.002)	0.122 (7.38e-04)	0.050 (2.14e-04; 2.5e-02)
BOLT:MoG	1.93 (0.010)	1.012 (0.002)	0.126 (7.80e-04)	**0.051** (1.84e-04; 0.0e+00)
Regenie	1.87 (0.009)	1.224 (0.017)	0.122 (6.84e-04)	**0.076** (1.97e-03; 0.0e+00)
Regenie+PCA	1.86 (0.009)	1.002 (0.002)	0.121 (6.73e-04)	0.050 (1.83e-04; 1.8e-01)
fastGWA	1.89 (0.009)	1.073 (0.007)	0.126 (8.86e-04)	**0.059** (8.03e-04; 0.0e+00)
fastGWA+PCA	1.83 (0.008)	1.003 (0.002)	0.119 (6.56e-04)	0.050 (1.85e-04; 2.8e-02)
Lin Reg	1.91 (0.010)	1.093 (0.008)	0.129 (1.03e-03)	**0.060** (1.01e-03; 0.0e+00)
LinReg+PCA	1.83 (0.008)	1.007 (0.002)	0.120 (6.48e-04)	**0.051** (1.87e-04; 0.0e+00)

Table B.5: Results for the EUR sample with 1% polygenicity. Description as in Table B.1, with the difference that this case involves population stratification.

	Mean χ^2	Mean χ^2		
	at all causal	at non-causal	Power	Type I error (se; pval)
FMA:C1	1.57 (0.007)	1.013 (0.002)	0.107 (4.30e-04)	**0.052** (2.28e-04; 0.0e+00)
FMA:C8	1.56 (0.006)	1.003 (0.002)	0.105 (3.68e-04)	0.051 (2.22e-04; 1.4e-02)
FMA:C16	1.56 (0.006)	1.004 (0.002)	0.106 (4.02e-04)	0.051 (2.17e-04; 1.0e-02)
FMA:AprLoco	1.54 (0.005)	0.971 (0.003)	0.105 (4.44e-04)	0.047 (2.78e-04; 0.0e+00)
FMA:Pruned	1.56 (0.007)	0.993 (0.002)	0.105 (4.00e-04)	0.049 (2.09e-04; 1.2e-03)
BOLT:Inf	1.56 (0.006)	1.008 (0.002)	0.107 (4.14e-04)	0.051 (2.25e-04; 1.3e-02)
BOLT:MoG	1.57 (0.006)	1.010 (0.002)	0.108 (4.24e-04)	**0.051** (2.18e-04; 1.0e-03)
Regenie	1.57 (0.008)	1.315 (0.021)	0.107 (6.01e-04)	**0.086** (2.41e-03; 0.0e+00)
Regenie+PCA	1.55 (0.006)	1.001 (0.002)	0.105 (4.10e-04)	0.050 (2.05e-04; 3.0e-01)
fastGWA	1.62 (0.013)	1.102 (0.011)	0.113 (1.23e-03)	**0.062** (1.36e-03; 0.0e+00)
fastGWA+PCA	1.53 (0.006)	1.002 (0.002)	0.104 (4.07e-04)	0.050 (2.20e-04; 4.2e-01)
Lin Reg	1.65 (0.015)	1.130 (0.014)	0.117 (1.53e-03)	**0.065** (1.71e-03; 0.0e+00)
LinReg+PCA	1.54 (0.006)	1.006 (0.002)	0.104 (4.10e-04)	**0.051** (2.18e-04; 2.1e-03)

Table B.6: Results for the EUR sample with 5% polygenicity. Description as in Table B.1, with the difference that this case involves population stratification and 19,081 causal variants.

