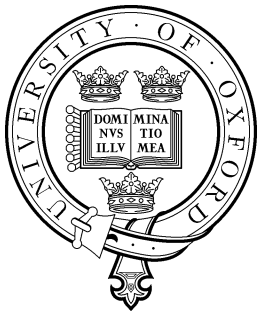
Biobank-Scale Ancestral Recombination Graphs

Inference and Applications to the Analysis of Complex Traits



Brian C. Zhang
St. John's College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy

Hilary 2022

Abstract

Across living species, DNA is transmitted from generation to generation via the processes of inheritance, mutation, and recombination. The history of these processes can be recorded using genome-wide gene genealogies. Accurate inference of gene genealogies from genetic data has the potential to facilitate a wide range of analyses, but is computationally challenging. In this thesis, we introduce a scalable method, called ARG-Needle, that uses genotype hashing and a coalescent hidden Markov model to infer genome-wide genealogies from sequencing or genotyping array data in modern biobanks. We develop strategies that utilise the inferred genome-wide genealogies within linear mixed models to perform association and other analyses of biomedical traits.

We validate the accuracy and scalability of ARG-Needle through extensive coalescent simulations, and use ARG-Needle to build genome-wide genealogies from genotypes of 337,464 UK Biobank individuals. We perform genealogy-based association analysis of 7 complex traits, detecting more rare and ultra-rare signals ($N = 133$, frequency range $0.0004\% - 0.1\%$) than genotype imputation from ∼65,000 sequenced haplotypes ($N = 65$). We validate these signals using exome sequencing data from 138,039 individuals. ARG-Needle associations strongly tag (average $r = 0.72$) underlying sequencing variants that are enriched for missense ($2.3\times$) and loss-of-function ($4.5\times$) variation. Compared to imputation, inferred genealogies also capture additional signals for higher frequency variants. These results demonstrate that biobank-scale inference of gene genealogies may be leveraged in the analysis of complex traits, complementing approaches that require the availability of large, population-specific sequencing panels.

Contents

List of Figures

Chapter 1

Introduction

The completion of the Human Genome Project in 2003, seen in its own right as a scientific milestone, was only the starting point for the exciting questions that could be explored and answered using human DNA. While the goal of the project was to decode the base pair sequence of a reference human genome, the full picture of human genetics lies not in a single representative sequence, but rather in the statistical diversity that exists compared to this norm. Since the first human genome was sequenced, the cost of sequencing has dropped exponentially [Hayden, 2014], enabling large genetic datasets of close to a million individuals [Gaziano et al., 2016, Bycroft et al., 2018] and creating a demand for efficient computer algorithms to comb through these datasets in search of patterns. The development, analysis, and use of these algorithms fall under the scientific field of **statistical genetics**.[1]

Broadly speaking, statistical genetics of human datasets can be divided into two halves, depending on the types of questions asked. The first set of questions seeks to understand the processes that have shaped modern genomes. Past events such as evolutionary adaptation, migration of ancestral populations, and contact between civilisations have all left signatures in the genomes of humans, and may be modelled and studied using the tools of **population genetics**. The applications enabled by such work range from taking a genetic ancestry test to complementing and perhaps

[1]We note that statistical genetics encompasses all living organisms, and that many technologies, algorithms, and ideas have proved useful across different species. However, in this thesis, we will focus on human genetics.

revising theories of the past developed by archaeologists and historians. The second set of questions within human statistical genetics combines DNA datasets with information about observable **phenotypes**—biomedical traits like height and weight, diseases like schizophrenia and melanoma, or even sociological traits like income and education. This direction of **medical genetics** or **genetic epidemiology** explores how personal genetics does or does not affect one's physiology, health, and lived experience. Additionally, there is a focus on how such knowledge can be used to improve human health, ranging from a recommendation for someone with at-risk genetics to adopt a certain diet or exercise program, to developing new treatments such as pharmaceuticals or gene therapy that draw on genetic discoveries. We could summarise human population genetics as asking "Where does genetic diversity come from?" and medical genetics as asking "What are the effects of genetic diversity on health and behaviour?"

This thesis presents a novel approach to combine the two streams of population genetics and medical genetics. In our proposed approach, we first infer genome-wide genealogies from DNA datasets, an analysis typically belonging to population genetics. Then, we combine these genome-wide genealogies with phenotypes to perform analyses within medical genetics. This roundabout route improves the **statistical power** of the resulting analyses—the amount of relevant signal that a statistical test is able to find, or in our case, the number of genes that can be determined to be linked with a phenotype. To demonstrate these improvements in statistical power, we performed extensive analyses using simulated data, as well as a real data study of 7 phenotypes within the UK Biobank, which comprises hundreds of thousands of individuals [Bycroft et al., 2018]. Using our methods, researchers may be able to make new genetic discoveries without needing to collect additional data. We plan to release the implementation of our algorithms as an open source software package to enable others to build on our approach.

This thesis proceeds as follows. In Chapter 2, we provide relevant background for understanding the contributions made in the remainder of the thesis. This includes a definition of the ancestral recombination graph, which is the more technical term for

the genome-wide genealogies central to our approach. We review previous methods in ancestral recombination graph simulation and inference. We also review three other themes in which population genetics has been utilised in the service of medical genetics: genotype imputation, the linear mixed model, and identity by descent detection and mapping. The methods we develop for genealogy-based analysis of phenotypes are intimately related to yet extend each of these themes.

In Chapter 3, "A Genealogical Perspective of Mixed-Model Analysis for Complex Traits," we begin the main portion of this thesis by demonstrating the potential of genealogy-based approaches to phenotype analysis. We propose two complementary innovations to increase the ability to detect phenotypic associations. First, branches of the genealogy can be tested for association with phenotypes using a linear mixed model. Second, the genealogy can be used to model relatedness, which increases association power and can also be leveraged to estimate heritability and predict polygenic risk.[2] An Appendix at the end of this thesis provides additional mathematical derivations supporting this method. Using simulations that assume perfectly accurate genealogies, we demonstrate that both of these approaches increase association power compared to association of genotyping array data, and that the two approaches can be combined to yield further improvements.

Chapter 3 leaves open the question of how these methods would perform if true genealogies were not available. Beyond a few generations into the past, we do not expect to have perfectly accurate genealogical records, and must infer plausible genealogies from data. Because the methods of Chapter 3 demonstrate a marked improvement compared to association of genotyping array data, and because large biobanks of array data are being collected around the world, it is desirable to be able to infer genealogies from large collections of array data. In Chapter 4, "A Scalable Method for Inferring Genome-Wide Genealogies from Array or Sequencing Data," we introduce such a method which we call ARG-Needle. While previous methods for genealogical inference exist, ours is the first scalable method that specifically models array data. We provide a description of the ARG-Needle algorithm, as well as a slower

[2]See Chapter 2 for an overview of heritability estimation and polygenic risk prediction.

method we call ASMC-clust which we recommend for smaller datasets, and analyse the theoretical properties and computational requirements for both algorithms.

In Chapter 5, "Performance of Genealogical Inference in Simulations," we assess the ARG-Needle and ASMC-clust algorithms in simulations. We first compare ARG-Needle and ASMC-clust to the algorithms `tsinfer` [Kelleher et al., 2019] and Relate [Speidel et al., 2019] in terms of computational requirements and inference accuracy. We find that ARG-Needle is a computationally efficient method that performs better than or as well as other methods across most accuracy metrics. Next, we revisit the strong results of Chapter 3, substituting the perfectly accurate genealogies for genealogies inferred with ARG-Needle. We find that compared to using array data directly, mixed-model association power and heritability estimates can be improved by applying our genealogy-based methods on genealogies inferred from array data. These improvements hold even if a moderate number of reference sequences are available for genotype imputation. Our last set of simulations uses ARG-Needle to build genealogies on a combination of sequencing and genotyping array data, showing that we can reliably impute missing genotypes in the array samples using the genealogies.

In Chapter 6, "Inference and Analysis of Ancestral Recombination Graphs in the UK Biobank," we evaluate the methods of Chapters 3 and 4 in a real data setting. Using ARG-Needle, we inferred genome-wide genealogies consisting of all 22 autosomal chromosomes from genotyping array data of 337,464 unrelated White British individuals. We then matched the samples in the genealogies with individual-level information for 7 phenotypes, including height, HDL and LDL cholesterol, and 4 other molecular traits. We performed a genome-wide association scan of each of these phenotypes using the genealogies, and compared our results to association scans using genotyping array data and data imputed from around 65 thousand references [Huang et al., 2015, McCarthy et al., 2016]. Genealogy-based association detected complementary signals to imputation for common variant association of height, and detected more rare and ultra-rare variant associated regions than array or imputed data. This included 5 associations with aspartate aminotransferase in the *GOT1* gene and 3 associations with alkaline phosphatase in the *ALPL* gene that were missed

by imputation from a within-cohort exome sequencing panel [Barton et al., 2021], and which we validated using exome sequencing data from 138,039 individuals. To our knowledge, our work represents the first example of performing genome-wide association using inferred genealogies in a modern biobank.

Finally, in Chapter 7, we summarise the key contributions of this thesis.[3] We compare our paradigm to three themes reviewed in Chapter 2: genotype imputation, the linear mixed model, and identity by descent detection and mapping. We argue that each of these methodologies can be more holistically understood when viewed from the perspective of genome-wide genealogies. In this sense, in addition to the innovations developed herein, this thesis also serves to strengthen the connections between population genetics and medical genetics, which could clarify existing approaches and inspire future developments. We conclude by discussing limitations of the present work which we see as exciting avenues for future research.

[3]Many of the results of this thesis have been submitted for publication and made available as a preprint; see [Zhang et al., 2021].

Chapter 2

Background

In this chapter, we introduce relevant theoretical background and literature that will be referenced in the remainder of the thesis. In Section 2.1, we offer an introduction to human genetic variation and the setting of genome-wide association studies (GWAS). In Section 2.2, we define the ancestral recombination graph and review existing methods for inference. In Section 2.3, we review three extensions to the basic GWAS paradigm: genotype imputation, identity-by-descent mapping, and the linear mixed model. These approaches each draw on population genetics in some way to improve GWAS power, and will therefore serve as points of comparison with our methods in the remainder of the thesis, culminating in a concluding discussion in Chapter 7.

2.1 Introduction to GWAS

The promise of using genetic discoveries to diagnose and treat disease has led to a boom in scientific research and wider interest in human genetics over the past two decades. The dominant statistical design used for analysing phenotypes is the **genome-wide association study (GWAS)** (*"jee-wahs"*). In a GWAS, participants are recruited and given medical tests and/or surveys to collect phenotype information. From blood samples, their DNA is **genotyped** using a **microarray chip** or similar device, which measures the DNA information at anywhere from thousands to over

millions of fixed **base pair** locations in the genome. These locations are given the technical term "**single nucleotide polymorphisms**" or **SNPs** (*"snips"*), which once measured can be encoded numerically with a value of 0, 1, or 2. While not an exhaustive readout of the DNA sequence (the entire human genome contains around three billion base pairs), the set of genotyped SNPs can end up capturing most of the significant variation between humans.

The most common statistical technique applied on GWAS data has been **single-SNP association testing**. This method goes one SNP at a time through the data, and looks for SNPs whose correlation with the phenotype is statistically significant. Significant SNPs may highlight regions with mutations that affect the trait, which can be further followed up in the laboratory or clinic. In this section, we briefly review some of the key ideas behind GWAS. We refer the reader to [Claussnitzer et al., 2020] for a complementary historical perspective.

2.1.1 Human DNA variation

In humans, hereditary information is transmitted through three mechanisms: nuclear DNA, mitochondrial DNA, and epigenetic marks. Of these, nuclear DNA is the most significant in terms of its effects on the individual, and will be our focus in what follows. Short for **deoxyribonucleic acid**, **DNA** is found in all living organisms. It is structured as a double helix made up of two complementary strands, with each strand built up from four molecular "letters": adenine (A), thymine (T), guanine (G), and cytosine (C). Across the strands, A pairs with T and G pairs with C; therefore knowing one strand is enough to completely determine both strands.

Within the human cell, nuclear DNA is typically divided into 46 **chromosomes**. Of these, 44 of the chromosomes are made up of 22 **diploid** pairs, and numbered accordingly from 1 to 22.[1] The last 2 chromosomes are typically X and X for females, or X and Y for males.[2] This difference of the sexes is but one of the many possible

[1]Pairs of chromosomes are called diploid, while single unpaired chromosomes are called **haploid**.

[2]Individuals may be transgender or intersex, or may have other combinations of X and Y chromosomes; see aneuploidy, defined below.

ways in which two individuals may differ in their DNA.

On the chromosomal level, there are occasional cases of humans with a different number of chromosomes—the general condition is known as **aneuploidy**, with Down's syndrome or trisomy 21 being a particular example. Such mismatch can also happen with a part of a chromosome being duplicated or loss, known as **copy number variation (CNV)**. At a finer level, **insertions** and **deletions** can slightly alter the number of base pairs in a chromosome. Lastly, **point mutations**, **inversions**, and **substitutions** typically change a sequence while preserving the number of base pairs.

Current **whole-genome sequencing (WGS)** techniques, which attempt to read out all three billion base pairs in a genome, are able to capture more or less all the variation carried by an individual, including **de novo mutations** unique to that individual.[3] However, whole-genome sequencing is currently one to two orders of magnitude more expensive than genotyping a sample using a microarray chip [Schwarze et al., 2020, Wasik et al., 2021]. In this tradeoff of quality versus cost, many GWAS prefer to increase sample size while using genotyping arrays, rather than spend the same budget on whole-genome sequencing of fewer samples. An intermediate approach is to perform **whole-exome sequencing (WES)**, in which full readouts are performed of the approximately 2% of the genome that codes for proteins.

We now explain how genetic variation, measured by a technology such as a genotyping microarray, is represented as data. Current microarray chips typically focus on the most common type of variation: point mutations. These are the SNPs mentioned earlier. If we focus on only the **autosomal chromosomes** (not X or Y), then each SNP occurs twice, once for each chromosome in the diploid pair. If we furthermore assume that the SNP is **bi-allelic**, meaning it shows up in one of two ways, then the result of genotyping can be encoded as a value of 0, 1, or 2. This comes from picking a strand to read out and then choosing a default value. For instance, if a SNP takes values A and C on the forward strand, and A is chosen as the default value, then a

[3]Long-read sequencing may be necessary in some cases to resolve the more complex types of variation such as copy number variation.

0 would correspond to the genotype AA, a 1 would correspond to AC or CA, and a 2 would correspond to CC.

Genotyping microarray chips are sometimes also used to measure **short indels**—insertions or deletions of up to around 50 base pairs. In this case, one might have two possible values of A and ATGACAGG for an allele. This can either be seen as an insertion of the base pairs TGACAGG after the base pair "A", or as a deletion of the base pairs TGACAGG from the sequence. In any case, this variation would likewise be encoded using the values 0, 1, or 2. For simplicity, we refer to only SNPs in what follows, as the concepts and equations are the same regardless of the type of variation being measured.

2.1.2 Single-SNP association testing

We introduce some notation. Assume our GWAS contains N samples (individuals) and M SNPs. We index samples by n and SNPs by m. We can assume that every SNP m comes with some metadata information defining the two possible alleles. Let $x_{nm} \in \{0, 1, 2\}$ be the value for the mth SNP of sample n. Denote our phenotype of interest as y_n. One typically standardises the genotypes and phenotype so that the mean and variance across samples are 0 and 1. Let $\widetilde{x}_{nm}$ and $\widetilde{y}_n$ denote the standardised versions of x_{nm} and y_n.

Phenotypes can either be **case-control traits**, taking one of two values as in the presence or absence of a disease, or **quantitative traits**, taking a continuous range of values. Typically, quantitative traits are analysed using the linear model and its extensions, while case-control traits are analysed using the logistic regression model and its extensions. However, when cases and controls are relatively balanced, it is also possible to interpret the phenotype as a quantitative trait and apply a linear model. In this thesis, we focus on analysis of quantitative traits.[4]

The simplest setup for GWAS is single-SNP association testing, which in the case of quantitative traits, simply performs a univariate linear regression per SNP. For

[4]See Sections 2.3.3.3 and 7.3.1 for remarks regarding case-control traits.

SNP m, the linear model can be written as[5]

$$\widetilde{y}_n = \widetilde{x}_{nm}\beta_m + \epsilon_{nm},$$

where β_m represents the effect size of SNP m on the phenotype, and ϵ_{nm} represents the error compared to a linear fit. If we assume normally distributed errors, $\epsilon_{nm} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_m^2)$, we can estimate β_m using maximum likelihood as

$$\hat{\beta}_m = \frac{1}{N}\sum_{n=1}^{N} \widetilde{x}_{nm}\widetilde{y}_n.$$

We see that the estimated effect is proportional to the correlation between the genotypes for SNP m and the phenotype. Whether the effect is significant depends on the (estimated) standard error of $\hat{\beta}_m$, which can be computed as[6]

$$\hat{se}(\hat{\beta}_m) = \sqrt{\frac{\hat{\sigma}_m^2}{N}}, \quad \text{where } \hat{\sigma}_m^2 = \frac{1}{N-2}\sum_{n=1}^{N}(\widetilde{y}_n - \widetilde{x}_{nm}\hat{\beta}_m)^2.$$

For small samples, the proper hypothesis test for whether β_m is significant is a two-sided t-test. However, for GWAS-size samples we can practically always assume asymptotic normality. If we form the z-score as

$$z_m = \frac{\hat{\beta}_m}{\hat{se}(\hat{\beta}_m)}$$

then we can either perform a two-sided test of z_m against a standard normal distribution, or a one-sided test of z_m^2 against a χ_1^2 distribution, to obtain a p-value for SNP m. The values $\hat{\beta}_m$, $\hat{se}(\hat{\beta}_m)$, and z_m^2 are often termed the GWAS **summary statistics**, and consist of the estimated effect sizes, estimated standard errors, and chi-squared statistics. A full genome-wide scan can be done in $O(NM)$ time, outputting summary statistics for each of the M SNPs.

One way to interpret GWAS summary statistics is using a **Manhattan plot** (see

[5]There is no need for an intercept term as we are dealing with centred x_{nm} and y_n.
[6]See for instance [Casella and Berger, 2002].

Fig. 2.1), a scatter plot of SNPs with position along each chromosome on the x-axis and $-\log_{10}(p)$ on the y-axis. Higher values of y correspond to more significant p-values, and indicate potentially causal biological signal. Furthermore, where there is a significant p-value, other significant p-values are often found in the vicinity, creating towers that resemble a city skyline. Typically, one sets a **genome-wide significance threshold** for which p-values less than the threshold are interpreted as significant; a commonly used value is 5×10^{-8} [Pe'er et al., 2008]. The regions crossing this significance threshold are then prioritised for further analyses and experiments in search of causal genes.

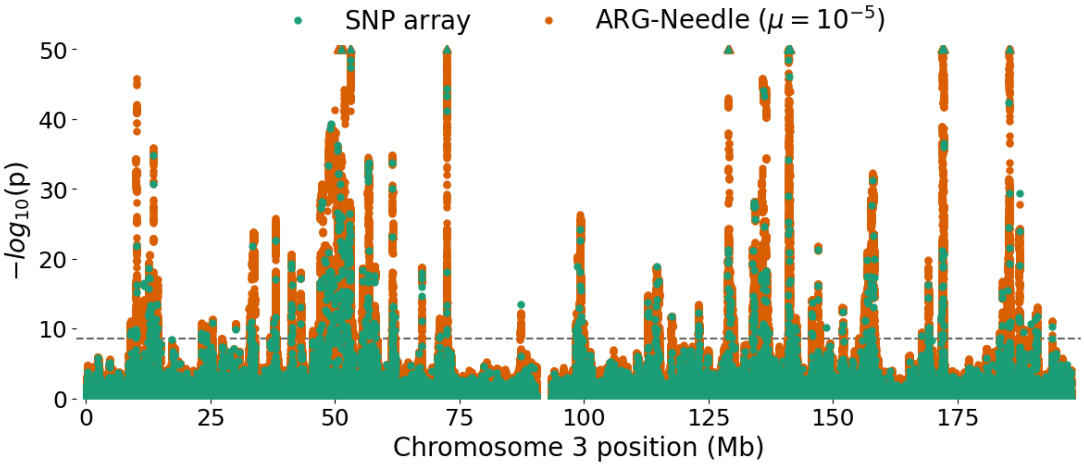


Figure 2.1: **Example Manhattan plot.** Each green point is a SNP, with chromosome 3 position on the x-axis and $-\log_{10}(p)$ on the y-axis. Triangles indicate associations with $p < 10^{-50}$. Orange points correspond to our ARG-Needle association method, outlined in Chapters 3-6. Fig. 6.4a contains a full description of this figure.

The summary statistics from a GWAS can also be utilised in other downstream analyses, including exciting applications in **heritability estimation** [Bulik-Sullivan et al., 2015, Finucane et al., 2015] (quantifying the extent to which a trait is due to nature versus nurture) and **polygenic prediction** [Vilhjálmsson et al., 2015, Ge et al., 2019] (predicting individuals' phenotypes from genotypes).

2.1.3 Linkage disequilibrium, tagging, and chip design

Genotyping arrays only measure a fraction of the mutations found in human genomes.[7] In an array-based GWAS, we test all sites measured by the array for association with a phenotype. But doesn't this pass over many possibly causal mutations, which are never genotyped by the array?

Indeed, GWAS do not directly measure many of the possibly causal mutations for a phenotype. However, the array is able to indirectly account for many more mutations using a principle called **linkage disequilibrium (LD)** [Risch and Merikangas, 1996, Slatkin, 2008].

Linkage disequilibrium refers to the fact that mutations occur in a correlated manner across the genome, with SNPs close together on the same chromosome often having a strong correlation. The details of LD are an entire research area in itself [Reich et al., 2001, Myers et al., 2005], but roughly, these correlations arise due to the fact that **genetic recombination** during meiosis is rare. Because of this, multiple variants may accumulate within genomic regions while remaining highly correlated due to a lack of recombination events.

Within a set of genetic samples, one way of computing the LD between two SNPs m and m' is as the r^2 between the genotype vectors $X_{\cdot,m}$ and $X_{\cdot,m'}$ [Slatkin, 2008]. In the case of GWAS, if SNP m is included in the array, and SNP m' is not included in the array but has a high LD with SNP m, then SNP m acts as a **tag SNP** for SNP m'. If SNP m' is biologically causal for a phenotype y, then an association test of y using either SNPs m or m' will have a chance of detecting a significant association, because the two are highly correlated.

It therefore suffices to include as many SNPs on the array that will sufficiently tag neighboring SNPs of interest. To reiterate, the array SNPs are not expected to be causal themselves, but rather to highlight nearby causal variation through LD tagging. When designing a genotyping microarray chip, the LD structure of the population is usually considered, and SNPs are selected for inclusion on the chip based on their

[7] While genotyping arrays measure anywhere from thousands to over millions of variants, the human genome contains three billion base pairs, each representing a possible site for a mutation.

ability to collectively tag the genetic variation of the population.

When a GWAS reveals a significantly associated SNP, this can either be due to the SNP's own influence on the trait, or due to tagging the causal effect of a neighboring SNP which may not be present on the array.[8] Further **fine-mapping** methods [Schaid et al., 2018] and biological experiments may then be needed to pinpoint the exact causal gene(s).

2.1.4 Medical genetics and UK Biobank

As a field, medical genetics originally focused on conditions that were highly heritable and could be explained by Mendelian laws of inheritance [Claussnitzer et al., 2020]. These include haemophilia and cystic fibrosis, and are called **monogenic or Mendelian traits**. Such conditions were easy to predict using pedigree information and could be studied using small sample sizes. By contrast, **polygenic or complex traits**, like height and schizophrenia, arise from a large number of genetic signals [Yang et al., 2010, Loh et al., 2015a]. For complex traits, each **genetic locus** is likely to have a smaller effect; thus large sample sizes are needed to fully understand their genetic makeup. This has motivated the design of larger and larger GWAS efforts.

Some of the largest GWAS to date have been made possible by the **UK Biobank**, a **prospective cohort study** of around 500,000 individuals throughout the United Kingdom, aged 40 to 69 years at enrollment [Bycroft et al., 2018]. UK Biobank includes genotyping of all participants; rich phenotype collection from electronic medical records, surveys, and tests; and survey questions to elucidate environmental / non-genetic factors. Participants are also asked to perform follow-up tests every few years to track onset of disease. The data are made available to researchers worldwide in an anonymised format, and participants may ask to have their data removed from future analyses at any time.

After an interim release in May 2015, the first full UK Biobank release was distributed in July 2017 and March 2018 and consisted of 488,377 samples, 805,426

[8] Confounding effects and spurious associations are also possible.

genotyped markers, and 96 million markers after genotype imputation[9] [Bycroft et al., 2018]. Exome sequencing of the UK Biobank was completed with a release of 454,787 exomes in October 2021 [Backman et al., 2021], following earlier releases of 49,960 (March 2019) [Van Hout et al., 2020], 200,643 (October 2020) [Szustakowski et al., 2021], and 300,000 exome sequences (September 2021). Efforts are currently under way to perform whole-genome sequencing of the entire UK Biobank [Halldorsson et al., 2021], with 200,000 sequences released in November 2021 and the remaining sequences expected in early 2023 [Kaiser, 2021].

2.1.5 Ethics of GWAS in humans

We conclude this introductory overview of GWAS with a discussion about ethics.

The societal benefits of genetics research are generally understood. Finding "the gene" responsible for a Mendelian disease can lead scientists to develop a cure, saving or improving the well-being of many lives. For complex traits, although genetic risk is distributed across many causal variants, polygenic risk scores may be used to identify at-risk patients in the clinic and suggest lifestyle changes [Khera et al., 2018]. Individually identified risk variants may also illuminate links between genetics and disease, contributing to progress in pharmaceutical interventions [Nelson et al., 2015, Visscher et al., 2021].

However, the field of human genetics also presupposes differences among individuals, and carries inherent risks depending on how findings are translated to society. A century ago, many of the world's leading statisticians and biologists—including such names as Francis Galton and Karl Pearson—were deeply invested in creating the modern eugenics movement [MacKenzie, 1976]. These scientists lent their intellectual and institutional credibility to a set of ideas that came to be associated with forced sterilisation, anti-Semitism and the Holocaust, ableism, and racism. This historical precedent should alert the genetics community to consider any potential negative consequences of research being performed today.

Particularly as genetic technologies become more powerful and widely available,

[9] See Section 2.3.1.

it is important for the genetics community to engage in ethical introspection, dialogue, and partnership with other disciplines. We outline six particular challenges and opportunities arising from the field of GWAS.

1. *Biological determinism.* Although GWAS have the potential to both confirm and disprove the role of genetics in human differences, the public may gravitate towards interpretations of GWAS that exaggerate the role of genetics and downplay the role of culture, environment, and chance [Schork et al., 2022].

2. *Selective reproduction.* GWAS have enabled more accurate predictions of individuals' phenotypes from their genes, which government or individual actors may use to define desirable or undesirable life in reproductive policies and decisions [Turley et al., 2021].

3. *Personalised medicine.* GWAS' applications in the clinic may include not just broadly available pharmaceuticals, but also actionable insights and treatments tailored to the individual patient [Schork, 2015, Joyner and Paneth, 2015].

4. *Racial disparities.* There is concern about racial disparities surrounding GWAS, from the demographics of the researchers in the field, to the datasets that have been collected [Popejoy and Fullerton, 2016], to expectations that clinical applications may benefit some populations more than others [Martin et al., 2019].

5. *Consent and privacy.* As genetic data becomes more valuable for the discoveries and commercial possibilities that may result, care must be taken to ensure those who provide their genetic data understand how the data is likely to be used [Claw et al., 2018, Fox, 2020].

6. *Economic ecosystem.* The presence of corporate interests in the genetics research and development space, with early access to some datasets such as the UK Biobank whole exome and whole genome sequencing data [Szustakowski et al., 2021, Kaiser, 2021], raises questions about how to align public and private interests.

We hope that this list may inspire the reader to play a role in encouraging bioethics discussions, whether among personal friends or on a larger scale.

2.2 Ancestral Recombination Graphs and Their Inference

In Section 2.1, we quickly surveyed the key concepts behind GWAS. In that introduction, we treated genotypes as a two-dimensional matrix of values 0, 1, or 2 (see Section 2.1.2). One axis of the genotype matrix represents the N samples (individuals), while the other axis represents the M SNPs along the genome. This genotype matrix can be thought of as a snapshot of the N genetic samples at the present time.

A richer picture of genetic variation, which turns out to be useful in many genetic analyses, can be obtained by adding a third dimension: that of time. We might recall that each individual's DNA is formed from combining two halves of DNA from each parent, through meiosis followed by the fusing of a sperm and egg. That parental DNA is in turn inherited from a total of four grandparents, eight great-grandparents, and so on. If perfect information about this inheritance history were available, then each mutation carried by an individual could be pinpointed to an ancestor in whom the mutation first occurred. This is a first sketch of what it is like to incorporate the historical dimension when working with genetic data.

As the two-dimensional object for genetic analysis is the genotype matrix, the canonical three-dimensional object is the **ancestral recombination graph (ARG)** [Hudson et al., 1990, Griffiths and Marjoram, 1996, Griffiths and Marjoram, 1997]. ARGs capture three types of historical events that give rise to modern genetic samples. First, ancestral lineages of samples can coalesce into common ancestors in the past, giving rise to **coalescent trees** at each position of the genome. Second, genetic recombination events can modify the trees as one moves across the genome.[10] Third, mutations occur on branches of the trees and are inherited by future samples. The combination of these three processes—coalescence, recombination, and mutation—

[10]See Section 2.1.3 for an earlier reference to recombination.

is termed the **coalescent with recombination** [Kingman, 1982, Hudson, 1983], a probabilistic model over the space of ARGs.

For the remainder of this section, we deal with three questions in order. First, what exactly is an ARG? Second, how does one simulate ARGs? Third, given genetic data, what algorithms exist to infer plausible ARGs? A fourth important question, how ARGs are relevant for GWAS, will be addressed briefly in the context of prior works, and will be taken up in more detail in Section 2.3 and Chapter 3.

2.2.1 The ancestral recombination graph

In this subsection, we build up to a definition of the ancestral recombination graph. We first consider an example in detail to provide intuition for the definition.

2.2.1.1 Motivating example

As mentioned above, the ARG captures three genetic processes that shape genomes: coalescence, recombination, and mutation. Let us consider a set of 12 modern haploid samples with an illustration of these three processes at play (see Fig. 2.2). The region of interest is a short segment of one chromosome, say chromosome 3. By haploid samples, we refer to the simple case where we only take one copy of chromosome 3 from each of 12 individuals. Within this region, we have three coalescent trees, shown in Fig. 2.2a-c, which can be summarised by a single ARG, shown in Fig. 2.2d. We will start with the first coalescent tree shown in Fig. 2.2a and explain the intuition behind the coalescent process.

The coalescent process is simply a probabilistic model of inheritance, except that inheritance is usually modelled forward in time, while coalescence is modelled backwards. Within this region, each of the 12 haploid samples was inherited through a parent, and further back in time through a grandparent, and so on. Following this pattern of inheritance backwards in time gives the ancestral lineage of a sample. Two ancestral lineages may coalesce if they meet at a shared ancestor in the past, who passed down this genomic segment to both modern samples; such a coalescence

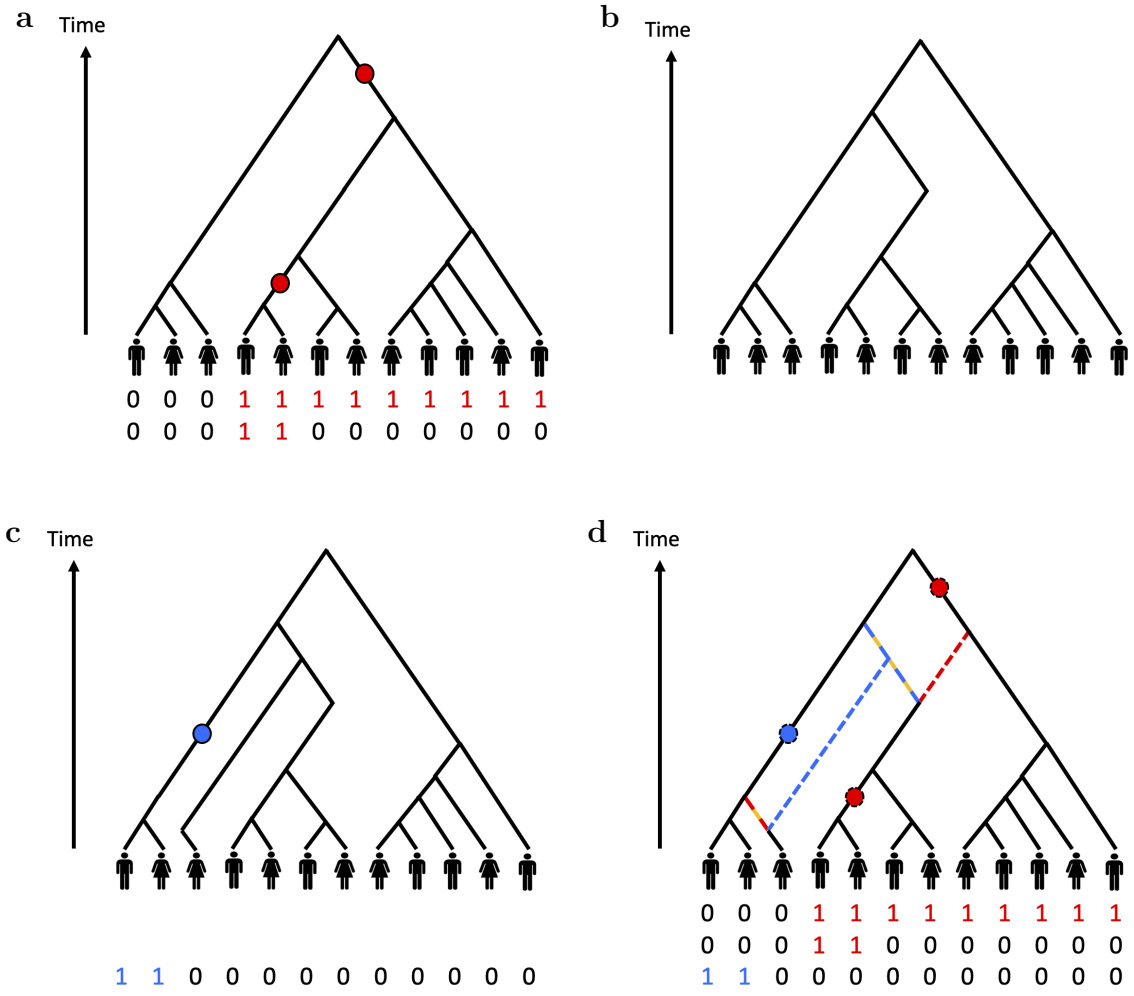


Figure 2.2: **Illustration of an ARG for 12 haploid samples. a-c**. Panels showing three local coalescent trees as one moves along the genome. **a**. A first coalescent tree with two mutations (red). **b**. Further along the genome, a recombination event produces a new coalescent tree, in this case with zero mutations. **c**. After another recombination event, we have a third coalescent tree with one mutation (blue). **d**. The ARG is a graph that concisely represents all three coalescent trees. Here solid black lines refer to edges of the ARG that persist across all three trees, while the coloured dashed lines refer to edges that only persist partially. By keeping only the events coloured red, yellow, or blue, the ARG reduces to one of the three trees **a**, **b**, or **c** respectively. The ARG implicitly represents all genotypes without needing to store a genotype matrix.

event reduces the total number of ancestral lineages by 1. Given enough time, it is a fact that with probability one, all ancestral lineages will have coalesced into a single ancient ancestor [Kingman, 1982, Griffiths and Marjoram, 1997]. We can represent the data of the coalescent process as a coalescent tree relating the samples, shown in Fig. 2.2a. The root of this tree is called the **most recent common ancestor (MRCA)** for the samples. Each coalescence event (where two ancestral lineages join up) is pinned to a time in the past, shown by the time axis on the left—this can be measured in either years or generations. The time of the root node is called the **time to most recent common ancestor (TMRCA)** of the 12 samples.

On top of this coalescence tree, we have shown two mutations as red circles. Each of these mutations is pinned to a certain time in the past, as we have shown. This time represents the ancestor (along the ancestral lineage) that first acquired this mutation. Samples that inherit from this ancestral lineage in the coalescent tree become carriers of the mutation, represented by the red 1's at bottom. Here the first row of 0's and 1's corresponds to the top mutation (generating 9 carriers), and the second row corresponds to the bottom mutation (generating 2 carriers). Along with a time, each mutation is also pinned to a certain position along the genome (not shown), which determines its base pair location.

Next we turn to Fig. 2.2b. Here we have moved further along the chromosome, and the coalescent tree from Fig. 2.2a is no longer valid. Recall that each local portion of our genome is inherited from one of our four grandparents, but that this grandparent is altered by genetic recombination as we move along each chromosome. In this case, one of the ancestral lineages—the one ancestral to the middle four samples—has experienced a recombination event, so that instead of coalescing with the rightmost ancestral lineage, it coalesces with the leftmost ancestral lineage instead. This creates a different local tree structure. On top of this tree, there are no mutations.

Moving further along the chromosome, we have yet another recombination event, where the ancestral lineage of the third-to-left sample recombines to coalesce with a different lineage (Fig. 2.2c). On this tree is a single mutation, shown as a blue circle, and generating 2 carriers shown with blue 1's. Note that if this mutation had

occurred on the same lineage but before the recombination event occurred, we would instead have 3 carriers rather than 2. This would look like a mutation on the same lineage but in the second tree, or Fig. 2.2b. Recombination events thus alter the possible mutational patterns that can be observed.

The ARG provides a composite picture of all three trees, as if they were superimposed on each other (Fig. 2.2d). As mentioned earlier, the ARG is a three-dimensional object, taking the two dimensions of samples and time present in a coalescent tree and adding a third dimension of genomic position. Each lineage in the ARG is given a genomic range. In the example, solid black lines refer to lineages that persist across all three trees, while the coloured dashed lines refer to lineages that only persist partially. We use the colours red, yellow, and blue to mark lineages that persist through the first, second, and third trees, while lineages carrying two colours (red and yellow, or yellow and blue) persist for the combined extent of the two corresponding trees. In addition to these lineages, the ARG in this example also contains the three mutations, each "living" on a lineage at a particular time and genomic position.

Given a particular genomic position, we can select all data in the ARG that crosses that genomic position to obtain a single coalescent tree. For instance, by selecting all the black lineages as well as only the red events, the ARG simplifies to Fig. 2.2a, and likewise for the yellow and blue events. Given a mutation on the ARG, we can deduce the carriers of that mutation by considering only the lineages of the ARG that overlap the position of the mutation. Because this perfectly generates the mutation's carriers while storing the mutation's position, the ARG provides a compact representation of all genotypes without needing to store a genotype matrix.

Before we give a formal definition of the ARG, we note two caveats for this particular example. First, the illustration we have shown in Fig. 2.2d suggests that the ARG is a planar graph, in that it can be drawn on a piece of paper with no intersections of lineages. This is usually not the case, as with increasing recombinations and sample size the ARG is very likely to be a non-planar graph. Second, the illustration pictured each haploid sample as an individual. As we mentioned earlier, this could come from choosing one copy of say chromosome 3 from each individual. In practice,

given N diploid individuals, one would usually consider the collection of $2N$ copies of chromosomes across all individuals, and obtain an ARG over these $2N$ haploid chromosomes. The diploid genotypes of one individual can be obtained by summing the genotypes of two constituent haploid chromosomes.

2.2.1.2 ARG definition

We now formalise the definition of the ARG based on the example shown in Fig. 2.2d. The idea is to create a directed graph where edges represent the ancestral lineages and nodes represent the coalescence events between lineages.

In our formulation, an ARG over a range of genomic positions $[s, t) \subset \mathbb{R}$ consists of the following data:

- A collection of *nodes*, where each node n consists of a height $h(n)$ and a subrange $R(n) \subset [s, t)$ representing the positions over which the node is "live." Of these nodes, N nodes represent haploid *sample nodes* with $h(n) = 0$ and $R(n) = [s, t)$, and the remaining nodes are *internal nodes* with $h(n) > 0$.

- A collection of directed *edges*, where each edge e points from a *child node* $n_c(e)$ to a *parent node* $n_p(e)$ and additionally consists of a positional interval $[x_1(e), x_2(e)) \subset [s, t)$ representing the positions over which the edge is "live."

- A collection of *mutations*, where each mutation m consists of an edge $e(m)$, a height $h(m)$, and a position $x(m)$.

with the following properties:

1. Child nodes are always of lesser height than parent nodes.[11]

2. If an edge is "live," then its parent and child nodes must be "live."[12]

3. Every internal node n is the parent of at least one live edge for every $x \in R(n)$.[13]

[11] i.e., $h(n_c(e)) < h(n_p(e))$ for every e.

[12] i.e., $[x_1(e), x_2(e)) \subset R(n_c(e))$ and $[x_1(e), x_2(e)) \subset R(n_p(e))$ for every e.

[13] If for some $x \in R(n)$, n is the parent for exactly one edge, we say that n is a *unary node*. Otherwise, every internal node n is the parent of at least two edges for every $x \in R(n)$, and we say that the ARG has no unary nodes.

4. For every position $x \in [s, t)$, all but one of the live nodes are the child of exactly one live edge. The exception is the *root node* at x, which is not the child of any live edges.

5. A mutation's height is bounded by the height of the parent and child nodes.[14]

6. A mutation's position is bounded by the start and end of the edge.[15]

The ARG (without mutations) is a directed acyclic graph, as edges always go from nodes of lesser height to nodes of greater height (property 1). By taking only the nodes, edges, and mutations of an ARG that intersect with a position $x \in [s, t)$, we obtain the coalescent tree at x, where there is exactly one path from each sample node to the root node (properties 3 and 4). Referring to the example in Fig. 2.2d, each of the points where two lineages coalesce is an ancestral node, each of the bottom points of the ARG is a sample node, and each of the line segments joining two nodes is an edge pointing upwards. The range that each edge is live for is given by its colours, and the range that each node is live for is given by the union of the ranges of its adjoining edges (property 2).

The ARG is meant to represent the genomic history of the N haploid samples within up to one chromosome, as recombination operates within chromosomes but not across chromosomes. For genetic data of N human individuals, we would expect $2N$ samples in an ARG, and 22 ARGs for each of the autosomal chromosomes. The genomic range $[s, t)$ for each ARG could represent the positions contained in that chromosome. The height dimension corresponds to time in the past, and could be in units of years or generations.

Each of the internal nodes corresponds to a portion of a haploid sample in the past from which modern samples were inherited. These are also called ancestral **haplotypes**. The subrange $R(n)$ of each of these haplotypes does not need to be a contiguous interval, and could instead consist of the union of multiple disjoint intervals. In this case, an ancestral node contains portions of ancestral material,

[14] i.e., $h(n_c(e(m))) \leq h(m) < h(n_p(e(m)))$ for every m.
[15] i.e., $x_1(e(m)) \leq x(m) < x_2(e(m))$ for every m.

along with trapped non-ancestral material in between. However, for the purposes of being able to reconstruct each of the coalescent trees, our definition could just as well require the live extents $R(n)$ to all be contiguous intervals, thus splitting nodes with non-contiguous extents into multiple nodes.

We briefly address other terminology that exists in the literature.[16] In our definition, the ARG only contains information about relevant ancestral material. Some authors refer to this as Hudson's ARG [Hudson, 1983, Hudson et al., 1990] or the "small ARG" [Hein et al., 2004]. The more expansive ARG carries additional information for non-ancestral material, with coalescence continuing backwards in time until a single ancestral haplotype is reached across all positions [Griffiths and Marjoram, 1997]. However, because the added events are non-ancestral, it means that modern genomes carry no information about any of these events. As a focus of this thesis is on data-driven ARG inference, it is natural to only consider Hudson's ARG.[17]

Other authors use terms such as **genome-wide genealogies** [Speidel et al., 2019] or the **tree sequence** [Kelleher et al., 2016] where we refer to the ARG. These terms draw upon the idea that the ARG stores the same information as a series of local coalescent trees spanning the genome. However, it is important to realise that the graph-based representation we have outlined is more memory-efficient than a representation based on a collection of marginal trees, which requires storing the same ancestral haplotype (ARG node) multiple times in neighboring trees. The authors of [Kelleher et al., 2016] use the term "succinct tree sequence" to highlight the use of a more memory-efficient representation, which closely resembles our ARG definition[18] as well as the definitions in other recent works dealing with ARGs [Minichiello and Durbin, 2006, Rasmussen et al., 2014, Palamara, 2016]. We use the term "genome-wide genealogies" as a substitute for "ARG" in our less technical, introductory sections, and otherwise use the technical term "ARG."

[16] The remainder of this section is adapted from Supplementary Note 1 of [Zhang et al., 2021].

[17] Some formulations of Hudson's ARG distinguish between coalescent nodes and recombination nodes, which allows one to additionally localise the recombination events in time.

[18] In the succinct tree sequence, nodes are not annotated with the subrange $R(n)$ over which they are "live."

2.2.2 The coalescent with recombination process

The coalescent with recombination process [Kingman, 1982, Hudson, 1983] ("the coalescent" for short) is a probabilistic model that can be sampled to obtain an ARG.[19] It is convenient to describe the coalescent by means of the simulation algorithm that performs this sampling. In its simplest form,[20] backwards-in-time coalescent simulation starts with N modern haploid sample nodes and proceeds by sampling coalescence and recombination events. For each sampled event, new nodes are created and connected by new edges. Once the graph (the ARG) has been built, it is straightforward to simulate mutations on top of this graph. Backwards-in-time simulation was introduced by Hudson's `ms` simulator [Hudson, 2002]. Other simulators in this family include `msms` [Ewing and Hermisson, 2010], `cosi2` [Shlyakhter et al., 2014], and `msprime` [Kelleher et al., 2016], with the latter being widely used for its efficient implementation. It is also possible to formulate the coalescent with recombination as a sequential process along the genome. This viewpoint was introduced in [Wiuf and Hein, 1999], with the **SMC (sequentially Markovian coalescent)** [McVean and Cardin, 2005] and SMC' [Marjoram and Wall, 2006] models introducing a Markovian approximation that makes simulation and inference more tractable.

The coalescent with recombination takes in a list of parameters which guide the probabilistic process, with a set of parameters for each of the coalescence, recombination, and mutation components. First, coalescence is guided by a **demographic model**, or "demography" for short. In its basic form, the demographic model consists of a function giving the ancestral population size as a function of time into the past. When there are fewer ancestors in a population, ancestral lineages are more likely to match on the same ancestor, creating a coalescence event. One demography could have a constant population size with one parameter, whereas another demography could be parameterised as a combination of piecewise linear functions. However, such single-population demographic models assume that mating occurs randomly and

[19] The beginning of this section is adapted from Supplementary Note 1 of [Zhang et al., 2021].

[20] ignoring e.g. mutations, which do not affect the shape of the ARG, as well as migration and gene conversion

uniformly throughout the population,[21] but this becomes less likely if ancestors are spread over broad geographic areas. More complex demographic models are possible in which multiple ancestral populations undergo migration or admixture events, along with changes in population size.

Second, the recombination aspect of the process is parameterised using a recombination rate map. Recombination rates are usually in units of recombination events per base pair per generation, with higher values increasing the likelihood that a recombination event is sampled. A recombination rate map specifies a varying rate as a function of genomic position.

Third, mutation is affected by a mutation model. Here two choices are usually available, each with a simpler and a more complex option. In the first place, either mutations can be sampled anywhere over a continuous stretch of genome, in what is called the **infinite-sites model**, or mutations are each given a particular base pair of the genome, in what is called the **finite-sites model**. The finite-sites model is more complex as multiple mutation events can occur at the same site. In the second place, either all point mutations can be viewed as equally likely, independent of the starting and end allele, or the different possible transitions between alleles can be given different rates. The latter choice can account for the fact that point mutations are more likely to preserve purine (A or G) and pyrimidine (C or T) status than to go between the two classes. If both of the simpler choices are made, we have an infinite-sites model with a single mutation rate parameter, usually given in units of mutation events per base pair per generation.

Here we have only given a high-level description of the parameters involved in the coalescent with recombination process. For further mathematical details about the process in terms of these parameters, we refer the reader to [Hein et al., 2004]. We also note that we have only considered a **neutral model** with no natural selection, and that additional modelling may incorporate natural selection [Durrett, 2008].

[21] The population is called **panmictic** in this case.

