Imputation and Fine-Mapping in Genetic Association Studies

by

Ketian Yu

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in the University of Michigan
2023

Doctoral Committee:

Professor Gonçalo Abecasis, Chair
Professor Jean Morrison
Professor Jennifer A. Smith
Professor Xiaoquan Wen

Ketian Yu
yukt@umich.edu
ORCID iD: 0000-0001-9994-3399

DEDICATION

To my family and friends.

ACKNOWLEDGMENTS

First and foremost, I would like to express my deepest gratitude to my adviser, Professor Goncalo Abecasis for providing continuous support and guidance for my PhD studies and sharing with me his brilliant insights into genetic research. I would like to express my gratitude to Prof. William Wen, for invaluable advice in statistics and also for encouragements when I felt down.

My sincere thanks also go to the entire Abecasis group. I especially want to acknowledge Sayantan Das, who, during my initial days as a fresh Ph.D. candidate, guided me through the basics of genotype imputation. I would also like to extend my gratitude to Jonathon Lefaive for his invaluable assistance with coding across several projects.

I also would like to thank my cat Darth for accompany during the most stressful and isolated days in pandemic, an thank my doctor for prescribing me this wonderful emotional support animal.

Lastly, but most importantly, thanks to my family and friends.

TABLE OF CONTENTS

LIST OF FIGURES

FIGURE

LIST OF TABLES

TABLE

LIST OF APPENDICES

LIST OF ACRONYMS

1000G 1000 Genomes

ASW African Ancestry in Southwest US

BVS Bayesian Variable Selection

eQTL Expression Quantitative Loci

GC genomic control

GTEx Genotype-Tissue Expression

GWAS Genome-Wide Association Studies

HGDP Human Genome Diversity Project

HMM hidden Markov Model

HRC Haplotype Reference Consortium

LD linkage disequilibrium

LOO leave-one-out

MAF minor allele frequency

MIS Michigan Imputation Server

NGS next-generation sequencing

PBWT positional Burrows-Wheeler transform

PCA principal component analysis

RNA-seq RNA sequencing

SIS Sanger Imputation Service

SNPs single nucleotide polymorphisms

TOPMed Trans-Omics for Precision Medicine

TWAS Transcriptome-Wide Association Study

WES whole-exome sequencing

WGS whole-genome sequencing

ABSTRACT

Increasing availability of large whole genome sequencing and genomics data have brought both opportunities and challenges in genetic research. Genotype imputation is an integral tool in genome-wide association studies, where it facilitates meta-analysis, increases power and enables fine-mapping. With access to a multitude of reference panel choices for genotype imputation, investigators start to explore ways of utilizing information from different panels for better accuracy. The successive increase in sample size and genotype density in sequencing projects also enables high-resolution fine mapping, which improves the understanding of the underlying mechanisms of complex diseases. However, there is a reasonable chance that the lead variants from fine-mapping are not causal but are detected simply due to linkage disequilibrium (LD) with true causal variants, so caution is required when interpreting the association signals. In this dissertation, we present improved methods for genotype imputation and gene expression imputation, explore challenges in fine-mapping that result from complex LD structure and provide potential remedies.

In Chapter 2, we described an efficient meta-imputation framework that enables researchers to merge imputed data generated from multiple reference panels without the need to access individual-level genotype data for the underlying reference samples. We first impute against different reference panels separately using our minimac4 imputation software with a new built-in leave-one-out (LOO) imputation feature, and then combine the imputed results into a consensus dataset using weights that are tailored to each individual and genome segment. The weights are dynamically estimated through a hidden Markov model utilizing individual-specific LOO results. In the scenarios we examined, meta-imputation consistently outperforms imputation using a single reference panel and achieves comparable accuracy to imputation using a combined reference panel.