	Mean χ^2 at all causal	Mean χ^2 at non-causal	Power	Type I error (se; pval)
FMA:C1	1.88 (0.011)	1.004 (0.002)	0.120 (7.55e-04)	0.051 (2.72e-04; 3.4e-02)
FMA:C8	1.86 (0.011)	0.988 (0.002)	0.118 (7.19e-04)	0.049 (2.52e-04; 0.0e+00)
FMA:C16	1.85 (0.011)	0.989 (0.002)	0.118 (7.65e-04)	0.049 (2.27e-04; 0.0e+00)
FMA:AprLoco	1.56 (0.035)	0.927 (0.037)	0.099 (2.72e-03)	0.045 (2.93e-03; 1.1e-01)
BOLT:Inf	1.87 (0.011)	0.992 (0.001)	0.119 (7.41e-04)	0.049 (1.65e-04; 0.0e+00)
BOLT:MoG	1.93 (0.012)	0.996 (0.001)	0.123 (7.40e-04)	0.049 (1.62e-04; 0.0e+00)
Regenie	17.19 (0.971)	19.674 (0.982)	0.493 (1.19e-02)	**0.526** (1.26e-02; 0.0e+00)
Regenie+PCA	1.88 (0.011)	0.996 (0.001)	0.119 (7.55e-04)	0.050 (1.71e-04; 2.4e-02)
fastGWA	2.12 (0.048)	1.286 (0.049)	0.146 (4.25e-03)	**0.082** (5.04e-03; 0.0e+00)
fastGWA+PCA	1.82 (0.020)	0.995 (0.002)	0.114 (2.44e-03)	0.048 (1.02e-03; 1.4e-01)
Lin Reg	13.74 (2.206)	13.348 (2.300)	0.418 (2.55e-02)	**0.392** (2.85e-02; 0.0e+00)
LinReg+PCA	1.87 (0.011)	1.012 (0.001)	0.120 (7.62e-04)	**0.051** (1.80e-04; 0.0e+00)

Table B.7: Results for the CSR sample with 1% **polygenicity.** Description as in Table B.1, with the difference that this case involves continental structure and relatedness.

	Mean χ^2 at all causal	Mean χ^2 at non-causal	Power	Type I error (se; pval)
FMA:C1	1.49 (0.005)	0.994 (0.002)	0.100 (4.88e-04)	0.049 (2.41e-04; 1.3e-02)
FMA:C8	1.50 (0.005)	1.004 (0.002)	0.101 (4.91e-04)	0.051 (2.08e-04; 9.7e-03)
FMA:C16	1.48 (0.004)	0.987 (0.002)	0.099 (4.45e-04)	0.049 (1.74e-04; 0.0e+00)
FMA:AprLoco	1.47 (0.027)	0.910 (0.026)	0.093 (2.27e-03)	0.042 (2.25e-03; 3.0e-04)
BOLT:Inf	1.49 (0.004)	0.993 (0.001)	0.101 (4.65e-04)	0.049 (1.64e-04; 0.0e+00)
BOLT:MoG	1.49 (0.005)	0.995 (0.001)	0.101 (4.74e-04)	0.049 (1.67e-04; 0.0e+00)
Regenie	20.75 (1.102)	23.390 (1.331)	0.520 (1.46e-02)	**0.559** (1.53e-02; 0.0e+00)
Regenie+PCA	1.50 (0.004)	0.997 (0.001)	0.101 (4.12e-04)	0.050 (1.64e-04; 7.3e-02)
fastGWA	1.74 (0.039)	1.263 (0.038)	0.128 (3.95e-03)	**0.080** (4.32e-03; 0.0e+00)
fastGWA+PCA	1.47 (0.005)	0.998 (0.002)	0.099 (5.01e-04)	0.050 (2.80e-04; 5.3e-01)
Lin Reg	12.56 (1.855)	12.049 (1.844)	0.403 (2.59e-02)	**0.385** (2.80e-02; 0.0e+00)
LinReg+PCA	1.50 (0.004)	1.013 (0.001)	0.102 (4.15e-04)	**0.051** (1.67e-04; 0.0e+00)

Table B.8: Results for the CSR sample with 5% **polygenicity.** Description as in Table B.1, with the difference that this case involves continental structure and relatedness, and 18,929 causal variants.