2.2.3 Review of ARG inference works

Let the three types of parameters in the neutral coalescent with recombination be denoted by θ_d (demography), θ_r (recombination rate map), and θ_m (mutation model), each of which may be lists of parameters. Let the symbol $\mathcal{A}$ refer to an ARG without mutations. We can refer to the coalescent with recombination (without mutations) as a likelihood:

$$p(\mathcal{A}|\theta_d, \theta_r)$$

With mutations, $\mathcal{A}$ determines a genotype matrix X.[22] This mutation model can be described as another likelihood:

$$p(X|\mathcal{A}, \theta_m)$$

Together, these form the joint likelihood that is sampled from by coalescent simulators:

$$p(\mathcal{A}, X|\theta_d, \theta_r, \theta_m) = p(\mathcal{A}|\theta_d, \theta_r) \cdot p(X|\mathcal{A}, \theta_m)$$

The problem of ARG inference is to infer $\mathcal{A}$ given X. If we assume parameters are available, this is the problem of posterior inference of the latent variable $\mathcal{A}$:

$$p(\mathcal{A}|X, \theta_d, \theta_r, \theta_m)$$

Often, the parameters $\theta_d, \theta_r, \theta_m$ are themselves subject to inference. They are usually estimated in other data using point estimation, though sometimes the demography θ_d is simultaneously estimated from X [Li and Durbin, 2011, Schiffels and Durbin, 2014, Terhorst et al., 2017].

Several algorithms to infer ARGs have been proposed.[23] These inference algorithms differ based on whether they infer only topologies or also branch lengths, whether they infer ARGs over multiple samples or only partial genealogical structures, and in terms of their scalability. The most scalable methods only loosely rely

[22] X may additionally contain metadata about the SNP positions, and may be subject to genotyping errors or SNP ascertainment as well.

[23] The remainder of this section is expanded and adapted from [Zhang et al., 2021].

on probabilistic modelling, instead focusing on heuristic techniques such as parsimony to achieve higher computational speed [Lyngsø et al., 2005, Minichiello and Durbin, 2006, Mirzaei and Wu, 2017, Kelleher et al., 2019, Schaefer et al., 2021]. A recent method in this category, `tsinfer` [Kelleher et al., 2019], has been applied to upwards of tens of thousands of human sequencing samples, with the potential for further scalability. `tsinfer` creates an ancestral haplotype node for each mutation in the data, and orders these based on allele frequency, with rare mutations being the most recent. It then quickly infers plausible copying paths among the haplotypes under certain heuristics. `tsinfer` often outputs polytomies in the ARG and does not seek to infer ARG branch lengths. SARGE [Schaefer et al., 2021] is another scalable method that utilises similar heuristics to `tsinfer` and was applied to hundreds of modern and ancient sequencing samples.

A second class of methods [Li and Durbin, 2011, Schiffels and Durbin, 2014, Sheehan et al., 2013, Terhorst et al., 2017, Palamara et al., 2018, Mahmoudi et al., 2022] performs probabilistic modelling but tends to focus on reconstructing the ARG of a small number of genomes (e.g. a single pair). The first of these methods, PSMC [Li and Durbin, 2011], performed accurate pairwise ARG inference under the SMC approximation to the coalescent with recombination. Follow-up works extended this approach to include joint modelling of samples via information such as the allele frequency of variants [Schiffels and Durbin, 2014, Sheehan et al., 2013, Terhorst et al., 2017, Palamara et al., 2018]. **ASMC (ascertained sequentially Markovian coalescent)** [Palamara et al., 2018] enables efficiently building the pairwise ARG for tens of thousands of samples and supports both SNP array and sequencing data via a custom **hidden Markov model (HMM)** [Simonsen and Churchill, 1997, Hobolth and Jensen, 2014]. While other HMMs do not model **SNP ascertainment bias**, the fact that SNPs used on a genotyping chip are non-uniformly sampled, ASMC can specifically account for such artefacts in the data. However, ASMC does not currently combine this information to produce the ARG of all analysed samples. Lastly, a recent method, ARGinfer [Mahmoudi et al., 2022], performs inference under the coalescent with recombination, without an SMC-like approximation, but is unable to

scale to more than tens of sequenced samples.

A third class of methods combines probabilistic modelling with heuristic strategies to produce the ARG of a set of sequenced genomes [Rasmussen et al., 2014, Speidel et al., 2019, Wohns et al., 2022]. Within this family, ARGweaver [Rasmussen et al., 2014] samples from an exact posterior under the SMC model assuming a constant population size and a discretised series of times, but is unable to scale to more than tens of sequenced samples. ARGweaver was subsequently extended to handle a user-defined demography [Hubisz et al., 2020]. Relate [Speidel et al., 2019], uses a heuristic approach to build tree topologies based on [Li and Stephens, 2003], followed by a Gibbs sampling step under the coalescent prior, and is able to scale to tens of thousands of sequenced samples. `tsdate` [Wohns et al., 2022] leverages the heuristic ARG inference step of [Kelleher et al., 2019], then adds additional steps based on coalescent modelling to date the nodes of the ARG.

Inferred ARGs have been used in a wide range of genomic analyses, including efficient storage and manipulation of genomic data [Kelleher et al., 2019], inference of demographic history [Li and Durbin, 2011, Schiffels and Durbin, 2014, Sheehan et al., 2013, Terhorst et al., 2017, Speidel et al., 2019], and analysis of natural selection at individual loci or on complex traits [Rasmussen et al., 2014, Palamara et al., 2018, Speidel et al., 2019]. Early methods also suggested that ARGs may be leveraged to improve phenotype association and fine-mapping [Templeton et al., 1992, Zöllner and Pritchard, 2005, Minichiello and Durbin, 2006]. Despite the wealth of potential applications, whole-genome genealogical inference has been difficult to scale to the hundreds of thousands of samples required to encompass modern datasets. `tsinfer` and SARGE are both highly scalable but rely on heuristics and only output topologies, while Relate combines heuristics with coalescent modelling but is less scalable. Furthermore, existing methods for ARG inference of many samples are not optimised for SNP array data, while ASMC only outputs the ARG for a pair of samples at a time. To fully leverage the potential of genealogy-based analysis of complex traits, it is thus desirable to develop scalable ARG inference algorithms that perform well in both genotyping and sequencing data sets.

2.3 Population Genetics and GWAS: Three Themes

We conclude our survey of past ideas and works by reviewing three additional themes: genotype imputation, identity-by-descent detection and mapping, and linear mixed models. Each of these themes draws on population genetics in some way to improve GWAS power. Because the coalescent with recombination process serves as a baseline model for population genetics in the absence of selection, we claim that the ARG is implicitly tied to these various methods. In the future chapters of this thesis, particularly Chapter 7, we will explore the connections between the ARG and these various themes more explicitly.

2.3.1 Genotype imputation

In Section 2.1.3, we explained that the genetic variants captured by genotyping microarray chips are much sparser than whole genome sequencing, but represent common variation in a population and are predictive of other SNPs through LD tagging. **Genotype imputation** algorithms seek to leverage this LD information to impute (i.e. "fill in") genotype values for other variants in the population. These algorithms combine microarray data for the samples of interest with whole-genome sequencing from other samples, which make up a so-called **reference panel**. A schematic is shown in Fig. 2.3. Given SNP array data with markers at only a few sites (second row), as well as a reference panel (bottom), imputation fills in an expected genotype for additional sites in the reference panel—the five sites shown with question marks. An association scan can then be performed using not only the original genotypes but also the imputed genotypes, expanding the number of variants that can be tested. Such designs can improve the power of an association study as the imputed markers may serve to better tag underlying causal signals. Imputation can also aid in localising or fine mapping a detected signal, as the imputed markers have a higher density along the genome. For example, if a study originally genotyped only 100,000 SNPs, genotype imputation could expand this information to 1-10 million SNPs.

Genotype imputation algorithms [Marchini et al., 2007, Howie et al., 2009]

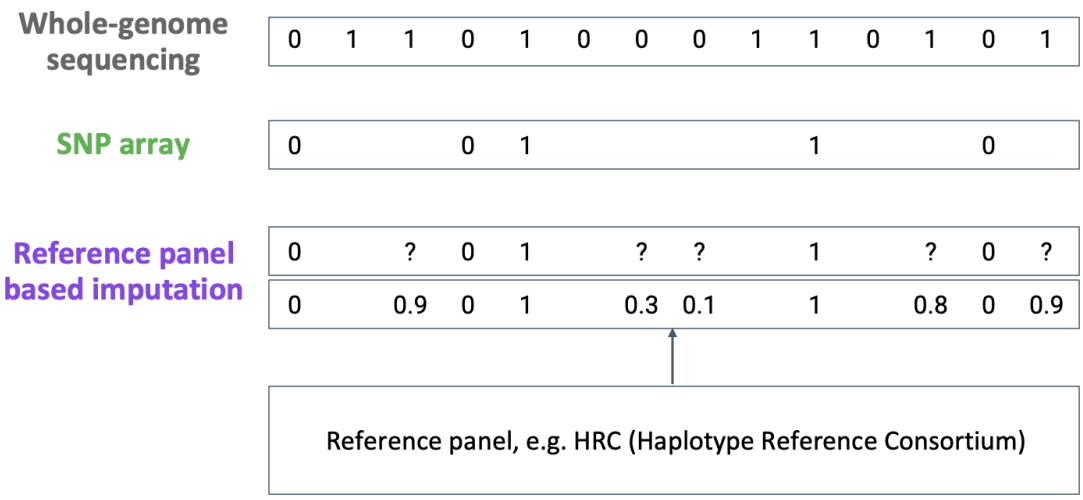


Figure 2.3: **Schematic of genotype imputation.** We show a haploid sample that has been whole-genome sequenced (top), genotyped with an array (second row), and imputed from a reference panel (third row). Whole-genome sequencing gives an entire readout of a genome, whereas SNP arrays only sample variation at a few locations throughout the genome. With a reference panel of whole genome sequences and an imputation algorithm, it is possible to predict the expected genotype at additional sites, shown as five question marks which get a value.

[Browning and Browning, 2016] are based on a "copying model" introduced by Li and Stephens [Li and Stephens, 2003], in which an array haplotype is modelled as being a composite from several of the sequencing haplotypes. During imputation, variants that are polymorphic in the reference panel are copied over to the array sample using a weighted average of the matched sequencing haplotypes. Additional innovations have dramatically improved the runtime of imputation algorithms, such as pre-phasing the array samples [Howie et al., 2012] and using efficient data structures to narrow down the possible sequencing haplotypes from which to copy [Browning et al., 2018, Rubinacci et al., 2020]. A recent work [Si et al., 2021] has also analysed the theoretical properties of imputation by performing imputation using simulated ARGs that contain both the array and sequencing samples.

A genealogical perspective can shed light on the basis for as well as the limitations of the Li-and-Stephens copying model, which approximates the coalescent with recombination. By introducing the dimension of time, it becomes clear that array samples are not directly inherited from the sequencing samples. Rather, ancestral

haplotypes, represented as internal nodes of the ARG, give rise to both the array samples and the sequencing samples, with possible mutation events hitting the intermediate branches. Rare mutations that lie on these branches and influence only array samples or sequencing samples will not be accurately imputed. In practice, rare mutations below a certain minor allele count in the reference panel are usually discarded, because of the difficulty of accurately imputing them. Knowledge of the clade structure of the ARG may aid in resolving these cases, and may improve rare variant imputation [Si et al., 2021].

Other rare mutations that only affect array samples but not sequencing samples will not be polymorphic in the reference panel; they do not show up as variants at all and have no chance of being imputed into the array samples. This is shown in Fig. 2.3, where genotype imputation (bottom) misses many of the variants that would be detected with whole-genome sequencing of all the array samples (top). To impute these rare variants, the reference panel must contain enough sequenced samples so that at least a few samples in expectation will contain the variant. This has motivated the collection of large, population-specific reference panels for performing genotype imputation [McCarthy et al., 2016, Turnbull et al., 2018, Halldorsson et al., 2021]. For the UK Biobank imputed data release, the reference panel consisted of a combination of ∼65,000 sequenced individuals from the **Haplotype Reference Consortium (HRC)** [McCarthy et al., 2016] and UK10K reference panels [Huang et al., 2015]. Samples were imputed using the IMPUTE4 algorithm. A follow-up work imputed exome sequencing variation from 49,960 UK Biobank exome sequences into the rest of the UK Biobank using a custom imputation algorithm [Barton et al., 2021].

In summary, genotype imputation performance is generally good for common variants, but accuracy decreases for variants on the rare end of the spectrum. Resolving this limitation within a genotype imputation design has motivated significant investment in additional sequencing efforts, which must represent the ancestral makeup of the samples of interest.

2.3.2 Identity-by-descent mapping

A second approach for improved association using SNP array data is **identity-by-descent (IBD) detection and mapping**. IBD haplotypes are those which are descended from an identical recent common ancestor [Browning and Browning, 2012, Thompson, 2013, Wakeley and Wilton, 2016]. The coalescent predicts that all genomes will meet at a single common ancestor if we go far enough back in time. Therefore, for the concept of IBD to be meaningful, the idea of *recent* common descent must be defined. Early works tended to require descent from an individual in an ancestral founder population.[24] A second approach is to define IBD by setting a time threshold, and requiring a common ancestor more recent than this time in the past [Wakeley and Wilton, 2016, Nait Saada et al., 2020]. A third approach for defining recent IBD observes that the average length of IBD segments is inversely related to the time of common descent [Palamara et al., 2012]. Therefore, IBD can also be defined as requiring segments of common descent longer than a length threshold [Browning and Browning, 2011, Wakeley and Wilton, 2016]. These various approaches are implemented in IBD detection methods, which we do not review here.

IBD detection has been applied to a variety of analyses of demography [Palamara et al., 2012, Palamara and Pe'er, 2013, Ralph and Coop, 2013], natural selection [Albrechtsen et al., 2010, Gusev et al., 2012, Nait Saada et al., 2020], and heritability [Browning and Browning, 2013]. A particular application termed IBD mapping or **population-based linkage** [Purcell et al., 2007] enables association of phenotypes using detected IBD [Houwen et al., 1994, Te Meerman et al., 1995, Purcell et al., 2007, Gusev et al., 2011, Browning and Thompson, 2012, Nait Saada et al., 2020]. In this framework, each detected IBD haplotype is used to designate carriers versus non-carriers of the haplotype, with this carrying status used as a stand-in genotype for association. The IBD mapping approach can reveal associated unseen low-frequency and rare variants which are inherited by carriers of the IBD haplotype. Although the

[24] For instance, [Wakeley and Wilton, 2016] quotes [Whitlock and Barton, 1997]: "The probability of identity by descent is defined as the chance that two genes are descended from the same gene in some ancestral population."

unseen rare variants are not present in the original array data used for IBD detection, they can be localised to somewhere within the haplotype. The power to detect these rare variants arises from modelling the LD and haplotype structure between the array SNPs. IBD mapping can also complement genotype imputation, which misses rare variants that are not polymorphic in the reference panel.

The ideas of IBD mapping have been reframed and applied using the perspective of the ARG [Zöllner and Pritchard, 2005, Minichiello and Durbin, 2006, Wu, 2008, Adhikari et al., 2012, Burkett et al., 2014]. In this approach, either a local coalescent tree or a local ARG is inferred from array data. Each of the branches of the tree or ARG are then scanned for association with the phenotype. This approach generalised IBD mapping because the branches that are tested correspond to ancestral haplotypes, but without a time threshold restricting what counts as IBD. [Minichiello and Durbin, 2006] developed methods in this area involving an ensemble of inferred ARGs and permutation testing to calibrate significance thresholds [Churchill and Doerge, 1994]. The ARG association approach is also closely related to the inference and testing of haplotype clades [Templeton et al., 1992, Durrant et al., 2004, Browning and Browning, 2007a, Gusev et al., 2011], with different terminology mainly due to whether an ARG is explicitly inferred.

2.3.3 Linear mixed models

Linear mixed models (LMMs) and their extensions represent a state-of-the-art approach for performing heritability estimation [Yang et al., 2010, Evans et al., 2018b], polygenic prediction [Wray et al., 2013], and association [Kang et al., 2008, Yang et al., 2014] of complex traits.[25] In their classical formulation, mixed models compute an N by N **kinship matrix** or **relatedness matrix** between samples. This matrix is coupled with available phenotype and covariate information in the various possible forms of analysis. The kinship matrix represents an aggregate of genetic **random effects** throughout the genome, while the covariates and any SNPs being tested for association are treated as **fixed effects**. The combination of random and fixed effects

[25]These terms were previously referenced in Section 2.1.2.

in the model is the source of the term "mixed models."

The three types of analysis enabled by linear mixed models form a chain of increasing sophistication. In heritability estimation, algorithms such as **restricted maximum likelihood (REML)** [Patterson and Thompson, 1971, Gilmour et al., 1995] are used to obtain parameter estimates for the underlying mixed model. These parameters describe the heritability of the trait as well as other aspects of the **genetic architecture**—what sorts of variants are more or less responsible for the heritability of the trait. Once these parameters have been estimated, they can be plugged into the model and used to provide phenotype predictions of samples from their genetic data. The third type of analysis in the chain, mixed model association, depends on both heritability estimation and polygenic prediction. In this setting, when testing a SNP on a certain chromosome, say chromosome 6, for association, the other chromosomes (1 to 5 and 7 to 22) are used to provide a combined polygenic prediction of the phenotype. This prediction is then subtracted from the true phenotype, and the remaining residual is tested for association with the SNP [Svishcheva et al., 2012, Loh et al., 2015b]. This procedure removes components of the phenotype arising from other genetic sources, thus yielding a more "pristine" phenotype in which associations can be more easily detected [Listgarten et al., 2012, Yang et al., 2014].

In Sections 2.3.3.2 and 2.3.3.3, we review the various models and algorithms that have been developed for mixed model analysis in human datasets. Innovations in this domain have typically focused either on scalability or on achieving more accurate estimates and increased association power through better modelling. First, however, we discuss the various techniques for computing kinship matrices, and the role of a genealogical viewpoint in this development.

2.3.3.1 Estimating kinship matrices: from pedigrees to markers

Before it became possible to measure DNA at high throughput across the genome, genetic inheritance was typically measured using pedigrees. A **pedigree** is a family tree of a set of samples, showing each sample's parents, grandparents, and so on. Pedigrees can either be available from family records, or can arise in an experiment

through breeding of plants or animals. If a ground truth ARG of a set of samples is available, the ARG provides a more detailed record than a pedigree because it tracks ancestral lineages for each genetic locus, adapting with recombination events. Pedigrees, on the other hand, list all possible ancestors, but may not be able to pinpoint the lineage along which a gene is inherited.

Nevertheless, pedigrees can predict the expected amount of genetic material inherited in common (or identity-by-descent[26]) by two samples. For instance, two full siblings who are not identical twins are expected to inherit 1/2 of their material in common from their parents. Two first cousins are expected to inherit 1/8 of their material in common from their common grandparents. In genetic studies where breeding experiments or close families are involved, pedigrees can be used to construct a kinship matrix between samples. This is a symmetric N by N matrix where the diagonal consists of ones, and the ijth entry is the expected fraction of material inherited in common by samples i and j: 1/2 for full siblings, 1/8 for first cousins, and encompassing values like 1/16 and 3/32 as well depending on the depth and complexity of the pedigree. The constructed pedigree kinship matrix can be used downstream in methods for heritability estimation and polygenic prediction [Lynch et al., 1998].

The pedigree kinship gives the expected amount of IBD between two samples, but the true fraction of IBD can vary based on the randomness inherent in recombination and meiosis [Hill and Weir, 2011]. As genome-wide markers became available in plant and animal genetics and subsequently human genetics, many works began to transition from pedigree-based kinships to marker-based kinship estimation [Visscher et al., 2006, VanRaden, 2008, Hayes et al., 2009]. These works obtained better estimates of true genetic similarity through capturing the realised IBD segments using the observed markers, rather than working only with expected IBD based on pedigrees. In parallel to this transition from pedigrees to markers in prediction and heritability estimation, the need to properly account for population structure in GWAS led to the development of mixed-model association methods [Yu et al., 2006, Zhao et al., 2007, Kang et al., 2008, Zhang et al., 2010, Kang et al., 2010]. These developments

[26]See Section 2.3.2.

culminated in two highly influential papers, published in 2010, which applied mixed models with marker-based kinships to human datasets of thousands of individuals for heritability estimation [Yang et al., 2010] and association [Kang et al., 2010].

A typical way of computing the marker-based kinship matrix is as the covariance matrix between samples given standardised SNP data [Yang et al., 2010] (for example, working with values $\widetilde{x}_{nm}$ described in Section 2.1.2). This became known as a **SNP-estimated kinship** or **genomic relatedness matrix (GRM)** [Yang et al., 2011a].

2.3.3.2 Improvements in mixed-model heritability estimation and prediction

Given an N by N GRM and a length N phenotype vector, software packages such as GCTA [Yang et al., 2011a] are able to perform REML estimation and return estimates for the two variance parameters σ_g^2 and σ_e^2 in the mixed model, representing genetic and environmental variance. Typically, this comes in the form of estimates $\hat{\sigma}_g^2$ and $\hat{\sigma}_e^2$, along with estimated standard errors $\hat{se}(\hat{\sigma}_g^2)$ and $\hat{se}(\hat{\sigma}_e^2)$. The narrow sense heritability h^2 of the linear mixed model is defined as

$$h_g^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2},$$

which can be estimated using $\hat{\sigma}_g^2$ and $\hat{\sigma}_e^2$. Once σ_g^2 and σ_e^2 have been estimated (a step referred to as "model fitting"), the model can also be used to estimate the **best linear unbiased predictor (BLUP)** for polygenic prediction under the mixed model. By increasing the density of markers included in the mixed model, for example through imputed [Yang et al., 2015] and whole-genome-sequencing data [Wainschtein et al., 2022], more phenotype variance is explained, leading to increased heritability estimates and better polygenic predictors. Polygenic prediction can also be improved through larger sample sizes [Wray et al., 2013].

The original GRM proposed by [Yang et al., 2010] included all available markers and formed a covariance matrix after standardising the genotypes. This setting assumes that all markers have an equal prior distribution effect size on the phenotype,

regardless of **minor allele frequency (MAF)**.[27] If we model our phenotype as

$$\widetilde{y}_n = \sum_{m=1}^{M} \widetilde{x}_{nm}\beta_m + \epsilon_n,$$

where β_m corresponds to standardised SNP effect sizes and ϵ_n is an environmental variance, the mixed model and resulting GRM of [Yang et al., 2010] corresponds to a consistent normal prior $\beta_m \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_g^2/M)$ on effect sizes. Future works allowed the normal prior on β_m to vary based on the MAF and local LD of SNP m [Speed et al., 2012, Speed et al., 2017]. For instance, [Speed et al., 2012] introduced a parameter α such that the variance of the standardised effect size β_m is proportional to $[p_m(1-p_m)]^{\alpha+1}$.[28] A value of $\alpha = -1$ corresponds to the GRM of [Yang et al., 2010] and represents strong negative selection on large effect variants as a result of stabilising selection on the trait [Zeng et al., 2018, Schoech et al., 2019], while $\alpha = 0$ represents no selection. [Speed et al., 2017] found that $\alpha = -0.25$ was a typical value for most traits, whereas later works [Zeng et al., 2018, Schoech et al., 2019] determined average α estimates of -0.37 and -0.38 across several UK Biobank traits.

The incorporation of the α parameter is an example of investigating different possible genetic architectures of a trait. Using a misspecified genetic architecture can lead to biased heritability estimates and hampered prediction performance. Therefore, many works have continued to explore extensions to the genetic architecture, including MAF, LD, and functional annotation dependence as well as non-infinitesimal architectures. One approach is to parameterise the architecture and then infer these parameters using maximum likelihood [Speed et al., 2017, Schoech et al., 2019], fully Bayesian approaches [Zhou et al., 2013, Zeng et al., 2018], or validation set predictive performance [Vilhjálmsson et al., 2015, Loh et al., 2015b]. An alternative approach is to create multiple GRMs spanning different MAF, LD, and functional annotation categories. A mixed model is then fit with these multiple variance components, allowing for heritability to be partitioned across categories of interest. The advantage

[27] Given raw genotypes $x_{nm} \in \{0, 1, 2\}$, we can define the allele frequency of SNP m as $p_m = \sum_{n=1}^{N} x_{nm}/2N$ and the minor allele frequency as $\min(p_m, 1-p_m)$.

[28] See Section 3.2.1 for the expression of the resulting GRM as a function of α.

of the multi-GRM approach is that it does not assume a particular functional form for the dependence, and instead infers a best fit in a non-parametric way. The disadvantage is that standard errors increase with the number of different components fit. The multi-component approach has been broadly applied in studying heritability and genetic architecture [Yang et al., 2011b, Lee et al., 2012, Lee et al., 2013, Gusev et al., 2014, Yang et al., 2015, Loh et al., 2015a, Evans et al., 2018b] as well as in polygenic prediction [Speed and Balding, 2014, Márquez-Luna et al., 2021].

Besides the many works that sought to more accurately model genetic architecture, there have also been improvements in computational efficiency. BOLT-REML [Loh et al., 2015a] observed that instead of explicitly forming and inverting the GRM, it is possible to perform REML using conjugate gradient-based optimisation instead. BOLT-REML supports multiple variance components and was applied to datasets of tens of thousands of samples, although it also scales to hundreds of thousands of samples. In the area of prediction, cvBLUP [Mefford et al., 2020] developed an efficient way to calculate leave-one-out polygenic predictions, which is useful for easily evaluating prediction accuracy in a sample.

2.3.3.3 Improvements in mixed-model association

As previously noted, multiple papers that used mixed-model association methods to control for population structure were published in 2006-2010, focusing on plant and animal genetics [Yu et al., 2006, Zhao et al., 2007, Kang et al., 2008, Zhang et al., 2010]. In 2010, the EMMAX method [Kang et al., 2010] applied these methods to human GWAS, with some algorithmic runtime improvements over previous methods. The FaST-LMM [Lippert et al., 2011] and GEMMA [Zhou and Stephens, 2012] papers thereafter introduced the strategy of performing a single eigendecomposition of the GRM, and subsequently expressing operations in terms of this eigendecomposition. This enabled scaling to even larger human GWAS.

Follow-up methods provided further algorithmic improvements to increase power and better account for confounding [Listgarten et al., 2012, Svishcheva et al., 2012, Yang et al., 2014]. [Listgarten et al., 2012] observed that mixed-model association

is able to improve power compared to univariate linear regression of SNPs. This is because the SNPs included as random effects not only represent potential sources of population structure confounding, but they may also tag causal SNPs for the polygenic phenotype. Including these SNPs reduces noise in the phenotype, making associations easier to detect. However, the SNPs included as random effects should not be in linkage disequilibrium with the SNP being tested, or power may be reduced, an effect the authors termed **proximal contamination**. This was explored further by [Yang et al., 2014], which introduced a simple procedure to guard against proximal contamination: **LOCO (leave one chromosome out)**. In this setup, when testing a SNP for association, the SNPs from that chromosome are excluded from the random effects, whereas SNPs from all other chromosomes are included.

Many further substantial improvements to mixed-model association were introduced with the BOLT-LMM method [Loh et al., 2015b, Loh et al., 2018], which remains the state of the art in terms of association power. BOLT-LMM combined modelling and algorithmic innovations, offering both an increase in power over previous methods as well as orders of magnitude of speedup. On the modelling side, BOLT-LMM allowed for a mixture of Gaussians prior on SNP effect sizes within the mixed model, extending the standard assumption of a single Gaussian prior. This non-infinitesimal model is fit using a variational Bayes algorithm, and offers increased association power compared to the standard infinitesimal model, which is implemented as BOLT-LMM-inf within the same package.

On the algorithmic side, BOLT-LMM rewrites the mixed-model association test statistic in a clever way to facilitate much faster computation.[29] The SNPs included as random effects are first used to build an in-sample polygenic prediction of the phenotype, which is then subtracted from the phenotype. The remaining residual is then tested for association with the SNP in question using the univariate linear regression model (see Section 2.1.2), with the residual values as a stand-in phenotype. The chi-squared statistics from this test are then divided by a globally estimated

[29]The viewpoint used builds off observations in the GRAMMAR [Aulchenko et al., 2007] and GRAMMAR-Gamma [Svishcheva et al., 2012] methods.

calibration constant. Because of the LOCO setup, 22 separate LOCO polygenic predictions and residuals are formed, with each LOCO residual being used for testing all SNPs on one chromosome. After the predictions and residuals are fit, testing can then be performed as quickly as univariate linear regression.

BOLT-LMM introduced many additional insights to improve computational efficiency. Similar to BOLT-REML, conjugate gradient operations were used to solve linear systems instead of inverting matrices directly. Other facets of the implementation that provided substantial constant-factor memory or speed improvements included operating on raw genotypes with 2 bits per base pair per individual, utilising level 3 BLAS and SSE instructions, and multithreading.

One limitation of BOLT-LMM is that it is designed for quantitative traits, although it can also be applied to relatively balanced case-control traits. SAIGE [Zhou et al., 2018] built on BOLT-LMM's innovations, particularly the use of conjugate gradient operations, to efficiently perform generalised mixed-model association of unbalanced case-control traits.

Two additional methods of note have achieved faster runtime than BOLT-LMM-inf in limited settings. REGENIE [Mbatchou et al., 2021] quickly estimates polygenic prediction effect sizes by considering SNPs in windows at a time, then creating LOCO residuals which are tested for association. REGENIE also contains modes to handle both quantitative and case-control traits. However, REGENIE does not support non-infinitesimal modelling, and thus yields decreased association power compared to BOLT-LMM for quantitative traits. fastGWA [Jiang et al., 2019] is another method which achieves fast runtime by using a sparse approximation to the GRM. However, fastGWA focuses on controlling for population structure, as the sparse approximation limits the potential of the mixed model to improve association power. fastGWA was subsequently extended to handle case-control traits [Jiang et al., 2021].

In summary, as of this writing, BOLT-LMM remains the method of choice for maximising mixed-model association power in analyses of quantitative traits, due to its non-infinitesimal model. Additional methods such as SAIGE, REGENIE, and fastGWA may be used to analyse imbalanced case-control traits.

Chapter 3

A Genealogical Perspective of Mixed-Model Analysis for Complex Traits

3.1 Overview

We concluded Chapter 2 by reviewing three themes that have enhanced genome-wide association studies. Each of the perspectives of genotype imputation, identity by descent mapping, and linear mixed models draws inspiration from population genetics to formulate methods that can be used in medical genetics. As further improvements in GWAS methodology are sought, a genealogical perspective may continue to provide a fruitful line of inquiry. The ARG, generated by the coalescent with recombination process, captures the full extent of genealogical processes studied by population genetics. ARGs may therefore serve as an ideal foundation for future developments, providing both a framework for reasoning about new methods as well as a data structure with which to implement these methods.

In this chapter, we introduce two such novel methods for ARG-based complex trait analysis within a mixed-model framework. Both methods leverage the correspondence between mutations and branches of the ARG, such that if the true ARG of a set

of samples is known, every mutation can be placed on one of the ARG branches. First, we propose to construct **ARG-GRMs**, which are GRMs built using the ARG, and show that these GRMs can be used in mixed model analyses such as heritability estimation, polygenic prediction, and mixed-model association. ARG-GRMs consider all possible mutations that can occur on the ARG, weighting the possible outcomes by their likelihood according to uniform sampling of mutations. Second, we develop a framework to test branches of the ARG for association with a phenotype while using a mixed model to improve power, which we call **ARG-MLMA (ARG-based mixed linear model association)**. In this setting, the ARG is used to propose variants which can be tested for association, while the mixed model controls for population stratification and improves power by explaining polygenic signal. ARG-MLMA builds on previous work in ARG-based association and IBD mapping, reviewed in Section 2.3.2, with the addition of utilising the state-of-the-art framework of mixed-model association, reviewed in Section 2.3.3.

In realistic simulations, we find that our ARG-based methods significantly outperform SNP array data and are equivalent to having access to sequencing data. We also demonstrate that the ARG-MLMA and ARG-GRM methods can be combined to further improve power. Crucially, both methods only require information about the ARG nodes and edges, without mutation information. Our results demonstrate an upper bound on performance if perfect ARGs can be reconstructed from available data, for instance sparse array data. In reality, the performance of these methods will depend on the accuracy of available ARGs, a subject that will be explored in later chapters.[1]

[1]The remainder of this chapter is expanded and adapted from [Zhang et al., 2021].

3.2 Methods

3.2.1 Construction of ARG-GRMs

We consider N haploid individuals, M sites, and genotypes x_{ik} for individual i at site k, where variant k has mean p_k. We assume an additive infinitesimal genetic architecture so that the genetic component of a trait is given by $g_i = \sum_k \beta_k x_{ik}$, where β_k is drawn with mean zero and MAF-dependent variance proportional to $(p_k(1 - p_k))^\alpha$, where α captures the strength of negative selection [Yang et al., 2010, Speed et al., 2012]. Using available markers, a common estimator for the ij-th entry of the $N \times N$ genomic relatedness matrix (GRM [Yang et al., 2011a]) may be computed as

$$K_\alpha(i,j) = \frac{1}{M} \sum_{k=1}^{M} \frac{(x_{ik} - p_k)(x_{jk} - p_k)}{[p_k(1 - p_k)]^{-\alpha}}. \tag{3.1}$$

Given an ARG, we compute the ARG-GRM as the expectation of the marker-based GRM that would be obtained using sequencing data, i.e. when all variants are observed, assuming that mutations are sampled uniformly over the area of the ARG. The rationale of this approach is that when sequencing data is not available but an accurate ARG can be estimated from an incomplete set of markers, the ARG-GRM may provide a good estimate for the sequence-based GRM. We briefly describe how ARG-GRMs are derived from the ARG for the special case of $\alpha = 0$. We discuss the more general case and provide further derivations in Appendix A.

By applying a series of normalising transformations prior to using the GRM, we derived a family of invariances under which related GRMs produce equivalent results (see Appendix A). Under these invariances and assuming $\alpha = 0$, (3.1) is equivalent to the matrix containing the Hamming distance between the sequences of pairs of samples, given by

$$K_H(i,j) = \sum_{k=1}^{M} x_{i,k} \oplus x_{j,k},$$

where $\oplus$ refers to the XOR function. Assume there are L total base pairs in the

genome, a constant mutation rate per base pair and generation of μ, and denote the TMRCA of i and j at base pair k with t_{ijk} (in generations). The ij-th entry of the ARG-GRM is obtained as the expected number of mutations that are carried by only one of the two individuals, which corresponds to the expected Hamming distance for sequences i and j:

$$\begin{aligned} K_{ARG}(i,j) &= \mathbb{E}\left[K_H(i,j)|ARG\right] \\ &= \sum_{k=1}^{L} P(\text{Poisson}(2\mu t_{ijk}) > 0) = \sum_{k=1}^{L} 1 - \exp(-2\mu t_{ijk}) \approx \sum_{k=1}^{L} 2\mu t_{ijk}. \end{aligned}$$

One can compute the above ARG-GRM by iterating over the entire ARG and summing pairwise TMRCAs across the genome (Fig. 3.1a). Note that in this case of $\alpha = 0$, the ARG-GRM may be estimated using pairwise TMRCA estimates alone [Palamara et al., 2016], e.g. using ASMC. For increased efficiency, one can compute a Monte Carlo ARG-GRM by uniformly sampling new mutations on the ARG with a high mutation rate and using these mutations to build the ARG-GRM, using (3.1). We used simulations to verify that Monte Carlo ARG-GRMs converge to exactly computed ARG-GRMs for large mutation rates (Fig. 3.1b-c), leading us to choose a default large value of $\mu = 1.65 \times 10^{-7}$ for ARG-GRM computations. Stratified Monte Carlo ARG-GRMs may also be computed by partitioning the sampled mutations based on e.g. allele frequency, LD, or allele age [Lee et al., 2013, Yang et al., 2015, Gazal et al., 2017, Evans et al., 2018b]. For simulations adopting MAF-stratification (e.g. Fig. 3.2g), we used MAF boundaries given by $\{0, 0.01, 0.05, 0.5\}$, normalised genotypes using $\alpha = -1$, and then provided the MAF-stratified ARG-GRMs in input to GCTA [Yang et al., 2011a].

3.2.2 Simulation of ARGs and genetic data

We used the `msprime` coalescent simulator [Kelleher et al., 2016] to simulate ARGs and genetic data. For each run, we first simulated sequence data with given physical length L for N haploid individuals. For this chapter, our simulations used a mutation

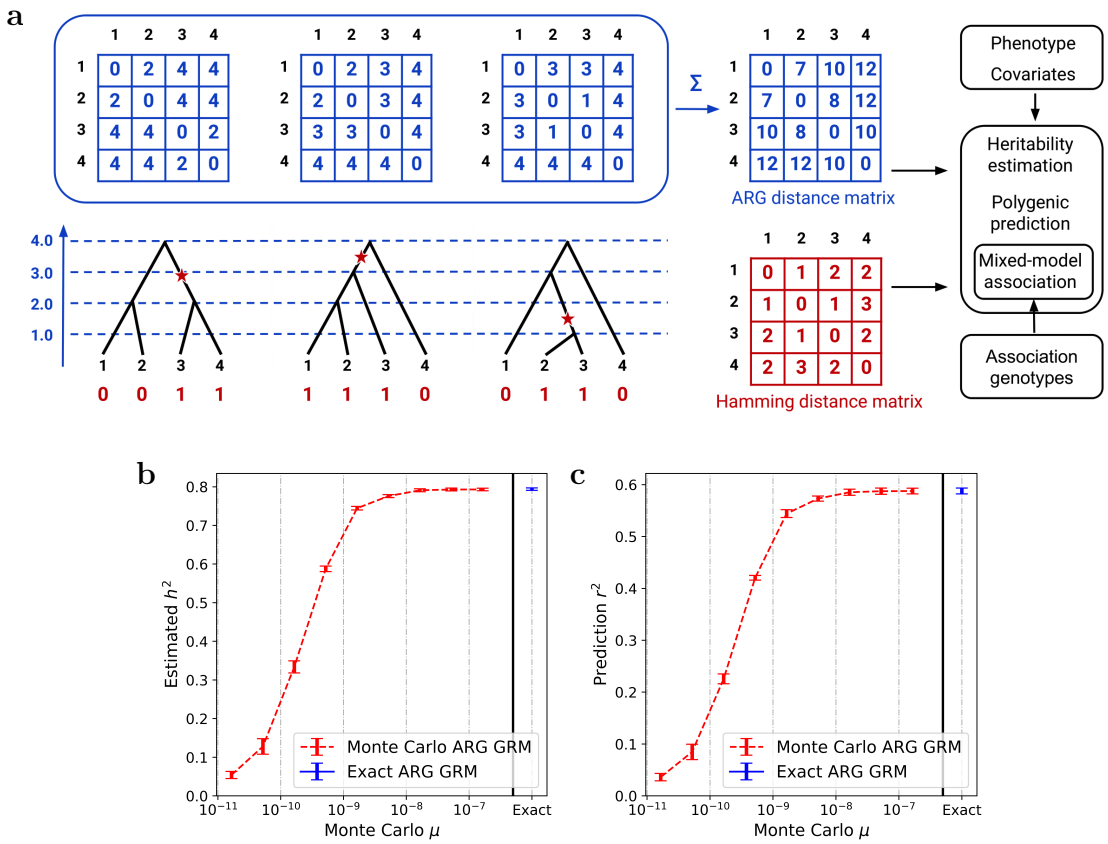


Figure 3.1: **Overview of ARG-GRM definition and Monte Carlo estimator. a**. Schematic of ARG-GRMs. Given an ARG between samples, we can compute the TMRCA matrix at each site and sum this over the genome to obtain the $\alpha = 0$ ARG distance matrix (top, in blue). This equals a scaled version of the expected Hamming distance matrix (bottom, in red), which is formed by counting the number of differences between the genotypes of samples. By applying a series of simple matrix transformations to the ARG distance matrix (see Appendix A), we obtain the ARG-GRM, which can subsequently be used in complex trait analysis just like genotype-based GRMs. **b**,**c**. We compare the use of an exact $\alpha = 0$ ARG-GRM to Monte Carlo $\alpha = 0$ ARG-GRMs for heritability estimation (**b**) and polygenic prediction (**c**). As we increase the mutation rate for the Monte Carlo ARG-GRMs (rightmost value of $\mu = 1.65 \times 10^{-7}$), we approach results from using the exact ARG-GRM. Simulations use $N = 2{,}000$ haploid samples, $h^2 = 0.8$, $\alpha = 0$, and 10 Mb. We show mean values over 5 runs. Error bars represent 2 s.e. from meta-analysis.

rate of $\mu = 1.65 \times 10^{-8}$ per base pair per generation, a constant recombination rate of $\rho = 1.2 \times 10^{-8}$ per base per generation, and a demographic model inferred using SMC++ on CEU (European demography) 1,000 Genomes samples [Terhorst et al., 2017]. These simulations also output the simulated genealogies, which we refer to as "ground-truth ARGs" or "true ARGs". To obtain realistic SNP data, we subsampled the simulated sequence sites to match the genotype density and allele frequency spectrum of UK Biobank SNP array markers (chromosome 2, with density defined using 50 evenly spaced MAF bins).

3.2.3 ARG-GRM simulation experiments

We simulated polygenic traits from haploid sequencing samples for various values of h^2 and α. We varied the number of haploid samples N but fixed the ratio L/N throughout experiments, where L is the genetic length of the simulated region. For heritability and polygenic prediction experiments, we adopted $L/N = 5 \times 10^{-3}$ Mb/individuals. For association experiments, we simulated a polygenic phenotype from 22 chromosomes, with each chromosome consisting of equal length $L/22$ and $L/N = 5.5 \times 10^{-3}$ Mb/individuals. Mixed-model prediction r^2 and association power may be roughly estimated as a function of h^2 and the ratio N/M, where M is the number of markers [Daetwyler et al., 2008, Wray et al., 2013, Loh et al., 2015b]. We thus selected values of M and L such that the N/M ratio is kept close to that of the UK Biobank ($L = 3 \times 10^3$ Mb, $N \approx 6 \times 10^5$).[2]

We computed GRMs using ground-truth ARGs, SNP data, and sequencing data and provided them in input to GCTA with the simulated phenotype. For heritability estimation, we ran GCTA with flags `--reml-no-constrain` and `--reml-no-lrt`. For polygenic prediction, we ran leave-one-out prediction using cvBLUP [Mefford et al., 2020] within GCTA, then computed r^2 between the resulting predictions and the phenotype. For ARG-GRM association experiments (Fig. 3.2c,f), we performed

[2]This approach assumes that M is proportional to L, independent of N. Although this may not always hold depending on the markers included under the count of M, it is accurate when restricting to variants above a minimum MAF.

MLMA of array data SNPs, testing each chromosome while using a LOCO GRM built on the other 21 chromosomes. We measured power improvement as the relative increase of mean $-\log_{10}(p)$ for MLMA compared to linear regression of array data SNPs and compared ARG-GRMs to GRMs of array and sequencing data.

3.2.4 ARG-MLMA

We also developed an approach to perform mixed linear model association of variants extracted from the ARG, which we refer to as ARG-MLMA. In this approach, we sample mutations from a given ARG using a specified rate μ and apply a mixed model association test to these variants. The choice of μ may be used to decrease the number of tests performed, sampling mutations based on their likelihood to be generated by the ARG. In the case where an ARG is inferred from data (see Chapters 5-6), branches that are less certain may have smaller area and will be less likely to be sampled.[3] The MLMA framework can be performed using GRMs built using array markers, sequencing GRMs, or ARG-GRMs, which model polygenicity and account for population structure.

For the simulation experiments in this chapter (Fig. 3.3) and Chapter 5, we traversed the ARG and wrote out all possible mutations to disk, which is equivalent to adopting a large value of μ, and used GCTA to perform LMM testing of these mutations.[4] When comparing against ARG-based linear regression, we tested the same written mutations using PLINK. We used sequencing variants from chromosomes 2-22 to form a polygenic background with narrow-sense heritability $h^2 = 0.8$ and negative selection parameter [Speed et al., 2012] $\alpha = -0.25$. In detail, we drew effects $\beta_i \sim \mathcal{N}(0, [p_i(1-p_i)]^\alpha)$, computed $y_g = \sum_i \beta_i x_i$ using unnormalised haploid genotypes x_i, and scaled y_g to have variance h^2. We then added a single causal sequencing variant on chromosome 1 (chosen at random from those with allele frequency $p = 0.0025$) with effect size β and added independent normally-distributed noise for the remaining

[3] See Appendix A for our definition of an ARG "branch" and its area.

[4] In larger UK Biobank analyses we used a specified rate of μ. We also developed a separate pipeline instead of GCTA for more computationally efficient association. See Chapter 6 for details.

environmental variance. We varied the value of β and measured association power for each method as the fraction of runs (out of 200) detecting a significant association on chromosome 1. Significance thresholds for each method were calibrated to yield a family-wise error rate of 0.05 under the null condition $\beta = 0$ based on 200 runs. We compared between association of array data and the true ARG (Fig. 3.3). For each method, we tested using both linear regression and MLMA with a leave one chromosome out LOCO GRM built from array markers, sequencing markers, or ARGs from chromosomes 2-22.

3.3 Results

3.3.1 Exact versus Monte Carlo ARG-GRMs

We performed experiments involving heritability estimation (Fig. 3.1b) and polygenic prediction (Fig. 3.1c) and compared between exact ARG-GRMs and Monte Carlo ARG-GRMs. In both experiments, a trait was simulated with $N = 2{,}000$ haploid samples, $h^2 = 0.8$, $\alpha = 0$, and 10 Mb. Both types of GRMs were computed using $\alpha = 0$. The exact ARG-GRMs enabled accurate heritability estimation of $\hat{h}^2 = 0.8$ and a polygenic prediction r^2 of around 0.6. Monte Carlo ARG-GRMs matched this performance for high values of μ, with performance saturating at $\mu \approx 1.65 \times 10^{-7}$. For lower values of μ, estimated heritability and prediction r^2 both decreased, yielding values approaching 0 for $\mu \approx 1.65 \times 10^{-11}$.

In our implementation, Monte Carlo and exact ARG-GRMs require $O(N^2)$ time for each mutation or branch considered (see Appendix A for our definition of an ARG "branch"). We observed that for $\mu \approx 1.65 \times 10^{-7}$, the number of mutations is significantly less than the number of branches in the ARG. We therefore chose this value of μ for all following experiments with ARG-GRMs, improving computational speed compared to exact ARG-GRMs without compromising accuracy.

3.3.2 ARG-GRM performance when α is known

We performed analyses of heritability estimation, polygenic prediction, and mixed-model association across values $h^2 \in \{0.8, 0.3\}$ and $\alpha \in \{0, -0.5, -1\}$. We simulated polygenic phenotypes using sequencing data, then built GRMs using the true value of α from either sequencing variants, ARGs, or sampled SNP array variants. In heritability estimation, ARG-GRMs performed as well as sequencing GRMs, which were both able to accurately estimate the heritability $h^2 = 0.8$ or $h^2 = 0.3$ of the trait (Fig. 3.2a,d). By contrast, SNP array GRMs underestimated the heritability, with the gap becoming more severe for more negative values of α, where there is a larger contribution to heritability from rare variants.