In Chapter 3, we presented a comprehensive exploration of the trade-offs associated with statistical fine-mapping strategies. We particularly focused on the impacts of the choice of data type (summary statistics versus individual-level data) and the algorithmic approach (greedy versus multiple starting-point strategy). Our evaluations revealed that using summary statistics typically resulted in decreased power and coverage in fine-mapping. We also highlighted the issues of non-identifiability in the presence of complex LD structures, a scenario where a greedy search strategy might overlook alternate model configurations, leading to false discoveries. To address this, we proposed a multiple starting-point strategy to improve the calibration of posterior probabilities,

albeit at an increased computational cost.

In Chapter 4, we systematically compared models for gene expression imputation based on TOPMed RNAseq data, and revealed a positive correlation between imputation accuracy and both reference sample size and degree of ancestry matching between reference and target samples. The study demonstrates that a large, diverse reference panel can achieve accuracy comparable to that of a smaller, ancestry-specific panel. This finding obviates the need to classify target samples into ancestry groups and carry out imputations using the corresponding ancestry-matching subpanels, thereby enhancing processing efficiency. Moreover, we have crafted gene expression imputation models based on DAP-G, leveraging TOPMed RNAseq data, to support transcriptome-wide association studies. This feature will soon be integrated into the TOPMed imputation server, creating a unified platform where users can access both imputed gene expressions and genotypes.

CHAPTER 1

Introduction

1.1 Background

Genome-Wide Association Studies (GWAS) are an essential tool in modern genetics research where they significantly advanced our understanding of the genetic architecture of many complex traits and diseases. As of March 2023, over 6,300 GWAS have been conducted on more than 5,000 human traits [72]. These have successfully identified numerous risk loci associated with a wide range of diseases including coronary artery disease [80], Type 2 Diabetes [70, 74], cancers [42, 63, 69], autoimmune disorders [25, 43, 79], psychiatric diseases [64, 75, 92] and many others. These findings have provided insights into the disease mechanisms and potential therapeutic targets, which are valuable for improving diagnostic accuracy and guiding the development of personalized treatments. For example, the discovery of the fat mass and obsesity-associated gene (FTO) has spurred further research into therapeutic targets for weight management [29, 36]. The identification of the interleukin-23 (IL-23) pathway through GWAS has led to the development of induction therapies such as risankizumab and ustekinumab for plaque psoriasis and Crohn's disease [24, 27, 35, 79]. Moreover, the finding that BCL11A acts as a repressor of fetal hemoglobin levels has led to the development of gene therapies to treat sickle cell disease [26, 28]. These targeted therapies have demonstrated significant clinical benefits for patients, enhancing disease management and overall quality of life.

The evolution of GWAS has been facilitated by advances of genotyping technologies. Early

genotyping approaches such as PCR-based assays were relatively expensive and lowthroughput, which restricted their application to linkage analysis and candidate-gene studies. The development of microarray technology in the early 2000s revolutionized genotyping by allowing the simultaneous analysis of hundreds of thousands of single nucleotide polymorphisms (SNPs), which was instrumental in the initial wave of GWAS. Klein et al. (2005) [47] published the very first GWAS study. Utilizing the Affymetrix GeneChip 100K Mapping Array Set, they conducted a genome-wide scan of 103,611 SNPs across 96 cases and 50 controls and identified a strong association in the complement factor H gene (CFH) with age-related macular degeneration (AMD). In 2007, the Wellcome Trust Case Control Consortium (WTCCC) [15] published a genome-wide association study of seven major diseases, involving a total of 17,000 samples genotyped with the GeneChip 500K Mapping Array Set. This landmark study not only pioneered the use of shared controls but also provided valuable methodological insights into study design and significance thresholds, laying the foundation for future GWAS analyses. As genotyping technologies progressed, higher-density arrays with increased resolution and coverage facilitated more comprehensive GWAS, enabling the identification of additional risk loci. Exome chips, which target rare and low-frequency protein-coding variants, aid in pinpointing functional genes associated with complex traits and diseases [56, 88]. Customized arrays were also developed for specific research areas. Examples include the ImmunoChip [16] and MetaboChip [83], which focused on immunological and metabolic diseases, respectively.