Sample	Methods	Mean χ^2 (top)	p-value
328k	BOLT:MoG vs BOLT:Inf	253.54 vs 235.37	6.6e-03
328k	BOLT:MoG vs FMA:C8	253.54 vs 235.39	6.7e-03
328k	BOLT:MoG vs Regenie	253.54 vs 238.36	7.0e-03
328k	**BOLT:MoG** vs fastGWA	253.54 vs 198.25	2.2e-03
328k	**BOLT:MoG** vs LinReg	253.54 vs 210.32	2.3e-03
328k	BOLT:Inf vs FMA:C8	235.37 vs 235.39	4.9e-01
328k	BOLT:Inf vs Regenie	235.37 vs 238.36	1.0e-02
328k	**BOLT:Inf** vs fastGWA	235.37 vs 198.25	1.5e-03
328k	**BOLT:Inf** vs LinReg	235.37 vs 210.32	1.1e-03
328k	FMA:C8 vs Regenie	235.39 vs 238.36	3.6e-02
328k	**FMA:C8** vs fastGWA	235.39 vs 198.25	1.4e-03
328k	**FMA:C8** vs LinReg	235.39 vs 210.32	1.2e-03
328k	**Regenie** vs fastGWA	238.36 vs 198.25	1.6e-03
328k	**Regenie** vs LinReg	238.36 vs 210.32	1.3e-03
328k	fastGWA vs LinReg	198.25 vs 210.32	8.8e-03

Table B.9: Pairwise comparisons of χ^2 statistics for 20 real phenotypes and $N = 328$k. For each pair of methods, I compare the distribution of average test statistic at top variants, considering all the 20 phenotypes. The set of top variants was defined as the intersection of `BOLT:MoG`'s, `fastGWA`'s, and `REGENIE`'s 1000 most significant associations. I used paired t-tests to assess if any averages are significantly different, reporting the corresponding p-values in the last column. Bold fonts indicate statistically significant differences, using the threshold of $0.05/15 = 0.0033$ to control for multiple testing.

scenario			`ARG-RHE`		`ARG-GCTA`	
h^2	pol	N	estimate	SE	estimate	SE
0.25	100%	2500	0.2702	0.0570	0.2563	0.0438
0.25	100%	5000	0.2541	0.0291	0.2488	0.0252
0.25	100%	10000	0.2635	0.0163	0.2560	0.0156
0.25	100%	25000	0.2451	0.0105	0.2499	0.0089
0.5	100%	2500	0.5042	0.0636	0.5107	0.0425
0.5	100%	5000	0.5028	0.0348	0.4964	0.0248
0.5	100%	10000	0.5100	0.0211	0.5031	0.0154
0.5	100%	25000	0.4927	0.0166	0.4947	0.0090
0.25	5%	2500	0.2710	0.0561	0.2537	0.0434
0.25	5%	5000	0.2515	0.0310	0.2422	0.0251
0.25	5%	10000	0.2405	0.0179	0.2449	0.0155
0.25	5%	25000	0.2351	0.0110	0.2441	0.0089
0.5	5%	2500	0.5205	0.0618	0.4981	0.0423
0.5	5%	5000	0.5006	0.0370	0.4920	0.0248
0.5	5%	10000	0.4801	0.0243	0.4938	0.0155
0.5	5%	25000	0.4728	0.0173	0.4890	0.0090

Table B.10: Evaluation of h^2 estimation using genealogies. This table reports the estimates obtained by either `ARG-RHE` or `ARG-GCTA`, as in figure 6.2, but this time reporting standard errors (SE), averaged for the 20 phenotypes. Pol refers to polygenicity.

References

[1] International Human Genome Sequencing Consortium. “Initial sequencing and analysis of the human genome”. In: *Nature* 409.6822 (2001), pp. 860–921.

[2] International HapMap Consortium et al. “A haplotype map of the human genome”. In: *Nature* 437.7063 (2005), p. 1299.

[3] 1000 Genomes Project Consortium et al. “A map of human genome variation from population scale sequencing”. In: *Nature* 467.7319 (2010), p. 1061.

[4] Clare Bycroft et al. “The UK Biobank resource with deep phenotyping and genomic data”. In: *Nature* 562.7726 (2018), pp. 203–209.

[5] Genevieve L Wojcik et al. “Genetic analyses of diverse populations improves discovery for complex traits”. In: *Nature* 570.7762 (2019), pp. 514–518.

[6] Ruth Dolly Johnson et al. “Leveraging genomic diversity for discovery in an HR-linked biobank: the UCLA ATLAS Community Health Initiative”. In: *medRxiv* (2021).

[7] John Novembre et al. “Genes mirror geography within Europe”. In: *Nature* 456.7218 (2008), p. 98.

[8] G Stamatoyannopoulos et al. “Genetics of the peloponnesean populations and the theory of extinction of the medieval peloponnesean Greeks.” In: *European Journal of Human Genetics* 25.5 (2017), pp. 637–645.