For polygenic prediction, ARG-GRMs enabled predictions that were as accurate as from sequencing variants, again outperforming SNP array data (3.2b,e). Based on this result, we expected mixed-model association using ARG-GRMs to outperform SNP GRMs, because the power improvement in MLMA is due to predicting and adjusting for polygenic signal from the other 21 chromosomes. We simulated a polygenic trait from 22 chromosomes, then performed association of SNP array variants on chromosome 1, while using a GRM built from chromosomes $2 - 22$ within the mixed model. We computed the mean $-\log_{10}(p)$ of the association p-values and measured improvement relative to the mean $-\log_{10}(p)$ from linear regression of these same SNP array variants. For $h^2 = 0.8$, MLMA using sequencing GRMs yielded a mean fractional improvement over linear regression of 53% for $\alpha = 0$, 27% for $\alpha = -0.5$, and 1.3% for $\alpha = -1$ (Fig. 3.2c). This is consistent with the decreasing prediction r^2 for more negative values of α. For $h^2 = 0.8$, ARG-GRMs yielded MLMA association power numbers that were statistically indistinguishable from sequencing GRMs across 50 random seeds, while significantly improving over SNP GRMs (Fig. 3.2c). For $h^2 = 0.3$, the error bars (2 standard errors) for sequencing GRMs, ARG-GRMs, and SNP array GRMs overlapped for $\alpha = -1$ with 500 random seeds, due to the nature of the small improvements in association power ($\approx 0.07\%$ more power than linear regression for sequencing GRMs). Otherwise, ARG-GRMs again matched

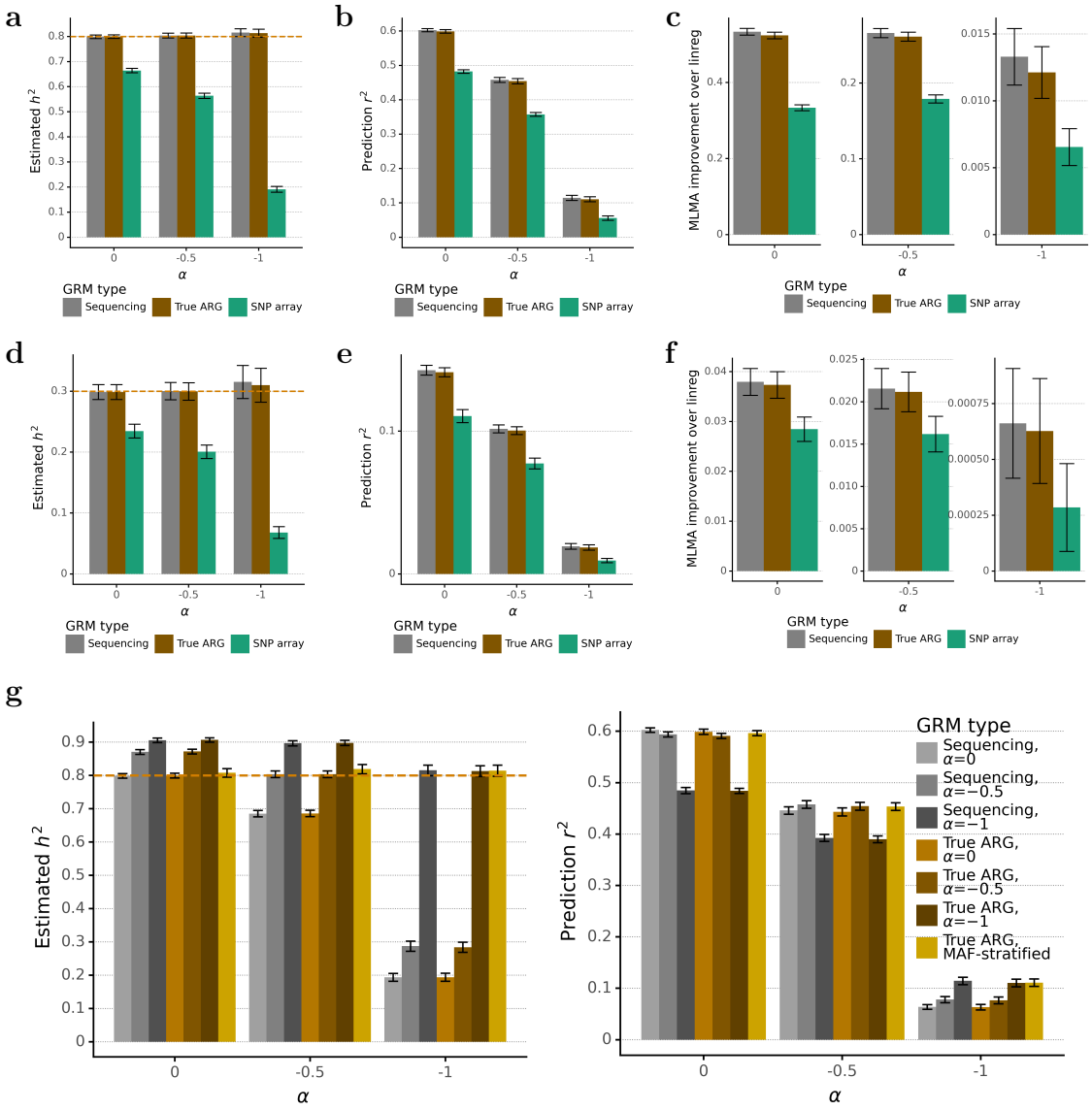


Figure 3.2: **Performance of ground-truth ARG-GRMs in simulations. a-f.** All simulations contain $N = 10{,}000$ haploid samples with varying $h^2 \in \{0.8, 0.3\}$ and $\alpha \in \{0, -0.5, -1\}$. **a**,**d**. Heritability estimation for a 50 Mb region for $h^2 = 0.8$ (**a**) and $h^2 = 0.3$ (**d**). **b**,**e**. Polygenic prediction for a 50 Mb region for $h^2 = 0.8$ (**b**) and $h^2 = 0.3$ (**e**). **c**,**f**. Mixed-model association for 22 chromosomes of 2.5 Mb each for $h^2 = 0.8$ (**c**) and $h^2 = 0.3$ (**f**). We show the relative improvement in mean $-\log_{10}(p)$ of MLMA compared to linear regression (see Section 3.2.3). **g**. Heritability estimation and polygenic prediction using ARG-GRMs and sequencing GRMs with various values of (possibly misspecified) α, as well as using MAF-stratified ARG-GRMs ($N = 10{,}000$ haploid samples, 50 Mb, $h^2 = 0.8$). For all panels, heritability and prediction experiments involve 5 simulations per bar, and most association experiments involve 50 simulations per bar, except for the $h^2 = 0.3, \alpha = -1$ condition in **f**, which involved 500 simulations. Error bars represent 2 s.e. (from meta-analysis in the case of heritability estimation).

sequencing GRMs and outperformed SNP array GRMs (Fig. 3.2d).

In summary, when α is known, we found that ARG-GRMs built using the known value of α and ground-truth ARGs performed as well as sequencing GRMs and outperformed SNP array GRMs across heritability estimation, polygenic prediction, and mixed-model association.

3.3.3 ARG-GRM performance when α is unknown

The above results assume that α is known, and used the true value of α to construct the sequencing, SNP, and ARG-GRMs. When α is unknown, previous works have built multiple GRMs, stratifying by MAF, yielding a flexible nonparametric approach to heritability estimation and polygenic prediction (see Section 2.3.3.2). We developed methods to partition the ARG by MAF and build separate GRMs for each MAF bin, then perform multi-GRM analyses (see Section 3.2.1).

Our results confirm the flexibility and robustness of such an approach, which performs comparably to using a single GRM built with the true value of α (Fig. 3.2g). When building only a single GRM, if the value of α is misspecified, heritability estimation is biased and prediction r^2 is hampered. This is true both for ARG-GRMs and sequencing GRMs. MAF-stratified ARG-GRMs, on the other hand, perform as well as using the correct value of α, matching the optimal estimates for both ARG-GRMs and sequencing GRMs. In summary, MAF-stratification for ARG-GRMs enables robust heritability estimation and polygenic prediction when α is unknown.

3.3.4 ARG-MLMA outperforms linear regression of ARG and SNP variants

While ARG-GRMs use variants within an ARG to build a GRM, ARG-MLMA tests these variants for association using a separately defined mixed model. In a first instance, we tested for variants within the ARG of chromosome 1, and used SNP array variants from chromosomes 2-22 as random effects in the mixed model. Our phenotype consisted of polygenic effects from chromosomes 2-22, plus a single rare

variant with effect size β and an allele frequency of 0.25%, or an allele count of 5. For each value of β, we performed 200 independent simulations and measured association power as the fraction of runs that detected a significant association on chromosome 1, with significance thresholds calibrated using null simulations (see Section 3.2.4). We compared against linear regression of the ARG of chromosome 1 and linear regression of SNP array variants from chromosome 1.

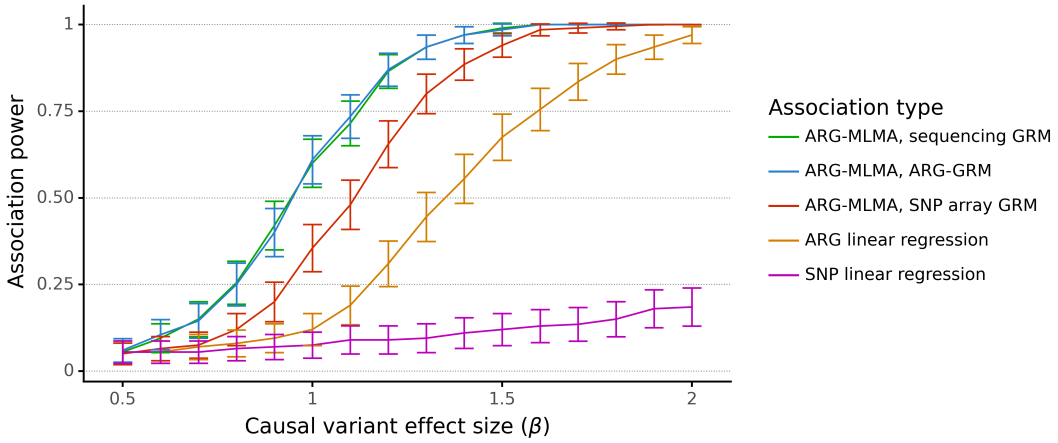


Figure 3.3: **Performance of ARG-MLMA using ground-truth ARGs and ARG-GRMs in simulations.** Power to detect a low-frequency causal variant (MAF = 0.25%) in simulations of a polygenic phenotype. We compare ARG-MLMA of ground-truth ARGs with linear regression of ground-truth ARG branches and linear regression of SNP array variants as we vary the effect size β (200 independent simulations of $h^2 = 0.8, \alpha = -0.25$, $N = 2{,}000$ haploid samples, and 22 chromosomes of 0.5 Mb each, see Section 3.2.4). For ARG-MLMA, we include three types of LOCO GRMs, built from sequencing variants, ground-truth ARGs, and SNP array variants on the other 21 chromosomes. Error bars represent 2 s.e.

Across all values of β, ARG-MLMA using a SNP array GRM (plotted in red) outperformed linear regression of ARG variants and of SNP array variants (Fig. 3.3). Because the causal variant on chromosome 1 is selected from those with MAF = 0.25%, SNP array variants struggle to tag the causal variation, and SNP linear regression (plotted in light purple) therefore performs poorly. Linear regression of ARG variants offers an improvement, as the ARG contains all possible variants, including the true causal variant. However, due to the large number of tests performed, one for each possible variant in the ARG, linear regression using the ARG (plotted in yellow) does not always yield a significant result, especially for small values of β. If instead a SNP

GRM is included, polygenic effects from the other 21 chromosomes can be modelled and accounted for. This approach of ARG-MLMA improves upon classical ARG association by modelling polygenicity, making it easier to detect the causal effect on chromosome 1. The ARG-MLMA and classical ARG association methods both test the same variants for association, and thus obtain similar significance thresholds, which are calculated using null model simulations. This threshold is much stricter than that of linear regression of SNP array variants. Despite this stricter threshold, the ARG-based methods benefit from capturing the true causal variant, while SNP array association must tag this variant using LD.

3.3.5 ARG-MLMA and ARG-GRMs combine to improve association power

ARG-MLMA can be further improved by using an ARG-GRM to better model polygenicity. In Fig. 3.3, we also include power as a function of β for ARG-MLMA using a sequencing GRM (plotted in green) and an ARG-GRM (plotted in blue), with both being built from chromosomes 2-22. All GRMs are built using the true value of $\alpha = -0.25$, which is a realistic value for human phenotypes [Speed et al., 2017]. We observe that ARG-MLMA using the ARG-GRM performs as well as ARG-MLMA using the sequencing GRM, and outperforms ARG-MLMA using the SNP array GRM. These findings build off our results in Fig. 3.2b,c,e,f, which demonstrated improved polygenic prediction and mixed-model association when using an ARG-GRM compared to SNP array GRMs.

3.4 Discussion

In this chapter, we demonstrated two novel approaches that extend state-of-the-art mixed model methods to leverage the rich information contained in the ARG. Our ARG-MLMA methodology is suitable for association analyses, while our ARG-GRM approach can be applied to perform heritability estimation, polygenic prediction,

and mixed-model association. We introduced novel computational methods such as an efficient Monte Carlo estimator for ARG-GRMs and ARG-based MAF-stratified, multi-GRM analyses.

In the context of association, we showed that ARG-MLMA and ARG-GRM can be combined to greatly improve association power compared to linear regression of SNP array variants. In this setting, the increases in power come from three streams. First, by testing variants of the ARG rather than SNP array variants, we no longer rely on SNPs to tag low-frequency or rare causal variation, and can instead observe these variants directly. Second, by modelling polygenicity from other chromosomes using the LMM, we condition away effects that are not explained by the chromosome of interest. Third, by using an ARG-GRM instead of a SNP array GRM, we capture more of the polygenic variance due to the other chromosomes. Each of these streams increases power on its own, and they provide additional improvements when used together.

Our simulations sought to incorporate many aspects of genotyping data and complex traits that would be found in large modern biobanks. These include representative constant mutation and recombination rates, a CEU (European) demographic model, SNP array density matching that of the UK Biobank array, and realistic values of h^2 and α. However, additional modelling extensions could also be considered, such as sparse non-infinitesimal genetic architectures. Although our simulations went up to a maximum scale of $N = 10{,}000$ samples due to the many replicates that were run, our ratio of N to genome length L was chosen to match a genome-wide analysis comprising $N \approx 6 \times 10^5$ samples (see Section 3.2.3). Previous works have derived that mixed-model prediction r^2 stays roughly constant if an increase in N is met by a proportional increase in L [Daetwyler et al., 2008, Wray et al., 2013], a result we observed in exploratory analyses as well. Furthermore, mixed-model association improves power compared to linear regression in a way that depends on in-sample prediction r^2 [Loh et al., 2015b]. Therefore, we can expect our ARG-GRM results for polygenic prediction and mixed-model association to roughly translate to a genome-wide analysis with $N \approx 6 \times 10^5$ samples. Future work can increase the scale of these

simulations to verify how the ARG-GRM and ARG-MLMA methods perform as a function of N and L.

The results in this chapter represent an upper bound in the performance of ARG-GRMs and ARG-MLMA assuming true ARGs are available. Notably, we only require node and edge information in the ARGs, without mutation information. We showed that ARG-GRMs are able to perform as well without mutation information as GRMs built from sequencing data, which can be thought of as ARGs with mutation information. For ARG-MLMA, we expect that knowing the true mutations in addition to the true ARG would enable association of the resulting sequencing data, which would yield fewer tested variants, a less stringent significant threshold, and therefore increased power to detect association. However, we still observed an excellent upper bound if true ARGs are known. In applications where ARGs are inferred from data, the performance of both ARG-GRMs and ARG-MLMA will be lower than these upper bounds. We will therefore revisit these complex trait analyses in the context of inferred ARGs in Chapters 5 and 6, after introducing two new methods for ARG inference in the next chapter.

Chapter 4

A Scalable Method for Inferring Genome-Wide Genealogies from Array or Sequencing Data

4.1 Overview

In the previous chapter, we showed that the ARG of a set of samples, if known, can improve association power and modelling of relatedness compared to utilising array datasets. It is natural to ask how these methods would perform if they instead used ARGs inferred from genetic data. Large biobanks have generated some of the most exciting findings in medical genetics in the past few years, with genotyping arrays representing the typical point of entry for genetic measurement [Chen et al., 2011, Nagai et al., 2017, Bycroft et al., 2018, Kurki et al., 2022]. A scalable algorithm for inferring ARGs from array biobanks would enable ARG-based complex trait analysis to be applied in these data sets. It is also desirable that the inference algorithm output branch lengths, which are leveraged by our ARG-GRM and ARG-MLMA methods. More broadly, being able to construct ARGs from large array or sequencing data sets would enable a variety of applications in statistical genetics, reviewed in Section 2.2.3. To meet these needs, in this chapter we introduce **ARG-Needle**, an ARG inference

method we developed which (1) can accommodate SNP array data, (2) scales to biobank-scale sample sizes, and (3) provides branch length estimates in addition to topologies.[1]

ARG-Needle infers the ARG for large genotyping array or sequencing data sets by iteratively "threading" [Rasmussen et al., 2014] one haploid sample at a time to an existing ARG, as depicted in Fig. 4.1a for a single marginal coalescent tree. Given an existing ARG, which we initialise to contain a single sample, we randomly select the next sample to be added (or threaded) to the ARG. We then compute a *threading instruction*, which at each genomic position provides the index of a sample in the ARG that is estimated to be most closely related to the target sample, as well as their time to most recent common ancestor (TMRCA) (see Fig. 4.1b). We use this threading instruction to add the target sample to the ARG and continue iterating these steps until all samples are in the ARG.

To compute the threading instruction of a sample, ARG-Needle performs genotype hashing [Gusev et al., 2009, Nait Saada et al., 2020] to rapidly detect a subset of candidate closest relatives within the ARG, then uses the ASMC algorithm [Palamara et al., 2018] to compute pairwise TMRCA values and to select the closest sample at each genomic position (see Fig. 4.1c). ARG-Needle is guaranteed to recover the true ARG if the closest relatives and pairwise TMRCA values are correctly inferred (see Section 4.3.2). When all samples have been threaded to the ARG, ARG-Needle uses a fast post-processing step, which we call ARG normalisation, that applies a monotonic correction to the node times of the ARG while leaving the topology unchanged. ARG-Needle builds the ARG in time approximately linear in sample size, depending on hashing parameters (see Section 4.2.2.2).

We also introduce an extension of ASMC [Palamara et al., 2018], called **ASMC-clust**, that builds genome-wide genealogies by forming a tree at each site using hierarchical clustering on pairwise TMRCAs output by ASMC. This approach scales quadratically with sample size, but the joint modelling of all pairs of samples may improve accuracy in certain scenarios. ASMC-clust also includes ARG normalisation

[1]The remainder of this chapter is expanded and adapted from [Zhang et al., 2021].

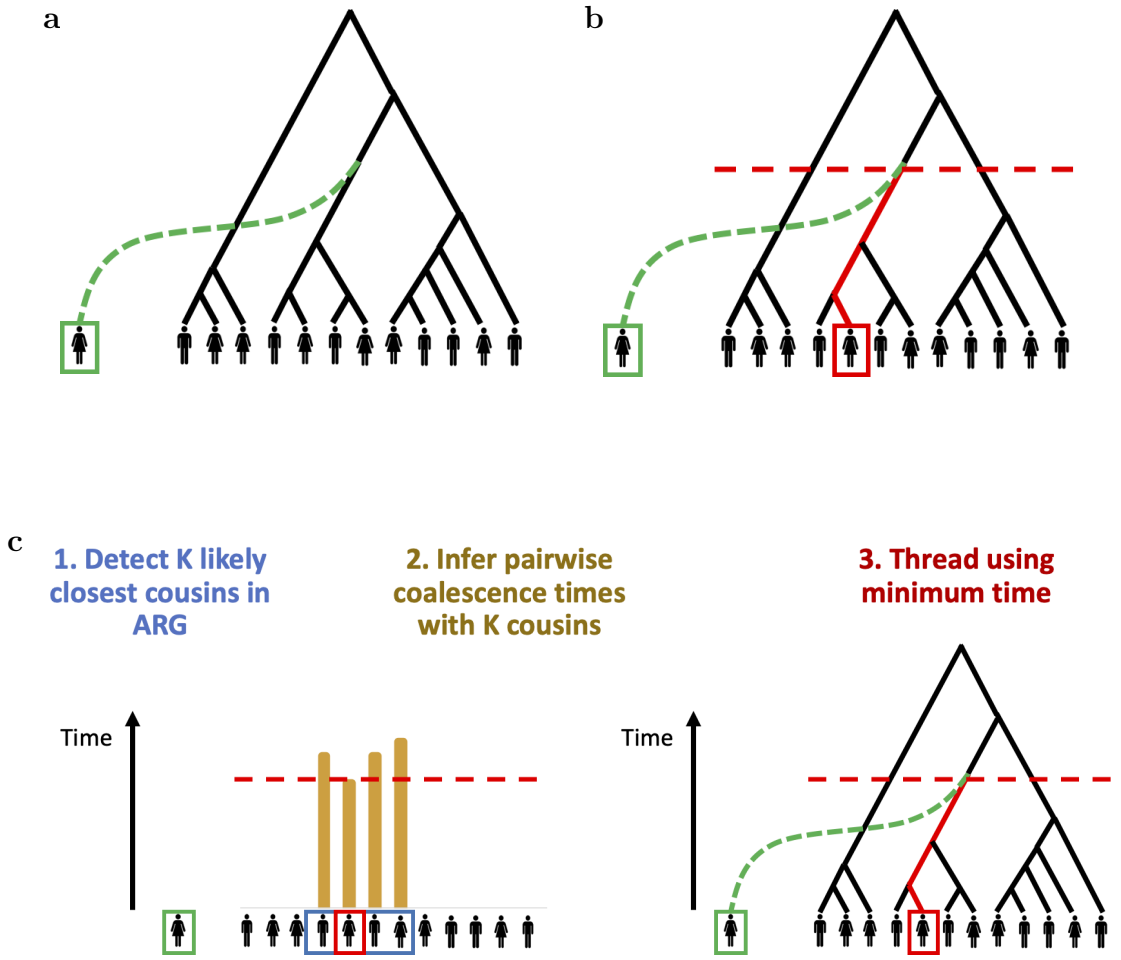


Figure 4.1: **Illustration of ARG-Needle threading for one marginal coalescent tree. a**. ARG-Needle iteratively constructs an ARG by "threading" [Rasmussen et al., 2014] one haploid sample at a time to an existing ARG. Here the threading is shown in green, focusing on a single tree within the ARG. **b**. To thread the new sample, we specify a sample already in the ARG and a time in the past, shown here in red. The new sample is then connected to the ARG along the ancestral lineage of the specified sample at the specified time. **c**. To compute the threading instruction, we first use genotype hashing to select K genetically similar "cousins" in the ARG, shown in blue. We then estimate pairwise coalescence times between the sample to be threaded and each of these K cousins, shown in yellow. The cousin with minimum coalescence time is selected and threaded to at the minimum time. For a more detailed illustration showing how the threading instruction can vary as a function of genomic position, see Fig. 4.2.

as a post-processing step.

4.2 Methods

Both ARG-Needle and ASMC-clust leverage output from the ASMC algorithm [Palamara et al., 2018], which takes as input a pair of genotyping array or sequencing samples and outputs a posterior distribution of the time to most recent common ancestor (TMRCA) across the genome. These pairwise TMRCAs are equivalent to an ARG between two samples, which ARG-Needle and ASMC-clust use to assemble the ARG for all individuals.

4.2.1 ASMC-clust algorithm

ASMC-clust runs ASMC on all pairs of samples and performs hierarchical clustering of TMRCA matrices to obtain an ARG. At every site, we apply the **unweighted pair group method with arithmetic mean (UPGMA)** clustering algorithm [Sneath and Sokal, 1973] on the $N \times N$ posterior mean TMRCA matrix to yield a marginal tree. We combine these marginal trees into an ARG, using the midpoints between sites' physical positions to decide when one tree ends and the next begins. By using an $O(N^2)$ implementation of UPGMA [Gronau and Moran, 2007, Müllner, 2013], we achieve a total runtime and memory complexity of $O(N^2M)$, assuming N samples and M sites.

We also implemented an option for ASMC-clust to instead use $O(N^2 + NM)$ memory, at the cost of more runtime. In the default algorithm, we run ASMC on all $\binom{N}{2}$ pairs of samples across M sites, and store the results in an $\binom{N}{2}$ by M matrix. We then iterate through the TMRCA values at each of the M sites and perform UPGMA hierarchical clustering. The resulting ARG consists of M trees, each over N nodes, which can be represented in $O(NM)$ memory. Therefore, the memory bottleneck comes from the storage of the TMRCA results, which takes up $O(N^2M)$ memory.

To reduce the memory required from storing the TMRCA results, we iterate through groups of M_{max} sites, each time computing and storing the TMRCA re-

sults for those sites, performing UPGMA clustering, and saving the clustered trees. This results in $O(N^2 M_{max} + NM)$ memory usage. As described earlier, however, we pad additional sites (by default, 1 cM) on either end of the region to provide the ASMC HMM with some context. Using this approach, each site will therefore be processed multiple times by ASMC, leading to a time complexity somewhere between $O(N^2 M)$ and $O(N^2 M^2)$, depending on the choice of M_{max} and the number of SNPs provided as context in each ASMC run. In our simulations we fixed $M_{max} = 300$ for SNP data and $M_{max} = 2000$ for sequencing data. This led to memory usage that scaled quadratically as a function of N.

4.2.2 ARG-Needle algorithm

4.2.2.1 Overview of ARG-Needle three steps

ARG-Needle starts with an empty ARG and repeats three steps to add additional samples to the ARG: (1) detecting a set of closest genetic relatives via hashing, (2) running ASMC, and (3) "threading" the new sample into the ARG (Figs. 4.1-4.2). Given a new sample, the first step performs a series of hash table queries to determine the candidate closest samples already in the ARG [Gusev et al., 2009]. We divide up the sites present in the genetic data into non-overlapping "words" of S sites and store hash tables mapping from the possible values of the ith word to the samples that carry that word. We use this approach to rapidly detect samples already in the ARG that share words with the target sample and return the top K samples with the most consecutive matches. A tolerance parameter T controls the number of mismatches allowed in an otherwise consecutive stretch. We also allow the top K samples to vary across the genome due to recombination events, by partitioning the genome into regions of genetic distance L. Assuming this results in R regions, the hashing step outputs a matrix of $R \times K$ sample IDs containing the predicted top K related samples over each region.

The sample IDs output by the first step inform the second step of ARG-Needle, in which ASMC is run over pairs of samples. In each of the R regions, ASMC computes

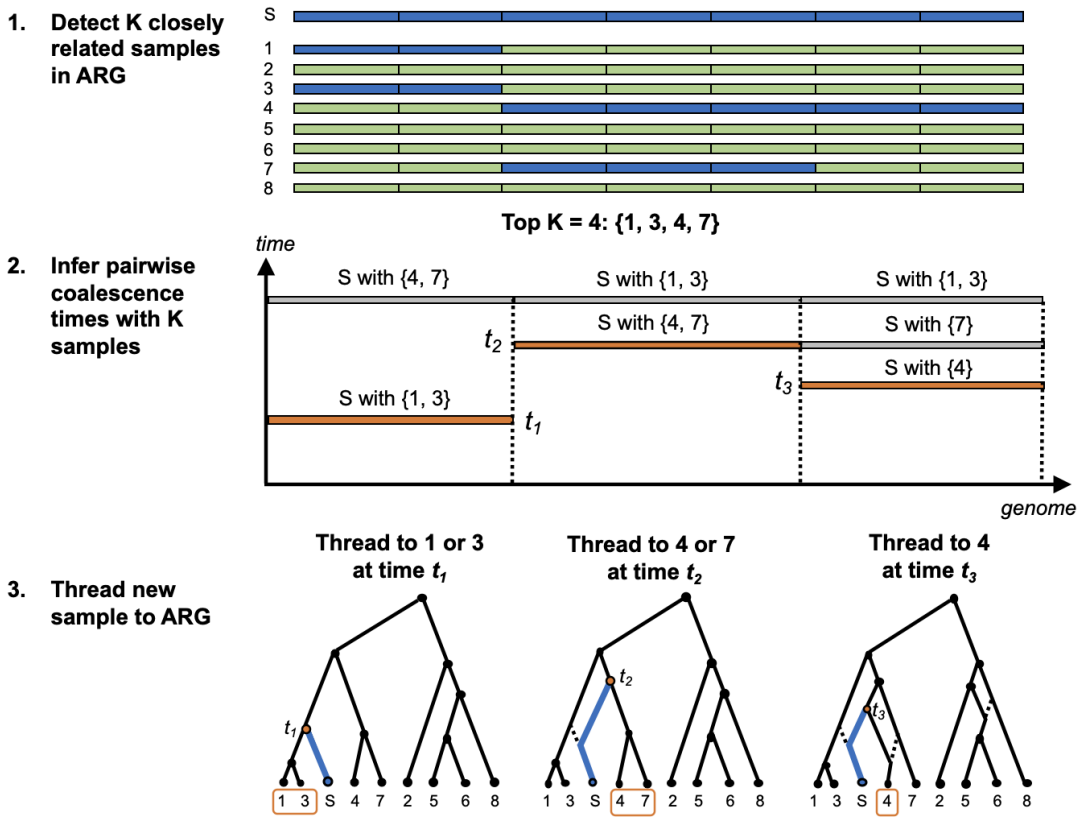


Figure 4.2: **Detailed overview of the ARG-Needle algorithm.** ARG-Needle adds one haploid sample at a time to an existing ARG, each time performing three steps: 1. shortlisting a subset of most related samples already in the ARG through genotype hashing, 2. obtaining pairwise coalescence time estimates with these samples using ASMC [Palamara et al., 2018], and 3. using the ASMC output to "thread" [Rasmussen et al., 2014] the new sample to the ARG. We depict an example of adding sample S to an ARG, focusing on one genomic region. Step 1 divides the genome into "words" and checks for identical matches with sample S. Based on these matches (shown in blue), samples 1, 3, 4, and 7 are output as the $K = 4$ candidate most related samples already in the ARG. Step 2 computes pairwise coalescence time estimates between sample S and each of the samples 1, 3, 4, and 7. The minimum time for each position is highlighted. Step 3 uses these minimum times and samples to define a "threading instruction" that is performed to add sample S to the ARG. Threading connects the new sample to the ancestral lineage of each chosen sample at the chosen time. Dotted lines indicate past ARG edges that are inactive due to recombination. When all samples have been threaded, ARG-Needle performs a final post-processing step called ARG normalisation (see Section 4.2.3).

the posterior mean and maximum a posteriori (MAP) TMRCA between the sample being threaded and each of the K candidate most related samples. We add up to 1.0 cM on either side of the region, to provide additional context for the ASMC model.

In the third step, ARG-Needle finds the minimum posterior mean TMRCA among the K candidates at each site of the genome. The corresponding IDs determine which sample in the ARG to thread to at each site. Because the posterior mean assumes continuous values and changes at each site, we average the posterior mean over neighboring sites where the ID to thread to and the associated MAP remain constant. This produces piecewise constant values which determine how high above the sample to thread, with changes corresponding to inferred recombination events. The sample is then efficiently threaded into the existing ARG, utilising custom data structures and algorithms.

Of the three steps of ARG-Needle, the second step (ASMC) dominates runtime for small N and the first step (hashing) dominates runtime for large N. The parameters K, L, and S each carry an accuracy versus runtime tradeoff, with large K, small L, and small S leading to a more accurate but slower algorithm. We recommend default parameters of $K = 64$ and $L = 0.5$ cM for array data and $K = 64$ and $L = 0.1$ cM for sequencing data. In experiments with simulated genotypes (see Chapter 5), we fixed the hash word size at $S = 16$. As threading proceeds, increasingly close relationships and thus increasingly long shared haplotypes are detected between the sample and other individuals in the ARG. In real data analysis (see Chapter 6) we therefore increase S as threading proceeds, which reduces computational cost without a significant loss in accuracy. We also implemented a larger primary hash word size S_1, leveraged for increased speed, and a smaller "backup hash word size" S_2, used for samples where more fine-grained hashing was needed (see Section 4.2.2.2). We set $S_1 = 16$ and $S_2 = 8$ when threading the first 50K individuals and set $S_1 = 64$ and $S_2 = 16$ for the remaining 287K individuals. We set $T = 1$ in simulations and real data analyses to enable robustness to a single genotyping error or recent mutation event in otherwise closely related samples.

In summary, ARG-Needle starts with an empty ARG, and iteratively adds one

new haploid sample at a time to the ARG. Adding the first sample generates a single ARG node at the present time spanning the whole genomic region. For each additional sample, three steps are used to generate a threading instruction and add the new sample to the ARG: (1) perform hash table queries; (2) run ASMC [Palamara et al., 2018] to form the threading instruction; (3) thread the new sample to the ARG. We now describe each of these three steps in greater detail.

4.2.2.2 ARG-Needle step 1: shortlisting of closest relatives via genotype hashing

Let the haploid data consist of N samples and M sites. Let the genetic distances of the sites be $g_1, \ldots, g_M$, and let the genotypes be $x_{ij} \in \{0, 1\}$, for $1 \leq i \leq N$ and $1 \leq j \leq M$. The parameters of the hash table operations are a hash word size S, a hash region size L, a hash tolerance T, and a hashing output size K. The units of the hash region size L are in genetic distance, while all other parameters are integers. S and L are used to partition the sites into words and regions. S simply denotes the number of sites in each word. For $1 \leq k \leq \lceil M/S \rceil$, word k consists of sites $a(k) = (k-1) \times S + 1$ to $b(k) = \max(k \times S, M)$ inclusive, resulting in a total of $W = \lceil M/S \rceil$ words. For $1 \leq i \leq N$ and $1 \leq j \leq W$, let w_{ik} be the integer resulting from interpreting the W bits of the kth word for sample i as a binary number, i.e. $w_{ik} = (x_{ib(k)} \ldots x_{ia(k)})_2$.

The words are further grouped into non-overlapping regions such that each region spans a genetic distance of at least L, with the exception of when $L > g_M - g_1$, in which case all words are placed in the same region. Starting with word 1, we find the first index k such that $g_{b(k)} - g_1 \geq L$. This makes up the first region, spanning words 1 to k inclusive. For the second region, we start at word $k+1$ and find the first index k' such that $g_{b(k')} - g_{a(k)} \geq L$. The second region then spans words $k+1$ to k'. We continue in a similar fashion. When we reach the end, if we perfectly have just formed a last region covering all words, the algorithm terminates. Otherwise, we combine any remaining words on to the previous region so that all regions are of genetic distance approximately or just over L. Let the number of regions resulting

from this procedure be R, and for $1 \leq r \leq R$, let $c(r)$ and $d(r)$ be the start and end words, inclusive, $1 \leq c(r) \leq d(r) \leq W$.

The hashing query takes a new sample in input and outputs the top K candidate closest relatives to this sample out of those already in the ARG, for each of the R regions. The purpose of having the hash region size L be in genetic distance is to control for the expected number of recombination events within each region. Recombination will alter the set of closest relatives, but by keeping L sufficiently low it is possible to obtain closest relatives that remain consistent through the region. Within each region, the hashing query selects the top K candidate closest relatives by a score which we call the T-tolerant IBS score (for identical-by-state). Given two samples i and i' and a hash tolerance T, we say that i and i' are T-tolerant IBS over $[f, g]$, $1 \leq f \leq g \leq W$, if

$$|\{f \leq k \leq g | w_{ik} \neq w_{i'k}\}| \leq T.$$

The T-tolerant IBS score of samples i and i' over a region r is the number of words in the longest T-tolerant IBS stretch with at least one word of overlap with region r. In notation, we consider the set

$$\{[f, g] | 1 \leq f \leq g \leq W, f \leq d(r), g \geq c(r), |\{f \leq k \leq g | w_{ik} \neq w_{i'k}\}| \leq T\}$$

and take the maximum value of $g - f + 1$ over this set. For each region r, we then take the top K samples with the largest T-tolerant IBS score over this region. In the case of ties, we select the sample(s) that have been added to the ARG more recently, though it is also possible to break ties randomly. If the number of samples with a nonzero score is less than K, only this many results are returned, or in other words, samples scoring zero are not returned. In summary, the hashing query returns a list of at most K top-scoring sample IDs for each of the R regions.

Our hashing data structure consists of a vector of W hash tables, one for each of the W words. The kth hash table maps from possible values for the kth word, guaranteed to be a number between 0 and $2^S - 1$ inclusive, to a vector of sample

IDs which contain that value for the kth word. Every time a sample i' is added to the ARG, we perform $O(W)$ operations to update this data structure. For each k, $1 \leq k \leq W$, we access the entry in the kth hash table at $w_{i'k}$, initializing an empty vector if no entry exists, and appending i' to this vector. Similarly, when we want to query a new sample i', we can index $w_{i'k}$ into the kth hash table to find the samples which match sample i' at the kth word. In this way, we only need to visit the values (i, k) for which there is a match, $w_{ik} = w_{i'k}$.

Suppose there are N' samples already in the ARG, and we want to query the closest relatives for a new sample i'. Our hashing-based implementation runs in $O\left(\overline{N}_{overlap}W\right)$, where $0 \leq \overline{N}_{overlap} \leq N'$ is the average number of samples already in the ARG which match sample i' per word, i.e. the average size of $\{0 \leq i \leq N' | w_{ik} = w_{i'k}\}$ as k varies from 1 to W. As one increases the word size S, $W = \lceil M/S \rceil$ decreases and $\overline{N}_{overlap}$ also decreases because longer words are less likely to be shared. The pattern of this latter trend as a function of S is data set-specific due to factors such as the demographic history of the population, but in summary, S can be used as a parameter to increase hashing speed at the cost of coarser hashing-based IBS detection.

In building ARGs on 337K UK Biobank samples, the hashing step dominates runtime. We implemented a form of dynamic hashing that uses a primary hash word size S_1 and falls back to a backup hash word size S_2 in some cases, with $S_1 > S_2$. The criterion for falling back to the backup hash word size is parameterised by a factor F. For each region, we compute the sum of the top K T-tolerant IBS scores in that region, and check that this is greater than F times the number of words in the region (in notation, $F \times (d(r) - c(r) + 1)$). If the check fails for any of the regions, then we fall back to the backup hash word size, repeating the entire hash query with S_2 instead of S_1. This has the effect of performing the initial hashing queries with S_2, then transitioning to more and more queries done with S_1 as the ARG-Needle algorithm proceeds. For our run on 337K diploid individuals, we started with $S_1 = 16, S_2 = 8, F = 4$ for adding the first 50K individuals, then progressed to $S_1 = 64, S_2 = 16, F = 8$ for the remaining individuals. Other parameters were set to

$T = 1$, $K = 64$, and $L = 0.5$ cM.

4.2.2.3 ARG-Needle step 2: ASMC queries

In step 2 of the ARG-Needle algorithm, the candidate closest relatives output by hashing are validated using ASMC to yield a more certain estimate of true closest relatives and their TMRCA to the target sample. Each pairwise ASMC comparison takes $O(MD)$ time to run, where D is the number of discretised time bins used to represent the posterior distribution of pairwise coalescence time (we use $D = 69$, the default for ASMC v1.0). With no hashing, each new sample being threaded would need to be compared against all existing samples in the ARG using ASMC, yielding $O(N^2MD)$ runtime overall. By only selecting K candidate closest relatives for each sample being threaded, the total runtime is instead linear in N. Step 1 of ARG-Needle thus functions as a heuristic pre-selection strategy, similar to the use of genotype hashing or the positional Burrows-Wheeler transform [Durbin, 2014] to speed up genotype imputation [Browning et al., 2018, Delaneau et al., 2019, Rubinacci et al., 2020], phasing [Loh et al., 2016], and identity-by-descent (IBD) detection [Nait Saada et al., 2020].

For each of the R regions (partitioned based on genetic distance, see above), step 1 of ARG-Needle returns up to K IDs representing the candidate closest relatives to this sample within the region. Using our earlier definitions, region r contains words with indices from $c(r)$ to $d(r)$ inclusive, with $1 \leq c(r) \leq d(r) \leq W$, which means it contains SNPs with indices from $a(c(r))$ to $b(d(r))$ inclusive, $1 \leq a(c(r)) \leq b(d(r)) \leq M$. We use ASMC (run in array or sequencing mode, depending on the experiment) to predict the pairwise coalescence time posterior distribution between the new sample and each of the K samples at each of these SNP positions, resulting in an output of size $K \times (b(d(r)) - a(c(r)) + 1) \times D$. Because the ASMC Hidden Markov Model (HMM) takes into account sequential context when making a prediction, we pad the input region by adding 1 cM on either side. This increases accuracy at the cost of slightly increased runtime. The ASMC output for each pair and site is a posterior distribution, given as probabilities over the D discretised time bins. We use it to

extract the posterior mean and the mode, or maximum a posteriori (MAP) values. This does not affect the runtime complexity, as for each site and pair computing these summaries takes $O(D)$ time. Before proceeding to step 3, we aggregate these summaries so that at each site we have at most K triplets, which represent the ID of the candidate closest relative, the posterior mean of the TMRCA with the target sample, and the MAP.

4.2.2.4 ARG-Needle Step 3: processing the ASMC output and performing threading

In the final step of the ARG-Needle algorithm, we use the ASMC output to add the new sample to the ARG. This takes place using a "threading" operation, named after the terminology coined by ARGweaver [Rasmussen et al., 2014]. We first describe the ARG-Needle threading operation, then explain how we perform threading using the ASMC output.

We define a "threading instruction" to contain the sample ID of a closest relative and their TMRCA to the target sample, at each position along the genome. If we assume a genomic extent $[s, t)$ and sample IDs $\{1, \ldots, N'\}$ already in the ARG, a threading instruction may be represented as a function $f : [s, t) \to \{1, \ldots, N'\} \times (0, \infty)$, assigning to each position x a pair $f(x) = (i(x), T(x))$, where $i(x)$ is a sample ID and $T(x) > 0$ is a time.

We now describe the threading operation in terms of how it affects each marginal tree. (Our actual implementation considers threading as an ARG-wide problem and is more efficient, see Section 4.3.1; we take this perspective for simplicity.) Consider a position x with a marginal tree over N' samples, and suppose the threading instruction has $f(x) = (i, T)$. First, a new node u for sample $N' + 1$ is created. Next, let v be the oldest ancestor node of sample i with a time less than or equal to T. One of three conditions holds:

1. If v has time less than T and v has a parent node v' (necessarily with time greater than T), create a new node w with time T. Delete the edge between v and v' and add three new edges between pairs (u, w), (v, w), and (w, v').

2. If v has time less than T but does not have a parent node, then v is the root node of the marginal tree at x. Create a new node w with time T and add two new edges between pairs $(u,\ w)$ and $(v,\ w)$. (Node w becomes the new root node.)

3. If v has time T, then create an edge going from u to v.

These three operations describe how to perform threading on marginal trees. The third operation creates polytomies: nodes with more than two children at a position. We assume this operation exists in our theoretical results (see Section 4.3.2), but in practice, it is extremely unlikely as the time must perfectly match. Our inferred ARGs always choose a different time and never contain polytomies.

Given the output of step 2, we generate a threading instruction as follows. At each site, we select the sample $i(x)$ to be the one with smallest ASMC posterior mean TMRCA among the K candidate closest relatives (ties are arbitrarily broken). To compute the time $T(x)$, we select regions where both $i(x)$ and the TMRCA MAP with $i(x)$ remain constant, and compute the average posterior mean TMRCA to $i(x)$ within each such region. This has several benefits over using the raw value of either the MAP or the posterior mean at each site. Although changes in the MAP better reflect IBD segment breakpoints, the MAP output by ASMC only takes one of D possible values, which would lead to a large number of polytomies in the ARG. The MAP also has a higher RMSE for the TMRCA compared to the posterior mean. The posterior mean, on the other hand, takes different values at each site, which would reduce haplotype sharing across marginal trees and produce a less compact ARG (as in the case of ASMC-clust). By combining the MAP and posterior mean in the threading instruction we thus obtain better estimates for the location of recombination breakpoints without sacrificing accuracy or compactness of the inferred ARG.

4.2.3 ARG normalisation

Although the posterior mean has a lower RMSE than the MAP in estimating TMRCAs, it tends to be more biased towards the average coalescent time induced by the

demographic prior, particularly in sparser array data. For this reason, we observed that ARGs inferred using the above definition of threading instruction tend to overestimate the time of recent coalescence events and underestimate the height of root nodes. We thus developed a procedure, which we call ARG normalisation, to further leverage the demographic prior to reduce the bias due to the use of posterior mean estimates, while preserving the inferred ordering of coalescent events (see Fig. 5.6). ARG normalisation performs a quantile normalisation of the heights of inferred ARG nodes, rescaling them to match the quantiles observed in 1,000 independent trees sampled from the demographic prior (after accounting for the span of inferred ARG nodes).

In more detail, assuming the 1,000 simulations generate Q non-leaf nodes with times that can be ordered increasingly from t_1 to t_Q, we compute quantiles from the simulated trees by assigning quantile $(2i-1)/(2Q)$ to time t_i. We also assign quantile 0 to time 0 and quantile 1 to time $1.05 \times t_Q$ to ensure that the mapping is strictly increasing. Given an inferred ARG, we sought to compute an analogous quantile distribution of node times that was sensitive to differences between nodes spanning short vs. long segments of ancestral material and sensitive to polytomies. For each edge in the inferred ARG, we recorded the time of its parent node and the distance in base pairs it spanned. We aggregated these time-distance pairs to obtain Q' distinct parent node times, ordered increasingly from t_1 to $t_{Q'}$, and corresponding aggregate distances spanned by edges with that parent node time, d_1 to $d_{Q'}$. We assigned quantile value $(d_1 + d_2 + \ldots + d_{i-1} + d_i/2)/(d_1 + \ldots + d_{Q'})$ to time t_i. We then matched the inferred node time quantile distribution with the target node time quantile distribution using linear interpolation on the quantiles. We used this mapping to rewrite all nodes of time t_i to the new corresponding time.

4.3 Results

4.3.1 ARG-Needle implementation

ARG-Needle is coded in C++ and Python. In an ARG-Needle ARG, nodes represent ancestral haplotypes, storing metadata such as their age (in generations). Edges connect pairs of nodes implying ancestor/descendant relationships, and store start and end coordinates for the inherited chromosomal region. Access to a node's ancestor or children at a particular position requires $O(\log K)$ computation, where K is the total number of ancestors or children proceeding from the node.

Each of the three steps of ARG-Needle involves a C++ routine to speed up the key computational steps. For step 1, genotyping hashing is performed using hash tables provided in the `std::unordered_map` data structure. For step 2, we leverage the ASMC software, written in C++ with Python bindings, to query pairwise TMRCAs. ASMC utilises AVX operations such that requesting a batch of TMRCAs can be parallelised. For step 3, we have implemented the threading procedure as a set of efficient transformations to the ARG data structure. Every threading operation can be expressed via creating new nodes and edges and rewriting the existing graph. The number of new nodes created is equal to the number of intervals in the threading instruction with distinct closest relative and/or TMRCA. Because of the efficiency of the ARG representation, step 3 in aggregate represents a negligible part of runtime compared to steps 1 and 2.

The ARG-Needle inference algorithm is part of a wider package we are developing for manipulating ARGs, including various algorithms for complex trait analysis (see Chapter 3). ARG-Needle ARGs can be imported from and exported to `tskit` format. We plan to release this package in the near future.

4.3.2 Theoretical guarantees of ARG-Needle and ASMC-clust algorithms

4.3.2.1 Setting of correct pairwise coalescence times

In this section, we will prove the following theorem regarding the threading instructions used by step 3 of ARG-Needle.

Theorem. Let $\mathcal{A}$ be an ARG over N samples, and let $\mathcal{B}$ be an ARG constructed via threading from an initially trivial ARG with one sample. Suppose that for threading sample N', one uses the threading instruction $f(x) = (i(x), T(x))$ consisting of $i(x) = \text{argmin}_{1 \le i' < N'}(\text{tmrca}_{\mathcal{A},x}(N', i'))$ and $T(x) = \min_{1 \le i' < N'}(\text{tmrca}_{\mathcal{A},x}(N', i'))$. Then at each x, the marginal trees of $\mathcal{A}$ and $\mathcal{B}$ will be identical (not considering the possible creation or deletion of unary nodes).

Applying the theorem, we can think of $\mathcal{A}$ as the true ARG and $\mathcal{B}$ as the ARG inferred by ARG-Needle (omitting ARG normalisation). The theorem assumes that for every sample being threaded, we thread to its "closest cousin" in the true ARG $\mathcal{A}$, given by $i(x)$, using the true TMRCA $T(x)$ to this cousin.[2] The theorem then states that the resulting "inferred ARG" will contain the same marginal trees as the true ARG. Essentially, the theorem says that out of all possible samples to thread to, one should thread to the sample of minimum TMRCA using that minimum time, which is exactly what we do in step 3.

ARG-Needle step 3 selects the sample to thread to by taking the minimum of the inferred TMRCAs from step 2. Therefore, if (for each position) step 1 finds and returns a sample that is a true "closest cousin" of the sample being threaded, and if the posterior mean TMRCAs from step 2 are accurate (for each position), then the conditions surrounding $i(x)$ in the theorem will be satisified. The conditions surrounding $T(x)$ will be satisfied if the posterior mean TMRCAs are correct and the MAP TMRCA changes along with this value, with the second condition due to the averaging we perform. In other words, if correct pairwise coalescence times and closest cousins are returned by steps 1 and 2, the ARG-Needle threading procedure

[2] In the case of multiple closest cousins, we can select any.

outlined in step 3 leads to recovering the true marginal trees.

To prove this theorem, we first establish some lemmas. As in our earlier description, consider a position x with a marginal tree over N' samples, and suppose the threading instruction for sample $N'+1$ has $f(x) = (i, T)$, joining to sample i at time T. Let $\text{tmrca}(a, b)$ denote the pairwise TMRCA between samples a and b (position x is implicit). We begin with two trivial observations:

Claim 1. For samples $1 \leq a < b \leq N'$ (samples already in the ARG), $\text{tmrca}(a, b)$ is the same before and after the above threading operation.

Claim 2. After the threading operation, $\text{tmrca}(N'+1, i) = T$.

The remaining pairwise TMRCAs after threading can also be determined:

Claim 3. Let j be any sample in the ARG other than i. After the threading operation, we have

$$\text{tmrca}(N'+1, j) = \max\left(T, \text{tmrca}(i, j)\right).$$

Proof of Claim 3. Let v be the pairwise MRCA node of samples i and j, and w be the pairwise MRCA node of samples $N'+1$ and i after threading (having time T). There are now three cases:

1. If $\text{tmrca}(i, j) < T$, then v lies below w, and there is a path from the node for sample i to w which must pass through v. Therefore w is the MRCA node of j and $N'+1$, so $\text{tmrca}(N'+1, j) = T = \max\left(T, \text{tmrca}(i, j)\right)$.