[9] Pier Francesco Palamara et al. “High-throughput inference of pairwise coalescence times identifies signals of selection and enriched disease heritability”. In: *Nature Genetics* 50.9 (2018), pp. 1311–1317.

[10] Abdel Abdellaoui et al. “Genetic correlates of social stratification in Great Britain”. In: *Nature human behaviour* 3.12 (2019), pp. 1332–1342.

[11] Juba Nait Saada et al. “Identity-by-descent detection across 487,409 British samples reveals fine scale population structure and ultra-rare variant associations”. In: *Nature Communications* 11.1 (2020), pp. 1–15.

[12] Wellcome Trust Case Control Consortium. “Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls”. In: *Nature* 447.7145 (2007), pp. 661–678.

[13] Mark I McCarthy et al. “Genome-wide association studies for complex traits: consensus, uncertainty and challenges”. In: *Nature Reviews Genetics* 9.5 (2008), p. 356.

[14] Teri A Manolio et al. “Finding the missing heritability of complex diseases”. In: *Nature* 461.7265 (2009), pp. 747–753.

[15] William S Bush and Jason H Moore. “Genome-wide association studies”. In: *PLoS computational biology* 8.12 (2012), p. 28.

[16] Peter M Visscher et al. “10 years of GWAS discovery: biology, function, and translation”. In: *The American Journal of Human Genetics* 101.1 (2017), pp. 5–22.

[17] Emil Uffelmann et al. “Genome-wide association studies”. In: *Nature Reviews Methods Primers* 1.1 (2021), pp. 1–21.

[18] Abdel Abdellaoui et al. “15 years of GWAS discovery: Realizing the promise”. In: *The American Journal of Human Genetics* (2023).

[19] Matthew R Nelson et al. “The support of human genetic evidence for approved drug indications”. In: *Nature Genetics* 47.8 (2015), p. 856.

[20] Melina Claussnitzer et al. “A brief history of human disease genetics”. In: *Nature* 577.7789 (2020), pp. 179–189.

[21] Saori Sakaue et al. “Dimensionality reduction reveals fine-scale structure in the Japanese population with consequences for polygenic risk prediction”. In: *Nature Communications* 11.1 (2020), pp. 1–11.

[22] Patrick Turley et al. “Problems with using polygenic scores to select embryos”. In: *New England Journal of Medicine* 385.1 (2021), pp. 78–86.

[23] William J Astle et al. “The allelic landscape of human blood cell trait variation and links to common complex disease”. In: *Cell* 167.5 (2016), pp. 1415–1429.

[24] Loïc Yengo et al. “A saturated map of common genetic variants associated with human height”. In: *Nature* 610.7933 (2022), pp. 704–712.

[25] Anubha Mahajan et al. “Multi-ancestry genetic study of type 2 diabetes highlights the power of diverse populations for discovery and translation”. In: *Nature Genetics* 54.5 (2022), pp. 560–572.

[26] Alkes L Price et al. “New approaches to population stratification in genome-wide association studies”. In: *Nature Reviews Genetics* 11.7 (2010), pp. 459–463.

[27] Alexander I Young et al. “Deconstructing the sources of genotype-phenotype associations in humans”. In: *Science* 365.6460 (2019), pp. 1396–1400.

[28] Greg Gibson. “Rare and common variants: twenty arguments”. In: *Nature Reviews Genetics* 13.2 (2012), pp. 135–145.

[29] Gundula Povysil et al. “Rare-variant collapsing analyses for complex traits: guidelines and applications”. In: *Nature Reviews Genetics* (2019), pp. 1–13.

[30] Nadav Brandes, Omer Weissbrod, and Michal Linial. “Open problems in human trait genetics”. In: *Genome Biology* 23.1 (2022), pp. 1–32.

[31] Jonathan Marchini and Bryan Howie. “Genotype imputation for genome-wide association studies”. In: *Nature Reviews Genetics* 11.7 (2010), p. 499.

[32] Shane McCarthy et al. “A reference panel of 64,976 haplotypes for genotype imputation”. In: *Nature Genetics* 48.10 (2016), p. 1279.

[33] Iain Mathieson and Gil McVean. “Differential confounding of rare and common variants in spatially structured populations”. In: *Nature Genetics* 44.3 (2012), p. 243.

[34] Alicia R Martin et al. “Clinical use of current polygenic risk scores may exacerbate health disparities”. In: *Nature Genetics* 51.4 (2019), pp. 584–591.

[35] Roseann E Peterson et al. “Genome-wide association studies in ancestrally diverse populations: opportunities, methods, pitfalls, and recommendations”. In: *Cell* 179.3 (2019), pp. 589–603.

[36] Po-Ru Loh et al. “Efficient Bayesian mixed-model analysis increases association power in large cohorts”. In: *Nature Genetics* 47.3 (2015), p. 284.

[37] Wei Zhou et al. “Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies”. In: *Nature Genetics* 50.9 (2018), pp. 1335–1341.

[38] Longda Jiang et al. “A resource-efficient tool for mixed model association analysis of large-scale data”. In: *Nature Genetics* 51.12 (2019), pp. 1749–1755.

[39] Joelle Mbatchou et al. “Computationally efficient whole-genome regression for quantitative and binary traits”. In: *Nature Genetics* 53.7 (2021), pp. 1097–1103.

[40] Michael C Wu et al. “Rare-variant association testing for sequencing data with the sequence kernel association test”. In: *The American Journal of Human Genetics* 89.1 (2011), pp. 82–93.

[41] Wei Zhou et al. “Scalable generalized linear mixed model for region-based association tests in large biobanks and cohorts”. In: *Nature Genetics* 52.6 (2020), pp. 634–639.

[42] Seunggeung Lee et al. “Rare-variant association analysis: study designs and statistical tests”. In: *The American Journal of Human Genetics* 95.1 (2014), pp. 5–23.

[43] Han Chen et al. “Efficient variant set mixed model association tests for continuous and binary traits in large-scale whole-genome sequencing studies”. In: *The American Journal of Human Genetics* 104.2 (2019), pp. 260–274.

[44] Zhangchen Zhao et al. “UK biobank whole-exome sequence binary phenome analysis with robust region-based rare-variant test”. In: *The American Journal of Human Genetics* 106.1 (2020), pp. 3–12.

[45] Longda Jiang et al. “A generalized linear mixed model association tool for biobank-scale data”. In: *Nature Genetics* 53.11 (2021), pp. 1616–1621.

[46] Po-Ru Loh et al. “Mixed-model association for biobank-scale datasets”. In: *Nature Genetics* 50.7 (2018), p. 906.

[47] Jian Yang et al. “Advantages and pitfalls in the application of mixed-model association methods”. In: *Nature Genetics* 46.2 (2014), p. 100.

[48] Brian C Zhang et al. “Biobank-scale inference of ancestral recombination graphs enables genealogical analysis of complex traits”. In: *Nature Genetics* (2023), pp. 1–9.

[49] Shaun Purcell et al. “PLINK: a tool set for whole-genome association and population-based linkage analyses”. In: *The American Journal of Human Genetics* 81.3 (2007), pp. 559–575.

[50] Samuel Pattillo Smith et al. “Enrichment analyses identify shared associations for 25 quantitative traits in over 600,000 individuals from seven diverse ancestries”. In: *The American Journal of Human Genetics* 109.5 (2022), pp. 871–884.

[51] Alkes L Price et al. “Principal components analysis corrects for stratification in genome-wide association studies”. In: *Nature Genetics* 38.8 (2006), p. 904.

[52] Mashaal Sohail et al. “Polygenic adaptation on height is overestimated due to uncorrected stratification in genome-wide association studies”. In: *Elife* 8 (2019), e39702.

[53] Jeremy J Berg et al. “Reduced signal for polygenic adaptation of height in UK Biobank”. In: *Elife* 8 (2019), e39725.

[54] Nick Patterson, Alkes L Price, and David Reich. “Population structure and eigenanalysis”. In: *PLoS Genetics* 2.12 (2006), e190.

[55] Gil McVean. “A genealogical interpretation of principal components analysis”. In: *PLoS Genetics* 5.10 (2009).

[56] Jian Yang et al. “GCTA: a tool for genome-wide complex trait analysis”. In: *The American Journal of Human Genetics* 88.1 (2011), pp. 76–82.

[57] Bernie Devlin and Kathryn Roeder. “Genomic control for association studies”. In: *Biometrics* 55.4 (1999), pp. 997–1004.