2. If $\text{tmrca}(i, j) > T$, then w lies below v. The MRCA of w (a non-leaf node) and sample j is v, and sample $N'+1$ is a descendant of w, so $\text{tmrca}(N'+1, j) = \text{tmrca}(i, j) = \max\left(T, \text{tmrca}(i, j)\right)$.

3. If $\text{tmrca}(i, j) = T$, then when sample $N'+1$ was being threaded to the sample i, it would have found node v already at time T, leading to the second threading case above and adding a single edge between the node for sample $N'+1$ and v. Therefore $\text{tmrca}(N'+1, j) = T = \max\left(T, \text{tmrca}(i, j)\right)$.

□

Now we proceed with the proof of the theorem.

Proof of Theorem. It suffices to consider an arbitrary position x and show that the marginal trees of $\mathcal{A}$ and $\mathcal{B}$ are equivalent. We therefore rely on the fact that the full set of TMRCAs uniquely determines a rooted tree [Böcker and Dress, 1998], and show that the $\binom{N}{2}$ pairwise TMRCAs between any two samples are the same within $\mathcal{A}$ and $\mathcal{B}$. We proceed by induction, and assume that the pairwise TMRCAs between the first N' samples are the same in $\mathcal{A}$ and $\mathcal{B}$. The pairwise TMRCAs between $N'+1$ and the first N' samples are set when sample $N'+1$ is threaded to $\mathcal{B}$, and unchanged after (by Claim 1). Let i and T denote the chosen sample and time as defined above, then by Claim 2, $\text{tmrca}_{\mathcal{B}}(N'+1, i) = T = \text{tmrca}_{\mathcal{A}}(N'+1, i)$. It remains to show that $\text{tmrca}_{\mathcal{B}}(N'+1, j) = \text{tmrca}_{\mathcal{A}}(N'+1, j)$ for any $j \neq i$ among the first N' samples. We have

$$\begin{aligned}
\text{tmrca}_{\mathcal{B}}(N'+1, j) &= \max\left(T, \text{tmrca}_{\mathcal{B}}(i, j)\right) \\
&= \max\left(T, \text{tmrca}_{\mathcal{A}}(i, j)\right),
\end{aligned}$$

by Claim 3 and the inductive hypothesis. By the definition of T,

$$\begin{aligned}
\text{tmrca}_{\mathcal{B}}(N'+1, j) &= \max\left(T, \text{tmrca}_{\mathcal{A}}(i, j)\right) \\
&\leq T \\
&= \min_{1 \leq i' \leq N'} (\text{tmrca}_{\mathcal{A}}(N'+1, i')) \\
&\leq \text{tmrca}_{\mathcal{A}}(N'+1, j).
\end{aligned}$$

However, by the ultrametric property,

$$\begin{aligned}
\text{tmrca}_{\mathcal{B}}(N'+1, j) &= \max\left(T, \text{tmrca}_{\mathcal{A}}(i, j)\right) \\
&= \max\left(\text{tmrca}_{\mathcal{A}}(N'+1, i), \text{tmrca}_{\mathcal{A}}(i, j)\right) \\
&\geq \text{tmrca}_{\mathcal{A}}(N'+1, j).
\end{aligned}$$

Putting these together, $\text{tmrca}_{\mathcal{B}}(N'+1, j) = \text{tmrca}_{\mathcal{A}}(N'+1, j)$. □

4.3.2.2 Setting of ultrametric pairwise coalescence times

As mentioned in Section 4.2, ASMC outputs an ARG between two samples, which ARG-Needle and ASMC-clust extend to an ARG for all individuals. Consider the case in which the pairwise TMRCAs returned by ASMC are ultrametric for every position. It turns out that in this setting, ASMC-clust and ARG-Needle output the same ARG (assuming hashing is sufficiently accurate). Furthermore, the ARG output by ARG-Needle does not depend on the order in which samples are threaded.

To see why this is the case, first consider performing ASMC-clust using the TMRCAs to obtain an ARG $\mathcal{A}$. It is easily shown that performing UPGMA of an ultrametric distance matrix leads to perfect reconstruction of the pairwise distances. Therefore, the TMRCAs within $\mathcal{A}$ will match the TMRCAs as predicted by ASMC. Now, applying the earlier Theorem, we have that ARG-Needle inference using the ASMC TMRCAs, under sufficiently accurate hashing, will lead to reconstructing $\mathcal{A}$. Because the Theorem makes no reference to the order of threading, different orders of threading will all reconstruct $\mathcal{A}$. Hence ASMC-clust and ARG-Needle output identical ARGs[3] under this scenario.

This result provides intuition for the ARG-Needle algorithm as resembling ASMC-clust in terms of output, but more computationally efficient. Future analysis could quantify the discrepancy between ARG-Needle and ASMC-clust, or between ARG-Needle with different orders of threading, in terms of the degree to which ASMC TMRCAs diverge from being ultrametric.

4.4 Discussion

In this chapter, we introduced two ARG inference algorithms that leverage pairwise coalescence time estimation in genotyping array or sequencing data sets. ASMC-clust performs hierarchical clustering of pairwise TMRCA matrices, but scales quadratically in sample size. ARG-Needle scales faster than ASMC-clust through the use of a genotype hashing step to select closely related samples, such that the total amount

[3]In terms of branch lengths and topology.

of time spent on pairwise coalescent modelling is linear in sample size. We proved that if inferred pairwise coalescence times are ultrametric and the hashing step is accurate, ARG-Needle and ASMC-clust output ARGs with identical branch lengths and topology.

ARG-Needle relies on a threading operation that transforms inference into an iterative process. The threading operation may also be used to add additional samples to an existing ARG. For instance, one may first build an ARG using ARG-Needle on sequencing data, then thread additional array samples to this ARG to create a joint ARG of sequencing and array samples (see Section 5.3.3). One may also choose to run ASMC-clust or other inference algorithms on an initial set of samples, then thread the remaining samples using ARG-Needle.

For additional discussion about possible improvements to ARG-Needle, see Section 7.3.2. In the next chapter, we evaluate ARG-Needle and ASMC-clust in simulations.

Chapter 5

Performance of Genealogical Inference in Simulations

5.1 Overview

So far, we have introduced two classes of new methods related to ARGs. In Chapter 3, we showed that the ARG of a collection of samples could be used to perform ARG-based mixed linear model association (ARG-MLMA), as well as to construct ARG-GRMs which are then leveraged for mixed-model analyses. Both approaches are highly accurate and powerful, with ARG-GRMs performing as well as sequencing data (see Fig. 3.2). In Chapter 4, we introduced two novel methods for ARG inference from array or sequencing data sets—ARG-Needle and ASMC-clust—and analysed their theoretical properties. In this chapter, we evaluate our ARG inference methods in simulations and combine them with the ARG-GRM and ARG-MLMA approaches, illustrating that one can infer an ARG from array data and subsequently perform complex trait analyses. We also demonstrate additional examples involving ARGs built from sequencing data or a combination of sequencing and array data.

Specifically, we performed three types of experiments to evaluate ARG-Needle and ASMC-clust in simulations. First, we compared ARG-Needle and ASMC-clust to Relate [Speidel et al., 2019] and `tsinfer` [Kelleher et al., 2019] in terms of ARG inference accuracy and computational requirements. Second, we compared ARG-GRM

and ARG-MLMA complex trait analyses that were based on ARG-Needle inferred ARGs to corresponding analyses that used imputed or SNP array data. Third, we designed a small-scale experiment of ARG-based imputation using ARG-Needle and compared to IMPUTE4 [Bycroft et al., 2018] and Beagle 5 [Browning et al., 2018]. In the next chapter, we will further develop the second approach—involving ARG-Needle inference and ARG-MLMA—and use it to perform extensive analyses with the UK Biobank data set.[1]

5.2 Methods

5.2.1 ARG simulation conditions

We used the `msprime` coalescent simulator [Kelleher et al., 2016] to benchmark ARG inference algorithms. We simulated ground-truth ARGs and SNP array data in the same manner as in Chapter 3. For each run, we simulated sequence data with given physical length L for N haploid individuals. Our primary simulations used a mutation rate of $\mu = 1.65 \times 10^{-8}$ per base pair per generation, a constant recombination rate of $\rho = 1.2 \times 10^{-8}$ per base per generation, and a demographic model inferred using SMC++ on CEU (European demography) 1,000 Genomes samples [Terhorst et al., 2017]. To obtain realistic SNP data, we then subsampled the simulated sequence sites to match the genotype density and allele frequency spectrum of UK Biobank SNP array markers (chromosome 2, with density defined using 50 evenly spaced MAF bins).

Both ARG-Needle and ASMC-clust rely on ASMC to compute pairwise coalescence times. When running ASMC, we used decoding quantities for version 1.1, which were precomputed using a European demographic model and UK Biobank SNP array allele frequencies [Palamara et al., 2018]. ASMC and the hashing step of ARG-Needle also require a genetic map, which we computed based on the recombination rate used in simulations.

[1]The remainder of this chapter is expanded and adapted from [Zhang et al., 2021].

Besides our primary simulations, we included various additional simulation conditions where we modified one parameter in the primary simulations while keeping all else fixed. First, we varied the recombination rate to $\rho \in \{6 \times 10^{-9}, 2.4 \times 10^{-8}\}$ per base pair per generation, and passed this information to ARG-Needle and ASMC using the genetic map. Second, we used a constant demography of 15,000 diploid individuals, for which we generated new decoding quantities for ASMC. Third, we inferred ARGs using sequencing data, for which we ran ASMC in sequencing mode and modified the parameter L of ARG-Needle (see below). Fourth, we introduced genotyping errors into the array data. After sampling the array SNPs, we flipped each haploid genotype per SNP and individual with probability p, and varied the parameter p between a minimum of 10^{-7} and a maximum of 10^{-3}.

5.2.2 Comparisons of ARG inference methods

In comparisons of ARG inference methods, we used simulations with $L = 1$ Mb for sequencing data and $L = 5$ Mb for array data. We used ARG-Needle hashing parameters of $K = 64$ relatives chosen, hash word size $S = 16$ bits, tolerance $T = 1$ (allowing of one mismatch in an otherwise IBS stretch), and hashing region size $L = 0.5$ cM for array data and $L = 0.1$ cM for sequencing data (see Section 4.2.2). We ran Relate with the ground truth mutation rate, recombination rate, and demographic model. Relate includes a default option, which we kept, that limits the memory used for storing pairwise matrices to 5 GB. For each choice of sample size, we generated genetic data using either 5 or 25 random seeds and applied ARG-Needle, ASMC-clust, Relate, and `tsinfer` to infer ARGs. We included ARG normalisation by default in ARG-Needle and ASMC-clust. Due to scalability differences, we ran ASMC-clust and Relate in up to $N = 8{,}000$ haploid samples ($N = 4{,}000$ for sequencing) and ARG-Needle and `tsinfer` up to $N = 32{,}000$ haploid samples. We ran all analyses on Intel Skylake 2.6 GHz nodes on the Oxford Biomedical Research Computing cluster.

A commonly used metric for comparing topologies of trees is the Robinson-Foulds metric [Robinson and Foulds, 1981], which counts the number of unique mutations that can be generated by one tree but not the other (lower implies higher accuracy).

Because the presence of polytomies can skew this metric, we randomly break polytomies of `tsinfer`-inferred ARGs before comparing using Robinson-Foulds, as was done in [Kelleher et al., 2019]. We report a genome-wide average of the Robinson-Foulds metric, where we divide by the maximum possible value of $2N - 4$ to obtain a rescaled quantity between 0 (minimum) and 1 (maximum).

We generalised the Robinson-Foulds metric to measure mutational dissimilarity while considering branch lengths, to better capture the accuracy in predicting unobserved variants using an inferred ARG. To this end, we consider the probability distribution of mutations induced by uniform sampling over an ARG and compare the resulting distributions for the true versus inferred ARG using the total variation distance, a common metric for comparing probability measures. Polytomies do not need to be broken using this metric, as they simply concentrate the probability mass on fewer predicted mutations. We refer to this metric as **ARG total variation distance** (see Section 5.2.2.2 for further details).

For pairwise TMRCA comparisons, at each position we compute the TMRCA of each pair of samples in the true ARG and the inferred ARG, resulting in two vectors of length $N(N-1)/2$. We average the squared Euclidean distance between these vectors over all positions, then take a square root, resulting in a pairwise TMRCA root mean squared error (RMSE).

Because the branch lengths output by `tsinfer` are not calibrated, we did not include `tsinfer` in comparisons with the TMRCA RMSE metric. However, `tsinfer` ARGs still define a probability distribution of mutations, so we included `tsinfer` with a a dotted line in comparisons using the total variation distance.

We also considered the Kendall-Colijn (KC) topology-only distance [Kendall and Colijn, 2016] to compare ARG topologies, calculated between marginal trees of two ARGs and averaged over all positions. We observed that the performance of methods that output binary trees (Relate, ASMC-clust, and ARG-Needle) under this metric significantly improved when we selected inferred branches at random and collapsed them to create polytomies (Fig. 5.2c), suggesting that the KC distance tends to reward inferred ARGs that contain polytomies. We therefore developed a heuristic

approach to form polytomies in inferred ARGs by collapsing branches based on their size and height, which capture the confidence in the inferred tree branches. We applied this heuristic to all methods to compare their performance under the KC topology-only distance while accounting for different prevalence of polytomies in the inferred ARGs (Figs. 5.1d, 5.2d, 5.4d).

We describe additional details of our evaluation metrics below.

5.2.2.1 Evaluating metrics via stabbing queries

Most of the metrics we considered—Robinson-Foulds distance, pairwise TMRCA RMSE, and KC topology-only distance—are originally defined on two trees.[2] To generalise these metrics to ARGs, we consider the metrics as comparing two marginal trees at the same position and take a genome-wide average of the metrics over all positions.

Consider two ARGs $\mathcal{A}$ and $\mathcal{B}$, which may for instance be the true and inferred ARGs, with a common genomic extent $[s, t) \subset \mathbb{R}$.[3] Let Tree be a binary operator that takes an ARG and a position and returns the tree at that position, and suppose $d : \mathcal{T}_N \times \mathcal{T}_N \to \mathbb{R}$ is a metric of interest that operates on the space $\mathcal{T}_N$ of rooted trees on N leaves. Then the genome-wide average of d is

$$\frac{1}{t-s} \int_s^t d(\text{Tree}(\mathcal{A}, x), \text{Tree}(\mathcal{B}, x)) dx.$$

Although it is possible to compute this integral exactly, we found that such an approach is slowed down by needing to iterate over all marginal trees in both ARGs. In particular, the true ARG often contains many recombination events, each of which generates a new marginal tree. Therefore, we instead approximated the integral using sampling. Given a set of points $x_1, \ldots, x_n$ uniformly distributed among $[s, t)$, we have

$$\frac{1}{t-s} \int_s^t d(\text{Tree}(\mathcal{A}, x), \text{Tree}(\mathcal{B}, x)) dx \approx \frac{1}{n} \sum_{i=1}^{n} d(\text{Tree}(\mathcal{A}, x_i), \text{Tree}(\mathcal{B}, x_i)).$$

[2]The ARG total variation distance is defined directly between two ARGs and is treated separately.
[3]Note that we are modelling the genome as continuous.

We call each such $d(\text{Tree}(\mathcal{A}, x_i), \text{Tree}(\mathcal{B}, x_i))$ term a "stabbing query", because we are sampling two trees at a position along the ARG, which entails finding the edges of the ARG that overlap this position. The estimate is then an unweighted average over n stabbing queries. For choosing the points $\{x_i\}_{i=1}^n$, we set $x_i = s + (i \cdot \phi - \lfloor i \cdot \phi \rfloor) \times (t - s)$, where $\phi = (1 + \sqrt{5})/2$ is the golden ratio. For all evaluations, we used $n = 5{,}000$ stabbing queries.

5.2.2.2 ARG total variation distance as a generalisation of the Robinson-Foulds distance

Earlier, we defined the scaled Robinson-Foulds distance, which we average over $n = 5{,}000$ genome-wide stabbing queries to obtain a quantity between 0 (perfect match of present mutations) and 1 (complete mismatch of non-singleton mutations). The Robinson-Foulds distance weights all possible mutations in a marginal tree equally when considering dissimilarity. However, rare variants in these marginal trees tend to correspond to recent, short branches and are less likely to occur through random mutations. Furthermore, these short and recent branches tend to represent long haplotypes and thus appear in multiple neighboring trees, leading to a disproportionate contribution to the overall distance. The ARG total variation distance generalises the Robinson-Foulds distance to overcome these limitations when comparing ARGs.

To define the ARG total variation distance, we consider the probability distribution of possible mutations encoded by an ARG, and compute the total variation distance between two such distributions. More formally, given N samples, the possible non-trivial mutational patterns that can occur over these samples are given by the set $\{0,1\}^N \setminus \{(0,\ldots,0),(1,\ldots,1)\}$. We call each element of this set an N-bitset; there are $2^N - 2$ possible N-bitsets, which we denote as $\mathcal{S}_N$. We choose to represent each mutation using its position $x \in [s,t)$ and the N-bitset $b \in \mathcal{S}_N$. We thus consider the probability distribution corresponding to the mutations $(x,b) \in [s,t) \times \mathcal{S}_N$ that can be observed. Assuming a constant mutation rate, each simulated mutation is uniformly distributed over the area of the ARG, which consists of the physical distance times the height of each mutation-generating branch. We can capture such a distribution

via a probability density function (PDF) $f_{\mathcal{A}} : [s,t) \times \mathcal{S}_N \to \mathbb{R}$ which is induced by uniform sampling over the ARG. We can use two equivalent expressions for the total variation distance and apply them to compute the total variation distance between two ARGs $\mathcal{A}$ and $\mathcal{B}$ in terms of their PDFs:

$$TV_{ARG}(\mathcal{A},\mathcal{B}) = \frac{1}{2}\int_s^t \sum_{b\in\mathcal{S}_N} |f_{\mathcal{A}}(x,b) - f_{\mathcal{B}}(x,b)|dx \tag{5.1}$$

$$= 1 - \int_s^t \sum_{b\in\mathcal{S}_N} \min(f_{\mathcal{A}}(x,b), f_{\mathcal{B}}(x,b))dx. \tag{5.2}$$

The distance is 0 for identical distributions and 1 if the two distributions contain no overlapping support. To evaluate the ARG total variation distance in our simulations, we used the second expression, where we approximated the integral by sampling $n =$ 5,000 stabbing queries and performing a sum over all present mutations for each stabbing query.

To highlight the connection between the ARG total variation distance and Robinson-Foulds more explicitly, we first describe how we compute the PDF $f_{\mathcal{A}}(x,b)$ of an ARG $\mathcal{A}$. The value $f_{\mathcal{A}}(x,b)$ should be 0 if no branch yielding N-bitset b exists at the tree at position x. Otherwise, $f_{\mathcal{A}}(x,b)$ should be proportional to the length of the branch yielding b. Define a function $l_{\mathcal{A}} : [s,t) \times \mathcal{S}_N \to \mathbb{R}$ which will be an unnormalised version of $f_{\mathcal{A}}(x,b)$:

$$l_{\mathcal{A}}(x,b) = \begin{cases} 0, & \text{if no branch giving } b \text{ exists at Tree}(\mathcal{A},x), \\ \text{the length of a branch giving } b \text{ at Tree}(\mathcal{A},x), & \text{otherwise.} \end{cases} \tag{5.3}$$

We can then write

$$f_{\mathcal{A}}(x,b) = \frac{1}{Z_{\mathcal{A}}} l_{\mathcal{A}}(x,b) \tag{5.4}$$

for some scaling constant $Z_{\mathcal{A}}$. We would like all the properties of a PDF to hold, e.g.

$$\int_s^t \sum_{b \in \mathcal{S}_N} f_{\mathcal{A}}(x, b) dx = 1. \tag{5.5}$$

Combining (5.4) and (5.5), we obtain

$$Z_{\mathcal{A}} = \int_s^t \sum_{b \in \mathcal{S}_N} l_{\mathcal{A}}(x, b) dx. \tag{5.6}$$

For a given ARG $\mathcal{A}$, equations (5.3), (5.4), and (5.6) provide a way to evaluate its PDF. First, we compute $Z_{\mathcal{A}}$ using stabbing queries. Then, to get the value of $f_{\mathcal{A}}(x, b)$, we look up $l_{\mathcal{A}}(x, b)$ using the marginal tree at x, and divide by $Z_{\mathcal{A}}$.

Consider again the definition of $l_{\mathcal{A}}(x, b)$. We can obtain a topology-only version of the ARG total variation distance by rewriting $l_{\mathcal{A}}(x, b)$ to not use branch lengths:

$$l'_{\mathcal{A}}(x, b) = \begin{cases} 0, & \text{if no branch giving } b \text{ exists at } \text{Tree}(\mathcal{A}, x), \\ 1, & \text{otherwise.} \end{cases}$$

If we use the same definitions for TV_{ARG} but replace $l_{\mathcal{A}}(x, b)$ with $l'_{\mathcal{A}}(x, b)$ and assume no polytomies in $\mathcal{A}$ or $\mathcal{B}$, the ARG total variation distance reduces to a scaled version of the Robinson-Foulds metric.

Note that in our definition, we considered the space of possible mutations as consisting of a position $x \in [s, t)$ and an N-bitset $b \in \mathcal{S}_N$. Including the position enables us to measure the ability to correctly localise mutations in the genome. It may be worthwhile to also consider the time (height) of the mutations, which would measure the ability to correctly localise the time of mutation events.

5.2.2.3 KC distance with breaking and formation of polytomies

The Kendall-Colijn topology-only distance [Kendall and Colijn, 2016] (henceforth "KC distance" for short) has been used to evaluate inferred ARGs (e.g. in [Kelleher et al., 2019]). As a topology-only metric, it does not consider branch length or

coalescence time information. Instead, it counts the number of nodes on the path from a desired MRCA node to the root node, and compares these counts for all possible MRCA pairs. Given two marginal trees T_1 and T_2, each with the same set of samples labeled 1 to N, the KC distance computation first calculates a vector of length $N(N-1)/2$ for each tree as follows. For one of the trees T, let $\text{root}(T)$ denote the root node of the tree, and for any two nodes a and b in the tree, let $\text{mrca}_T(a, b)$ denote the most recent common ancestor of a and b and let $\text{dist}_T(a, b)$ denote the tree-path distance between nodes a and b in the tree, defined as the number of edges in the tree that need to be traversed to travel between a and b. For any two distinct samples $1 \leq i < j \leq N$, let $\text{leaf}_T(i)$ and $\text{leaf}_T(j)$ denote the corresponding leaf nodes. We consider the MRCA of the two leaf nodes and count the distance from the root of the tree, computing

$$m_T(i, j) = \text{dist}_T\left(\text{root}(T), \text{mrca}_T(\text{leaf}_T(i), \text{leaf}_T(j))\right).$$

We then form vectors

$$\begin{aligned}\mathbf{m}_{T_1} &= \left(m_{T_1}(1,2), m_{T_1}(1,3), \ldots, m_{T_1}(N-1, N)\right), \\ \mathbf{m}_{T_2} &= \left(m_{T_2}(1,2), m_{T_2}(1,3), \ldots, m_{T_2}(N-1, N)\right).\end{aligned}$$

(Note that in [Kendall and Colijn, 2016] the $\mathbf{m}_T$ vectors are augmented with additional entries, which are however irrelevant in our case.) Finally, we define the KC distance as the L2 norm between these two length $N(N-1)/2$ vectors:

$$\text{KC}(T_1, T_2) = ||\mathbf{m}_{T_1} - \mathbf{m}_{T_2}||_2 = \left(\sum_{1 \leq i < j \leq N} \left(m_{T_1}(i, j) - m_{T_2}(i, j)\right)^2\right)^{1/2}.$$

One option to combine tree-wise KC distances to get a comparison between ARGs is to weight the KC metric by the genomic distance spanned by each tree, as done in [Kelleher et al., 2019]. For efficiency, this can be approximated by taking an unweighted average of the KC metric for several stabbing queries. In this work we

opted to instead average the squared KC distance over the stabbing queries, then perform the square root (similar to our TMRCA RMSE calculations). A benefit of this approach is that it preserves the interpretation of the (ARG) KC distance as an L2 norm. (Suppose $\mathbf{m}_{1x}$ and $\mathbf{m}_{1y}$ are vectors from the first ARG at two locations x and y, and $\mathbf{m}_{2x}$ and $\mathbf{m}_{2y}$ are vectors from the second ARG at two locations. Assume both locations receive a weight of 1/2. Then our metric is equivalent to computing $||((\mathbf{m}_{1x}, \mathbf{m}_{1y}) - (\mathbf{m}_{2x}, \mathbf{m}_{2y}))/2||_2$, whereas the other approach results in $(||\mathbf{m}_{1x} - \mathbf{m}_{2x}||_2 + ||\mathbf{m}_{1y} - \mathbf{m}_{2y})||_2)/2$.)

The KC topology-only distance is affected whenever two close coalescence events are joined to form a polytomy, or when polytomies are broken to form strictly bifurcating trees, since these operations create or remove nodes and thus alter the distances from internal nodes to the root. To randomly resolve polytomies in marginal trees produced by `tsinfer` we replicated the approach used in [Kelleher et al., 2019]. At each polytomy with k child edges coalescing, a random binary tree with k leaves was generated and was substituted in place of the polytomy.

We performed additional experiments where instead of breaking polytomies into bifurcations, we collapsed branches in a marginal tree to create polytomies. Because our goal is to measure similarity to the true ARG, we applied this operation only on inferred ARGs and not the true ARG. The merging operation takes a real parameter f between 0 and 1 corresponding to the fraction of branches that are collapsed in each marginal tree of the inferred ARG. We implemented two types of merging: random merging and heuristic merging. In random merging, for each branch we sample a uniform random real number between 0 and 1, and if it is less than f we collapse that branch. In heuristic merging, we aim to instead first collapse branches that are predicted with least confidence. We order the branches in the tree by computing the ratio of the branch length divided by the height of the parent node (to take into account that coalescent events in recent time tend to have short branches). We select the first fraction f of the branches under the ordering and merge these to form polytomies. For either method of merging, we use a single merged tree per site and evaluate the KC distance against the marginal tree in the true ARG.

5.2.3 Simulations of complex trait analysis

As in Chapter 3, we simulated polygenic traits from haploid sequencing samples for various values of h^2 and α. We adopted a ratio of L/N of 5×10^{-3} Mb/individuals for heritability experiments and of $L/N = 5.5 \times 10^{-3}$ Mb/individuals for association experiments, where L is the total length across 22 chromosomes in the case of association. We performed heritability estimation using ARGs inferred by ARG-Needle from SNP array data, and compared against heritability estimates from SNP array data and imputed data. To obtain the imputed data, we simulated 500 or 1,000 additional haploid references in the same ARG simulation and used IMPUTE4 [Bycroft et al., 2018] to impute variation onto the array samples. We built MAF-stratified GRMs for ARG-Needle ARGs and imputed data, using the same MAF boundaries of $\{0, 0.01, 0.05, 0.5\}$ and a value of $\alpha = -1$ within each bin, as was done in Chapter 3. When we attempted this approach for SNP data, GCTA did not converge, so we instead built a single GRM using the true value of α.

We repeated the ARG-MLMA experiment from Chapter 3, this time adding an ARG inferred by ARG-Needle from SNP array data, as well as association of imputed data. We traversed the true or inferred ARG and wrote out all possible mutations to disk, then used GCTA or PLINK to perform LMM or linear regression testing. For mixed-model association, we used a LOCO GRM built from array markers on chromosomes 2-22.[4] We compared between association of array data, the true ARG, an ARG-Needle inferred ARG, and data imputed from 500 or 1,000 haploid references using IMPUTE4 [Bycroft et al., 2018]. The complex trait was simulated as in Chapter 3, with a polygenic background on chromosomes 2-22 with narrow-sense heritability $h^2 = 0.8$ and negative selection parameter [Speed et al., 2012] $\alpha = -0.25$, and a single causal allele on chromosome 1 (this time with allele frequency $p \in \{0.0025, 0.005, 0.01\}$) with varying effect size β. We measured association power as the fraction of runs (out of 200) detecting a significant association on chromosome 1, with significance thresholds calibrated to yield a family-wise error

[4]Note that Chapter 3 also included LOCO GRMs using sequencing data or ARGs, which we did not include here.

rate of 0.05.

We also extended the above ARG-MLMA experiment to a setting with diploid individuals and phasing error. For association with N individuals, we simulated an ARG with $2N$ haploid samples as well as additional samples representing a reference panel. We merged pairs of the $2N$ haploids to obtain diploid genotypes. To incorporate realistic switch errors, we phased these diploid genotypes with Beagle 5.1 [Browning et al., 2018], using samples from the reference panel. We used ARG-Needle to infer ARGs on either the $2N$ phased samples or on the original $2N$ haploid samples where the true phase is known. We performed association using the branches of the ARG to propose diploid genotypes, writing all genotypes out to disk. Trait simulation used the same parameters as above but with diploid genotypes. For imputed data, we did not introduce any phasing errors, but rather used the original $2N$ haploid samples as well as reference samples, with the assumption that true phase is known.

5.2.4 ARG-based genotype imputation

Given a collection of sequencing and array samples, we perform ARG-based imputation as follows (see Fig. 5.9a). We use ARG-Needle in sequencing mode to first thread the sequencing samples, then thread the array samples using array mode. For each sequencing site not genotyped by the array samples, we consider the marginal tree at that position in the inferred ARG. We select the branches in the marginal tree for which an unseen mutation best explains the observed sequencing data in terms of Hamming distance. Each of these branches implies genotypes of 0 or 1 for the array samples, so we output a weighted average of the implied genotypes, weighting branches by their length in the marginal tree. We applied this framework to perform genotype imputation using ground-truth ARGs and ARG-Needle inferred ARGs in 10 Mb of simulated data and compared to IMPUTE4 [Bycroft et al., 2018] and Beagle 5 [Browning et al., 2018] imputation using the binned aggregate r^2 metric [McCarthy et al., 2016] (Fig. 5.9b-c).

5.3 Results

To evaluate ARG-Needle and ASMC-clust, we performed three types of experiments outlined in Section 5.1: ARG inference, ARG-based complex trait analysis, and ARG-based imputation.

5.3.1 Comparison of ARG inference methods

We used extensive coalescent simulations to compare the accuracy and scalability of ARG-Needle, ASMC-clust, Relate, and `tsinfer` (Fig. 5.1 and Figs. 5.2-5.6). We generated synthetic array data sets of up to 32,000 haploid samples using a European demographic model (see Section 5.2.1) and measured ARG reconstruction accuracy. To this end, we considered several metrics used in the past to compare ARGs, including: the Robinson-Foulds distance [Robinson and Foulds, 1981], which reflects dissimilarities between the possible mutations that can be generated by two ARGs; the root mean squared error (RMSE) between true and inferred pairwise TMRCAs, which captures the accuracy in predicting allele sharing between individuals; and the Kendall-Colijn (KC) topology-only distance [Kendall and Colijn, 2016]. We found that the KC distance is systematically lower for trees containing polytomies (Fig. 5.2b-d, see Section 5.3.1.1), which are not output by Relate, ASMC-clust, or ARG-Needle by default. We therefore applied a simple heuristic to allow all methods to output polytomies (see Section 5.2.2.3). Although these three metrics capture the similarity between marginal trees and are in some cases interpretable in terms of accuracy in downstream analyses, they are not specifically developed for applications related to ARGs. We therefore also developed a new metric, called the ARG total variation distance, which generalises the Robinson-Foulds distance to better capture the ability of a reconstructed ARG to predict mutation patterns that may be generated by the true underlying ARG (see Section 5.2.2.2). Across these metrics, ARG-Needle and ASMC-clust achieved best performance for our primary array data simulations (Fig. 5.1a-d).

We next measured the speed and memory footprint of these methods. ARG-Needle

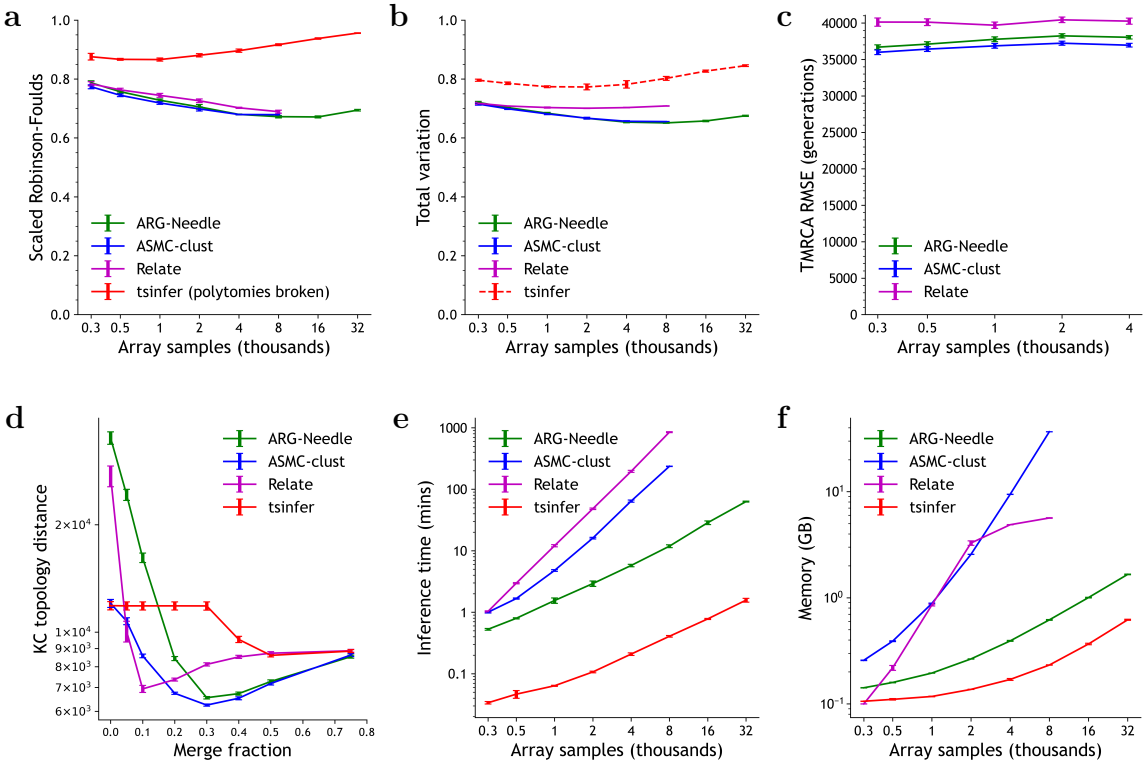


Figure 5.1: **Comparison of ARG inference algorithms in array data simulations.** We benchmark ARG inference performance for ARG-Needle, ASMC-clust, Relate, and `tsinfer` in realistic array data simulations across a variety of metrics related to accuracy and computational resources (lower values indicate better performance for all metrics), including **a.** the Robinson-Foulds distance, **b.** the ARG total variation distance (see Section 5.2.2.2), **c.** pairwise TMRCA root mean squared error, **d.** the Kendall-Colijn topology-only metric, **e.** runtime, and **f.** peak memory. In **c**, we only run up to $N = 4{,}000$ samples. In **d**, we fix $N = 4{,}000$ samples and vary the fraction of branches that are merged to form polytomies, using a heuristic that preferentially merges branches that are less confidently inferred (see Section 5.2.2.3). Both **c** and **d** involve 25 random seeds. All other examples use 5 random seeds and run up to 32,000 samples for ARG-Needle and ASMC-clust, and 8,000 samples for ASMC-clust and Relate due to runtime or memory constraints. Error bars represent 2 s.e. We plot `tsinfer` using a dotted line in **b** and omit it from **c** (see Section 5.2.2). Relate's default settings cap the memory for intermediate computations at 5 GB (see **f**). For additional simulations, see Figs. 5.2-5.3.

requires lower computation and memory than Relate and ASMC-clust, which both scale quadratically with respect to sample size (Fig. 5.1e,f and Fig. 5.2a). It runs slower than `tsinfer` but with a similar (approximately linear) scaling (see Section 4.2.2).

Finally, we performed simulations in a variety of additional settings, including a constant demographic history, varying recombination rates, genotyping error, and with sequencing data (Figs. 5.2-5.5). ARG-Needle tended to achieve best performance across all accuracy metrics in array data, sometimes tied or in close performance with ASMC-clust or Relate (Figs. 5.2-5.3). In sequencing data, ASMC-clust performed best on the ARG total variation and TMRCA RMSE metrics, with ARG-Needle and Relate close in performance, while Relate and `tsinfer` performed better on the Robinson-Foulds metric (Fig. 5.4). In the presence of genotyping error, ARG-Needle exhibited robust performance for error rates up to 0.1% per base pair per sample (Fig. 5.5).

We elaborate on two areas of findings: the effect of polytomies on the KC distance, and the effect of ARG normalisation on ARG node times.

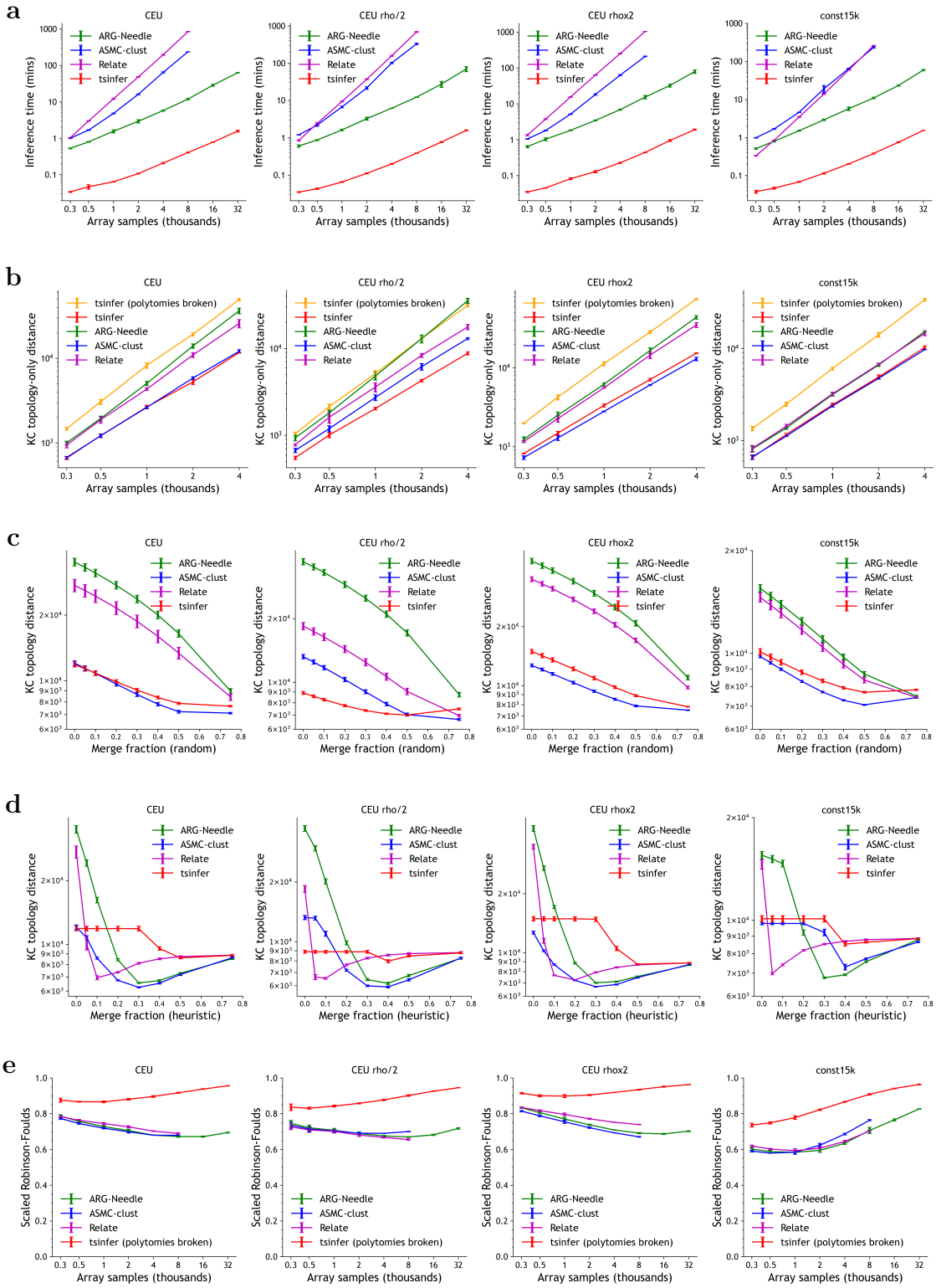

a
CEU
CEU rho/2
CEU rhox2
const15k
Inference time (mins)
Array samples (thousands)
ARG-Needle
ASMC-clust
Relate
tsinfer
b
KC topology-only distance
tsinfer (polytomies broken)
c
KC topology distance
Merge fraction (random)
d
Merge fraction (heuristic)
e
Scaled Robinson-Foulds

Figure 5.2: **Additional comparison of ARG inference methods with array data and topology-only metrics.** We compare methods on runtime and topology-only metrics, as in Fig. 5.1 but with additional simulation conditions. All columns are for 5 Mb of CEU demography array data, and individual columns represent standard parameters (see Section 5.2.2), a factor of 2 smaller recombination rate ($\rho = 6 \times 10^{-9}$), a factor of 2 larger recombination rate ($\rho = 2.4 \times 10^{-8}$), and a constant population size demography of 15,000 individuals. **a**. Inference time as a function of the number of samples N. **b**. KC topology-only distance as a function of N, additionally showing the results of `tsinfer` with randomly resolved polytomies. **c**. KC topology-only distance for $N = 4{,}000$ samples, showing performance as branches in marginal inferred trees are randomly collapsed to form polytomies. **d**. The same as **c**, except using a heuristic to preferentially merge branches that are least certain (see Section 5.2.2.3). **e**. Robinson-Foulds distance as a function of N, where values are scaled to lie between 0 and 1, and with randomly resolved polytomies. Error bars represent 2 s.e.

5.3.1.1 KC distance favours ARGs with polytomies

We performed extensive experiments which indicate that the KC distance yields systematically lower values for methods that generate polytomies. First, we compared methods on the KC distance with randomly broken `tsinfer` polytomies, where we sample the breaking procedure 10 times for each polytomy and average the KC results. As in [Kelleher et al., 2019], we observed that randomly resolving polytomies in `tsinfer` leads to reduced KC distance performance.

We next allowed for branches in inferred ARGs to be collapsed to increase the presence of polytomies. We developed a random merging strategy as well as a strategy that merges less confident branches based on a heuristic (see Section 5.2.2.3). For both strategies, we merged a fraction of branches $f \in \{0, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.75\}$, and measured the KC distance after applying merging for ARGs inferred by `tsinfer`, Relate, ASMC-clust, and ARG-Needle (Figs. 5.2c-d and 5.4c-d). We used 4,000 sequences and 1 Mb for inference from sequences, and we used 4,000 array samples and 5 Mb for inference from array data. Relate, ASMC-clust, and ARG-Needle achieved lower KC distance when nodes were merged to form polytomies, including when the random merging strategy was used, suggesting that the KC distance is systematically lower for inferred trees that contain polytomies. Using heuristic merging in array data led to improvements for all methods, with Relate, ASMC-clust, and ARG-Needle

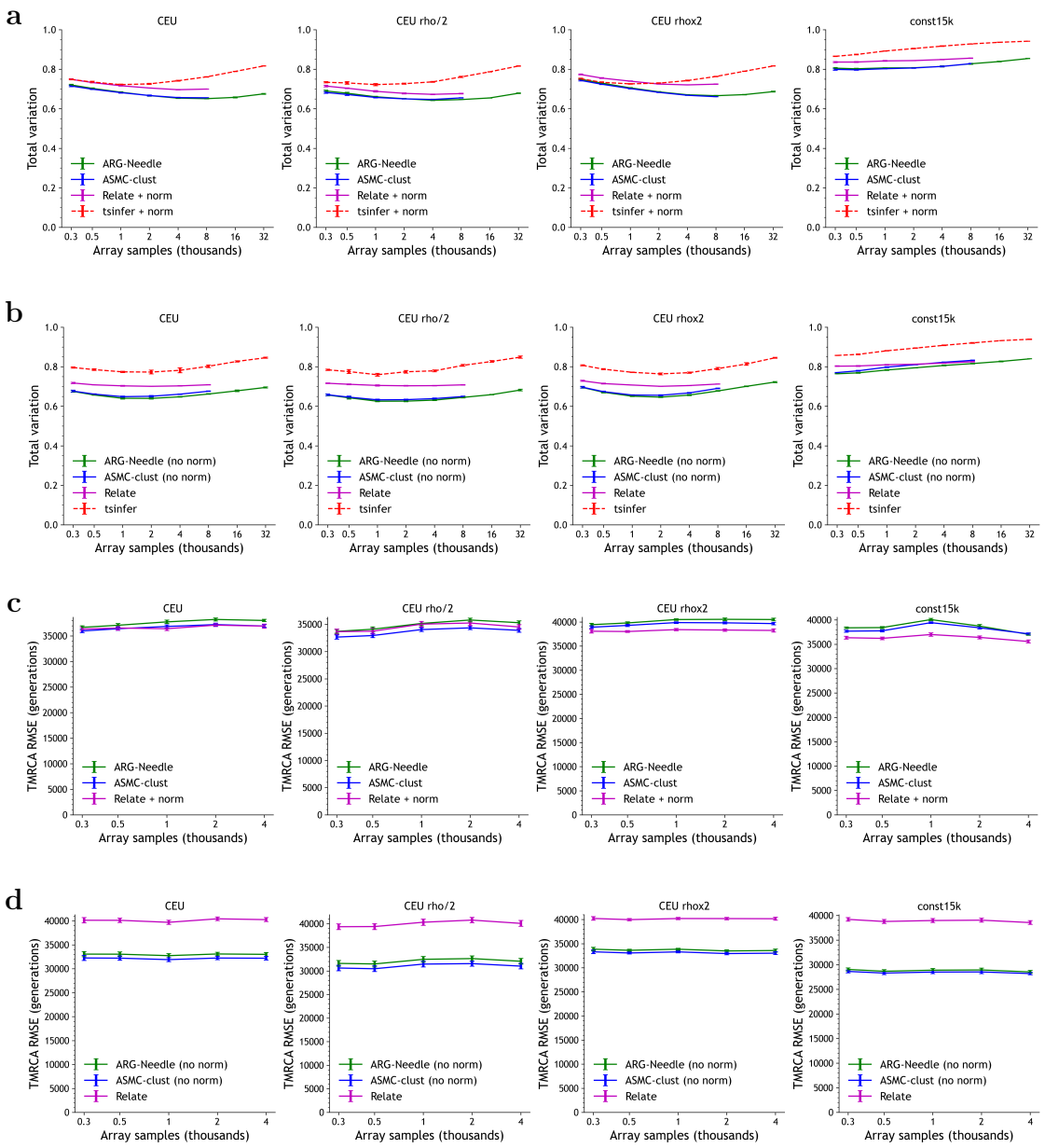


Figure 5.3: **Additional comparison of ARG inference methods with array data and metrics that take into account branch length.** As in Fig. 5.1b-c but with additional simulation conditions. All columns are for 5 Mb of CEU demography array data, and individual columns represent standard parameters (see Section 5.2.2), a factor of 2 smaller recombination rate ($\rho = 6 \times 10^{-9}$), a factor of 2 larger recombination rate ($\rho = 2.4 \times 10^{-8}$), and a constant population size demography of 15,000 individuals. We show results for the ARG total variation distance (**a**,**b**) and pairwise TMRCA RMSE (**c**,**d**) across different conditions with and without ARG normalisation for the various methods, as these metrics are sensitive to branch length. `tsinfer` is shown with dotted lines for total variation as it focuses on topologies. Error bars represent 2 s.e.