[58] Benjamin F Voight and Jonathan K Pritchard. “Confounding from cryptic relatedness in case-control association studies”. In: *PLoS genetics* 1.3 (2005).

[59] Hyun Min Kang et al. “Efficient control of population structure in model organism association mapping”. In: *Genetics* 178.3 (2008), pp. 1709–1723.

[60] Hyun Min Kang et al. “Variance component model to account for sample structure in genome-wide association studies”. In: *Nature Genetics* 42.4 (2010), pp. 348–354.

[61] Gabriel E Hoffman. “Correcting for population structure and kinship using the linear mixed model: theory and extensions”. In: *PloS one* 8.10 (2013), e75707.

[62] Jennifer Listgarten, Christoph Lippert, and David Heckerman. “FaST-LMM-Select for addressing confounding from spatial structure and rare variants”. In: *Nature Genetics* 45.5 (2013), p. 470.

[63] David Heckerman et al. “Linear mixed model for heritability estimation that explicitly addresses environmental variation”. In: *Proceedings of the National Academy of Sciences* 113.27 (2016), pp. 7377–7382.

[64] Yiqi Yao and Alejandro Ochoa. “Limitations of principal components in quantitative genetic association models for human studies”. In: *bioRxiv* (2022).

[65] Peter Armitage. “Tests for linear trends in proportions and frequencies”. In: *Biometrics* 11.3 (1955), pp. 375–386.

[66] Wei-Min Chen and Gonçalo R Abecasis. “Family-based association tests for genomewide association scans”. In: *The American Journal of Human Genetics* 81.5 (2007), pp. 913–926.

[67] Ronald Christensen. *Advanced linear modeling*. Springer, 2019.

[68] Ronald A Fisher. “The correlation between relatives on the supposition of Mendelian inheritance.” In: *Earth and Environmental Science Transactions of the Royal Society of Edinburgh* 52.2 (1919), pp. 399–433.

[69] Joel Mefford et al. “Efficient estimation and applications of cross-validated genetic predictions to polygenic risk scores and linear mixed models”. In: *Journal of Computational Biology* 27.4 (2020), pp. 599–612.

[70] Jennifer Listgarten et al. “Improved linear mixed models for genome-wide association studies”. In: *Nature methods* 9.6 (2012), pp. 525–526.

[71] Montgomery Slatkin. “Linkage disequilibrium—understanding the evolutionary past and mapping the medical future”. In: *Nature Reviews Genetics* 9.6 (2008), pp. 477–485.

[72] Gulnara R Svishcheva et al. “Rapid variance components–based method for whole-genome association analysis”. In: *Nature Genetics* 44.10 (2012), pp. 1166–1170.

[73] Ismo Strandén and Martin Lidauer. “Solving large mixed linear models using preconditioned conjugate gradient iteration”. In: *Journal of Dairy Science* 82.12 (1999), pp. 2779–2787.

[74] Itsik Pe’er et al. “Estimation of the multiple testing burden for genomewide association studies of nearly all common variants”. In: *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society* 32.4 (2008), pp. 381–385.

[75] Jian Yang et al. “Genomic inflation factors under polygenic inheritance”. In: *European Journal of Human Genetics* 19.7 (2011), pp. 807–812.

[76] Brendan K Bulik-Sullivan et al. “LD Score regression distinguishes confounding from polygenicity in genome-wide association studies”. In: *Nature Genetics* 47.3 (2015), pp. 291–295.

[77] Shuang Song et al. “Leveraging LD eigenvalue regression to improve the estimation of SNP heritability and confounding inflation”. In: *The American Journal of Human Genetics* 109.5 (2022), pp. 802–811.

[78] Bingshan Li and Suzanne M Leal. “Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data”. In: *The American Journal of Human Genetics* 83.3 (2008), pp. 311–321.

[79] Alexander Gusev et al. “Whole population, genome-wide mapping of hidden relatedness”. In: *Genome research* 19.2 (2009), pp. 318–326.

[80] Pier Francesco Palamara et al. “Length distributions of identity by descent reveal fine-scale demographic history”. In: *The American Journal of Human Genetics* 91.5 (2012), pp. 809–822.

[81] Alexander Gusev et al. “DASH: a method for identical-by-descent haplotype mapping uncovers association with recent variation”. In: *The American Journal of Human Genetics* 88.6 (2011), pp. 706–717.

[82] Sharon R Browning and Elizabeth A Thompson. “Detecting rare variant associations by identity-by-descent mapping in case-control studies”. In: *Genetics* 190.4 (2012).

[83] Matthew D Rasmussen et al. “Genome-wide inference of ancestral recombination graphs”. In: *PLoS Genetics* 10.5 (2014), e1004342.

[84] Leo Speidel et al. “A method for genome-wide genealogy estimation for thousands of samples”. In: *Nature Genetics* 51.9 (2019), pp. 1321–1329.

[85] Jerome Kelleher, Alison M Etheridge, and Gilean McVean. “Efficient coalescent simulation and genealogical analysis for large sample sizes”. In: *PLoS computational biology* 12.5 (2016).

[86] Pouria Salehi Nowbandegani et al. “Extremely sparse models of linkage disequilibrium in ancestrally diverse association studies”. In: *bioRxiv* (2022), pp. 2022–09.

[87] Vivian Link et al. “Tree-based QTL mapping with expected local genetic relatedness matrices”. In: *bioRxiv* (2023), pp. 2023–04.

[88] Jerome Kelleher et al. “Inferring whole-genome histories in large population datasets”. In: *Nature Genetics* 51.9 (2019), pp. 1330–1338.

[89] Cristopher V Van Hout et al. “Exome sequencing and characterization of 49,960 individuals in the UK Biobank”. In: *Nature* 586.7831 (2020), pp. 749–756.

[90] Jonathan K Pritchard and Molly Przeworski. “Linkage disequilibrium in humans: models and data”. In: *The American Journal of Human Genetics* 69.1 (2001), pp. 1–14.

[91] Gleb Kichaev et al. “Leveraging polygenic functional enrichment to improve GWAS power”. In: *The American Journal of Human Genetics* 104.1 (2019), pp. 65–75.

[92] Christian Gieger et al. “New gene functions in megakaryopoiesis and platelet formation”. In: *Nature* 480.7376 (2011), pp. 201–208.

[93] Masahiro Kanai et al. “Genetic analysis of quantitative traits in the Japanese population links cell types to complex human diseases”. In: *Nature Genetics* 50.3 (2018), pp. 390–400.

[94] William McLaren et al. “The ensembl variant effect predictor”. In: *Genome biology* 17.1 (2016), pp. 1–14.

[95] Liesbeth Van Wesenbeeck et al. “Six novel missense mutations in the LDL receptor-related protein 5 (LRP5) gene in different conditions with an increased bone density”. In: *The American Journal of Human Genetics* 72.3 (2003), pp. 763–771.

[96] Arslan A Zaidi and Iain Mathieson. “Demographic history mediates the effect of stratification on polygenic scores”. In: *Elife* 9 (2020), e61548.

[97] Eugene J Gardner et al. “Reduced reproductive success is associated with selective constraint on human genes”. In: *Nature* 603.7903 (2022), pp. 858–863.

[98] Gillian M Belbin et al. “Toward a fine-scale population health monitoring system”. In: *Cell* 184.8 (2021), pp. 2068–2083.

[99] Elizabeth T Cirulli et al. “Genome-wide rare variant analysis for thousands of phenotypes in over 70,000 exomes from two cohorts”. In: *Nature Communications* 11.1 (2020), pp. 1–10.

[100] Alison R Barton et al. “Whole-exome imputation within UK Biobank powers rare coding variant association and fine-mapping analyses”. In: *Nature Genetics* 53.8 (2021), pp. 1260–1269.

[101] Joshua D Backman et al. “Exome sequencing and analysis of 454,787 UK Biobank participants”. In: *Nature* 599.7886 (2021), pp. 628–634.

[102] Bjarni V Halldorsson et al. “The sequences of 150,119 genomes in the UK biobank”. In: *Nature* 607.7920 (2022), pp. 732–740.

[103] Jianming Yu et al. “A unified mixed-model method for association mapping that accounts for multiple levels of relatedness”. In: *Nature Genetics* 38.2 (2006), pp. 203–208.

[104] Zhiwu Zhang et al. “Mixed linear model approach adapted for genome-wide association studies”. In: *Nature Genetics* 42.4 (2010), pp. 355–360.