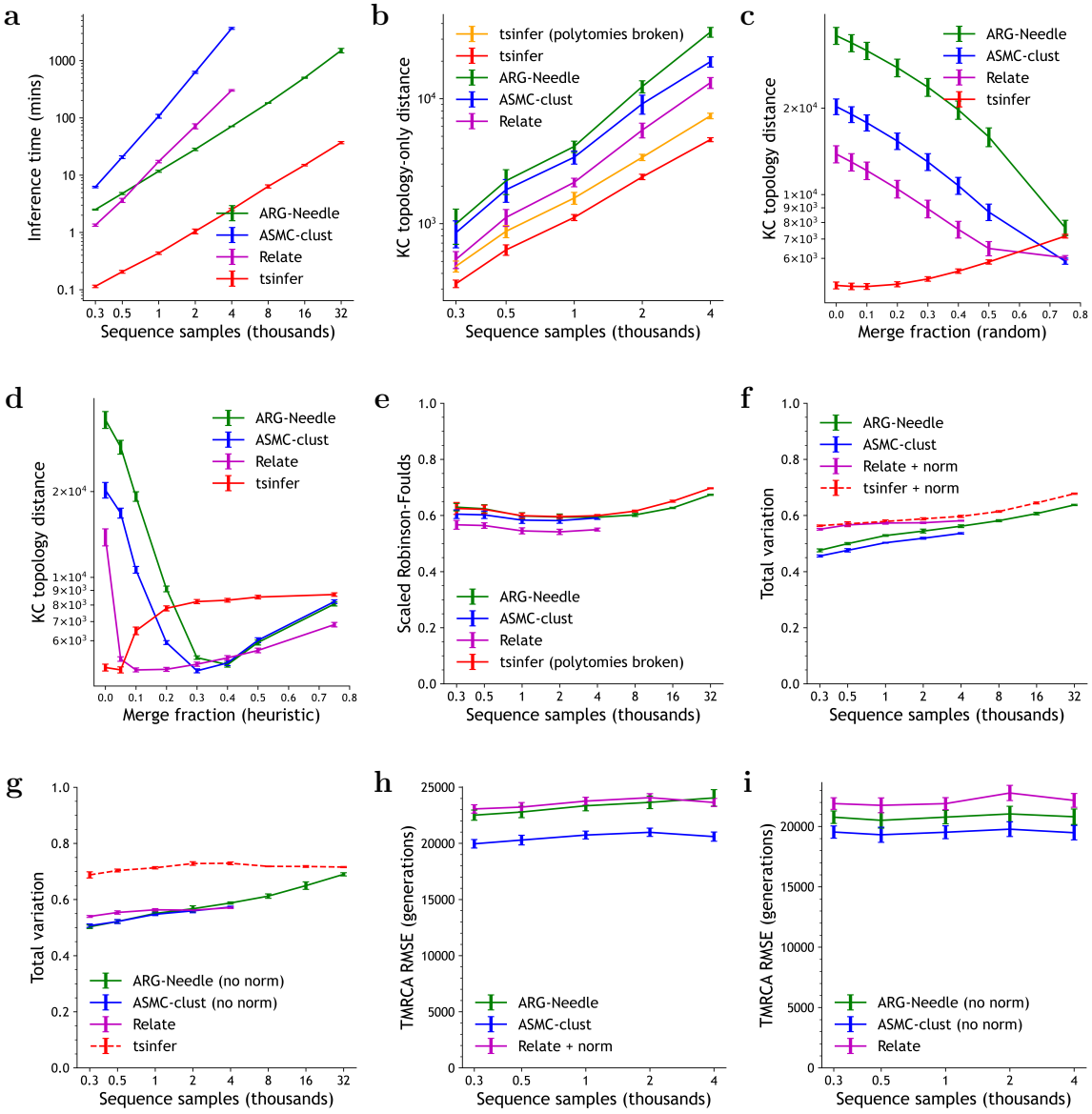


Figure 5.4: **Comparison of ARG inference methods with sequencing data.** Simulations use 1 Mb of CEU sequencing data and otherwise standard parameters (see Section 5.2.2). Individual panels correspond to rows of Figs. 5.2a-e and 5.3a-d, in that order, with the same metrics used.

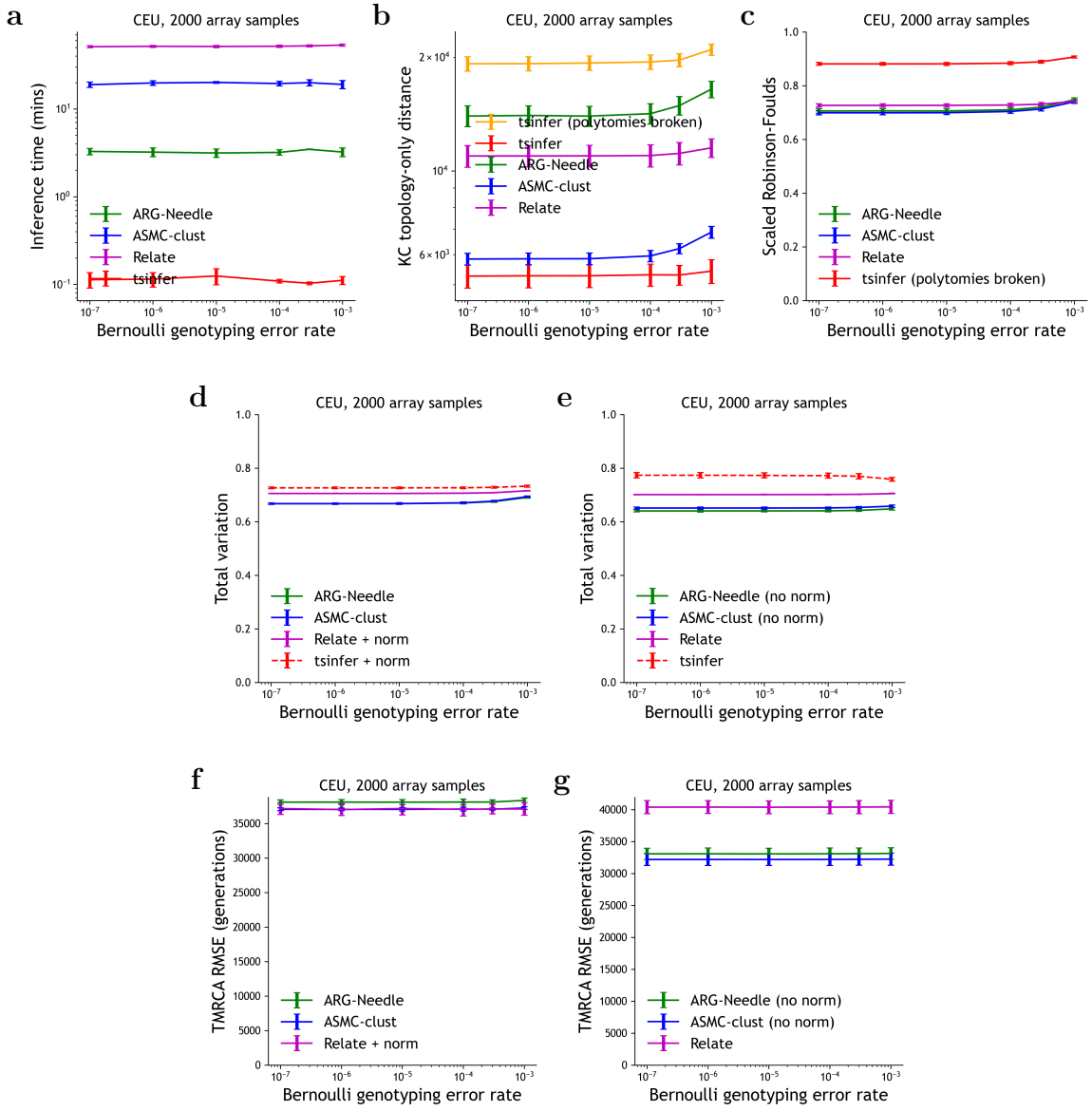


Figure 5.5: **Comparison of ARG inference methods for array data with genotyping error.** Simulations fix $N = 2{,}000$ haploid array samples, vary the genotyping error rate, and use otherwise standard parameters (see Section 5.2.2). Individual panels correspond to rows of Figs. 5.2a,b,e and 5.2a-d, in that order, with the same metrics used. Because the KC merging experiments in Fig. 5.2c,d would require varying both the genotyping error rate and the merge fraction f, we chose to omit these experiments. The first column of Fig. 5.2c,d with data point $N = 2{,}000$ corresponds to KC merging performance in the case of no genotyping error. TMRCA plots in **f** and **g** average over 5 random seeds rather than the usual 25 random seeds (see Fig. 5.1 caption).

performing better than `tsinfer` at the optimal merging fraction. In sequencing data, the KC performance of `tsinfer` was not improved by heuristic merging, and was matched by Relate, ASMC-clust, and ARG-Needle as f was varied.

5.3.1.2 Effect of ARG normalisation

In Chapter 3, we introduced ARG normalisation as a way to adjust the node times of an inferred ARG to be more consistent with a demographic prior. We observed that ARG normalisation improves the pairwise TMRCA RMSE of Relate in array data, though not in sequencing data (Figs. 5.3c,d and 5.6e-f). This suggests that ARG normalisation provides a reasonable branch length estimation heuristic when branch lengths are not modelled or inferred under model misspecification, as in the case of Relate on array data. For ARG-Needle and ASMC-clust, ARG normalisation improves the overall calibration of TMRCAs by making the range of predicted coalescence times closer to that expected from the demographic prior (Fig. 5.6a-d). Interestingly, however, ARG normalisation decreases the TMRCA RMSE performance of ARG-Needle and ASMC-clust in both array and sequencing data (Fig. 5.3c,d), such that ARG-Needle and ASMC-clust without ARG normalisation consistently achieve the best TMRCA RMSE across methods (Fig. 5.3d). This is likely linked to our use of ASMC's posterior mean TMRCA estimator, which is biased towards the average prior TMRCA but leads to good RMSE performance due to the weight placed by an L2 norm on outliers. ARG normalisation is also likely to reduce the accuracy in the root node height for ARGs built using ARG-Needle and ASMC-clust. This may have a large impact on TMRCA RMSE as a large fraction of pairwise coalescence events involve the root.

5.3.2 ARG-Needle and ARG-based complex trait analysis

ARG-Needle is uniquely suited for inference from large array datasets. In Chapter 3, we demonstrated that the ARG-GRM and ARG-MLMA strategies greatly outperform comparable methods using array data if true ARGs are available. To further

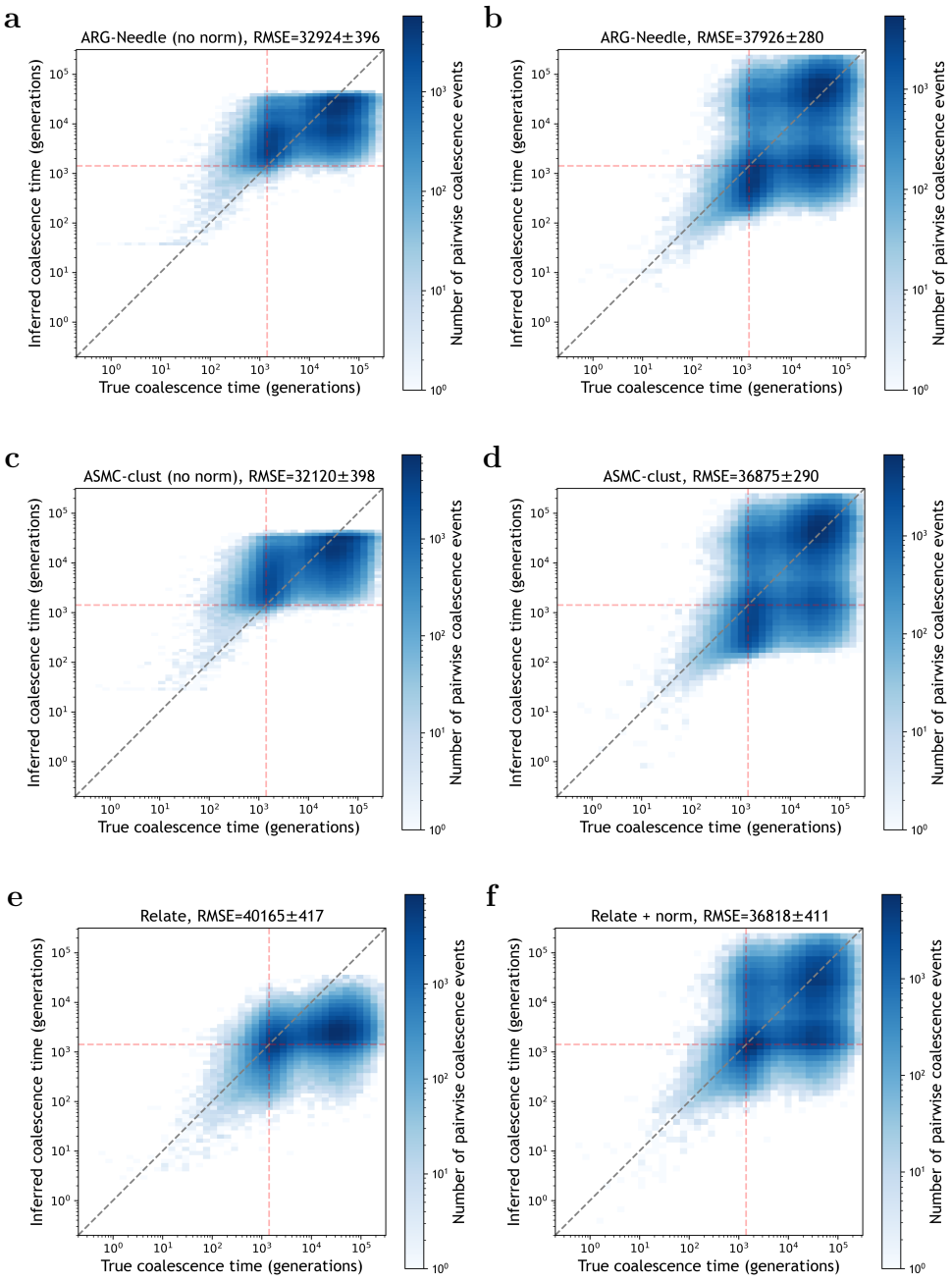


Figure 5.6: **Scatter plots of true versus inferred pairwise TMRCAs.** We show scatter plots for ARG-Needle (**a**-**b**), ASMC-clust (**c**-**d**), and Relate (**e**-**f**), for $N = 4{,}000$ array samples, showing an aggregate over 20,000 randomly sampled pairs for each of 25 simulations. The left column corresponds to no ARG normalisation and the right column corresponds to including ARG normalisation. Titles display TMRCA RMSE with 2 s.e. from bootstrap sampling of the 25 simulations used. Removing ARG normalisation decreases the pairwise TMRCA RMSE for ARG-Needle and ASMC-clust, but skews the distribution of pairwise TMRCAs towards the center. Dotted red lines show the time of the CEU demography population bottleneck, when many pairs of lineages are expected to coalesce.

validate these methods, we sought to test if an ARG inferred from array data using ARG-Needle could be combined with the ARG-GRM and ARG-MLMA strategies for heritability estimation and mixed-model association.

We performed simulations similar to those in Chapter 3, but with ARG-Needle inferred ARGs instead of true ARGs. We compared our methods against analyses that directly used the array data as well as imputed data. First, we used ARG-Needle and MAF-stratified ARG-GRMs to perform heritability estimation of a polygenic trait with $N = 2{,}000$ and $h^2 \in \{0.3, 0.5, 0.8\}$ (Fig. 5.7). ARG-based heritability estimates, despite requiring no sequencing data, were more accurate than those from imputed data with 500 or 1,000 reference samples. Furthermore, the heritability estimates were stable as α was varied, suggesting that the inferred ARGs capture a full allele frequency spectrum of variation which can be leveraged by the MAF-stratified GRMs.

Second, we used ARG-Needle and our ARG-MLMA strategy to perform genealogy-wide association in a setting of $N = 2{,}000$ haploid individuals. In simulations, we observed that genealogy-wide association and ARG-MLMA can achieve higher statistical power to detect signals that are linked to low frequency causal variants (MAF = 0.25%) than testing based on SNP array variants or variants imputed from a sequenced reference panel (Fig. 5.8a). For more common causal variation (MAF $\in \{0.5\%, 1\%\}$), genealogy-wide association performed comparably to imputed data (Fig. 5.8b-c). Across all conditions, ARG-MLMA increased association power compared to linear regression of ARG variants (Fig. 5.8d-e).

Third, we modified the preceding ARG-Needle and ARG-MLMA experiment to a setting of $N = 2{,}000$ diploid individuals (Fig. 5.8f). It is difficult to directly measure the effect of phasing errors on ARG inference metrics as switch errors make it difficult to match samples in the true and inferred ARGs. However, diploid ARG-based association combines the genotypes from the two corresponding haploid samples, making it simple to measure the effect of phasing error. We first ran ARG-Needle and ARG-MLMA using the known phase of haplotypes. We then combined pairs of haploid samples to form diploid genotypes, phased the diploid genotypes into haploid pairs using Beagle 5.1 and 500 diploid references, and applied ARG-Needle and

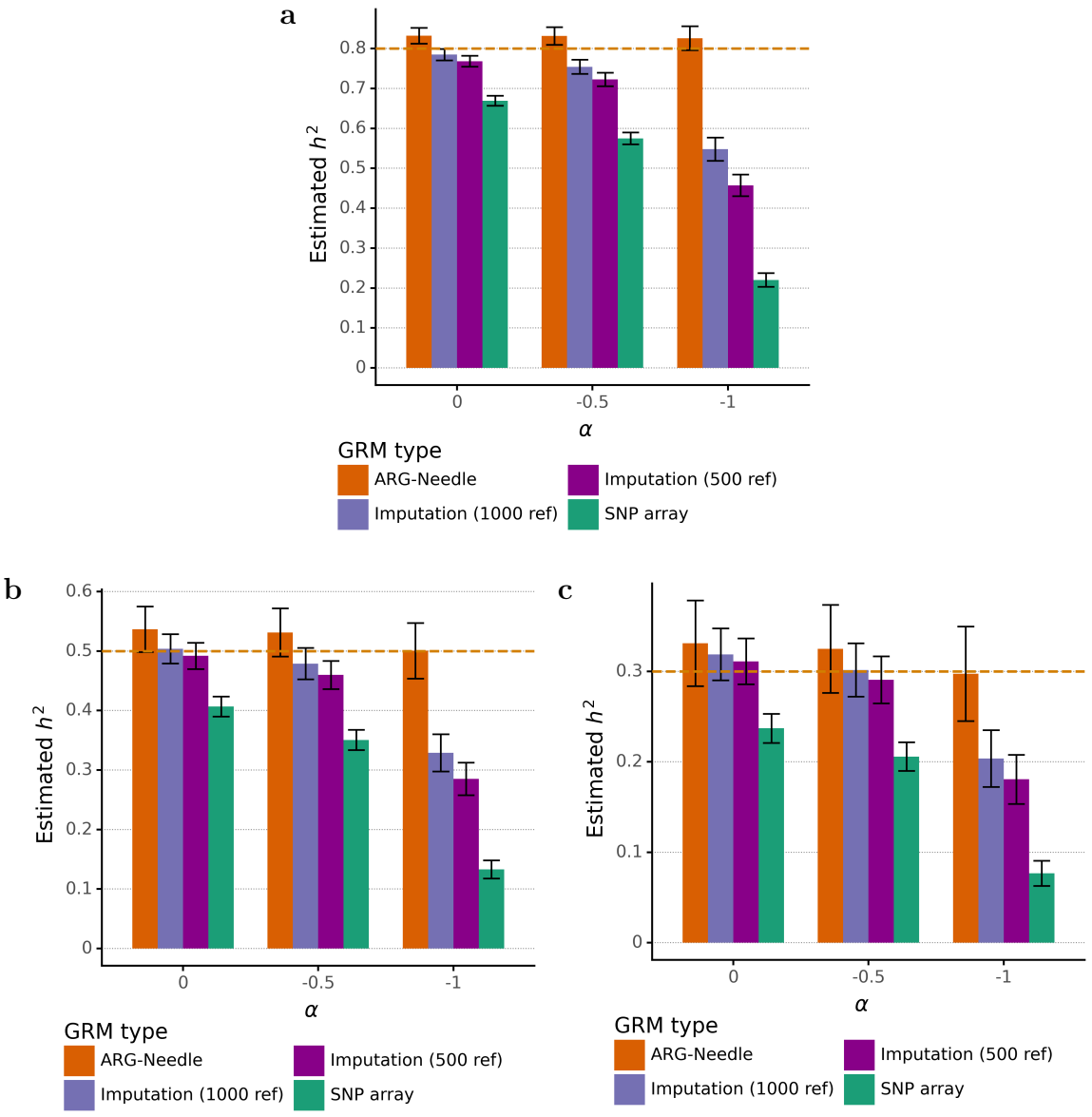


Figure 5.7: **Heritability estimation using ARG-Needle and ARG-GRMs. a**. Heritability estimation using ARG-GRMs from ARG-Needle inference on SNP array data, compared to using imputed or array SNPs (5 simulations of 25 Mb, 5,000 haploid samples, $h^2 = 0.8$, and $\alpha \in \{0, -0.5, -1\}$). **b-c**. As in **a** but with $h^2 = 0.5$ (**b**) and $h^2 = 0.3$ (**c**). Error bars represent 2 s.e. from meta-analysis. *ref* indicates the number of haploid reference samples used for imputation.

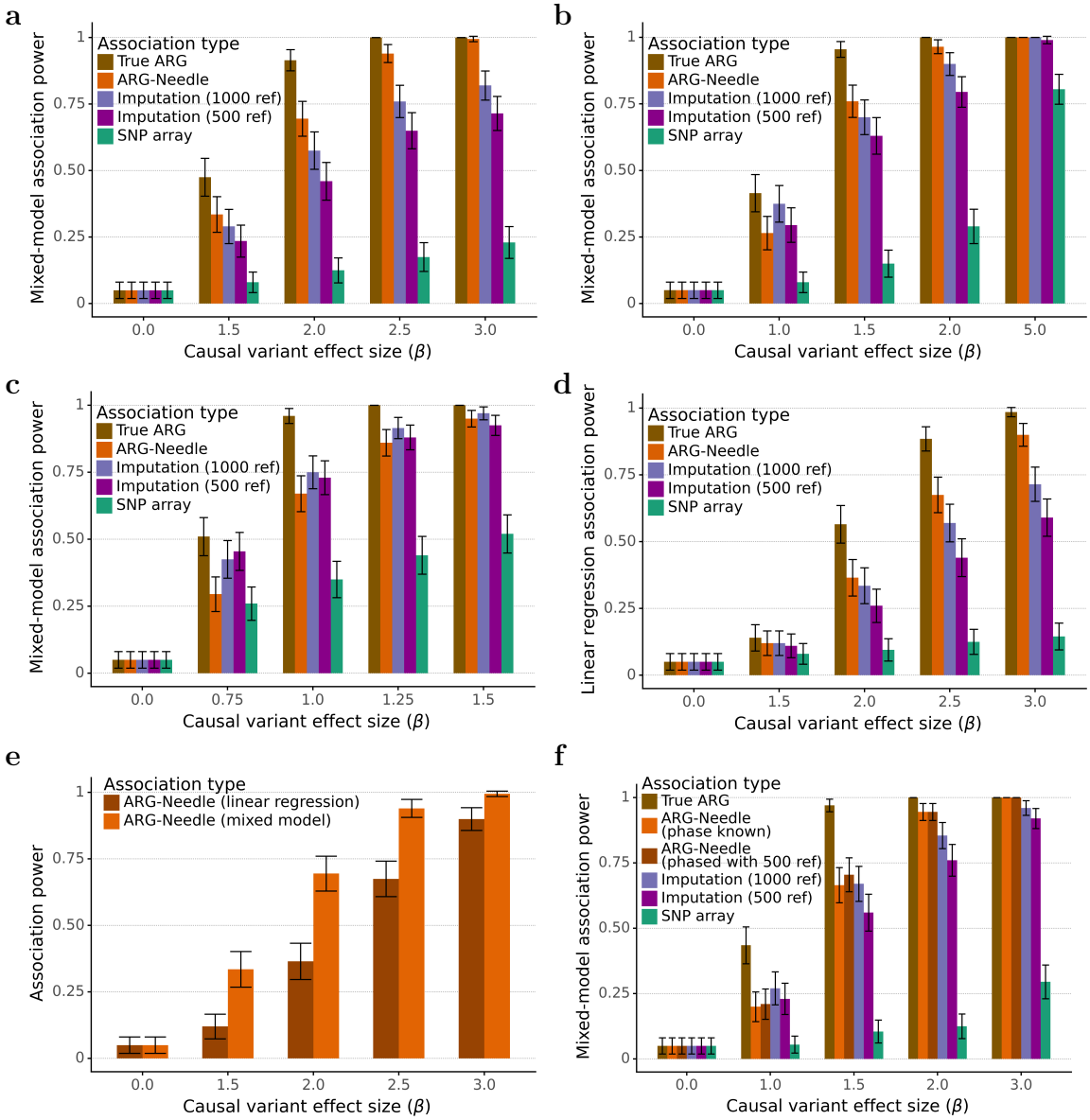


Figure 5.8: **Simulations of ARG-MLMA genealogy-wide association power. a**. Power to detect a low-frequency causal variant (MAF = 0.25%) in simulations of a polygenic phenotype. We compare ARG-MLMA of ground-truth ARGs and ARG-Needle inferred ARGs with MLMA of imputed and SNP array variants as we vary the effect size β (200 independent simulations of $h^2 = 0.8, \alpha = -0.25$, $N = 2{,}000$ haploid samples, and 22 chromosomes of 0.5 Mb each, see Section 5.2.3). **b**,**c**. Similar to **a**, except with the causal variant MAF chosen to be 0.5% (**b**) and 1% (**c**) instead of 0.25%. **d**. Similar to **a**, except using linear regression instead of the linear mixed model to test for association. In **a**-**d**, *ref* indicates the number of haploid references used for imputation. **e**. We combine the association power results of ARG-Needle association from **a** and **d**. highlighting the improvement of ARG-MLMA compared to directly testing ARG clades using linear regression. **f**. As in **a**, but with diploid instead of haploid individuals (*ref* indicates the number of diploid references used for phasing or imputation). ARG-Needle is run with the true phase known and with genotypes phased using 500 references; note that ARG-MLMA yielded a slightly more stringent significance threshold when phase is known.

ARG-MLMA on this data. ARG-MLMA power was not affected when we compared between the conditions with and without the Beagle phasing procedure. Both conditions performed similarly to imputation from 500 or 1,000 diploid references, where phase was assumed to be known during imputation.

Overall, these experiments suggest that accurate genealogical inference combined with linear mixed models allows increasing association power by testing variants that are not well tagged using available markers while modelling polygenicity. The ARG may also be utilised to obtain improved estimates of genomic similarity, to the benefit of additional LMM-based complex trait analyses.

5.3.3 ARG-Needle and ARG-based imputation

ARG-Needle can be used to build an ARG using a combination of sequencing and array samples. To evalute this feature, we performed a preliminary study of ARG-based imputation (Fig. 5.9). We compared imputation with ARG-Needle to IMPUTE4, Beagle 5, and imputation using the true ARG (see Section 5.2.4). All methods used the same data, consisting of a 10 Mb region, 1000 haploid array samples, and either 300 or 1000 haploid sequencing samples. ARG-Needle outperformed Beagle 5 for the case of 300 haploid references, though this may be due to approximations used in Beagle 5 that limit accuracy in small data sets. ARG-Needle showed particular promise for rare variant imputation, a setting where imputation is less accurate (see discussion in Section 7.2.1).

5.4 Discussion

In this chapter, we performed extensive analyses of the ARG-Needle and ASMC-clust methods in simulation. First, we briefly discuss our ARG-based analyses, which we will summarise and discuss more extensively in Chapter 7. We demonstrated that ARGs inferred from array data or a combination of array and sequencing data could be used to perform heritability estimation (Fig. 5.7), mixed-model association (Fig. 5.8), and genotype imputation (Fig. 5.9). Our results indicate that ARG-Needle

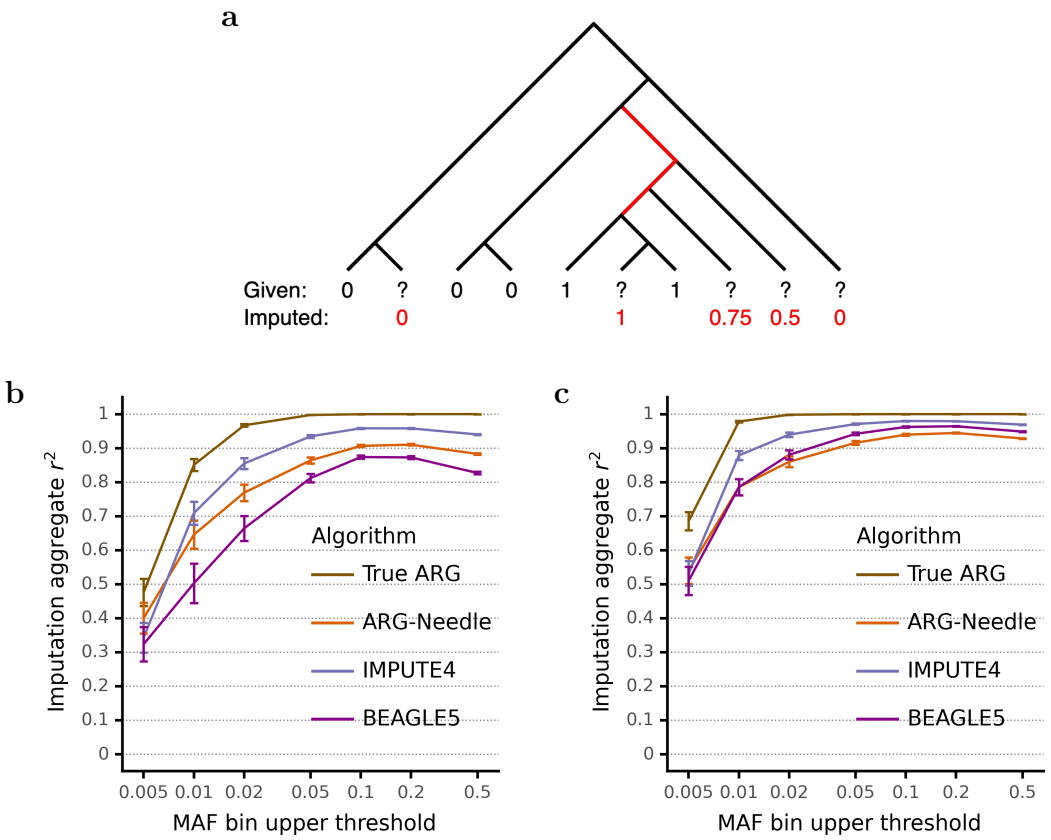


Figure 5.9: **Joint array and sequencing ARG inference for genotype imputation. a**. Given a polymorphic sequenced site containing sequenced samples, unobserved genotypes for array samples, and a marginal coalescent tree relating all samples, we perform genotype imputation as follows. We first identify all branches in the tree for which a mutation on that branch best explains the observed sequencing data in terms of Hamming distance (red branches in the example). Each branch implies genotypes of 0 or 1 for the array samples, and we weight by branch length to produce a weighted predicted dosage for each array sample. In this example, the three branches have lengths in ratio 1:1:2, resulting in the predicted dosages shown in red. **b-c**. We perform ARG-based imputation using the true ARG and an ARG-Needle inferred ARG and compare to IMPUTE4 and Beagle 5. Simulations use a 10 Mb region, 300 (**b**) or 1000 (**c**) haploid sequencing samples, and 1000 haploid array samples. For IMPUTE4 and Beagle 5, we input a genetic map corresponding to the recombination rate used for simulation, and otherwise use default parameters. Variants are binned by MAF in the sequencing samples, and we report the aggregate r^2 within each bin (mean over 25 runs), with each bin represented by its maximal MAF. Error bars represent 2 s.e.

discovers clades capturing genetic sharing of variants not included in the data used for inference. For instance, ARG-GRMs built using ARGs inferred by ARG-Needle in array data captured more phenotypic variance than GRMs built using array data, consistent with results observed for ARG-MLMA. We also demonstrated that ARG-Needle and ARG-MLMA can be used for diploid genotypes, including when data must first be phased (Fig. 5.8f). Although we designed our simulations to realistically capture aspects of complex traits in large biobanks (as discussed in Chapter 3, see Section 3.4), the simulations in this chapter remain of a relatively small scale. In the next chapter, we focus on our ARG-Needle and ARG-MLMA methods, scaling them up to perform real data analyses in a subset of 337,464 UK Biobank samples.

Turning to our comparisons with other ARG inference methods, Relate [Speidel et al., 2019] and `tsinfer` [Kelleher et al., 2019], we observed the following overarching themes. In terms of scalability, ARG-Needle and `tsinfer` both improve considerably upon ASMC-clust and Relate, which are quadratic in time complexity (Fig. 5.1e,f). ARG-Needle scaled similarly to `tsinfer` up to $N = 32{,}000$, and for even larger sample sizes, where hashing dominates runtime, the scalability of ARG-Needle can be maintained by gradually increasing the hash word size (see Section 4.2.2.2). Across a variety of accuracy metrics, ARG-Needle tended to achieve best performance in array data, sometimes tied or in close performance with ASMC-clust or Relate. While ARG-Needle threads one sample at a time using a fast coalescence-based approach, ASMC-clust performs joint modelling of all samples by clustering pairwise coalescence times, and tended to slightly outperform ARG-Needle on the TMRCA RMSE metric (Figs. 5.1c and 5.3c,d). We therefore recommend ASMC-clust for smaller array data sets where scalability is less important.

We relied on several existing metrics in our comparisons: Robinson-Foulds [Robinson and Foulds, 1981], KC distance [Kendall and Colijn, 2016], and pairwise TMRCA RMSE. We also developed a new metric called the ARG total variation distance that generalises Robinson-Foulds to take into account branch lengths. As we are interested in performing complex trait analysis using inferred ARGs, we desired a metric that captured the notion of mutational similarity between two ARGs. Furthermore, both

our applications of ARG-GRMs and ARG-MLMA take into account the length of branches.[5] The ARG total variation distance combines both these considerations.

Every metric has its limitations. For instance, the ARG total variation distance relies on a hard 0-1 loss in its definition: it does not consider whether two mutations are correlated (with small Hamming distance) or occur at close but disjoint positions in two ARGs, a limitation it shares with the Robinson-Foulds distance. It may be possible to instead use a Wasserstein distance, which generalises the total variation distance with the aid of a metric on the probability space, thus allowing "margin for error" with a smooth loss. As another example, we noted that while ARG normalisation improved the calibration of node times, it hampered the pairwise TMRCA RMSE performance of ARG-Needle and ASMC-clust in array data (Fig. 5.6a-d). It may be worthwhile to develop additional metrics that incorporate pairwise TMRCA values without being as influenced by the root event or large values (see Section 5.3.1.2), for instance using an L1 norm or measuring RMSE using log TMRCA or the square root of TMRCA values. We note, however, that accuracy on the pairwise TMRCA RMSE is connected to several analyses discussed in this work in the context of ARG-GRMs. In particular, increased performance for pairwise TMRCA RMSE reflects increased similarity (under the Frobenius norm) between true and inferred ARG-GRMs (when we assume $\alpha = 0$, see Section 3.2.1).

Lastly, we found that the KC distance favours trees with polytomies, such that comparisons between different methods should take care in how polytomies are handled. We randomly resolved polytomies as done by [Kelleher et al., 2019], and also developed two strategies for collapsing branches to form polytomies. We briefly speculate on why collapsing branches to form polytomies may improve performance on the KC distance. The KC distance compares trees by computing the L2 norm between the two vectors $\mathbf{m}_{T_1}$ and $\mathbf{m}_{T_2}$, which will penalise large differences (see Section 5.2.2.3). Take T_1 to be the true tree and T_2 to be the inferred tree. In our coalescent simulations, T_1 is a relatively well-balanced binary tree, so the values of $\mathbf{m}_{T_1}$ will

[5] In the case of ARG-MLMA, we devised a strategy to sample mutations for testing with a rate μ, which we applied in our real data analysis (see Chapter 6).

range from 0 (a pair of samples that coalesces at the root) to $O(\log N)$, with a mode and mean of $O(1)$ (as a large fraction of pairs coalesce at the root). The introduction of polytomies may result in shrinkage of the values of $\mathbf{m}_{T_2}$, because the paths from MRCA nodes to the root contain fewer intermediate nodes, introducing bias but reducing variance. This may reduce the overall MSE, as seen in other shrinkage estimators such as ridge regression. Additional work may be needed to further understand the connections between the KC distance and properties of inferred ARGs that perform well under this metric, including their performance in downstream analyses.

Chapter 6

Inference and Analysis of Ancestral Recombination Graphs in the UK Biobank

6.1 Overview

Whereas Chapter 5 evaluated the methods introduced in Chapters 3-4 in simulations, in this chapter we describe an application of our ARG inference and mixed-model analysis methods to real data from the UK Biobank.[1] Using ARG-Needle, we built the genome-wide ARG from genotyping array data for 337,464 unrelated White British individuals in the UK Biobank. We applied our ARG-MLMA framework genealogy-wide to all 22 autosomal chromosomes, scanning for associations with 7 complex traits across common, rare, and ultra-rare allele frequencies. In this chapter, we describe extensions to the methods of Chapter 3 to enable these analyses at scale. We report extensive results comparing our approach to association of imputed genotypes from ∼65K haplotypes, where we used whole exome sequencing data of ∼138K individuals in the UK Biobank to validate our associations. We also report additional comparisons to GIANT Consortium summary statistics of ∼700K samples [Yengo et al., 2018] and

[1] Analyses in this chapter were performed in collaboration with Pier Palamara and Arjun Biddanda; see Section 6.5 for a contribution statement.

association of variants imputed from whole exome sequencing of ∼50K UK Biobank individuals [Barton et al., 2021].[2]

6.2 Methods

6.2.1 ARG-Needle inference in the UK Biobank

We performed ARG-Needle inference in the UK Biobank as follows. Starting from 488,337 samples and 784,256 available autosomal array variants (including SNPs and short indels), we removed 135 samples (129 withdrawn, 6 due to missingness $> 10\%$) and 57,126 variants (missingness $> 10\%$). We phased the remaining variants and samples using Beagle 5.1 [Browning and Browning, 2007b] and extracted the subset of 337,464 unrelated White British samples reported in [Bycroft et al., 2018]. We built the ARG of these samples using ARG-Needle, using hashing parameters of $K = 64$ relatives chosen, tolerance $T = 1$ (allowing one mismatch in an otherwise IBS stretch), hashing region size $L = 0.5$ cM, and a flexible primary and second hash word size S_1 and S_2 (see Section 4.2.2). We set $S_1 = 16$ bits and $S_2 = 8$ bits when threading the first 50K individuals and set $S_1 = 64$ bits and $S_2 = 16$ bits for the remaining 287K individuals. We parallelised the ARG inference by splitting phased genotypes into 749 non-overlapping "chunks" of approximately equal numbers of variants. We added 50 variants on either side of each chunk to provide additional context for inference and independently applied ARG normalisation on each chunk.

6.2.2 ARG-MLMA methods for real data

Our ARG-MLMA analyses in the UK Biobank differed from our simulation analyses (see Section 3.2.4) in two ways. First, instead of testing all mutations in the ARG, we sampled variants using a mutation rate $\mu = 10^{-5}$, also adding variants sampled with $\mu = 10^{-3}$ to locus-specific Manhattan plots to gain further insights. This improves computational efficiency while prioritising branches of the ARG that are more certain

[2]The remainder of this chapter is expanded and adapted from [Zhang et al., 2021].

(see Section 3.2.4). Second, to achieve greater scalability and power than through using GCTA, we leveraged the BOLT-LMM software package [Loh et al., 2015b, Loh et al., 2018] (see Section 2.3.3.3) as follows. For each phenotype, we first regressed out covariates, then ran BOLT-LMM on SNPs from all chromosomes to extract BOLT-LMM's calculated calibration factor. We additionally ran BOLT-LMM 22 times, once with each chromosome excluded, and extracted estimated prediction effects (`--predBetasFile` flag) to form LOCO polygenic predictors. We then obtained LOCO residuals by subtracting these LOCO predictions from the processed phenotype. We finally used ARG-Needle to test clades of the ARG for association against these residuals, traversing the ARG and calculating BOLT-LMM's non-infinitesimal test statistics for each clade of interest. For this last step, we include runtime optimisations for sparse clades, so that a sparse clade can be tested in fewer than $O(N)$ operations by only accessing the phenotype values for the samples in the clade. We incorporate the calibration factor which corrects for inflated or deflated statistics, and we use BOLT-LMM's option to perform non-infinitesimal modelling for increased power.

6.2.3 Computation of permutation-based significance thresholds

We used a permutation-based approach to establish genome-wide significance thresholds corresponding to a family-wise error rate of 0.05 (see [Zhang et al., 2021] Supplementary Table 1). In detail, we ran 1,000 null simulations using random phenotypes drawn from a standard normal distribution, performed univariate linear regression against imputed or ARG data [Churchill and Doerge, 1994, Minichiello and Durbin, 2006, Kanai et al., 2016], and computed the minimum p-value. We then estimated the genome-wide significance threshold using the most significant 5% quantile of these minimum p-values. We verified that performing this analysis using either the entire genome, chromosome 1, or the first chunk of the genome produced compatible results in a limited number of settings. To reduce computational costs, we thus performed these analyses using a subset of the genome and extrapolated to the whole genome.

When a MAF cut-off was applied, significance thresholds remained compatible across several choices of sample size. We thus used $50,000$ haploid samples to estimate significance thresholds when MAF filtering was applied. Significance thresholds for several choices of filtering parameters are reported in [Zhang et al., 2021] Supplementary Table 1; specific thresholds used in individual analyses are detailed below.

6.2.4 Data pre-processing and filtering

To process phenotypes (standing height, alkaline phosphatase, aspartate aminotransferase, LDL/HDL cholesterol, mean platelet volume, and total bilirubin) we first stratified by sex and performed quantile normalisation. We then regressed out age, age squared, genotyping array, assessment centre, and the first 20 genetic principal components computed in [Bycroft et al., 2018]. We built a non-infinitesimal BOLT-LMM mixed model using SNP array variants, then tested HRC+UK10K imputed data [Huang et al., 2015, McCarthy et al., 2016, Bycroft et al., 2018] and variants inferred using the ARG (ARG-MLMA, described above). For association of imputed data (including SNP array) we restricted to variants with Hardy-Weinberg equilibrium $p > 10^{-12}$, missingness < 0.05, and info score > 0.3 (matching the filtering criteria adopted in [Bycroft et al., 2018]). For all analyses we did not test variants with a minor allele count (MAC) < 5 and used minor allele frequency (MAF) thresholds detailed below.

6.2.5 Rare and ultra-rare variant association analysis

Using filtering criteria above and no additional MAF cutoff, we obtained genome-wide permutation significance thresholds of $p < 4.8 \times 10^{-11}$ (95% CI: $[4.06 \times 10^{-11}, 5.99 \times 10^{-11}]$) for ARG and $p < 1.06 \times 10^{-9}$ (95% CI: $5.13 \times 10^{-10}, 2.08 \times 10^{-9}]$) for imputation. After performing genome-wide MLMA for the 7 traits, we selected genomic regions harboring rare ($0.01\% \leq \text{MAF} < 0.1\%$) or ultra-rare ($\text{MAF} < 0.01\%$) variant associations. We then formed regions by grouping any associated variants within 2 Mb of each other and adding 1 Mb on either side of the leftmost and rightmost variant

in each of these groups. We next performed several filtering and association analyses to extract sets of approximately independent signals, using a procedure similar to that of [Barton et al., 2021], using PLINK (v1.90b6.21) with specified flags for several of these steps. For each region, we extracted hard-called raw genotypes for all genome-wide significant signals from either ARG or imputed data, tested for association (`---assoc` flag), and performed two-stage LD-clumping of the variants. The first clumping step used parameters `--clump-p1 0.0001 --clump-p2 0.0001 --clump-r2 0.5 --clump-kb 10`; the second used same parameters except for `--clump-kb 100000`. For each variant i, we considered each other variant j and computed the approximate chi-squared statistic that would be obtained by including j as covariate [Yang et al., 2012, Barton et al., 2021]:

$$\chi^2_{i|j} = \chi^2_i \left(1 - \text{sign}(\beta_i\beta_j) r_{ij} \sqrt{\chi^2_j / \chi^2_i}\right)^2,$$

where χ^2_i and χ^2_j denote the respective chi-square statistics and $\text{sign}(\beta_i\beta_j)$ is 1 if the effect sizes for the two variants have the same sign, -1 otherwise. LD was computed using the `--r` flag. We only retained variants i such that $\chi^2_{i|j}$ remained significant for all choices of j. We refer to the set of variants remaining after these filtering steps as approximately independent ("independent" for short, reported in [Zhang et al., 2021] Supplementary Tables 2-3). Of the 7 phenotypes, total bilirubin did not yield any rare or ultra-rare independent signals and height did not yield any independent ultra-rare signals.

We next leveraged the UK Biobank whole exome sequencing (WES) data to validate and localise independent associations. We extracted 138,039 exome sequenced samples that overlap with the analysed set of White British individuals and performed lift-over of exome sequencing positions from genome build hg38 to hg19. We then computed pairwise LD (`--r` flag) between the set of independent associated variants and the set of all WES variants. The "WES partner" of an independent variant was selected to be the WES variant with largest r^2 to it. For each pair of independent ARG or imputation signals with their WES partners, we computed the distance to

the WES partner, the LD with the WES partner, the values of association $\hat{\beta}$ (with standard error), and the confusion matrix of genotype overlap between the independent signal and the WES partner, which we used to determine precision and recall of predicting the carriers in the WES variant (reported in [Zhang et al., 2021] Supplementary Tables 2 and 3). Because the BOLT-LMM non-infinitesimal model does not produce $\hat{\beta}$ estimates, we instead obtained them using PLINK (`---assoc` flag).

Variant annotations for the WES partners used in Fig. 6.1c were obtained using the Ensembl Variant Effect Predictor (VEP) tool [McLaren et al., 2016]; variants reported by VEP as either of `frameshift_variant`, `splice_acceptor_variant`, `splice_donor_variant`, `stop_gained` were classed as loss-of-function (LoF) and variants reported as either of `stop_lost`, `start_lost`, `missense_variant`, `inframe_deletion`, `inframe_insertion` were classed as missense. Gene annotations for each WES variant were also obtained from VEP, where we allow for each WES partner to be annotated with multiple genes if they overlap the WES variant position. We used the WES variant position to check against the results of [Barton et al., 2021] to assess whether the variants we detected were present in the summary statistics, marginally significant ($p < 5 \times 10^{-8}$), or reported as likely causal (defined in [Barton et al., 2021]).

6.2.6 Association analysis for higher frequency variants

For genome-wide association analyses of higher frequency variants and height, we matched filtering criteria used in [Bycroft et al., 2018], retaining imputed variants that satisfy the basic filters listed above, as well as MAF $\geq 0.1\%$. Using these criteria, we estimated a permutation-based genome-wide significance threshold of 4.5×10^{-9} (95% CI: [2.2×10^{-9}, 9.6×10^{-9}], see [Zhang et al., 2021] Supplementary Table 1). To facilitate direct comparison, we aimed to select parameters that would result in a comparable significance threshold for the ARG-MLMA analysis. Two sets of parameters satisfied this requirement: 3.4×10^{-9} (95% CI: [2.4×10^{-9}, 5×10^{-9}]), obtained for $\mu = 10^{-5}$, MAF $\geq 1\%$; and 4×10^{-9} (95% CI: [3.1×10^{-9}, 5.3×10^{-9}]), obtained for $\mu = 10^{-6}$, MAF $\geq 0.1\%$. We selected the former set of parameters, as

a low sampling rate of $\mu = 10^{-6}$ leads to worse signal-to-noise and lower association power. We thus used a significance threshold of $p < 3 \times 10^{-9}$ for all analyses of higher frequency variants. We formed Manhattan plots for all chromosomes (Figs. 6.4a-b and 6.5) as well as for significant loci of interest. For the loci of interest, we presented the results of association over ARG clades sampled with $\mu = 10^{-3}$ in addition to those sampled with $\mu = 10^{-5}$ to leverage additional information about the haplotype structure in the region (Fig.6.4e-f).

To perform **conditional and joint (COJO) analyses** [Yang et al., 2012] (Fig. 6.4c-d, Fig. 6.3e) we first performed LD clumping of associated variants using PLINK (v1.90b6.21) with flags `--clump-p1 0.0001 --clump-p2 0.0001 --clump-r2 0.5 --clump-kb 1000` for all data types. For ARG data we also pre-processed each of the 749 non-overlapping chunks by running LD-clumping with the same parameters but with a reduced `--clump-kb 100`. The GCTA software implementing the COJO procedure requires effect size estimates, which are not produced by the BOLT-LMM non-infinitesimal model. We therefore extracted genotype variants with MLMA $p < 5 \times 10^{-7}$ for array data and $p < 5 \times 10^{-8}$ for imputed and ARG data and performed further association analysis using PLINK (v2.00a3LM, `--glm` flag). To create merged data sets (e.g. ARG + SNP array data), we started from the variants that were selected from the LD clumping step and merged them using PLINK. Because imputed data contains genotype array markers, we first removed any overlapping markers from the set of LD-clumped imputed variants in cases where both imputed and array data were merged and separately considered. COJO analyses were performed using GCTA (v1.93.2) using the `--cojo-slct` flag and COJO $p < 3 \times 10^{-9}$.