[105] Ali Pazokitoroudi et al. “Efficient variance components analysis across millions of genomes”. In: *Nature Communications* 11.1 (2020).

[106] JK Haseman and RC Elston. “The investigation of linkage between a quantitative trait and a marker locus”. In: *Behavior Genetics* 2.1 (1972), pp. 3–19.

[107] Yue Wu and Sriram Sankararaman. “A scalable estimator of SNP heritability for biobank-scale data”. In: *Bioinformatics* 34.13 (2018), pp. 187–194.

[108] Po-Ru Loh et al. “Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis”. In: *Nature Genetics* 47.12 (2015), pp. 1385–1392.

[109] Steven Gazal et al. “Linkage disequilibrium–dependent architecture of human complex traits shows action of negative selection”. In: *Nature Genetics* 49.10 (2017), pp. 1421–1427.

[110] Hilary K Finucane et al. “Partitioning heritability by functional annotation using genome-wide association summary statistics”. In: *Nature genetics* 47.11 (2015), pp. 1228–1235.

[111] Pierrick Wainschtein et al. “Assessing the contribution of rare variants to complex trait heritability from whole-genome sequence data”. In: *Nature Genetics* 54.3 (2022), pp. 263–273.

[112] Doug Speed and David J Balding. “MultiBLUP: improved SNP-based prediction for complex traits”. In: *Genome Research* 24.9 (2014), pp. 1550–1557.

[113] Etienne J Orliac et al. “Improving GWAS discovery and genomic prediction accuracy in biobank data”. In: *Proceedings of the National Academy of Sciences* 119.31 (2022), e2121279119.

[114] Christopher C Chang et al. “Second-generation PLINK: rising to the challenge of larger and richer datasets”. In: *Gigascience* 4.1 (2015), s13742–015.

[115] Doug Speed et al. “Improved heritability estimation from genome-wide SNPs”. In: *The American Journal of Human Genetics* 91.6 (2012), pp. 1011–1021.

[116] S Hong Lee et al. “Estimation of SNP heritability from dense genotype data”. In: *The American Journal of Human Genetics* 93.6 (2013), pp. 1151–1155.

[117] Luke M Evans et al. “Comparison of methods that use whole genome data to estimate the heritability and genetic architecture of complex traits”. In: *Nature Genetics* 50.5 (2018), pp. 737–745.

[118] Saori Sakaue et al. “A cross-population atlas of genetic associations for 220 human phenotypes”. In: *Nature Genetics* 53.10 (2021), pp. 1415–1424.

[119] Noah Zaitlen et al. “Using extended genealogy to estimate components of heritability for 23 quantitative and dichotomous traits”. In: *PLoS Genetics* 9.5 (2013), e1003520.

[120] Jian Yang et al. “Common SNPs explain a large proportion of the heritability for human height”. In: *Nature Genetics* 42.7 (2010), pp. 565–569.

[121] Michael F Hutchinson. “A stochastic estimator of the trace of the influence matrix for Laplacian smoothing splines”. In: *Communications in Statistics-Simulation and Computation* 18.3 (1989), pp. 1059–1076.

[122] Frank MTA Busing, Erik Meijer, and Rien Van Der Leeden. “Delete-m jackknife for unequal m”. In: *Statistics and Computing* 9.1 (1999), pp. 3–8.

[123] Jonathan Terhorst, John A Kamm, and Yun S Song. "Robust and scalable inference of population history from hundreds of unphased whole genomes". In: *Nature genetics* 49.2 (2017), pp. 303–309.

[124] Sean J Jurgens et al. "Adjusting for common variant polygenic scores improves yield in rare variant association analyses". In: *Nature Genetics* 55.4 (2023), pp. 544–548.

[125] Wei Zhou et al. "SAIGE-GENE+ improves the efficiency and accuracy of set-based rare variant association tests". In: *Nature genetics* 54.10 (2022), pp. 1466–1469.

[126] Zachary R McCaw et al. "Operating characteristics of the rank-based inverse normal transformation for quantitative trait analysis in genome-wide association studies". In: *Biometrics* 76.4 (2020), pp. 1262–1272.

[127] Jae Hoon Sul and Eleazar Eskin. "Mixed models can correct for population structure for genomic regions under selection". In: *Nature Reviews Genetics* 14.4 (2013), pp. 300–300.