6.3 Results

6.3.1 Genealogy-wide association scan of rare and ultra-rare variants in the UK Biobank

Using ARG-Needle, we built the genome-wide ARG from genotyping array data for 337,464 unrelated White British individuals in the UK Biobank (see Section 6.2). We performed ARG-MLMA for standing height and 6 molecular traits, comprising alkaline phosphatase, aspartate aminotransferase, LDL/HDL cholesterol, mean platelet volume, and total bilirubin. To scale this analysis to the entire data set, we built on a recent method for large-scale MLMA [Loh et al., 2015b, Loh et al., 2018], which uses an array-based GRM to model polygenicity (see Section 6.2). We compared ARG-MLMA to standard MLMA testing of variants imputed using the combined Haplotype Reference Consortium (HRC) and UK10K reference panels (hereafter HRC+UK10K) [Huang et al., 2015, McCarthy et al., 2016, Bycroft et al., 2018], comprising ∼65K haploid samples. We filtered the imputed variants using standard criteria and focused on rare ($0.01\% \leq \text{MAF} < 0.1\%$) and ultra-rare ($\text{MAF} < 0.01\%$) genomic variants. We used a permutation-based approach [Kanai et al., 2016] to establish genome-wide significance thresholds of $p < 4.8 \times 10^{-11}$ for ARG variants (sampled with mutation rate $\mu = 10^{-5}$) and $p < 1.06 \times 10^{-9}$ for imputed variants and performed extensive LD-based filtering to extract a stringent set of approximately independent associations (hereafter "independent associations") resulting from each analysis (see Section 6.2). To aid the localisation and validation of these independent associations, we leveraged a subset of 138,039 individuals for whom whole exome sequencing (WES) data was also available (we refer to this data set as WES-138K). For each detected independent variant, we selected the exome sequenced variant with the largest correlation, which we refer to as its *WES partner*.

Applying this approach, we detected 133 independent signals using the ARG and 65 using imputation, implicating a set of 149 unique WES partners (see [Zhang et al., 2021] Supplementary Tables 2-3). Of these WES variants, 38 were implicated us-

ing both approaches, confirming a common underlying source for these associations (Fig. 6.1a, for region-level results see Fig. 6.2a). The fraction of WES partners uniquely identified using the ARG was larger among ultra-rare variants (84%) compared to rare variants (40%), reflecting a scarcity of ultra-rare variants in the sequenced HRC+UK10K panel. We observed a strong correlation between the phenotypic effects estimated in the 337K individuals using ARG-derived or imputed associations and those directly estimated for the set of WES partners in the WES-138K data set (Fig. 6.1b), with a stronger correlation (bootstrap $p = 0.002$) for ARG-derived variants (average $r^2_{ARG} = 0.93$) compared to imputed variants (average $r^2_{imp} = 0.79$). Only 73% of the WES partners for ARG-derived rare variant associations were significantly associated (at $p < 5 \times 10^{-8}$) in the smaller WES-138K data set, a proportion that dropped to 59% for ultra-rare variants. Variants detected using genealogy-wide association had a larger average phenotypic effect than those detected via imputation (bootstrap $p < 0.0001$; average $|\hat{\beta}_{ARG}| = 1.45$; average $|\hat{\beta}_{imp}| = 0.90$), reflecting larger effects observed in ultra-rare variants. In addition, the set of WES partners implicated by either ARG or imputation were $\sim$2.3$\times$ enriched for missense variation, and ARG-derived WES partners were $\sim$4.5$\times$ enriched for loss-of-function variation compared to exome-wide variants of the same frequency (bootstrap $p < 0.001$, Fig. 6.1c), supporting their likely causal role.

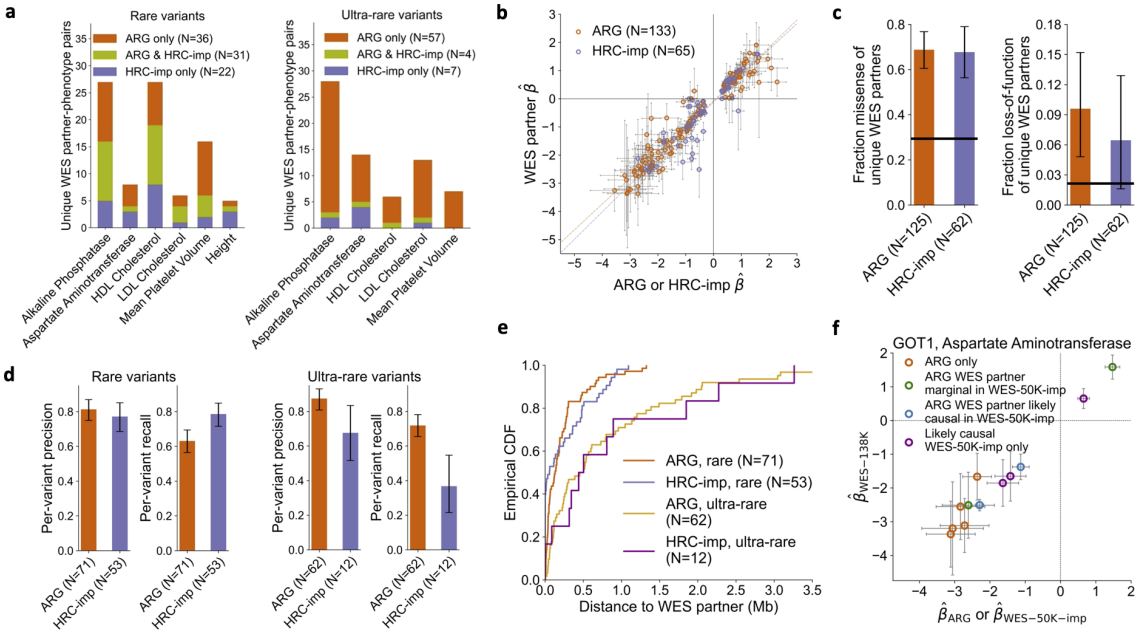


Figure 6.1: **Association of ARG-derived and imputed rare and ultra-rare variants with 7 quantitative traits in UK Biobank. a**. Counts of unique WES partners for ARG and HRC+UK10K imputed ("HRC-imp") independent associations, partitioned by traits and frequency and showing overlap. Total bilirubin was not associated at these frequencies. **b**. Scatter plot of $\hat{\beta}$ (estimated effect) for independent variants against $\hat{\beta}$ for their WES partners, with linear model fit. **c**. Fraction of missense and loss-of-function variants for the unique WES partners of independent variants. Horizontal black lines represent genome-wide averages. **d**. Average per-variant precision and recall of predicting WES carrier status, partitioned by frequency. **e**. Cumulative distribution function for the distance between independent variants and their WES partners. **f**. Scatter plot of $\hat{\beta}$ for ARG-derived independent variants with aspartate aminotransferase in the *GOT1* gene against $\hat{\beta}$ for their WES partners. We colour points based on whether the WES partner is likely causal in WES-50K-imp (imputation from WES-50K [Barton et al., 2021]), not likely causal but marginally significant in WES-50K-imp, or not marginally significant in WES-50K-imp ("ARG only"). We also plot the $\hat{\beta}$ for the additional likely causal variants in WES-50K-imp against the $\hat{\beta}$ in WES-138K. Error bars represent 1.96 s.e. in **b** and **f** and represent bootstrap 95% confidence intervals in **c** and **d**. Additional results are shown in Fig. 6.2.

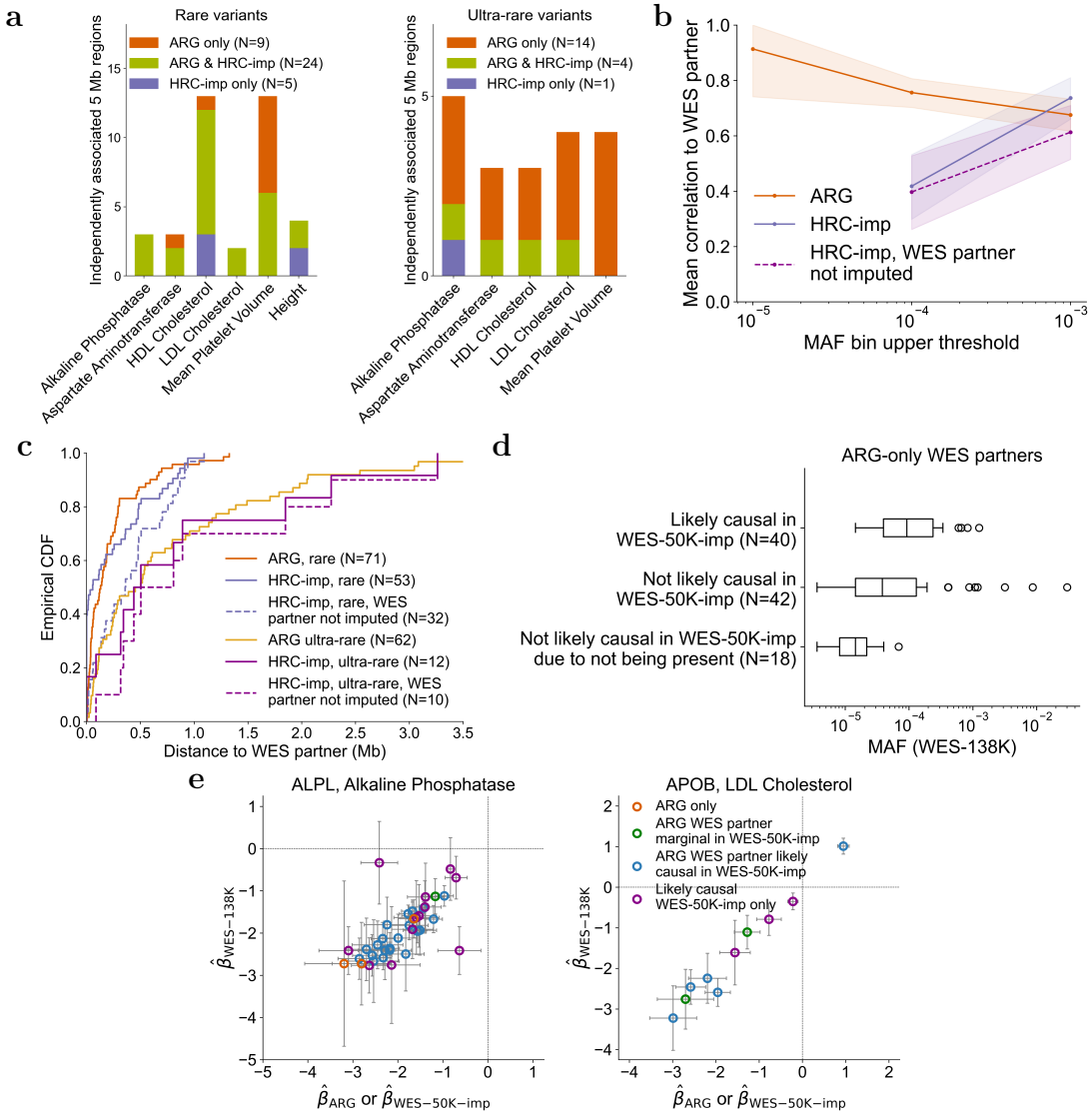

a
Rare variants
ARG only (N=9)
ARG & HRC-imp (N=24)
HRC-imp only (N=5)
Independently associated 5 Mb regions
Alkaline Phosphatase
Aspartate Aminotransferase
HDL Cholesterol
LDL Cholesterol
Mean Platelet Volume
Height
Ultra-rare variants
ARG only (N=14)
ARG & HRC-imp (N=4)
HRC-imp only (N=1)
b
Mean correlation to WES partner
ARG
HRC-imp
HRC-imp, WES partner not imputed
MAF bin upper threshold
c
Empirical CDF
ARG, rare (N=71)
HRC-imp, rare (N=53)
HRC-imp, rare, WES partner not imputed (N=32)
ARG ultra-rare (N=62)
HRC-imp, ultra-rare (N=12)
HRC-imp, ultra-rare, WES partner not imputed (N=10)
Distance to WES partner (Mb)
d
ARG-only WES partners
Likely causal in WES-50K-imp (N=40)
Not likely causal in WES-50K-imp (N=42)
Not likely causal in WES-50K-imp due to not being present (N=18)
MAF (WES-138K)
e
ALPL, Alkaline Phosphatase
APOB, LDL Cholesterol
ARG only
ARG WES partner marginal in WES-50K-imp
ARG WES partner likely causal in WES-50K-imp
Likely causal WES-50K-imp only

Figure 6.2: **Further results for rare and ultra-rare variant associations. a**. Counts of implicated 5 Mb regions containing ARG and HRC+UK10K imputation ("HRC-imp") independent associations, partitioned by traits and frequency and showing overlap. Total bilirubin was not associated at these frequencies. **b**. Average Pearson correlation between independent variants and their WES partners as a function of frequency, for ARG-derived variants, HRC+UK10K imputed variants, and HRC+UK10K imputed variants for which the WES partner was not the imputed variant. Dots represent the upper end of a frequency range. Shaded areas represent 95% bootstrap confidence intervals. **c**. Cumulative distribution function for the distance between independent variants and their WES partners, partitioned by frequency. As in Fig. 6.1b, but also showing HRC+UK10K imputed variants for which the WES partner was not the imputed variant. **d**. Box plots of MAF for WES partners found by ARG-derived but not HRC+UK10K imputed independent variants (centre line, median; box limits, upper and lower quartiles, whiskers, 1.5x interquartile range; points, outliers), stratifying by status in WES-50K-imp (imputation from WES-50K). **e**. Scatter plot of $\hat{\beta}$ (estimated effect) for ARG-derived independent variants against $\hat{\beta}$ for their WES partners, as in Fig. 6.1f but for associations with alkaline phosphatase in the *ALPL* gene and with LDL cholesterol in the *APOB* gene. We colour points based on whether the WES partner is likely causal in WES-50K-imp, not likely causal but marginally significant in WES-50K-imp, or not marginally significant in WES-50K-imp ("ARG only" in figure). We also plot the $\hat{\beta}$ for the additional likely causal variants in WES-50K-imp against the $\hat{\beta}$ in WES-138K. Error bars represent 1.96 s.e.

We also used the WES-138K data set to measure the extent to which carrying an associated ARG-derived or imputed variant is predictive of carrying the corresponding sequence-level WES partner variant (Fig. 6.1d). We quantified this using variant-level precision and recall statistics (see Section 6.2). ARG-derived and imputed rare variants had similar levels of variant-level precision, while imputation had higher recall (bootstrap $p = 0.0009$). For ultra-rare variants, ARG-derived signals performed better than imputed variants for both precision (bootstrap $p = 0.01$) and recall (bootstrap $p < 0.001$). Similarly, ARG-derived and imputed rare variants provided comparable tagging for their WES partners (Fig. 6.2b). ARG-derived ultra-rare variants, on the other hand, provided stronger tagging compared to imputed ultra-rare variants (average $r_{ARG} = 0.77$, average $r_{imp} = 0.42$, bootstrap $p < 0.001$; average $r_{ARG} = 0.72$ for combined rare and ultra-rare variants). Compared to ARG-derived variants, genotype imputation has the advantage that associated variants that are sequenced in the reference panel may be directly localised in the genome. We found that for 21/53 of rare and 2/12 of ultra-rare independent imputation signals the

WES partner had been imputed (having the same physical position as the associated variant), while the remaining signals likely provide indirect tagging for underlying variants. Comparing ARG-derived and imputed variants in terms of the distance to their WES partners, however, revealed similar distributions (Fig. 6.1e and Fig. 6.2c). This suggests that genealogy-wide associations have the same spatial resolution as associations obtained using genotype imputation in cases where the variant driving the signal cannot be directly imputed, unless sequencing data is available to further localise the signal, as done here using WES partners.

A recent study leveraged exome sequencing data from a subset of ∼50K participants (hereafter WES-50K) to perform genotype imputation for ∼459K European samples, achieving association power equivalent to exome sequencing of ∼250K samples [Barton et al., 2021]. We considered the set of WES partners implicated using ARG-derived independent signals and not using HRC+UK10K imputation, and found that, of these, 14/30 rare and 28/54 ultra-rare variants were also detected as likely-causal associations (at $p < 5 \times 10^{-8}$) in [Barton et al., 2021] (see [Zhang et al., 2021] Supplementary Table 2). For the remaining 42 independent associations that are detected using the ARG but are not reported in [Barton et al., 2021], WES partners are often very rare variants (median MAF $= 3.8 \times 10^{-5}$; Fig. 6.2d) of large phenotypic effect (median $|\beta| = 1.12$). These variants are difficult to impute, even when large reference panels are available: 18/42 such variants were absent or singletons in the WES-50K data set used for imputation or had poor imputation quality score. The set of associations uniquely detected using the ARG often extended allelic series at key genes linked to the analysed traits. For instance, restricting to loss-of-function or missense WES partners for independent ARG signals that are not present or marginally significant in [Barton et al., 2021], 5 novel variants with aspartate aminotransferase are mapped to the *GOT1* gene (Fig. 6.1f) and 3 with alkaline phosphatase are mapped to *ALPL* (Fig. 6.2e). A subset of strong independent associations uniquely detected by the ARG had weak correlation with their WES partners (e.g. a signal for aspartate aminotransferase with $p = 3.2 \times 10^{-39}$, ARG-MAF = 0.00053, WES partner $r = 0.2$, WES-138K MAC = 6, WES-50K MAC = 1). These signals may

tag variants, such as structural or regulatory variation, that are unobserved in the WES-138K data set.

In summary, a genealogy-wide association scan using an ARG inferred from common SNPs revealed more independent rare and ultra-rare associations than a scan performed using genotype data imputed from a reference panel comprising ∼65K sequenced haplotypes. Using genealogy-wide association, we also detected ultra-rare variant associations that were not detected using imputation from a subset of ∼50K exome sequenced participants from the same cohort. ARG-derived associations accurately predicted the effect of underlying sequencing variants as well as the subset of carrier individuals. Leveraging a subset of exome sequenced samples enabled further fine-mapping of several genomic regions implicated using the ARG.

6.3.2 Genealogy-wide association for low and high frequency variants

Lastly, we performed genealogy-wide association for low ($0.1\% \leq \text{MAF} < 1\%$) and high ($\text{MAF} \geq 1\%$) frequency variants. Variants within these frequencies are more easily imputed using reference panels that are not necessarily large and population-specific. Consistent with this, extending our previous analysis to consider low-frequency variants yielded a similar number of independent associations for ARG-derived and HRC+UK10K-imputed variants ($N_{ARG} = 102$, $N_{imp} = 100$, see [Zhang et al., 2021] Supplementary Tables 4-5, Fig. 6.3a-c). Associations detected using the ARG had slightly larger effects compared to those found using imputation (bootstrap $p = 0.045$; average $|\beta_{ARG}| = 0.31$; average $|\beta_{imp}| = 0.27$) but provided lower tagging to their WES partners (bootstrap $p < 0.001$; average $r_{ARG} = 0.56$; average $r_{imp} = 0.73$), reflecting the large fraction (42/100) of WES partners that could be directly imputed.

We hypothesised that although imputation of higher frequency variants is generally accurate, branches in the marginal trees of the ARG may tag underlying causal variants better than the set of available polymorphic markers, revealing complementary signal. To test this, we performed MLMA for height using HRC+UK10K im-

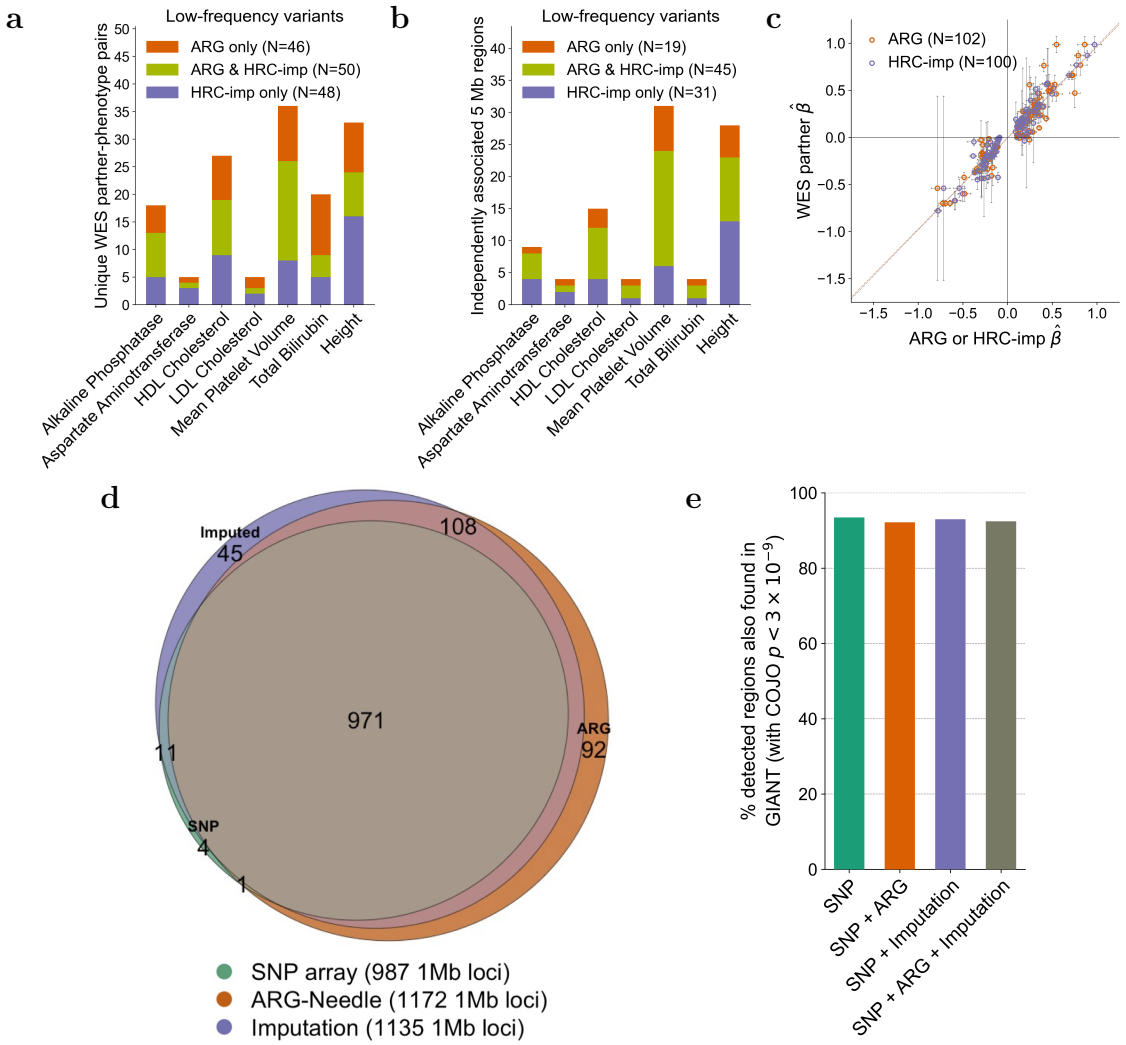


Figure 6.3: **Additional results for low (0.1% ≤ MAF < 1%) and high frequency (MAF ≥ 1%) variant associations. a-c**. Association of ARG-derived and imputed low-frequency variants with 7 quantitative traits. **a**. Counts of unique WES partners for ARG and HRC+UK10K imputed ("HRC-imp") independent associations, partitioned by traits and showing overlap. **b**. Counts of implicated 5 Mb regions containing ARG and HRC+UK10K imputation independent associations, partitioned by traits and showing overlap. **c**. Scatter plot of estimated effect ($\hat{\beta}$) for independent variants against estimated effect for their WES partners, with linear model fit. Error bars represent 1.96 s.e. **d-e**. Association of higher frequency variants with height. **d**. Venn diagram of number of 1 Mb regions containing a significant hit at $p < 3 \times 10^{-9}$ for ARG-Needle ($\mu = 10^{-5}$), HRC+UK10K imputed, and SNP array association. ARG-Needle association detected 971 out of 982 (98.9%) 1 Mb regions found by both imputation and array, 108 out of 153 (71%) 1 Mb regions found by imputation but not array, and an additional 92 (8% increase upon 1140) 1 Mb regions to those already found by imputation and array. **e**. Percent of 1 Mb regions containing independent associations (COJO $p < 3 \times 10^{-9}$) in association scans of 337,464 UK Biobank individuals that were also present in a GIANT consortium meta-analysis of ∼700K samples.

puted variants filtered using the criteria used in [Bycroft et al., 2018], including MAF $> 0.1\%$ and info score > 0.3 (see Section 6.2). For these variants, we established a permutation-based genome-wide significance threshold of 4.5×10^{-9} (95% CI: $[2.2 \times 10^{-9}, 9.6 \times 10^{-9}]$). To facilitate direct comparison, we selected ARG-MLMA parameters that correspond to a comparable genome-wide significance threshold (see Section 6.2). We thus ran ARG-MLMA by sampling mutations from the ARG at rate $\mu = 1 \times 10^{-5}$ and restricting to MAF $> 1\%$, for which we obtained a permutation-based threshold of 3.4×10^{-9} (95% CI: $[2.4 \times 10^{-9}, 5 \times 10^{-9}]$). In downstream analyses, we adopted a significance threshold of 3×10^{-9}.

We assessed the total number of 1 Mb regions that contain an association ($p < 3 \times 10^{-9}$) for either genotype array, imputed, or ARG-derived variants. We found that ARG-MLMA detected 98.9% of regions found by both SNP array and imputation, as well as 71% of regions found by imputation but not array data, and detected an additional 8% of regions not found by either imputation or array data (Fig. 6.3d). A significant fraction (54/92, permutation $p < 0.0001$) of regions identified using the ARG but not HRC+UK10K-imputation contained significant ($p < 3 \times 10^{-9}$) associations in a larger meta-analysis by the GIANT Consortium [Yengo et al., 2018] ($N \approx 700$K), which comprised the UK Biobank and additional cohorts. By inspecting associated loci, we observed that ARG-MLMA captures association peaks and haplotype structure found using genotype imputation but not array data (Figs. 6.4a-b,f, 6.6d-h, and 6.5) as well as association peaks uniquely identified using ARG-MLMA (Figs. 6.4e and 6.6a-c).

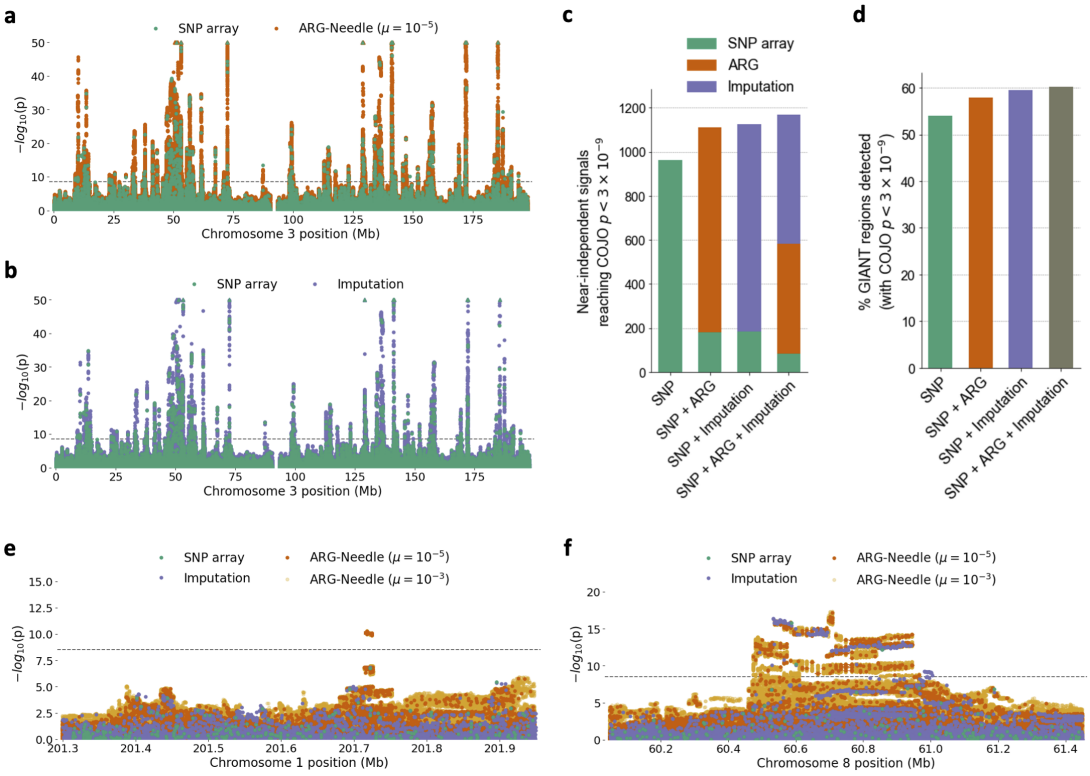
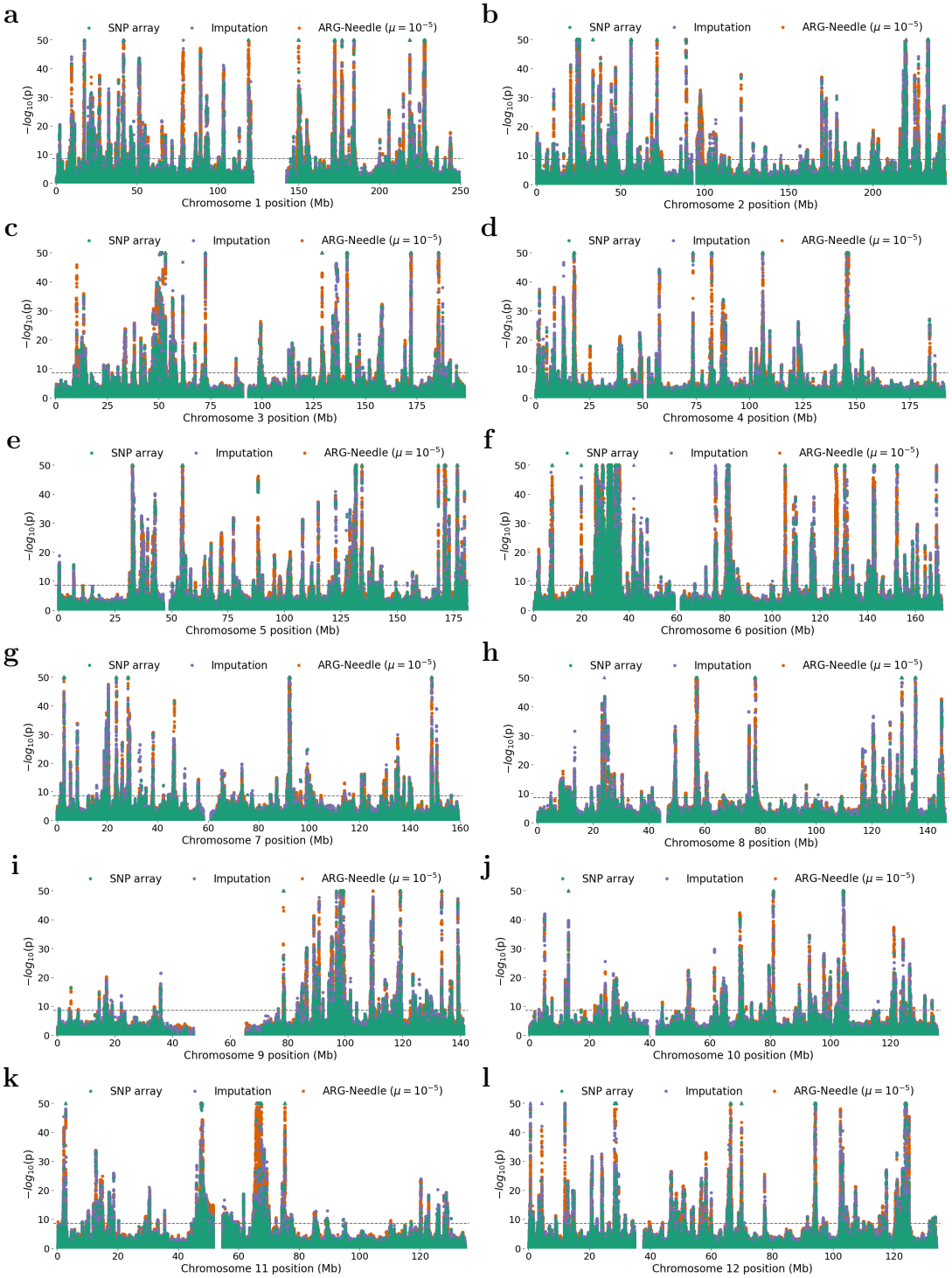


Figure 6.4: **Genealogy-wide association of higher frequency variants with height in UK Biobank. a-b**. Chromosome 3 Manhattan plots showing MLMA of ARG-Needle on SNP array data vs. array SNPs (**a**) and HRC+UK10K imputed variants vs. array SNPs (**b**). **c-d**. Near-independent associations (COJO $p < 3 \times 10^{-9}$) when considering array SNPs alone, array SNPs and ARG-Needle variants, array SNPs and imputed variants, and all three types of variants. **c**. Total number of independent variants found and attribution based on data type. **d**. Percent of 1 Mb regions containing COJO associations in a GIANT consortium meta-analysis of ∼700K samples that are detected using our methods. **e-f**. Manhattan plots of two example loci. **e**. An association peak found by ARG-MLMA that was significant ($p < 3 \times 10^{-9}$) in the GIANT meta-analysis. **f**. ARG-MLMA detects haplotype structure that is found using imputation, while indicating a new association peak. For the Manhattan plots, the order of plotting is ARG-Needle with $\mu = 10^{-3}$ (used for follow-up), then ARG-Needle with $\mu = 10^{-5}$ (used for discovery), then imputation, then SNP array variants on top. Dotted lines correspond to $p = 3 \times 10^{-9}$ (see Section 6.2) and triangles indicate associations with $p < 10^{-50}$. See also Figs. 6.5-6.6.


a
b
c
d
e
f
g
h
i
j
k
l
SNP array
Imputation
ARG-Needle ($\mu = 10^{-5}$)
$-log_{10}$(p)
Chromosome 1 position (Mb)
Chromosome 2 position (Mb)
Chromosome 3 position (Mb)
Chromosome 4 position (Mb)
Chromosome 5 position (Mb)
Chromosome 6 position (Mb)
Chromosome 7 position (Mb)
Chromosome 8 position (Mb)
Chromosome 9 position (Mb)
Chromosome 10 position (Mb)
Chromosome 11 position (Mb)
Chromosome 12 position (Mb)

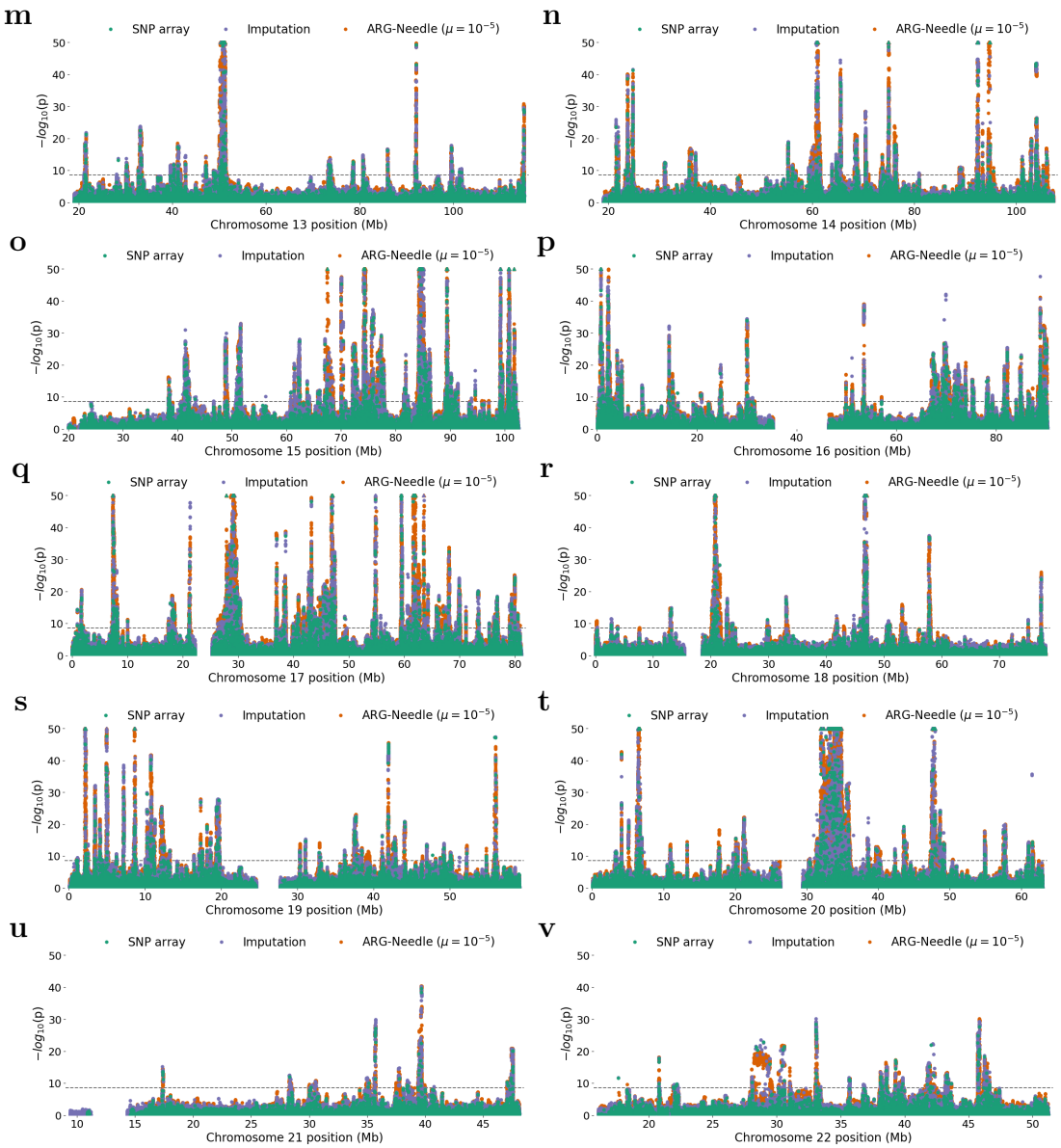
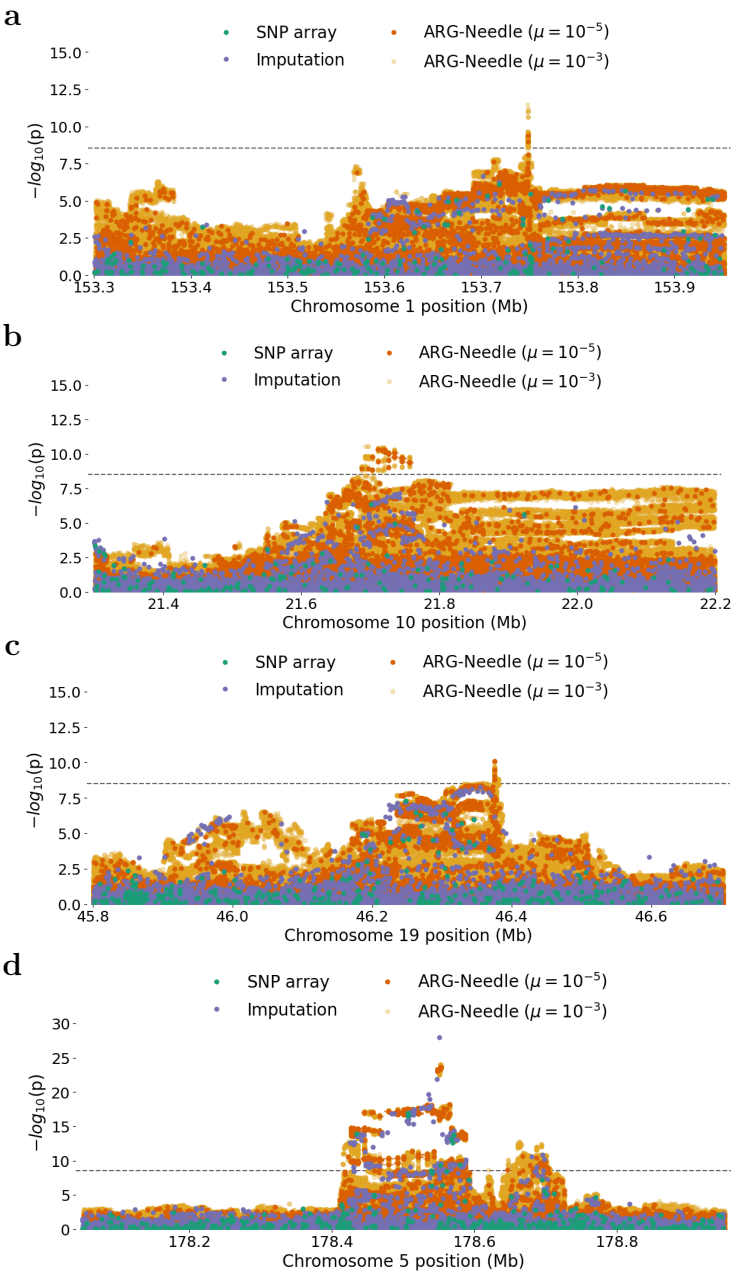


Figure 6.5: (Continued from previous page.) **Additional chromosome-wide Manhattan plots of mixed-model association of higher frequency variants with height.** Manhattan plots showing ARG-Needle, HRC+UK10K imputed variants, and SNP array association, as in Fig. 6.4a-b but with all methods on one plot and for all 22 chromosomes. Dotted lines correspond to $p = 3 \times 10^{-9}$ (see Section 6.2). Triangles indicate associations with $p < 10^{-50}$. The order of plotting is ARG-Needle with $\mu = 10^{-5}$, then imputation, then SNP array variants on top.

We performed LD-based filtering as well as conditional and joint (COJO) association analysis [Yang et al., 2012] (Fig. 6.4c, see Section 6.2) to obtain a set of approximately independent association signals. Analyses including either or both ARG-derived and imputed variants in addition to array markers resulted in an in-


a
SNP array
Imputation
ARG-Needle ($\mu = 10^{-5}$)
ARG-Needle ($\mu = 10^{-3}$)
$-log_{10}$(p)
Chromosome 1 position (Mb)
b
Chromosome 10 position (Mb)
c
Chromosome 19 position (Mb)
d
Chromosome 5 position (Mb)

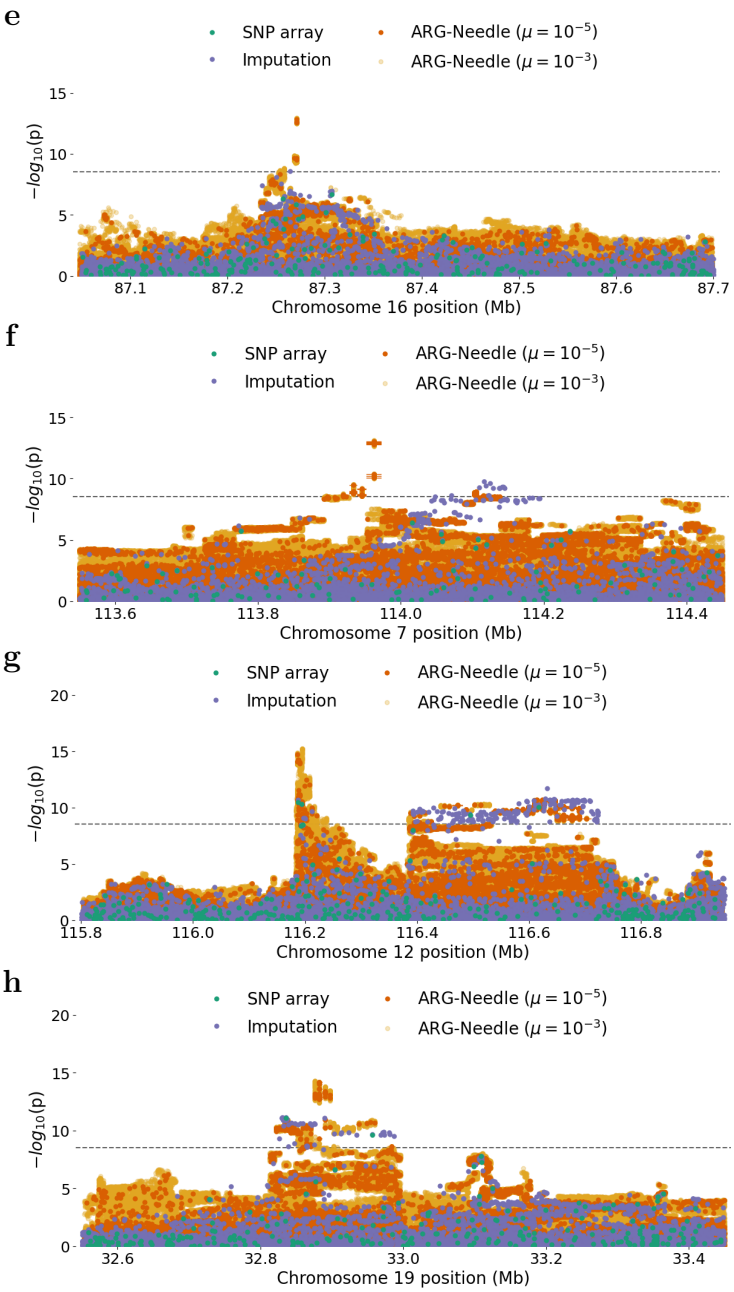


Figure 6.6: (Continued from previous page.) **Manhattan plots of higher frequency loci associated with height.** As in Fig. 6.4e-f but with eight additional loci of interest. **a-c**. Three regions where ARG-Needle ($\mu = 10^{-5}$) alone detects associations passing $p < 3 \times 10^{-9}$ significance. **d-e**. Two loci where ARG-Needle detects an association peak within 10 kb as HRC+UK10K imputation from ∼65K haploid references, despite only using SNP array data. **f-h**. Three loci where ARG-Needle detects a different primary association peak than SNP array or imputed data association. Dotted lines correspond to $p = 3 \times 10^{-9}$ (see Section 6.2). The order of plotting is ARG-Needle with $\mu = 10^{-3}$, then ARG-Needle with $\mu = 10^{-5}$, then imputation, then SNP array variants on top. For the $\mu = 10^{-5}$ ARG associations crossing significance (here and in Fig. 6.4e-f), we additionally plotted a horizontal line showing the extent of the corresponding ARG clade.

crease in the number of significant ($p < 3 \times 10^{-9}$) COJO variants (N_{SNP} = 964, $N_{SNP+ARG}$ = 1,110, $N_{SNP+imp}$ = 1,126, $N_{SNP+ARG+imp}$ = 1,161). The fraction of COJO-associated array markers was substantially reduced by the inclusion of ARG-derived or imputed variants, suggesting that both ARG and imputation provide better tagging of underlying signal than array markers alone. ARG-derived and imputed variants, on the other hand, resulted in comparable proportions of COJO associations when jointly analysed (Fig. 6.4c). We sought to validate this increase in the number of independent signals by leveraging COJO association summary statistics from the GIANT analysis [Yengo et al., 2018]. To this end, we considered the set of 1 Mb regions harboring significant COJO associations and observed that the additional COJO signals detected when including ARG-derived or imputed variants concentrated within regions that also harbor significant ($p < 3 \times 10^{-9}$) COJO signal in the GIANT analysis (Fig. 6.4d and Fig. 6.3e).

In summary, higher frequency variant analysis using the ARG inferred by ARG-Needle from SNP array data revealed associated haplotypes and peaks that were not found through association of array data alone, and complemented genotype imputation in detecting independent association signals.

6.4 Discussion

In this Chapter, we combined the methods of Chapters 3-4 to perform a genealogy-wide association analysis of 7 complex traits in the UK Biobank. We built the ARG of 337,464 genotyped samples from the UK Biobank using ARG-Needle. We leveraged the BOLT-LMM framework to efficiently perform genealogy-wide association using a non-infinitesimal mixed model. We also performed permutation testing to define appropriate genome-wide significance thresholds. To our knowledge, our results represent the first application of genome-wide ARG-based association in a large biobank. In the next and final chapter of this thesis, we will situate these results within the wider context of GWAS and interpret the utility of these methods.

6.5 Contribution Statement

Analyses in this chapter were performed through collaboration between myself (B.C.Z.), Pier Francesco Palamara (P.F.P.), and Arjun Biddanda (A.B.). B.C.Z. and P.F.P. designed and ran ARG inference in the UK Biobank. B.C.Z. and P.F.P performed data analysis on common and low-frequency variant associations. B.C.Z., P.F.P., and A.B. performed data analysis on rare and ultra-rare variant associations.

Chapter 7

Conclusion

7.1 Highlight of Key Contributions

We developed ARG-Needle, a method for accurately inferring genome-wide genealogies from genomic data that scales to large biobank data sets.[1] We performed extensive simulation-based benchmarks, showing that ARG-Needle is both accurate and scalable when applied to ascertained genotyping array and sequencing data. We also developed a framework that combines inferred genealogies with linear mixed models to increase statistical power to detect phenotypic associations for unobserved rare and ultra-rare variants, and showed that this strategy may be utilised in analyses of heritability and polygenic prediction. We applied these new tools to build genome-wide ARGs from genotyping array data for 337,464 UK Biobank individuals and performed a genealogy-wide association scan for height and 6 molecular phenotypes. Using the inferred ARG, we detected more associations to rare and ultra-rare variants of large effect than using genotype imputation from ∼65,000 ancestry-matched sequenced haplotypes, down to a frequency of $\sim 4 \times 10^{-6}$. We validated these signals using 138,039 exome sequencing samples, showing that they strongly tag underlying variants that are enriched for predicted missense and loss-of-function variation. Associations detected using the ARG overlap with and extend fine-mapped associations detected using genotype imputation based on the sequencing of a large fraction

[1] This section follows the Discussion of [Zhang et al., 2021].

of analysed individuals. Applied to the analysis of higher frequency variants, the ARG revealed haplotype structure and independent signals complementary to those obtained using genotype imputation.

7.2 Impact in Context: Three Themes Revisited

A unique feature of our ARG inference method, ARG-Needle, and the analyses we performed in the UK Biobank is that we have only required genotyping array data. While other ARG inference methods do not specially model SNP data, the ASMC Hidden Markov Model considers the distances between SNP variants as well as the ascertainment bias present in array datasets when constructing TMRCA predictions. This allowed us to infer accurate biobank-scale ARGs from array data and uncover associations using the ARG that were missed by association of the original array data and even of imputed variants. Although many whole-genome sequencing efforts are underway, with a few major datasets expected to be released soon [Turnbull et al., 2018, Taliun et al., 2021, Halldorsson et al., 2021], SNP array datasets will continue to be the entry point for studying genomic variation in many ethnic groups [Chen et al., 2011, Nagai et al., 2017, Kurki et al., 2022]. Genealogy-wide association based on ARGs inferred from incomplete genomic data may therefore offer new avenues to better utilise genomic resources for groups that are underrepresented in modern sequencing studies [Martin et al., 2019]. At present, genotype imputation, IBD mapping, and linear mixed models have also been applied to array biobanks to reveal more intricate information about complex traits than single-SNP association testing.[2] To fully examine the impact of the present work, it will be useful to compare our methods to these three themes that were previously reviewed in Section 2.3.

[2]In the case of genotype imputation, a sequenced reference panel is also necesary, but these can come from samples outside of the biobank.

7.2.1 Connections to imputation

Our analyses demonstrate that genealogical inference provides a complementary strategy to genotype imputation approaches, which rely on haplotype sharing between the analysed samples and a sequenced reference panel to extend the set of available markers.[3] Imputation has been successfully applied in the analysis of complex traits [Marchini and Howie, 2010, Yang et al., 2015], but its efficacy, particularly in the study of rare and ultra-rare variants, hinges on the availability of large, population-specific sequencing panels [McCarthy et al., 2016, Barton et al., 2021]. We compared our detected genealogy-wide associations with associations of imputed variants from ∼65,000 references in the HRC+UK10K reference panels, as well as from whole exome imputed variants from 49,960 UK Biobank exome sequences. For rare and ultra-rare variant association, we identified more signals than HRC+UK10K imputation while tagging missense and loss-of-function whole exome sequencing variants. The improvements afforded by our framework were particularly evident in ultra-rare variation (MAF $< 0.01\%$), where 84% of tagged whole exome sequencing (WES) partner variants were identified by the ARG and not HRC+UK10K imputation. For associations in the *GOT1* gene with aspartate aminotransferase, we detected 5 unique WES signals not reported by [Barton et al., 2021] (Fig. 6.1f). These novel signals were all ultra-rare loss-of-function or missense variants of large effect, extending the known allelic series at this locus. Similarly, we found 3 associations with alkaline phosphatase in the *ALPL* gene that were not reported by [Barton et al., 2021] (Fig. 6.2e).

These powerful results for rare and ultra-rare variants can be illuminated by visualising the genotype imputation process using the ARG. In Fig. 7.1a, we have illustrated a locus with a local genealogy; the mutations within this locus can be placed on branches of the genealogical tree (shown as gray dots). Some of these mutations may have biologically significant effects. When we genotype the samples (Fig. 7.1b), we observe common variant mutations (shown in green), but are unable to observe the more recent mutations. Instead, LD is used to tag these mutations using

[3]The beginning of this section is adapted from the Discussion of [Zhang et al., 2021].

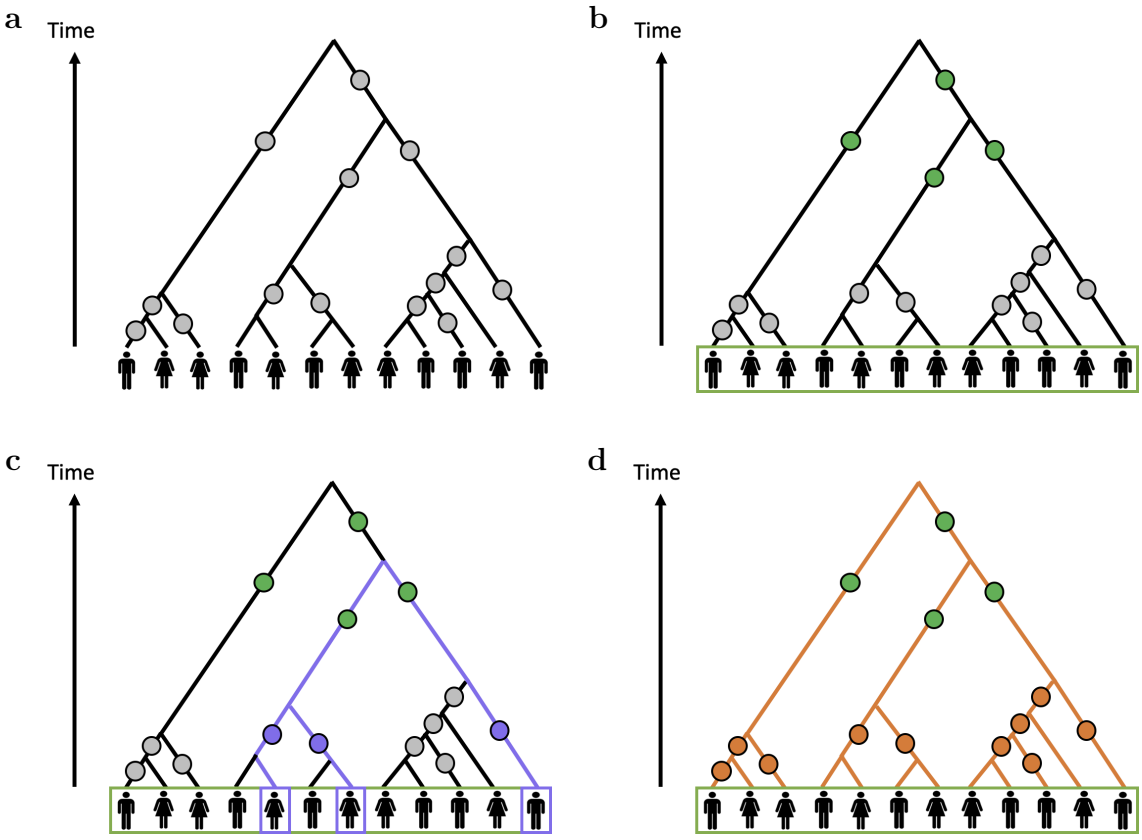


Figure 7.1: **Schematic comparing the SNP array, imputation, and ARG modalities for performing trait associations. a**. Several mutations on a local genealogical tree. **b**. Genotyping the samples will tend to reveal common variants (green), which may tag other variants using LD. **c**. Sequencing several samples (lavender) reveals the mutations they carry, which may be imputed to the array samples. **d**. The ARG allows for testing all branches as if they contained mutations (orange), with the ability to include variants that are not polymorphic in the sequenced reference panel.

the common variants, but this approach is only valid down to a certain frequency, missing many rare and ultra-rare causal variants.

Sequencing several samples and performing imputation adds resolution for rarer variants. In Fig. 7.1c, a schematic of imputation is given where mutations from the sequenced samples (lavender) are observed and added on to the ARG, enabling imputation on to the array samples. However, variants can only be imputed if carriers have been sequenced, and rarer variants will require larger reference panels to impute. Our approach, shown in Fig. 7.1d, is to infer an ARG from array data, for which each branch (orange) corresponds to a possible mutation. We can test these mutations for association with a trait as if they were observed. In particular, rare variants missed by imputation may be detected using this approach.

We note that the inferred ARG does not need to perfectly match the true ARG for this approach to succeed. Rather, the inferred ARG branches may serve as tags for causal unseen variants without being a perfect match. This may explain some of the common variant loci that were detected by genealogy-wide association of height but not by reference-based imputation (Figs. 6.4e and 6.6a-c). Although imputation of common variants is quite accurate, the ARG contains branches that may not be hit by a mutation event, and which may be captured using the inferred ARG. These branches will not manifest in the imputed data because there is no underlying mutation, but the branch may sufficiently tag causal rare variation so as to be significantly associated. Our common variant analysis also replicated haplotype structures and association peaks found through reference-based imputation (Figs. 6.4f and 6.6d-h), including many loci that were not detected using SNP array association alone. These results illustrate the potential for genome-wide ARG inference and genealogy-based association to complement approaches based on sequencing and imputation across both rare and common variants.

As a proof of concept, we also performed genotype imputation using an ARG inferred from sequencing and array samples, and compared against the IMPUTE4 [Bycroft et al., 2018] and BEAGLE5 [Browning et al., 2018] imputation algorithms as well as imputation from the true simulated ARG (Fig. 5.9). Genotype imputation us-

ing the true ARG was consistently better than IMPUTE4 and BEAGLE5, especially for rare variants. Our inferred ARGs also performed comparably to or better than reference-based imputation for the rarest variants in our small-scale experiments, suggesting that ARG imputation should be further explored for rare variant imputation.

7.2.2 Connections to identity-by-descent mapping

IBD mapping shares many of the same goals as our ARG-MLMA approach. Both IBD mapping and ARG-based association can test for unseen variation, which is localised according to the extent of the haplotype or ARG branch. However, association using inferred ARGs may be preferred to IBD mapping because it represents a more complete and rigorous framework. IBD detection approaches often set a time threshold for which the ancestral haplotype must be of a more recent time. With IBD mapping approaches, each modern haplotype is likely to be a carrier of at most one IBD haplotype at each position, whereas the ARG enables testing branches across the full spectrum of times and allele frequencies. IBD mapping approaches tend to focus on rare variants, whereas we demonstrated that there may be benefits in applying genealogy-wide association in low-frequency or common variants. In short, the ARG reveals more information than is possible using IBD detection methods.

The genealogy-wide association analysis we performed builds on prior works in cladistic mapping and ARG association (see Section 2.3.2), while demonstrating multiple advances. First, to our knowledge, our work is the first to test clades of inferred genome-wide ARGs within a biobank of hundreds of thousands of samples. This hinged on our investment in scalable methods, including performance optimisations for testing clades with few carriers. Second, we allowed for testing clades using not only a mixed model but a non-infinitesimal mixed model [Loh et al., 2015b], two extensions which compound to substantially improve power. Third, we sampled clades using a mutation rate, a procedure that tends to select more certain clades of larger area in the ARG. Combined with our permutation testing-based calibration of significance thresholds, this avoids testing all the branches of the ARG, which may create a more stringent significance threshold and reduce association power. Incorporating

the area of branches in our sampling approach leverages the uncertainty in the inferred ARG, where relatively shorter branches may represent an uncertain ordering of coalescent events and are less likely to be sampled by our procedure.

7.2.3 Connections to linear mixed models

As already discussed, the use of a non-infinitesimal linear mixed model to test branches of biobank-scale ARGs (ARG-MLMA) represents a novel application of mixed model methods. Here we further contextualise our methods on using ARG-GRMs for complex trait analysis.

The degree of improvement afforded by linear mixed models—whether in heritability estimation, polygenic prediction, and mixed-model association—is largely determined by two factors. First, the markers included in the mixed model must achieve enough density to tag underlying causal variants. Second, the model should have the flexibility to account for varying genetic architectures. If these conditions are not met, resulting heritability estimates will be deflated or otherwise biased, prediction accuracy will suffer, and the power improvements from mixed-model association compared to single-SNP testing will be decreased. For instance, earlier estimates of the narrow sense heritability of height [Yang et al., 2010], which relied on genotyping array data, have since been superseded by estimates from imputed and whole-genome sequencing data which explain more and more of the so-called "missing heritability" [Manolio et al., 2009, Yang et al., 2015, Wainschtein et al., 2022]. Likewise, many earlier examples of linear mixed models in complex trait analysis [Yang et al., 2010, Speed et al., 2012] have since been revised to more fully account for genetic architecture [Lee et al., 2013, Yang et al., 2015, Evans et al., 2018b].

Starting from SNP array datasets, our proposal is to first infer an ARG using ARG-Needle, then use ARG-GRMs from the inferred ARGs to perform mixed model analyses. We demonstrated that in realistic small-scale simulations of tens of thousands of samples, ARG-GRMs from the true ARGs performed as well as GRMs from whole-genome sequencing data, in all three application areas we considered (Fig. 3.2). These results provide an upper bound given perfect ARG inference and highlight the

potential of such an ARG-based approach. Using ARGs inferred from SNP data, we achieved close to unbiased heritability estimates in simulations, whereas GRMs from array variants or variants imputed from hundreds of reference samples led to underestimated heritability (Fig. 5.7). The gap in various estimates was particularly apparent for simulated traits with strong negative selection ($\alpha = -1$), where rare variants explain more of the trait heritability. This is consistent with the concentration of array SNPs within common variants and the difficulty of imputing rare variants due to their need to be polymorphic in the reference panel.[4]

Despite only requiring genotyping array data, the combined ARG inference and ARG-GRM approach is able to capture the effect of rare variants when building the GRM. Just as our ARG-MLMA framework enabled testing branches of the ARG which may contain unseen variants, the ARG-GRM approach captures unseen variants by computing an expected GRM over all branches of the ARG. The ARG-GRM approach can also be understood in relation to methods which have used IBD segments to construct GRMs [Browning and Browning, 2013, Zaitlen et al., 2013, Evans et al., 2018a]. The ARG-Needle inference algorithm implicitly identifies potential IBD while constructing the ARG, as IBD segments will share descent from an internal node in the ARG.[5] Each branch of the ARG implicitly represents the carriers who are IBD from that branch, with branches naturally weighted by their probability of receiving a mutation when computing the ARG-GRM. Therefore, ARG-GRMs can be seen as a more principled framework for constructing GRMs based on haplotype sharing that bypasses the heuristics used in IBD detection.[6]

A recent preprint also computed GRMs from ARGs, in a method which they call eGRMs (expected GRMs), and applied such eGRMs to analyses population genetics [Fan et al., 2021]. Besides our complementary focus on complex trait genetics, our ARG-GRMs include several more advanced modelling options. First, we investigated two options related to MAF-dependent trait architectures. When the MAF dependency parameter α is known, we allowed for building GRMs with that known value

[4]See Section 7.2.1.
[5]See [Nait Saada et al., 2020], which does this task explicitly using genotype hashing and ASMC.
[6]See Section 7.2.2.

of α, demonstrating that matching on the true value of α was necessary for accurate heritability estimation and optimal polygenic prediction. When α is unknown, we instead suggest building MAF-stratified GRMs and performing a multi-GRM analysis. MAF-stratified GRMs performed comparably to knowing the true value of α, indicating the flexibility of this approach. Second, we allowed for computing either exact ARG-GRMs or Monte Carlo ARG-GRMs. The exact ARG-GRM performs an exact integral over all branches of the ARG, whereas the Monte Carlo ARG-GRM simulates mutations on the ARG, which are used to then generate a GRM. We found that the Monte Carlo ARG-GRM approach is more computationally efficient without sacrificing accuracy. Furthermore, because it generates genetic data which is used to build a GRM, it is easy to understand and adapt. For instance, the generated mutations can be partitioned based on age, LD, and functional status, extending our current partitioning by MAF to more accurately model genetic architecture.

Additionally, we note several differences between our ARG-GRM paradigm and the use of pedigree kinship matrices.[7] Both ARG-GRMs and pedigree kinship matrices are computed based on a historical record of samples and patterns of inheritance. Pedigrees require detailed family records, whereas we have proposed to infer plausible ARGs based on genetic data, thus enabling analyses of cohorts which are not closely related. Furthermore, studying trait heritability in closely related cohorts using pedigrees often leads to overestimation due to shared environmental effects [Zaitlen and Kraft, 2012, Zaitlen et al., 2013]. Additionally, whereas the pedigree can predict the average genome-wide IBD between two samples, the true ARG contains true realised IBD within its representation. Explained differently, the pedigree stores all possible ancestral lineages of an individual, but the true ARG, if available, traces the exact ancestral lineage through which each gene is inherited. Therefore, GRMs based on the ARG are to be preferred to pedigree-based kinships.

In the past decades, the advent of high-throughput genomic measurement technologies have encouraged the field of genetics to move from considering pedigrees to analysing marker data directly. Now, however, may be an opportune moment

[7] See Section 2.3.3.

to revisit complex trait genetics within a genealogical viewpoint, armed with the richer modelling of the past afforded by the ARG and scalable ARG inference. We anticipate that the methods we have developed using the ARG—ARG-MLMA and ARG-GRMs—can be broadly applied and further extended. We also hope that the connections we have identified to imputation, IBD mapping, and linear mixed models will inspire future methods in genealogical complex trait analysis.

7.3 Future Work

We note several limitations and directions for future development regarding this work, some of which we are currently investigating.[8]

7.3.1 ARG-based methods for statistical genetics

First, although we have shown in simulation that ARG-GRMs built from true ARGs may be used to estimate heritability, perform polygenic prediction, and increase association power, we did not yet apply these methods to real datasets due to limits of scalability. Forming a GRM takes $O(N^2M)$ time, where N is the number of individuals and M is the number of markers, and inverting the GRM takes additional $O(N^3)$ time. This is impractical for the UK Biobank when M corresponds to genotyping array data, and is even more costly in our applications when M corresponds to mutations densely sampled in a Monte Carlo fashion on the ARG. Methods such as fastGWA [Jiang et al., 2019] and BOLT-REML [Loh et al., 2015a] have improved the scalability of mixed-model restricted maximum likelihood (REML) through the use of sparse GRMs and conjugate gradient operations. However, preliminary work we performed suggests that even with these ideas, ARG-GRMs will be difficult to use at the scale of the UK Biobank. We are currently working to combine an ARG-based framework with highly scalable methods in heritability estimation that do not rely on REML, such as LD score regression [Bulik-Sullivan et al., 2015, Finucane et al., 2015] and randomised Haseman-Elston regression [Haseman and Elston, 1972]

[8]This section is expanded and adapted from the Discussion of [Zhang et al., 2021].

[Wu and Sankararaman, 2018, Pazokitoroudi et al., 2020]. It will also be important to extend our approach of MAF-stratified ARG-GRMs to consider other factors influencing genetic architecture, such as LD and functional dependence [Finucane et al., 2015, Speed et al., 2017, Gazal et al., 2018, Gazal et al., 2019, Márquez-Luna et al., 2021] as well as environmental factors [Young et al., 2018].

Second, it would be worthwhile to explore additional analyses using ARG-Needle-inferred ARGs. For example, reconstructing biobank-scale ARGs will likely aid the study of additional evolutionary properties of disease-associated variants, including analyses of natural selection acting on complex traits [Palamara et al., 2018, Speidel et al., 2019, Yasumizu et al., 2020, Stern et al., 2021]. Whereas our work assumes a demographic model is available, previous methods in coalescent modelling have inferred demographies from data [Li and Durbin, 2011, Schiffels and Durbin, 2014, Terhorst et al., 2017], another possible application. Additionally, it may be possible to build on the developments of the SAIGE [Zhou et al., 2020] and REGENIE [Mbatchou et al., 2021] methods to support fast generalised mixed-model analysis of case-control traits.

7.3.2 ARG-Needle inference algorithm

Third, although we have focused on inferring the ARG from array or sequencing data, we plan to further explore the use of additional data types, such as imputed and low-coverage sequencing data. Generalising the ARG-Needle algorithm to other data types would be possible with two modifications. First, the step of finding candidate closest cousins could be extended to work with other data types. Second, the step of TMRCA prediction from two samples could likewise be extended. In the second of these areas, a recent workshop paper has demonstrated promise in inferring TMRCAs from imputed data using a so-called coalescent neural network [Nait Saada et al., 2021], which is trained using `msprime` simulations.

Fourth, the ARG-Needle inference algorithm could be made to run faster by improving the speed of closest cousin detection, which is currently the rate limiting step for large sample sizes. It may be possible to improve the efficiency of our genotype

hashing implementation by multithreading operations or utilising dynamic hashing (see the GERMLINE2 algorithm, [Nait Saada et al., 2020]). We also plan to explore closest cousin detection using the positional Burrows-Wheeler transform (PBWT) [Durbin, 2014] data structure.

Fifth, we could investigate ways to predict a distribution over possible ARGs, rather than a single inferred ARG (i.e., a point estimate). ARGweaver [Rasmussen et al., 2014] and the recent method ARGinfer [Mahmoudi et al., 2022] both output a posterior over ARGs via Markov chain Monte Carlo sampling, but only scale to tens of samples and assume sequencing data. We note two sources of stochasticity within ARG-Needle that can be used to output a distribution over ARGs. First, while we have relied on the posterior mean and MAP as summaries of the posterior TMRCA computed by ASMC, it could be possible to sample from this posterior instead. Second, preliminary work we performed indicates that the order in which samples are threaded changes the resulting ARG,[9] and thus building multiple ARGs under different orders would create a resulting ensemble. Samples could also be removed and re-threaded to the ARG, as is done in ARGweaver. It may be possible to average over multiple inferred ARGs to enable more accurate downstream analyses [Minichiello and Durbin, 2006], or to use the variation within the ensemble to calibrate confidence intervals [Rasmussen et al., 2014]. However, such a distribution is unlikely to be a true Bayesian posterior.

7.3.3 Real data applications

Sixth, although genealogy-wide association accurately detects the subset of individuals carrying independently associated variants, the signal is usually localised within a genomic region, whereas genotype imputation may implicate individual variants if they are present in the sequenced reference panel. However, when sequencing data is available, it may be utilised to further localise ARG-derived signals, as done using

[9] If the estimated TMRCAs are ultrametric, and the closest cousin selection step is accurate, ARG-Needle will output the same ARG independent of threading order (see Section 4.3.2.2). However, ASMC does not yield ultrametric TMRCAs.

WES partners in our analyses.

Finally, our analysis focused on the UK Biobank data set, which provides an excellent testbed due to the large volumes of high-quality data of different types that are available for validation. Future applications of our methods will involve analysis of cohorts that are less strongly represented in current sequencing studies. Nevertheless, we believe that the results described in this work represent an advance in large-scale data-driven genealogical inference and provide new tools for the analysis of complex traits.

Appendix A

Additional Details on ARG-GRMs

A.1 Overview

In Section 3.2.1 we introduced ARG-GRMs for the case of haploid samples and $\alpha = 0$, as well as Monte Carlo ARG-GRMs, which sample new mutations on the ARG and use those markers to construct GRMs. Monte Carlo ARG-GRMs enable easily taking into account diploid samples, modelling varying values of α, and working with stratified ARG-GRMs. In this Appendix, we describe computing an exact ARG-GRM for the cases of diploid samples, general α, and stratification. We also discuss various invariances of the GRM used in mixed-model analysis and show how all methods compute an expected version of the sequence-based GRM up to invariance.

We first provide notation for the various sequence-based GRMs which we seek to approximate using ARGs. In all cases we have M markers and N individuals. While we have focused on haploid individuals with genotypes $x_{ik} \in \{0, 1\}$, $1 \leq i \leq N$ and $1 \leq i \leq M$, we also describe the case of diploid individuals with genotypes $x_{ik} \in \{0, 1, 2\}$, $1 \leq i \leq N$ and $1 \leq i \leq M$.

We begin by discussing the sequence-based GRM for haploid genotypes and general α, with allele frequencies $p_k = \frac{1}{N} \sum_{i=1}^{N} x_{ik}$:

$$K_{\alpha,hap}(i, j) = \frac{1}{M} \sum_{k=1}^{M} \frac{(x_{ik} - p_k)(x_{jk} - p_k)}{[p_k (1 - p_k)]^{-\alpha}}. \tag{A.1}$$

We then describe stratified GRMs, which are obtained by partitioning the markers into various bins via MAF, LD, time intervals (to capture allele age), or other annotations. The haploid stratified GRM for a bin containing SNPs $B \subseteq \{1, \ldots, M\}$ using value α is given by:

$$K_{\alpha,hap,B}(i,j) = \frac{1}{|B|} \sum_{k=1}^{M} \mathbb{1}_{k \in B} \frac{(x_{ik} - p_k)(x_{jk} - p_k)}{[p_k(1-p_k)]^{-\alpha}}. \tag{A.2}$$

(We use $\mathbb{1}_A$ to represent the indicator function of event A, taking value 1 if A holds and 0 otherwise.)

Lastly, we consider the general α sequence-based GRM for diploid genotypes, with allele frequencies $p_k = \frac{1}{2N} \sum_{i=1}^{N} x_{ik}$:

$$K_{\alpha,dip}(i,j) = \frac{1}{M} \sum_{k=1}^{M} \frac{(x_{ik} - 2p_k)(x_{jk} - 2p_k)}{[2p_k(1-p_k)]^{-\alpha}}. \tag{A.3}$$

(We omit a discussion on diploid stratified GRMs, which are easily derived as a generalisation of the above two cases.)

A.2 From sequence-based GRMs to ARG-GRMs

The above GRM expressions assume we have access to a set of markers. Under the infinite sites assumption, each marker corresponds to an event that occurred at some time in the past. For simplicity, we refer to these events as mutations, though they may also consist of other variant types, such as short indels. Each such mutation occurs somewhere on the ARG, with position x and time t, and affects a set of descendants that will carry the derived allele. In the haploid case, if we let the descendants of a mutation m be $d(m) \subset \{1, \ldots, N\}$, we can also define the allele frequency $p(m)$ of a mutation as $p(m) = |d(m)|/N$.

Assuming a uniform mutation rate and the infinite-sites model, mutations arise uniformly over the area of the ARG, as a Poisson process characterised by the mutation rate μ. The area of the ARG consists of all its edges, but an ARG edge does

not always have a constant set of descendants, if a recombination event occurs on the path between the edge and its descendants. Therefore, we may further partition the edges of the ARG into "branches", such that each branch b is valid for only part of the genomic extent of an edge and carries a constant set of descendants. We let $A(b)$ denote the area of branch b, obtained by multiplying its extent in time (e.g. generations) by its extent along the genome, and assume that the set B enumerates all branches. Each branch b then also inherits a set of descendants $d(b) \subset \{1, \ldots, N\}$ and an allele frequency $p(b)$, just as for mutations. The area of the ARG is then the sum over all these branches, or $\sum_{b \in B} A(b)$.

In the case where we do not have access to all underlying markers needed to compute a GRM but have access to the ground truth ARG or an inferred ARG, we may compute the expectation for the GRM entries. For the haploid single-component ARG-GRM, we compute

$$K_{\alpha,hap}(i,j) = \mathbb{E}_{m \sim Uniform(\{m_1,\ldots,m_M\})} \frac{\left(\mathbb{1}_{i \in d(m)} - p(m)\right)\left(\mathbb{1}_{j \in d(m)} - p(m)\right)}{\left[p(m)\left(1-p(m)\right)\right]^{-\alpha}}. \tag{A.4}$$

Since mutations are uniformly distributed over the area of an ARG, we replace the set of known M mutations with the distribution over all possible mutations induced by uniform sampling over the ARG:

$$K_{\alpha,hap,ARG}(i,j) = \mathbb{E}_{m \sim Uniform(ARG)} \frac{\left(\mathbb{1}_{i \in d(m)} - p(m)\right)\left(\mathbb{1}_{j \in d(m)} - p(m)\right)}{\left[p(m)\left(1-p(m)\right)\right]^{-\alpha}}. \tag{A.5}$$

(A.5) reflects the expected value of (A.4) given the genealogical relationships of the ARG, without observing any markers.

$$K_{\alpha,hap,ARG}(i,j) = \mathbb{E}\left[K_{\alpha,hap}(i,j) | ARG\right]. \tag{A.6}$$

We can compute the ARG-wide expectation given by (A.5) using Monte Carlo, adopting a mutation rate μ to uniformly sample M' variants on the ARG. We then use these M' sampled variants to evaluate (A.5).

We can also take this expectation analytically, by weighting all possible branches

of the ARG by their area. Using our earlier notation,

$$K_{\alpha,hap,ARG}(i,j) = \mathbb{E}_{m\sim Uniform(ARG)} \frac{\left(\mathbb{1}_{i\in d(m)} - p(m)\right)\left(\mathbb{1}_{j\in d(m)} - p(m)\right)}{\left[p(m)\left(1-p(m)\right)\right]^{-\alpha}}$$
$$= \frac{1}{\sum_{b\in B} A(b)} \sum_{b\in B} \left[A(b) \cdot \frac{\left(\mathbb{1}_{i\in d(b)} - p(b)\right)\left(\mathbb{1}_{j\in d(b)} - p(b)\right)}{\left[p(b)\left(1-p(b)\right)\right]^{-\alpha}} \right] \quad \text{(A.7)}$$

This formulation, however, requires traversing all branches of the ARG and updating N^2 values for each branch, which is less efficient than using the Monte Carlo method with a sufficiently high mutation rate. The Monte Carlo method is also $O(N^2)$ per mutation, but fewer mutations are needed to be sampled compared to the exact ARG-GRM, which visits all branches.

The case of stratified GRMs is an extension of the above, where we assign each mutation to different GRMs according to the stratification criteria. In our experiments, this corresponded to selecting the appropriate GRM based on the allele frequency ($p(m)$ or $p(b)$) of each sampled mutation.

For diploid GRMs, each individual genotype x_{ik} is modelled as the sum of two haplotypes. We introduce notation where we consider an ARG of $2N$ haploid samples, numbered 1 to $2N$. Without loss of generality, each individual i consists of haplotypes $2i-1$ and $2i$, for $1 \leq i \leq N$. For shorthand, we define $i_1 = 2i-1$ and $i_2 = 2i$. Then we can rewrite (A.3) as

$$K_{\alpha,dip}(i,j) = \frac{1}{M} \sum_{k=1}^{M} \frac{\left(x_{i_1 k} + x_{i_2 k} - 2p_k\right)\left(x_{j_1 k} + x_{j_2 k} - 2p_k\right)}{\left[2p_k\left(1-p_k\right)\right]^{-\alpha}}. \quad \text{(A.8)}$$

If we denote descendants of a mutation in the ARG as $d(m) \subset \{1,\ldots,2N\}$, and $p(m) = |d(m)|/2N$, the diploid ARG-GRM can be computed using the ARG as

$$K_{\alpha,dip,ARG}(i,j)$$
$$= \mathbb{E}_{m\sim Uniform(ARG)} \frac{\left(\mathbb{1}_{i_1\in d(m)} + \mathbb{1}_{i_2\in d(m)} - 2p(m)\right)\left(\mathbb{1}_{j_1\in d(m)} + \mathbb{1}_{j_1\in d(m)} - 2p(m)\right)}{\left[2p(m)\left(1-p(m)\right)\right]^{-\alpha}}$$
$$= \frac{1}{\sum_{b\in B} A(b)} \sum_{b\in B} \left[A(b) \cdot \frac{\left(\mathbb{1}_{i_1\in d(b)} + \mathbb{1}_{i_2\in d(b)} - 2p(b)\right)\left(\mathbb{1}_{j_1\in d(b)} + \mathbb{1}_{j_2\in d(b)} - 2p(b)\right)}{\left[2p(b)\left(1-p(b)\right)\right]^{-\alpha}} \right]$$

The exact ARG-GRM is obtained by iterating over all branches, and a Monte Carlo ARG-GRM is obtained by sampling, as previously described.

In the remainder we derive simpler expressions for ARG-GRMs. We begin by describing three useful invariances of GRMs in mixed-model analysis.

A.3 Three GRM invariances

We discuss three invariances under which a GRM may be altered without changing downstream results, which we use to facilitate mixed model analysis using ARG-GRMs and in some of the proofs below. We refer to these invariances as scale invariance, data shift invariance, and constant shift invariance.

The two aspects of the ARG-LMM pipeline that are linked to these invariances are Gower centring and the inclusion of a centring covariate. We first describe Gower centring. Given an N by N GRM K, define the N by N identity matrix I_N, the size N column vector 1_N, and the projection matrix P_N (which is symmetric and idempotent and corresponds to projection onto the subspace orthogonal to 1_N) to be

$$P_N = I_N - \frac{1}{N} 1_N 1_N^T.$$

Then, the Gower centred version of K is defined as

$$C_{Gower}(K) = \frac{N-1}{\mathrm{Tr}(P_N K P_N)} K.$$

To obtain correct heritability estimates for sequence-based GRMs with general α (e.g. (A.1) above), we Gower centred ARG-GRMs provided in input to GCTA. (In the special case of $\alpha = 0$, the sequence-based GRM in (A.1) is approximately Gower centred already, and Gower centring merely multiplies by the factor $(N-1)/N$.) Consider the multiplication of K by any nonzero scalar constant γ:

$$C_{Gower}(\gamma K) = \frac{N-1}{\mathrm{Tr}(P_N(\gamma K)P_N)} \gamma K = \frac{N-1}{\gamma \mathrm{Tr}(P_N K P_N)} \gamma K = C_{Gower}(K).$$

Thus, multiplying a GRM by any nonzero scalar and then applying Gower centring will lead to identical downstream results, which we refer to as *scale invariance* of GRMs.

We next consider the inclusion of a centring covariate. Even when no covariates are specified in an analysis, most software packages for complex trait analysis (e.g. PLINK [Purcell et al., 2007], GCTA [Yang et al., 2011a], BOLT-LMM [Loh et al., 2015b, Loh et al., 2018], and BOLT-REML [Loh et al., 2015a]) implicitly or explicitly include a centring covariate. This can be thought of as a length N covariate vector consisting of all 1s. Including this covariate is equivalent to mean-centring the phenotype as well as the genotype vector for each marker. It can also be implemented by applying the projection operator P_N, defined above, to project out the component parallel to 1_N in the data.

In our analyses involving GCTA, we provided the GRM and a phenotype in input, rather than providing markers from which to compute the GRM. In this case, the phenotype is mean-centred, and although one does not have access to the underlying genotypes, the centring covariate is still applied implicitly during the relevant mixed-model calculations.

We also describe the inclusion of a centring covariate as a transformation on the GRM itself. In our analyses we performed this operation, which we call data centring, prior to Gower centring, and after both steps were completed, we provided the transformed GRM in input to GCTA. The data centred GRM is given by

$$C_{Data}(K) = P_N K P_N.$$

Right-multiplying by P_N corresponds to subtracting the mean column of a matrix from each column, and left-multiplying by P_N corresponds to subtracting the mean row of a matrix from each row. The transformations C_{Gower} and C_{Data} commute, so that the order in which they are applied does not matter, and that they are both idempotent, so that data centring outside of GCTA does not interfere with the application of the centring covariate inside.

We highlight the second invariance of GRMs we leveraged in these analyses using an example of the effects of data centring. Consider the GRM that would arise if we used raw genotypes instead of centred genotypes in (A.1):

$$\widetilde{K}_{\alpha,hap}(i,j) = \frac{1}{M}\sum_{k=1}^{M}\frac{(x_{ik})\,(x_{jk})}{[p_k\,(1-p_k)]^{-\alpha}}.$$

We verify that data centring gives the expression (A.1) with centred genotypes. Notice that the GRM $\widetilde{K}_{\alpha,hap}$ can be written as a product of matrices:

$$\widetilde{K}_{\alpha,hap} = XDX^T$$

where X is of size N by M with entries x_{ij}, and D is a diagonal matrix of size M by M with diagonal entries

$$d_{kk} = [p_k\,(1-p_k)]^{\alpha}/M.$$

The data centred version of $\widetilde{K}_{\alpha,hap}$ is then

$$C_{Data}\left(\widetilde{K}_{\alpha,hap}\right) = P_N\widetilde{K}_{\alpha,hap}P_N = P_N(XDX^T)P_N = (P_NX)D(P_NX)^T.$$

Also notice that

$$P_NX = \left(I_N - \frac{1}{N}1_N1_N^T\right)X = X - 1_N\left(\frac{1}{N}1_N^TX\right) = X - 1_N\mu^T$$

where $\mu^T = (p_1, p_2, \ldots, p_M)$ is a row vector consisting of the average row of X, or the collection of allele frequencies. (So that left-multiplying by P_N corresponds to subtracting the mean row of a matrix from each row.) Hence,

$$C_{Data}\left(\widetilde{K}_{\alpha,hap}\right) = (X - 1_N\mu^T)D(X - 1_N\mu^T)^T.$$

This is equivalent to forming a GRM using the centred data, and hence $C_{Data}(\widetilde{K}_{\alpha,hap}) = K_{\alpha,hap}$.

The invariance highlighted in this example, where centred data is replaced by raw genotypes, is more general. Rather than using raw genotypes, we can add a marker-specific constant to the genotypes for each marker before multiplying:

$$\widehat{K}_{\alpha,hap}(i,j) = \frac{1}{M}\sum_{k=1}^{M}\frac{(x_{ik}+c_k)(x_{jk}+c_k)}{[p_k(1-p_k)]^{-\alpha}},$$

where the c_k are any real scalars for $1 \leq k \leq M$. If we let $\rho^T = (c_1, \ldots, c_M)$, we see that

$$\widehat{K}_{\alpha,hap} = (X + 1_N\rho^T)D(X + 1_N\rho^T)^T.$$

We have

$$\begin{aligned} P_N(X + 1_N\rho^T) &= P_N X + P_N 1_N \rho^T \\ &= (X - 1_N\mu^T) + \left(I_N - \frac{1}{N}1_N 1_N^T\right)1_N\rho^T \\ &= (X - 1_N\mu^T) + \left(1_N - \frac{1}{N}1_N N\right)\rho^T \\ &= X - 1_N\mu^T. \end{aligned}$$

Hence

$$C_{Data}\left(\widehat{K}_{\alpha,hap}\right) = P_N\widehat{K}_{\alpha,hap}P_N = (X - 1_N\mu^T)D(X - 1_N\mu^T)^T = K_{\alpha,hap}.$$

We refer to this invariance as *data shift invariance*: by using a centring covariate, the markers used to construct a GRM can have a constant shift per marker applied to the genotypes before multiplying, without affecting downstream results. Although we have only described a detailed derivation in the case of haploid GRMs, the same considerations apply to other cases, starting from (A.2) and (A.3).[1]

Finally, *constant shift invariance* allows adding or subtracting a constant scalar to each entry of a GRM without affecting downstream results. This invariance also

[1]Note that this form of invariance was also discussed in [Dahl et al., 2020], Appendix B.

follows from the inclusion of a centring covariate. Assume data centring and consider adding a constant c to each entry of a GRM K:

$$\begin{aligned} C_{Data}(K + c1_N1_N^T) &= P_N(K + c1_N1_N^T)P_N \\ &= C_{Data}(K) + \left(I_N - \frac{1}{N}1_N1_N^T\right)(c1_N1_N^T)\left(I_N - \frac{1}{N}1_N1_N^T\right) \\ &= C_{Data}(K) + c\left(1_N1_N^T - 1_N1_N^T - 1_N1_N^T + 1_N1_N^T\right) \\ &= C_{Data}(K). \end{aligned}$$

In our experiments, we applied data centring directly on the GRM before passing into GCTA, which guaranteed constant shift invariance. We observed that if we omitted the data centring transformation, only relying on GCTA's implementation of a centring covariate, constant shift invariance still held for a range of shift values. However, some experiments involved adding a large negative shift to a GRM such that it is no longer positive semidefinite, and we observed that passing such a GRM directly into GCTA led to errors. On the other hand, we found that GCTA was robust to the earlier described data shift invariance, including when we omitted the data centring transformation, possibly because data shift transformations preserve positive definiteness.

In summary, by always applying Gower centring and what we referred to as data centring to our GRMs, before passing into GCTA, allowed us to guarantee three invariances in the GRM: scale invariance, where the GRM is multiplied by a non-zero scalar; data shift invariance, where a marker-specific shift is applied to each marker used to compute the GRM; and constant shift invariance, where a constant scalar is added to each entry of the GRM.

A.4 Exact ARG-GRM, haploid and general α

We use these invariances to describe simplified expressions for computing exact ARG-GRMs. Due to data shift invariance, we may replace the p_k terms in the numerator

of (A.1) with the value 1/2. Let $\equiv$ denote equivalence under invariance.[2] We can write

$$K_{\alpha,hap}(i,j) \equiv \frac{1}{M}\sum_{k=1}^{M}\frac{(x_{ik}-1/2)(x_{jk}-1/2)}{[p_k(1-p_k)]^{-\alpha}}. \tag{A.9}$$

Since $x_{ik}, x_{jk} \in \{0,1\}$,

$$\begin{aligned}(x_{ik}-1/2)(x_{jk}-1/2) &= \frac{1}{4}(2x_{ik}-1)(2x_{jk}-1)\\ &= \frac{1}{4}(1-2(x_{ik}\oplus x_{jk})),\end{aligned}$$

where $\oplus$ refers to the XOR of two binary values. Substituting into (A.9) and leveraging scale invariance and constant shift invariance,

$$\begin{aligned}K_{\alpha,hap}(i,j) &\equiv \frac{1}{4M}\sum_{k=1}^{M}\frac{(1-2(x_{ik}\oplus x_{jk}))}{[p_k(1-p_k)]^{-\alpha}}\\ &= \frac{1}{4M}\sum_{k=1}^{M}\frac{1}{[p_k(1-p_k)]^{-\alpha}} - \frac{1}{2M}\sum_{k=1}^{M}\frac{x_{ik}\oplus x_{jk}}{[p_k(1-p_k)]^{-\alpha}}\\ &\equiv \sum_{k=1}^{M}\frac{x_{ik}\oplus x_{jk}}{[p_k(1-p_k)]^{-\alpha}}.\end{aligned} \tag{A.10}$$

In the case of $\alpha = 0$, this simplifies to

$$K_{\alpha=0,hap}(i,j) \equiv \sum_{k=1}^{M} x_{ik}\oplus x_{jk}.$$

Note that this is the Hamming distance matrix between i and j. Under the infinite-sites model, mutations for which samples i and j differ occur on the area of the 2-sample ARG from samples i and j to their TMRCA along the genome. The number of such mutations is a Poisson random variable with rate μ (the per-site per-generation mutation rate) times the area (sites times generations) of the 2-sample ARG. We express the areas as $2 \times L \times \bar{t}_{ij}$, where $\bar{t}_{ij}$ is the genome-wide average TMRCA of i

[2]Note that with this notation we are referring to equivalence under invariance for the entire GRM, not for an individual entry.

and j and L is the physical extent of the ARG. Then,

$$\begin{aligned} K_{\alpha=0,hap,ARG}(i,j) &= \mathbb{E}\left[K_{\alpha=0,hap}(i,j)|ARG\right] \\ &\equiv \mathbb{E}\left[\sum_{k=1}^{M} x_{ik} \oplus x_{jk} \middle| ARG\right] \\ &= \mathbb{E}\left[\text{Poisson}(2 \times L \times \bar{t}_{ij})\right] \\ &= 2 \times L \times \bar{t}_{ij} \\ &\equiv \bar{t}_{ij}. \end{aligned}$$

Assuming Gower centring and data centring are used, we may therefore use the average TMRCA matrix between samples to compute the $\alpha = 0$ exact ARG-GRM.

For the case of general α, consider (A.10), then

$$\begin{aligned} K_{\alpha,hap,ARG}(i,j) &\equiv \mathbb{E}\left[\frac{1}{M}\sum_{k=1}^{M} \frac{x_{ik} \oplus x_{jk}}{\left[p_k\left(1-p_k\right)\right]^{-\alpha}} \middle| ARG\right] \\ &= \mathbb{E}_{m \sim Uniform(ARG)}\left[\frac{\mathbb{1}_{i \in d(m)} \oplus \mathbb{1}_{j \in d(m)}}{\left[p(m)\left(1-p(m)\right)\right]^{-\alpha}}\right] \\ &= \sum_{b \in B}\left[A(b) \cdot \frac{\mathbb{1}_{i \in d(b)} \oplus \mathbb{1}_{j \in d(b)}}{\left[p(b)\left(1-p(b)\right)\right]^{-\alpha}}\right] \end{aligned}$$

The expression $\mathbb{1}_{i \in d(b)} \oplus \mathbb{1}_{j \in d(b)}$ is 1 if exactly one of i and j is a descendant of branch b, and 0 otherwise. Therefore we may rewrite it as $\mathbb{1}_{|d(b) \cap \{i,j\}|=1}$, where we test whether the intersection of $d(b)$ and $\{i,j\}$ contains exactly one element. Thus:

$$\begin{aligned} K_{\alpha,hap,ARG}(i,j) &\equiv \sum_{b \in B}\left[A(b) \cdot \frac{\mathbb{1}_{|d(b) \cap \{i,j\}|=1}}{\left[p(b)\left(1-p(b)\right)\right]^{-\alpha}}\right] \\ &= \sum_{b \in B, |d(b) \cap \{i,j\}|=1}\left[\frac{A(b)}{\left[p(b)\left(1-p(b)\right)\right]^{-\alpha}}\right] \end{aligned} \tag{A.11}$$

The branches $b \in B$ with $|d(b) \cap \{i,j\}| = 1$ are those that lie in the 2-sample ARG containing samples i and j. Computing (A.11) is achieved by iterating over all these branches and summing their areas, weighted by a term involving $p(b)$ and α. When

$\alpha = 0$, this reduces to summing the areas of all branches in the 2-sample ARG:

$$\begin{aligned} K_{\alpha=0,hap,ARG}(i,j) &\equiv \sum_{b\in B,|d(b)\cap\{i,j\}|=1} A(b) \\ &= 2 \times L \times \bar{t}_{ij}, \end{aligned}$$

which coincides with the earlier derivation for $\alpha = 0$.

The MAF-stratified version of the haploid α GRM is similarly derived, except iteration is restricted to branches b with allele frequency $p(b)$ within a specified range. This generalises to other stratification criteria, e.g. allele age (see [Palamara et al., 2016]).

A.5 Exact ARG-GRM, diploid and general α

Finally, we describe the case of diploid genotypes and general α. We separate (A.8) into four terms:

$$\begin{aligned} K_{\alpha,dip}(i,j) &= \frac{1}{M}\sum_{k=1}^{M} \frac{(x_{i_1k} + x_{i_2k} - 2p_k)(x_{j_1k} + x_{j_2k} - 2p_k)}{[2p_k(1-p_k)]^{-\alpha}} \\ &= \frac{1}{M}\sum_{k=1}^{M} \frac{[(x_{i_1k} - p_k) + (x_{i_2k} - p_k)][(x_{j_1k} - p_k) + (x_{j_2k} - p_k)]}{[2p_k(1-p_k)]^{-\alpha}} \\ &= \frac{2^\alpha}{M}\left[\sum_{k=1}^{M} \frac{(x_{i_1k} - p_k)(x_{j_1k} - p_k)}{[p_k(1-p_k)]^{-\alpha}} + \sum_{k=1}^{M} \frac{(x_{i_1k} - p_k)(x_{j_2k} - p_k)}{[p_k(1-p_k)]^{-\alpha}} + \right. \\ &\qquad \left. \sum_{k=1}^{M} \frac{(x_{i_2k} - p_k)(x_{j_1k} - p_k)}{[p_k(1-p_k)]^{-\alpha}} + \sum_{k=1}^{M} \frac{(x_{i_2k} - p_k)(x_{j_2k} - p_k)}{[p_k(1-p_k)]^{-\alpha}}\right] \\ &= 2^\alpha \left[K_{\alpha,hap}(i_1,j_1) + K_{\alpha,hap}(i_1,j_2) + K_{\alpha,hap}(i_2,j_1) + K_{\alpha,hap}(i_2,j_2)\right], \end{aligned}$$

where terms such as $K_{\alpha,hap}(i_1,j_1)$ correspond to the haploid GRM over $2N$ samples given by (A.1). Using scale invariance, the factor of 2^α can be removed to write

$$K_{\alpha,dip}(i,j) \equiv K_{\alpha,hap}(i_1,j_1) + K_{\alpha,hap}(i_1,j_2) + K_{\alpha,hap}(i_2,j_1) + K_{\alpha,hap}(i_2,j_2).$$

A similar expression may be obtained for the ARG-GRMs:

$$\begin{aligned}
K_{\alpha,dip,ARG}(i,j) &= \mathbb{E}\left[K_{\alpha,dip}(i,j)|ARG\right] \\
&\equiv \mathbb{E}\left[K_{\alpha,hap}(i_1,j_1)|ARG\right] + \mathbb{E}\left[K_{\alpha,hap}(i_1,j_2)|ARG\right] + \\
&\quad \mathbb{E}\left[K_{\alpha,hap}(i_2,j_1)|ARG\right] + \mathbb{E}\left[K_{\alpha,hap}(i_2,j_2)|ARG\right] \\
&= K_{\alpha,hap,ARG}(i_1,j_1) + K_{\alpha,hap,ARG}(i_1,j_2) + \\
&\quad K_{\alpha,hap,ARG}(i_2,j_1) + K_{\alpha,hap,ARG}(i_2,j_2).
\end{aligned}$$

The diploid ARG-GRM is therefore obtained by first computing the exact ARG-GRM of size $2N$ by $2N$ over the haploid samples and then using terms involving (i_1, j_1), (i_1, j_2), (i_2, j_1), and (i_2, j_2) to compute the entry for pair (i, j).

Acknowledgements

First, I would like to thank my close collaborators: my supervisor Pier Palamara and postdoctoral researcher Arjun Biddanda. Pier, your extensive knowledge of various scientific fields—computer science, statistical learning, and genetics—and the connections you make between them has been a highlight of working together. Thank you for sharing your scientific vision with me and for supporting me through various personal and research challenges. Arjun, it has been a pleasure getting to know you through this work and witnessing your optimism and openness.

Thanks to all the members of the Palamara Lab for creating a vibrant and supportive environment for pursuing research. In addition, many thanks to Fergus Cooper, Sinan Shi, Juba Nait Saada, and Yiorgos Kalantzis for sharing code used for various parts of the analysis; and to Árni Gunnarsson, Romain Fournier, Zoi Tsangalidou, and Juba Nait Saada for reading and providing feedback on sections of my thesis.

I am grateful to Jonathan Marchini for supervising the first year of my D.Phil, Geoff Nicholls for providing secondary supervision, and my examiners—Simon Myers and Richard Durbin—for taking the time to assess my thesis. I have also benefited from wider contacts at Oxford who provided feedback on my work or sharpened my thinking through our discussions. To Simon Myers, Robert Davies, Chris Holmes, Leo Speidel, Yan Wong, Jerome Kelleher, Wilder Wohns, and Chris Cole—thank you for inspiring me and building up our research community. I am grateful for the Clarendon Scholarship, the European Research Council, and the Oxford COVID-19 Scholarship Extension Fund which funded my D.Phil. My research is indebted to the Oxford Biomedical Research Computing (BMRC) facility and the participants of the UK Biobank study.

In the midst of a years-long pandemic, several people in my department and college went above and beyond to provide support to myself and other students. I would

like to recognise the incredible efforts of Alison Etheridge, Beverley Lane, Hannah Harrison, Jonathan Whyman, Emma Bodger, Stuart McRobert, Mark Feasey, Simon Patchett, Caroline Lordan, and all whose names I have surely missed.

* * *

For conversations and mentorship that first exposed me to genomics—thanks to Eran Hodis, Adrian Veres, Po-Ru Loh, Yakir Reshef, Kevin Bu, and Albert Young. You each modelled a fascinating career path that I have in some way sought to follow.

For supervising my past research experiences—thanks to João Carreira, Andrew Zisserman, and Andriy Mnih at DeepMind; and to Elaine Angelino, Ryan Adams, Ben Feldman, Gilad Ben-Shach, Amir Yacoby, Takuya Kitagawa, and Eugene Demler at Harvard. I am a better researcher today because I have learned from all of you.

For teaching me to ask the big questions, to seek simplicity and universality, and to persevere as a scientist—thanks to my Dad, Shoucheng Zhang (1963–2018). You and Mom gifted me with an education few could dream of. Working towards this D.Phil. has helped me better understand your world, your struggles, and your achievements. I hope to continue to create in ways that would make you proud.

* * *

A bit over a year into my D.Phil., my father tragically ended his own life. I will always remember the friends who supported me as I walked "through the valley of the shadow of death," especially Sean Lau, Melissa and Jonathan Grant-Peters, Kirsten Mackerras, Kevin Zhang, the Gifford family, David Tang Quan, Paul White, Allan Jiang, Henry Li, Tania Loke, Ta-Wei Lin, Nate Otey, Matt Kerin, Chris Gill, Juba Nait Saada, Petya Kindalova, and Pier Palamara. Thank you for all the ways you reached out and made time to sit with me in my pain.

To Xiaoliang Qi, Biao Lian, Pam Davis, Steve Kivelson, and Diane Greene—thank you for your boundless love and dedication to my family. I am reminded of my

father's legacy through your presence in our lives. For friends that encouraged me across the finish line, thank you Kelly and Ty Bache, Kamélia Daudel, Yucan Chiu, Jenna Mauer, Mohamed Elmi, the Yeh family, and the Johnson family.

To Stephanie and Charlie, Chinlin and Tung, and Mom and Dad—thank you for always believing in me. Thank you Mom for your incredible love, wisdom, and faith that has guided our family through all these years. I know that you will do all you can to seek my good, no matter where I journey in the world. Lastly, to my wife, Ruth—thank you for faithfully being by my side, for the encouragement and love you have provided me, and for your example to me and others. Every chapter with you has been an adventure, and I look forward to our next one together.

Glossary of Terms

Bibliography

[Adhikari et al., 2012] Adhikari, K., AlChawa, T., Ludwig, K., Mangold, E., Laird, N., and Lange, C. (2012). Is it rare or common? *Genetic Epidemiology*, 36(5):419–429.

[Albrechtsen et al., 2010] Albrechtsen, A., Moltke, I., and Nielsen, R. (2010). Natural selection and the distribution of identity-by-descent in the human genome. *Genetics*, 186(1):295–308.

[Aulchenko et al., 2007] Aulchenko, Y. S., De Koning, D.-J., and Haley, C. (2007). Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis. *Genetics*, 177(1):577–585.

[Backman et al., 2021] Backman, J. D., Li, A. H., Marcketta, A., Sun, D., Mbatchou, J., Kessler, M. D., Benner, C., Liu, D., Locke, A. E., Balasubramanian, S., et al. (2021). Exome sequencing and analysis of 454,787 UK Biobank participants. *Nature*, 599(7886):628–634.

[Barton et al., 2021] Barton, A. R., Sherman, M. A., Mukamel, R. E., and Loh, P.-R. (2021). Whole-exome imputation within UK Biobank powers rare coding variant association and fine-mapping analyses. *Nature Genetics*, 53(8):1260–1269.

[Böcker and Dress, 1998] Böcker, S. and Dress, A. W. (1998). Recovering symbolically dated, rooted trees from symbolic ultrametrics. *Advances in Mathematics*, 138(1):105–125.

[Browning and Browning, 2007a] Browning, B. L. and Browning, S. R. (2007a). Efficient multilocus association testing for whole genome association studies using localized haplotype clustering. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society*, 31(5):365–375.

[Browning and Browning, 2011] Browning, B. L. and Browning, S. R. (2011). A fast, powerful method for detecting identity by descent. *The American Journal of Human Genetics*, 88(2):173–182.

[Browning and Browning, 2016] Browning, B. L. and Browning, S. R. (2016). Genotype imputation with millions of reference samples. *The American Journal of Human Genetics*, 98(1):116–126.

[Browning et al., 2018] Browning, B. L., Zhou, Y., and Browning, S. R. (2018). A one-penny imputed genome from next-generation reference panels. *The American Journal of Human Genetics*, 103(3):338–348.

[Browning and Browning, 2007b] Browning, S. R. and Browning, B. L. (2007b). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *The American Journal of Human Genetics*, 81(5):1084–1097.

[Browning and Browning, 2012] Browning, S. R. and Browning, B. L. (2012). Identity by descent between distant relatives: detection and applications. *Annual Review of Genetics*, 46:617–633.

[Browning and Browning, 2013] Browning, S. R. and Browning, B. L. (2013). Identity-by-descent-based heritability analysis in the Northern Finland Birth Cohort. *Human Genetics*, 132(2):129–138.

[Browning and Thompson, 2012] Browning, S. R. and Thompson, E. A. (2012). Detecting rare variant associations by identity-by-descent mapping in case-control studies. *Genetics*, 190(4):1521–1531.

[Bulik-Sullivan et al., 2015] Bulik-Sullivan, B. K., Loh, P.-R., Finucane, H. K., Ripke, S., Yang, J., Patterson, N., Daly, M. J., Price, A. L., and Neale, B. M. (2015). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics*, 47(3):291–295.

[Burkett et al., 2014] Burkett, K. M., McNeney, B., Graham, J., and Greenwood, C. M. (2014). Using gene genealogies to detect rare variants associated with complex traits. *Human Heredity*, 78(3-4):117–130.

[Bycroft et al., 2018] Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., et al. (2018). The UK Biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203–209.

[Casella and Berger, 2002] Casella, G. and Berger, R. L. (2002). *Statistical Inference*. Duxbury Resource Center, 2nd edition.

[Chen et al., 2011] Chen, Z., Chen, J., Collins, R., Guo, Y., Peto, R., Wu, F., and Li, L. (2011). China Kadoorie Biobank of 0.5 million people: survey methods, baseline characteristics and long-term follow-up. *International Journal of Epidemiology*, 40(6):1652–1666.

[Churchill and Doerge, 1994] Churchill, G. A. and Doerge, R. W. (1994). Empirical threshold values for quantitative trait mapping. *Genetics*, 138(3):963–971.

[Claussnitzer et al., 2020] Claussnitzer, M., Cho, J. H., Collins, R., Cox, N. J., Dermitzakis, E. T., Hurles, M. E., Kathiresan, S., Kenny, E. E., Lindgren, C. M.,

MacArthur, D. G., et al. (2020). A brief history of human disease genetics. *Nature*, 577(7789):179–189.

[Claw et al., 2018] Claw, K. G., Anderson, M. Z., Begay, R. L., Tsosie, K. S., Fox, K., and Garrison, N. A. (2018). A framework for enhancing ethical genomic research with indigenous communities. *Nature Communications*, 9(1):1–7.

[Daetwyler et al., 2008] Daetwyler, H. D., Villanueva, B., and Woolliams, J. A. (2008). Accuracy of predicting the genetic risk of disease using a genome-wide approach. *PLoS ONE*, 3(10):e3395.

[Dahl et al., 2020] Dahl, A., Nguyen, K., Cai, N., Gandal, M. J., Flint, J., and Zaitlen, N. (2020). A robust method uncovers significant context-specific heritability in diverse complex traits. *The American Journal of Human Genetics*, 106(1):71–91.

[Delaneau et al., 2019] Delaneau, O., Zagury, J.-F., Robinson, M. R., Marchini, J. L., and Dermitzakis, E. T. (2019). Accurate, scalable and integrative haplotype estimation. *Nature Communications*, 10(1):1–10.

[Durbin, 2014] Durbin, R. (2014). Efficient haplotype matching and storage using the positional Burrows–Wheeler transform (PBWT). *Bioinformatics*, 30(9):1266–1272.

[Durrant et al., 2004] Durrant, C., Zondervan, K. T., Cardon, L. R., Hunt, S., Deloukas, P., and Morris, A. P. (2004). Linkage disequilibrium mapping via cladistic analysis of single-nucleotide polymorphism haplotypes. *The American Journal of Human Genetics*, 75(1):35–43.

[Durrett, 2008] Durrett, R. (2008). *Probability models for DNA sequence evolution*, volume 2. Springer.

[Evans et al., 2018a] Evans, L. M., Tahmasbi, R., Jones, M., Vrieze, S. I., Abecasis, G. R., Das, S., Bjelland, D. W., de Candia, T. R., Yang, J., Goddard, M. E., et al. (2018a). Narrow-sense heritability estimation of complex traits using identity-by-descent information. *Heredity*, 121(6):616–630.

[Evans et al., 2018b] Evans, L. M., Tahmasbi, R., Vrieze, S. I., Abecasis, G. R., Das, S., Gazal, S., Bjelland, D. W., De Candia, T. R., Goddard, M. E., Neale, B. M., et al. (2018b). Comparison of methods that use whole genome data to estimate the heritability and genetic architecture of complex traits. *Nature Genetics*, 50(5):737.

[Ewing and Hermisson, 2010] Ewing, G. and Hermisson, J. (2010). MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus. *Bioinformatics*, 26(16):2064–2065.

[Fan et al., 2021] Fan, C., Mancuso, N., and Chiang, C. W. (2021). A genealogical estimate of genetic relationships. *bioRxiv*.

[Finucane et al., 2015] Finucane, H. K., Bulik-Sullivan, B., Gusev, A., Trynka, G., Reshef, Y., Loh, P.-R., Anttila, V., Xu, H., Zang, C., Farh, K., et al. (2015). Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nature Genetics*, 47(11):1228.

[Fox, 2020] Fox, K. (2020). The illusion of inclusion—the "All of Us" research program and indigenous peoples' DNA. *New England Journal of Medicine*, 383(5):411–413.

[Gazal et al., 2017] Gazal, S., Finucane, H. K., Furlotte, N. A., Loh, P.-R., Palamara, P. F., Liu, X., Schoech, A., Bulik-Sullivan, B., Neale, B. M., Gusev, A., et al. (2017). Linkage disequilibrium–dependent architecture of human complex traits shows action of negative selection. *Nature Genetics*, 49(10):1421–1427.

[Gazal et al., 2018] Gazal, S., Loh, P.-R., Finucane, H. K., Ganna, A., Schoech, A., Sunyaev, S., and Price, A. L. (2018). Functional architecture of low-frequency variants highlights strength of negative selection across coding and non-coding annotations. *Nature Genetics*, 50(11):1600–1607.

[Gazal et al., 2019] Gazal, S., Marquez-Luna, C., Finucane, H. K., and Price, A. L. (2019). Reconciling S-LDSC and LDAK functional enrichment estimates. *Nature Genetics*, 51(8):1202–1204.

[Gaziano et al., 2016] Gaziano, J. M., Concato, J., Brophy, M., Fiore, L., Pyarajan, S., Breeling, J., Whitbourne, S., Deen, J., Shannon, C., Humphries, D., et al. (2016). Million Veteran Program: A mega-biobank to study genetic influences on health and disease. *Journal of Clinical Epidemiology*, 70:214–223.

[Ge et al., 2019] Ge, T., Chen, C.-Y., Ni, Y., Feng, Y.-C. A., and Smoller, J. W. (2019). Polygenic prediction via Bayesian regression and continuous shrinkage priors. *Nature Communications*, 10(1):1–10.

[Gilmour et al., 1995] Gilmour, A. R., Thompson, R., and Cullis, B. R. (1995). Average information REML: an efficient algorithm for variance parameter estimation in linear mixed models. *Biometrics*, pages 1440–1450.

[Griffiths and Marjoram, 1996] Griffiths, R. and Marjoram, P. (1996). Ancestral inference from samples of DNA sequences with recombination. *Journal of Computational Biology: a Journal of Computational Molecular Cell Biology*, 3(4):479–502.

[Griffiths and Marjoram, 1997] Griffiths, R. C. and Marjoram, P. (1997). An ancestral recombination graph. *Institute for Mathematics and its Applications*, 87:257.

[Gronau and Moran, 2007] Gronau, I. and Moran, S. (2007). Optimal implementations of UPGMA and other common clustering algorithms. *Information Processing Letters*, 104(6):205–210.

[Gusev et al., 2011] Gusev, A., Kenny, E. E., Lowe, J. K., Salit, J., Saxena, R., Kathiresan, S., Altshuler, D. M., Friedman, J. M., Breslow, J. L., and Pe'er, I. (2011). DASH: a method for identical-by-descent haplotype mapping uncovers association with recent variation. *The American Journal of Human Genetics*, 88(6):706–717.

[Gusev et al., 2014] Gusev, A., Lee, S. H., Trynka, G., Finucane, H., Vilhjálmsson, B. J., Xu, H., Zang, C., Ripke, S., Bulik-Sullivan, B., Stahl, E., et al. (2014). Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. *The American Journal of Human Genetics*, 95(5):535–552.

[Gusev et al., 2009] Gusev, A., Lowe, J. K., Stoffel, M., Daly, M. J., Altshuler, D., Breslow, J. L., Friedman, J. M., and Pe'er, I. (2009). Whole population, genome-wide mapping of hidden relatedness. *Genome Research*, 19(2):318–326.

[Gusev et al., 2012] Gusev, A., Palamara, P. F., Aponte, G., Zhuang, Z., Darvasi, A., Gregersen, P., and Pe'er, I. (2012). The architecture of long-range haplotypes shared within and across populations. *Molecular Biology and Evolution*, 29(2):473–486.

[Halldorsson et al., 2021] Halldorsson, B. V., Eggertsson, H. P., Moore, K. H., Hauswedell, H., Eiriksson, O., Ulfarsson, M. O., Palsson, G., Hardarson, M. T., Oddsson, A., Jensson, B. O., et al. (2021). The sequences of 150,119 genomes in the UK Biobank. *bioRxiv*.

[Haseman and Elston, 1972] Haseman, J. and Elston, R. (1972). The investigation of linkage between a quantitative trait and a marker locus. *Behavior Genetics*, 2(1):3–19.

[Hayden, 2014] Hayden, E. C. (2014). The $1,000 genome. *Nature*, 507(7492):294.

[Hayes et al., 2009] Hayes, B. J., Visscher, P. M., and Goddard, M. E. (2009). Increased accuracy of artificial selection by using the realized relationship matrix. *Genetics Research*, 91(1):47–60.

[Hein et al., 2004] Hein, J., Schierup, M., and Wiuf, C. (2004). *Gene genealogies, variation and evolution: a primer in coalescent theory*. Oxford University Press.

[Hill and Weir, 2011] Hill, W. G. and Weir, B. S. (2011). Variation in actual relationship as a consequence of mendelian sampling and linkage. *Genetics Research*, 93(1):47–64.

[Hobolth and Jensen, 2014] Hobolth, A. and Jensen, J. L. (2014). Markovian approximation to the finite loci coalescent with recombination along multiple sequences. *Theoretical Population Biology*, 98:48–58.

[Houwen et al., 1994] Houwen, R. H., Baharloo, S., Blankenship, K., Raeymaekers, P., Juyn, J., Sandkuijl, L. A., and Freimer, N. B. (1994). Genome screening by

searching for shared segments: mapping a gene for benign recurrent intrahepatic cholestasis. *Nature Genetics*, 8(4):380–386.

[Howie et al., 2012] Howie, B., Fuchsberger, C., Stephens, M., Marchini, J., and Abecasis, G. R. (2012). Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nature Genetics*, 44(8):955–959.

[Howie et al., 2009] Howie, B. N., Donnelly, P., and Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genetics*, 5(6):e1000529.

[Huang et al., 2015] Huang, J., Howie, B., McCarthy, S., Memari, Y., Walter, K., Min, J. L., Danecek, P., Malerba, G., Trabetti, E., Zheng, H.-F., et al. (2015). Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. *Nature Communications*, 6(1):1–9.

[Hubisz et al., 2020] Hubisz, M. J., Williams, A. L., and Siepel, A. (2020). Mapping gene flow between ancient hominins through demography-aware inference of the ancestral recombination graph. *PLoS Genetics*, 16(8):e1008895.

[Hudson, 1983] Hudson, R. R. (1983). Properties of a neutral allele model with intragenic recombination. *Theoretical Population Biology*, 23(2):183–201.

[Hudson, 2002] Hudson, R. R. (2002). Generating samples under a wright–fisher neutral model of genetic variation. *Bioinformatics*, 18(2):337–338.

[Hudson et al., 1990] Hudson, R. R. et al. (1990). Gene genealogies and the coalescent process. *Oxford Surveys in Evolutionary Biology*, 7(1):44.

[Jiang et al., 2021] Jiang, L., Zheng, Z., Fang, H., and Yang, J. (2021). A generalized linear mixed model association tool for biobank-scale data. *Nature Genetics*, 53(11):1616–1621.

[Jiang et al., 2019] Jiang, L., Zheng, Z., Qi, T., Kemper, K. E., Wray, N. R., Visscher, P. M., and Yang, J. (2019). A resource-efficient tool for mixed model association analysis of large-scale data. *Nature Genetics*, 51(12):1749–1755.

[Joyner and Paneth, 2015] Joyner, M. J. and Paneth, N. (2015). Seven questions for personalized medicine. *JAMA*, 314(10):999–1000.

[Kaiser, 2021] Kaiser, J. (2021). 200,000 whole genomes made available for biomedical studies. *Science*, 374(6571):1036.

[Kanai et al., 2016] Kanai, M., Tanaka, T., and Okada, Y. (2016). Empirical estimation of genome-wide significance thresholds based on the 1000 Genomes Project data set. *Journal of Human Genetics*, 61(10):861–866.

[Kang et al., 2010] Kang, H. M., Sul, J. H., Service, S. K., Zaitlen, N. A., Kong, S.-y., Freimer, N. B., Sabatti, C., Eskin, E., et al. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics*, 42(4):348.

[Kang et al., 2008] Kang, H. M., Zaitlen, N. A., Wade, C. M., Kirby, A., Heckerman, D., Daly, M. J., and Eskin, E. (2008). Efficient control of population structure in model organism association mapping. *Genetics*, 178(3):1709–1723.

[Kelleher et al., 2016] Kelleher, J., Etheridge, A. M., and McVean, G. (2016). Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLoS Computational Biology*, 12(5):e1004842.

[Kelleher et al., 2019] Kelleher, J., Wong, Y., Wohns, A. W., Fadil, C., Albers, P. K., and McVean, G. (2019). Inferring whole-genome histories in large population datasets. *Nature Genetics*, 51(9):1330–1338.

[Kendall and Colijn, 2016] Kendall, M. and Colijn, C. (2016). Mapping phylogenetic trees to reveal distinct patterns of evolution. *Molecular Biology and Evolution*, 33(10):2735–2743.

[Khera et al., 2018] Khera, A. V., Chaffin, M., Aragam, K. G., Haas, M. E., Roselli, C., Choi, S. H., Natarajan, P., Lander, E. S., Lubitz, S. A., Ellinor, P. T., et al. (2018). Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nature Genetics*, 50(9):1219–1224.

[Kingman, 1982] Kingman, J. F. C. (1982). The coalescent. *Stochastic Processes and their Applications*, 13(3):235–248.

[Kurki et al., 2022] Kurki, M. I., Karjalainen, J., Palta, P., Sipilä, T. P., Kristiansson, K., Donner, K., Reeve, M. P., Laivuori, H., Aavikko, M., Kaunisto, M. A., et al. (2022). FinnGen: Unique genetic insights from combining isolated population and national health register data. *medRxiv*.

[Lee et al., 2012] Lee, S. H., DeCandia, T. R., Ripke, S., Yang, J., Sullivan, P. F., Goddard, M. E., Keller, M. C., Visscher, P. M., and Wray, N. R. (2012). Estimating the proportion of variation in susceptibility to schizophrenia captured by common snps. *Nature Genetics*, 44(3):247–250.

[Lee et al., 2013] Lee, S. H., Yang, J., Chen, G.-B., Ripke, S., Stahl, E. A., Hultman, C. M., Sklar, P., Visscher, P. M., Sullivan, P. F., Goddard, M. E., et al. (2013). Estimation of SNP heritability from dense genotype data. *The American Journal of Human Genetics*, 93(6):1151–1155.

[Li and Durbin, 2011] Li, H. and Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. *Nature*, 475(7357):493–496.

[Li and Stephens, 2003] Li, N. and Stephens, M. (2003). Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics*, 165(4):2213–2233.

[Lippert et al., 2011] Lippert, C., Listgarten, J., Liu, Y., Kadie, C. M., Davidson, R. I., and Heckerman, D. (2011). FaST linear mixed models for genome-wide association studies. *Nature Methods*, 8(10):833.

[Listgarten et al., 2012] Listgarten, J., Lippert, C., Kadie, C. M., Davidson, R. I., Eskin, E., and Heckerman, D. (2012). Improved linear mixed models for genome-wide association studies. *Nature Methods*, 9(6):525–526.

[Loh et al., 2015a] Loh, P.-R., Bhatia, G., Gusev, A., Finucane, H. K., Bulik-Sullivan, B. K., Pollack, S. J., de Candia, T. R., Lee, S. H., Wray, N. R., Kendler, K. S., et al. (2015a). Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. *Nature Genetics*, 47(12):1385.

[Loh et al., 2016] Loh, P.-R., Danecek, P., Palamara, P. F., Fuchsberger, C., Reshef, Y. A., Finucane, H. K., Schoenherr, S., Forer, L., McCarthy, S., Abecasis, G. R., et al. (2016). Reference-based phasing using the Haplotype Reference Consortium panel. *Nature Genetics*, 48(11):1443–1448.

[Loh et al., 2018] Loh, P.-R., Kichaev, G., Gazal, S., Schoech, A. P., and Price, A. L. (2018). Mixed-model association for biobank-scale datasets. *Nature Genetics*, 50(7):906–908.

[Loh et al., 2015b] Loh, P.-R., Tucker, G., Bulik-Sullivan, B. K., Vilhjalmsson, B. J., Finucane, H. K., Salem, R. M., Chasman, D. I., Ridker, P. M., Neale, B. M., Berger, B., et al. (2015b). Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nature Genetics*, 47(3):284.

[Lynch et al., 1998] Lynch, M., Walsh, B., et al. (1998). *Genetics and Analysis of Quantitative Traits*. Sinauer Sunderland, MA.

[Lyngsø et al., 2005] Lyngsø, R. B., Song, Y. S., and Hein, J. (2005). Minimum recombination histories by branch and bound. In *International Workshop on Algorithms in Bioinformatics*, pages 239–250. Springer.

[MacKenzie, 1976] MacKenzie, D. (1976). Eugenics in Britain. *Social Studies of Science*, 6(3-4):499–532.

[Mahmoudi et al., 2022] Mahmoudi, A., Koskela, J., Kelleher, J., Chan, Y.-b., and Balding, D. (2022). Bayesian inference of ancestral recombination graphs. *PLOS Computational Biology*, 18(3):e1009960.

[Manolio et al., 2009] Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., et al. (2009). Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753.

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

[Marchini et al., 2007] Marchini, J., Howie, B., Myers, S., McVean, G., and Donnelly, P. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature Genetics*, 39(7):906–913.

[Marjoram and Wall, 2006] Marjoram, P. and Wall, J. D. (2006). Fast "coalescent" simulation. *BMC Genetics*, 7(1):1–9.

[Márquez-Luna et al., 2021] Márquez-Luna, C., Gazal, S., Loh, P.-R., Kim, S. S., Furlotte, N., Auton, A., and Price, A. L. (2021). Incorporating functional priors improves polygenic prediction accuracy in UK Biobank and 23andMe data sets. *Nature Communications*, 12(1):1–11.

[Martin et al., 2019] Martin, A. R., Kanai, M., Kamatani, Y., Okada, Y., Neale, B. M., and Daly, M. J. (2019). Clinical use of current polygenic risk scores may exacerbate health disparities. *Nature Genetics*, 51(4):584–591.

[Mbatchou et al., 2021] Mbatchou, J., Barnard, L., Backman, J., Marcketta, A., Kosmicki, J. A., Ziyatdinov, A., Benner, C., O'Dushlaine, C., Barber, M., Boutkov, B., et al. (2021). Computationally efficient whole-genome regression for quantitative and binary traits. *Nature Genetics*, 53(7):1097–1103.

[McCarthy et al., 2016] McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A. R., Teumer, A., Kang, H. M., Fuchsberger, C., Danecek, P., Sharp, K., et al. (2016). A reference panel of 64,976 haplotypes for genotype imputation. *Nature Genetics*, 48(10):1279.

[McLaren et al., 2016] McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R., Thormann, A., Flicek, P., and Cunningham, F. (2016). The Ensembl Variant Effect Predictor. *Genome Biology*, 17(1):1–14.

[McVean and Cardin, 2005] McVean, G. A. and Cardin, N. J. (2005). Approximating the coalescent with recombination. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1459):1387–1393.

[Mefford et al., 2020] Mefford, J., Park, D., Zheng, Z., Ko, A., Ala-Korpela, M., Laakso, M., Pajukanta, P., Yang, J., Witte, J., and Zaitlen, N. (2020). Efficient estimation and applications of cross-validated genetic predictions to polygenic risk scores and linear mixed models. *Journal of Computational Biology*, 27(4):599–612.

[Minichiello and Durbin, 2006] Minichiello, M. J. and Durbin, R. (2006). Mapping trait loci by use of inferred ancestral recombination graphs. *The American Journal of Human Genetics*, 79(5):910–922.

[Mirzaei and Wu, 2017] Mirzaei, S. and Wu, Y. (2017). RENT+: an improved method for inferring local genealogical trees from haplotypes with recombination. *Bioinformatics*, 33(7):1021–1030.

[Müllner, 2013] Müllner, D. (2013). fastcluster: Fast hierarchical, agglomerative clustering routines for R and Python. *Journal of Statistical Software*, 53(9):1–18.

[Myers et al., 2005] Myers, S., Bottolo, L., Freeman, C., McVean, G., and Donnelly, P. (2005). A fine-scale map of recombination rates and hotspots across the human genome. *Science*, 310(5746):321–324.

[Nagai et al., 2017] Nagai, A., Hirata, M., Kamatani, Y., Muto, K., Matsuda, K., Kiyohara, Y., Ninomiya, T., Tamakoshi, A., Yamagata, Z., Mushiroda, T., et al. (2017). Overview of the BioBank Japan Project: study design and profile. *Journal of Epidemiology*, 27(3S):S2–S8.

[Nait Saada et al., 2021] Nait Saada, J., Hu, A., and Palamara, P. F. (2021). Inference of pairwise coalescence times and allele ages using deep neural networks. *Learning Meaningful Representations of Life (NeurIPS 2021)*.

[Nait Saada et al., 2020] Nait Saada, J., Kalantzis, G., Shyr, D., Cooper, F., Robinson, M., Gusev, A., and Palamara, P. F. (2020). Identity-by-descent detection across 487,409 British samples reveals fine scale population structure and ultra-rare variant associations. *Nature Communications*, 11(1):1–15.

[Nelson et al., 2015] Nelson, M. R., Tipney, H., Painter, J. L., Shen, J., Nicoletti, P., Shen, Y., Floratos, A., Sham, P. C., Li, M. J., Wang, J., et al. (2015). The support of human genetic evidence for approved drug indications. *Nature Genetics*, 47(8):856–860.

[Palamara, 2016] Palamara, P. F. (2016). ARGON: fast, whole-genome simulation of the discrete time Wright-Fisher process. *Bioinformatics*, 32(19):3032–3034.

[Palamara et al., 2012] Palamara, P. F., Lencz, T., Darvasi, A., and Pe'er, I. (2012). Length distributions of identity by descent reveal fine-scale demographic history. *The American Journal of Human Genetics*, 91(5):809–822.

[Palamara and Pe'er, 2013] Palamara, P. F. and Pe'er, I. (2013). Inference of historical migration rates via haplotype sharing. *Bioinformatics*, 29(13):i180–i188.

[Palamara et al., 2016] Palamara, P. F., Terhorst, J., Song, Y., and Price, A. (2016). Leveraging deep genealogical structure to estimate the phenotypic contribution of rare variants. Presented at the 66th Annual Meeting of The American Society of Human Genetics, Vancouver. Abstract 1931T.

[Palamara et al., 2018] Palamara, P. F., Terhorst, J., Song, Y. S., and Price, A. L. (2018). High-throughput inference of pairwise coalescence times identifies signals of selection and enriched disease heritability. *Nature Genetics*, 50(9):1311–1317.

[Patterson and Thompson, 1971] Patterson, H. D. and Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58(3):545–554.

[Pazokitoroudi et al., 2020] Pazokitoroudi, A., Wu, Y., Burch, K. S., Hou, K., Zhou, A., Pasaniuc, B., and Sankararaman, S. (2020). Efficient variance components analysis across millions of genomes. *Nature Communications*, 11(1):1–10.

[Pe'er et al., 2008] Pe'er, I., Yelensky, R., Altshuler, D., and Daly, M. J. (2008). Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. *Genetic Epidemiology*, 32(4):381–385.

[Popejoy and Fullerton, 2016] Popejoy, A. B. and Fullerton, S. M. (2016). Genomics is failing on diversity. *Nature*, 538(7624):161–164.

[Purcell et al., 2007] Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., De Bakker, P. I., Daly, M. J., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81(3):559–575.

[Ralph and Coop, 2013] Ralph, P. and Coop, G. (2013). The geography of recent genetic ancestry across Europe. *PLoS Biology*, 11(5):e1001555.

[Rasmussen et al., 2014] Rasmussen, M. D., Hubisz, M. J., Gronau, I., and Siepel, A. (2014). Genome-wide inference of ancestral recombination graphs. *PLoS Genetics*, 10(5):e1004342.

[Reich et al., 2001] Reich, D. E., Cargill, M., Bolk, S., Ireland, J., Sabeti, P. C., Richter, D. J., Lavery, T., Kouyoumjian, R., Farhadian, S. F., Ward, R., et al. (2001). Linkage disequilibrium in the human genome. *Nature*, 411(6834):199–204.

[Risch and Merikangas, 1996] Risch, N. and Merikangas, K. (1996). The future of genetic studies of complex human diseases. *Science*, 273(5281):1516–1517.

[Robinson and Foulds, 1981] Robinson, D. F. and Foulds, L. R. (1981). Comparison of phylogenetic trees. *Mathematical Biosciences*, 53(1-2):131–147.

[Rubinacci et al., 2020] Rubinacci, S., Delaneau, O., and Marchini, J. (2020). Genotype imputation using the Positional Burrows Wheeler Transform. *PLoS Genetics*, 16(11):e1009049.

[Schaefer et al., 2021] Schaefer, N. K., Shapiro, B., and Green, R. E. (2021). An ancestral recombination graph of human, Neanderthal, and Denisovan genomes. *Science Advances*, 7(29):eabc0776.

[Schaid et al., 2018] Schaid, D. J., Chen, W., and Larson, N. B. (2018). From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nature Reviews Genetics*, 19(8):491–504.

[Schiffels and Durbin, 2014] Schiffels, S. and Durbin, R. (2014). Inferring human population size and separation history from multiple genome sequences. *Nature Genetics*, 46(8):919–925.

[Schoech et al., 2019] Schoech, A. P., Jordan, D. M., Loh, P.-R., Gazal, S., O'Connor, L. J., Balick, D. J., Palamara, P. F., Finucane, H. K., Sunyaev, S. R., and Price, A. L. (2019). Quantification of frequency-dependent genetic architectures in 25 UK Biobank traits reveals action of negative selection. *Nature Communications*, 10(1):1–10.

[Schork et al., 2022] Schork, A. J., Peterson, R. E., Dahl, A., Cai, N., and Kendler, K. S. (2022). Indirect paths from genetics to education. *Nature Genetics*, pages 1–2.

[Schork, 2015] Schork, N. J. (2015). Personalized medicine: time for one-person trials. *Nature*, 520(7549):609–611.

[Schwarze et al., 2020] Schwarze, K., Buchanan, J., Fermont, J. M., Dreau, H., Tilley, M. W., Taylor, J. M., Antoniou, P., Knight, S. J., Camps, C., Pentony, M. M., et al. (2020). The complete costs of genome sequencing: a microcosting study in cancer and rare diseases from a single center in the United Kingdom. *Genetics in Medicine*, 22(1):85–94.

[Sheehan et al., 2013] Sheehan, S., Harris, K., and Song, Y. S. (2013). Estimating variable effective population sizes from multiple genomes: a sequentially Markov conditional sampling distribution approach. *Genetics*, 194(3):647–662.

[Shlyakhter et al., 2014] Shlyakhter, I., Sabeti, P. C., and Schaffner, S. F. (2014). Cosi2: an efficient simulator of exact and approximate coalescent with selection. *Bioinformatics*, 30(23):3427–3429.

[Si et al., 2021] Si, Y., Vanderwerff, B., and Zöllner, S. (2021). Why are rare variants hard to impute? Coalescent models reveal theoretical limits in existing algorithms. *Genetics*, 217(4):iyab011.

[Simonsen and Churchill, 1997] Simonsen, K. L. and Churchill, G. A. (1997). A Markov chain model of coalescence with recombination. *Theoretical Population Biology*, 52(1):43–59.

[Slatkin, 2008] Slatkin, M. (2008). Linkage disequilibrium—understanding the evolutionary past and mapping the medical future. *Nature Reviews Genetics*, 9(6):477–485.

[Sneath and Sokal, 1973] Sneath, P. H. and Sokal, R. R. (1973). *Numerical taxonomy. The principles and practice of numerical classification.* W. H.. Freeman and Co.

[Speed and Balding, 2014] Speed, D. and Balding, D. J. (2014). MultiBLUP: improved SNP-based prediction for complex traits. *Genome Research*, 24(9):1550–1557.

[Speed et al., 2017] Speed, D., Cai, N., Johnson, M. R., Nejentsev, S., and Balding, D. J. (2017). Reevaluation of snp heritability in complex human traits. *Nature Genetics*, 49(7):986–992.

[Speed et al., 2012] Speed, D., Hemani, G., Johnson, M. R., and Balding, D. J. (2012). Improved heritability estimation from genome-wide SNPs. *The American Journal of Human Genetics*, 91(6):1011–1021.

[Speidel et al., 2019] Speidel, L., Forest, M., Shi, S., and Myers, S. R. (2019). A method for genome-wide genealogy estimation for thousands of samples. *Nature Genetics*, 51(9):1321–1329.

[Stern et al., 2021] Stern, A. J., Speidel, L., Zaitlen, N. A., and Nielsen, R. (2021). Disentangling selection on genetically correlated polygenic traits via whole-genome genealogies. *The American Journal of Human Genetics*, 108(2):219–239.

[Svishcheva et al., 2012] Svishcheva, G. R., Axenovich, T. I., Belonogova, N. M., Van Duijn, C. M., and Aulchenko, Y. S. (2012). Rapid variance components-based method for whole-genome association analysis. *Nature Genetics*, 44(10):1166–1170.

[Szustakowski et al., 2021] Szustakowski, J. D., Balasubramanian, S., Kvikstad, E., Khalid, S., Bronson, P. G., Sasson, A., Wong, E., Liu, D., Wade Davis, J., Haefliger, C., et al. (2021). Advancing human genetics research and drug discovery through exome sequencing of the UK Biobank. *Nature Genetics*, 53(7):942–948.

[Taliun et al., 2021] Taliun, D., Harris, D. N., Kessler, M. D., Carlson, J., Szpiech, Z. A., Torres, R., Taliun, S. A. G., Corvelo, A., Gogarten, S. M., Kang, H. M., et al. (2021). Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature*, 590(7845):290–299.

[Te Meerman et al., 1995] Te Meerman, G., Van der Meulen, M., and Sandkuijl, L. (1995). Perspectives of identity by descent (IBD) mapping in founder populations. *Clinical & Experimental Allergy*, 25:97–102.

[Templeton et al., 1992] Templeton, A. R., Crandall, K. A., and Sing, C. F. (1992). A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping and DNA sequence data. III. cladogram estimation. *Genetics*, 132(2):619–633.

[Terhorst et al., 2017] Terhorst, J., Kamm, J. A., and Song, Y. S. (2017). Robust and scalable inference of population history from hundreds of unphased whole genomes. *Nature Genetics*, 49(2):303–309.

[Thompson, 2013] Thompson, E. A. (2013). Identity by descent: variation in meiosis, across genomes, and in populations. *Genetics*, 194(2):301–326.

[Turley et al., 2021] Turley, P., Meyer, M. N., Wang, N., Cesarini, D., Hammonds, E., Martin, A. R., Neale, B. M., Rehm, H. L., Wilkins-Haug, L., Benjamin, D. J., et al. (2021). Problems with using polygenic scores to select embryos. *New England Journal of Medicine*, 385(1):78–86.

[Turnbull et al., 2018] Turnbull, C., Scott, R. H., Thomas, E., Jones, L., Murugaesu, N., Pretty, F. B., Halai, D., Baple, E., Craig, C., Hamblin, A., et al. (2018). The 100 000 Genomes Project: bringing whole genome sequencing to the NHS. *BMJ*, 361.

[Van Hout et al., 2020] Van Hout, C. V., Tachmazidou, I., Backman, J. D., Hoffman, J. D., Liu, D., Pandey, A. K., Gonzaga-Jauregui, C., Khalid, S., Ye, B., Banerjee, N., et al. (2020). Exome sequencing and characterization of 49,960 individuals in the UK Biobank. *Nature*, 586(7831):749–756.

[VanRaden, 2008] VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *Journal of Dairy Science*, 91(11):4414–4423.

[Vilhjálmsson et al., 2015] Vilhjálmsson, B. J., Yang, J., Finucane, H. K., Gusev, A., Lindström, S., Ripke, S., Genovese, G., Loh, P.-R., Bhatia, G., Do, R., et al. (2015). Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *The American Journal of Human Genetics*, 97(4):576–592.

[Visscher et al., 2006] Visscher, P. M., Medland, S. E., Ferreira, M. A. R., Morley, K. I., Zhu, G., Cornes, B. K., Montgomery, G. W., and Martin, N. G. (2006). Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. *PLoS Genetics*, 2(3):e41.

[Visscher et al., 2021] Visscher, P. M., Yengo, L., Cox, N. J., and Wray, N. R. (2021). Discovery and implications of polygenicity of common diseases. *Science*, 373(6562):1468–1473.

[Wainschtein et al., 2022] Wainschtein, P., Jain, D., Zheng, Z., Cupples, L. A., Shadyab, A. H., McKnight, B., Shoemaker, B. M., Mitchell, B. D., Psaty, B. M., Kooperberg, C., et al. (2022). Assessing the contribution of rare variants to complex trait heritability from whole-genome sequence data. *Nature Genetics*, pages 1–11.

[Wakeley and Wilton, 2016] Wakeley, J. and Wilton, P. R. (2016). Coalescent and models of identity by descent.

[Wasik et al., 2021] Wasik, K., Berisa, T., Pickrell, J. K., Li, J. H., Fraser, D. J., King, K., and Cox, C. (2021). Comparing low-pass sequencing and genotyping for trait mapping in pharmacogenetics. *BMC Genomics*, 22(1):1–7.

[Whitlock and Barton, 1997] Whitlock, M. C. and Barton, N. H. (1997). The effective size of a subdivided population. *Genetics*, 146(1):427–441.

[Wiuf and Hein, 1999] Wiuf, C. and Hein, J. (1999). Recombination as a point process along sequences. *Theoretical Population Biology*, 55(3):248–259.

[Wohns et al., 2022] Wohns, A. W., Wong, Y., Jeffery, B., Akbari, A., Mallick, S., Pinhasi, R., Patterson, N., Reich, D., Kelleher, J., and McVean, G. (2022). A unified genealogy of modern and ancient genomes. *Science*, 375(6583):eabi8264.

[Wray et al., 2013] Wray, N. R., Yang, J., Hayes, B. J., Price, A. L., Goddard, M. E., and Visscher, P. M. (2013). Pitfalls of predicting complex traits from SNPs. *Nature Reviews Genetics*, 14(7):507–515.

[Wu, 2008] Wu, Y. (2008). Association mapping of complex diseases with ancestral recombination graphs: models and efficient algorithms. *Journal of Computational Biology*, 15(7):667–684.

[Wu and Sankararaman, 2018] Wu, Y. and Sankararaman, S. (2018). A scalable estimator of SNP heritability for biobank-scale data. *Bioinformatics*, 34(13):i187–i194.

[Yang et al., 2015] Yang, J., Bakshi, A., Zhu, Z., Hemani, G., Vinkhuyzen, A. A., Lee, S. H., Robinson, M. R., Perry, J. R., Nolte, I. M., van Vliet-Ostaptchouk, J. V., et al. (2015). Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nature Genetics*, 47(10):1114.

[Yang et al., 2010] Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., Madden, P. A., Heath, A. C., Martin, N. G., Montgomery, G. W., et al. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics*, 42(7):565.

[Yang et al., 2012] Yang, J., Ferreira, T., Morris, A. P., Medland, S. E., Madden, P. A., Heath, A. C., Martin, N. G., Montgomery, G. W., Weedon, M. N., Loos, R. J., et al. (2012). Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nature Genetics*, 44(4):369–375.

[Yang et al., 2011a] Yang, J., Lee, S. H., Goddard, M. E., and Visscher, P. M. (2011a). GCTA: a tool for genome-wide complex trait analysis. *The American Journal of Human Genetics*, 88(1):76–82.

[Yang et al., 2011b] Yang, J., Manolio, T. A., Pasquale, L. R., Boerwinkle, E., Caporaso, N., Cunningham, J. M., De Andrade, M., Feenstra, B., Feingold, E., Hayes, M. G., et al. (2011b). Genome partitioning of genetic variation for complex traits using common snps. *Nature Genetics*, 43(6):519–525.

[Yang et al., 2014] Yang, J., Zaitlen, N. A., Goddard, M. E., Visscher, P. M., and Price, A. L. (2014). Advantages and pitfalls in the application of mixed-model association methods. *Nature Genetics*, 46(2):100.

[Yasumizu et al., 2020] Yasumizu, Y., Sakaue, S., Konuma, T., Suzuki, K., Matsuda, K., Murakami, Y., Kubo, M., Palamara, P. F., Kamatani, Y., and Okada, Y. (2020). Genome-wide natural selection signatures are linked to genetic risk of modern phenotypes in the Japanese population. *Molecular Biology and Evolution*, 37(5):1306–1316.

[Yengo et al., 2018] Yengo, L., Sidorenko, J., Kemper, K. E., Zheng, Z., Wood, A. R., Weedon, M. N., Frayling, T. M., Hirschhorn, J., Yang, J., Visscher, P. M., et al. (2018). Meta-analysis of genome-wide association studies for height and body mass index in ∼700000 individuals of European ancestry. *Human Molecular Genetics*, 27(20):3641–3649.

[Young et al., 2018] Young, A. I., Frigge, M. L., Gudbjartsson, D. F., Thorleifsson, G., Bjornsdottir, G., Sulem, P., Masson, G., Thorsteinsdottir, U., Stefansson, K., and Kong, A. (2018). Relatedness disequilibrium regression estimates heritability without environmental bias. *Nature Genetics*, 50(9):1304–1310.

[Yu et al., 2006] Yu, J., Pressoir, G., Briggs, W. H., Vroh Bi, I., Yamasaki, M., Doebley, J. F., McMullen, M. D., Gaut, B. S., Nielsen, D. M., Holland, J. B., et al. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics*, 38(2):203–208.

[Zaitlen and Kraft, 2012] Zaitlen, N. and Kraft, P. (2012). Heritability in the genome-wide association era. *Human Genetics*, 131(10):1655–1664.

[Zaitlen et al., 2013] Zaitlen, N., Kraft, P., Patterson, N., Pasaniuc, B., Bhatia, G., Pollack, S., and Price, A. L. (2013). Using extended genealogy to estimate components of heritability for 23 quantitative and dichotomous traits. *PLoS Genetics*, 9(5):e1003520.

[Zeng et al., 2018] Zeng, J., De Vlaming, R., Wu, Y., Robinson, M. R., Lloyd-Jones, L. R., Yengo, L., Yap, C. X., Xue, A., Sidorenko, J., McRae, A. F., et al. (2018). Signatures of negative selection in the genetic architecture of human complex traits. *Nature Genetics*, 50(5):746.

[Zhang et al., 2021] Zhang, B. C., Biddanda, A., and Palamara, P. F. (2021). Biobank-scale inference of ancestral recombination graphs enables genealogy-based mixed model association of complex traits. *bioRxiv*.

[Zhang et al., 2010] Zhang, Z., Ersoz, E., Lai, C.-Q., Todhunter, R. J., Tiwari, H. K., Gore, M. A., Bradbury, P. J., Yu, J., Arnett, D. K., Ordovas, J. M., et al. (2010). Mixed linear model approach adapted for genome-wide association studies. *Nature Genetics*, 42(4):355–360.

[Zhao et al., 2007] Zhao, K., Aranzana, M. J., Kim, S., Lister, C., Shindo, C., Tang, C., Toomajian, C., Zheng, H., Dean, C., Marjoram, P., et al. (2007). An Arabidopsis example of association mapping in structured samples. *PLoS Genetics*, 3(1):e4.

[Zhou et al., 2018] Zhou, W., Nielsen, J. B., Fritsche, L. G., Dey, R., Gabrielsen, M. E., Wolford, B. N., LeFaive, J., VandeHaar, P., Gagliano, S. A., Gifford, A., et al. (2018). Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nature Genetics*, 50(9):1335–1341.

[Zhou et al., 2020] Zhou, W., Zhao, Z., Nielsen, J. B., Fritsche, L. G., LeFaive, J., Taliun, S. A. G., Bi, W., Gabrielsen, M. E., Daly, M. J., Neale, B. M., et al. (2020). Scalable generalized linear mixed model for region-based association tests in large biobanks and cohorts. *Nature Genetics*, 52(6):634–639.

[Zhou et al., 2013] Zhou, X., Carbonetto, P., and Stephens, M. (2013). Polygenic modeling with Bayesian sparse linear mixed models. *PLoS Genetics*, 9(2):e1003264.

[Zhou and Stephens, 2012] Zhou, X. and Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics*, 44(7):821–824.

[Zöllner and Pritchard, 2005] Zöllner, S. and Pritchard, J. K. (2005). Coalescent-based association mapping and fine mapping of complex trait loci. *Genetics*, 169(2):1071–1092.