The advent of next-generation sequencing (NGS) platforms in mid-2000s further expanded the capacity of GWAS and enabled large-scale whole-genome sequencing (WGS), whole-exome sequencing (WES) and RNA sequencing (RNA-seq) [33]. The 1000 Genomes (1000G) Project stands as one of the pioneering large-scale studies that employed genome-wide sequencing using high-throughput platforms [11]. This project established a comprehensive catalog of human genetic variation by reconstructing the genomes of 2,504 individuals from 26 populations [12], utilizing a combination of low coverage WGS, deep exome sequencing, and dense microarray genotyping. Larger variation catalogs, such as 1000G and the earlier HapMap Project [14], of-

fer improved coverage of the human genome and deliver valuable insights into the frequency and linkage disequilibrium (LD) patterns of genetic variations. This wealth of information has greatly enhanced the design and selection of variants for genotyping arrays, leading to more efficient and accurate genetic analyses. Additionally, these catalogs have facilitated genotype imputation, which will be discussed later in this chapter.

Over the past two decades, DNA sequencing cost has reduced dramatically with the cost per human genome falling to $562 as of August 2021, compared to $3 million per genome in January 2008 when sequencing centers transitioned from Sanger-based to NGS technologies [89]. This continuous reduction in cost spurred a rapid surge in the volume of DNA sequence data. Large-scale projects, such as the Trans-Omics for Precision Medicine (TOPMed) Program, which has generated WGS from more than 130,000 samples [78], and the UK Biobank, which has released WGS data for 150,119 participants [34] and WES data for 454,787 participants [2], exemplify this trend. These sequencing studies offer unprecedented opportunities to identify ultra-rare genetic variants and structural variants, which were previously inaccessible through microarray-based genotyping. For example, GWAS based on high-depth WGS have associated rare coding variants with circulating lipid levels [37], sleep-disordered breathing [7], among other common traits and diseases. Additionally, WGS have facilitated investigation of effects of structural variants on hematologic traits [90], cardiometabolic traits [9] and cancers [17, 51].

1.2 Genotype Imputation and Imputation Server

While WGS has certainly become more affordable over time due to the development in sequencing techniques, they can still be prohibitive for large-scale studies involving thousands of samples. In this context, genotyping arrays, when combined with genotype imputation, provide a far more cost-effective solution.

Development of Imputation Tools

Genotype imputation is a statistical technique to infer unobserved genetic variants from a subset of genotyped markers, utilizing information from a reference set of densely sequenced genomes. The basic intuition behind this approach is that two unrelated individuals may share short stretches of haplotype inherited from distant common ancestors, enabling the reconstruction of a study haplotype as an imperfect mosaic copy of short segments from the reference haplotypes.

Over the years, various genotype imputation tools have been proposed, with the most successful ones adopting a hidden Markov model (HMM) framework based on the Li and Stephens model [50]. As illustrated in Figure 1.1, the hidden state in the HMM represents the template haplotype for each marker, while the emission state corresponds to the observed genotypes (including missing data) at each marker for the target sample being imputed. The Li and Stephens model directly characterizes the underlying coalescent process through the emission and transition probabilities. Emission probabilities denote the likelihood of observing genotypes given the underlying haplotype template, considering potential mutations or genotyping errors. Transition probabilities represent the likelihood of switching from one haplotype template to another, effectively capturing the correlation between markers by accounting for the recombination rates along the genome. This idea was first implemented in PHASE v2.1.1 [77], a tool specifically designed for haplotype estimation, and was then extended for simultaneous inference of haplotypic phase and missing genotypes by IMPUTE [59] and MaCH [52], utilizing a diploid version of the Li and Stephens model that incorporates with a pair of unobserved copying states. Consider a set of N reference haplotypes $H = \{h_1, h_2, \ldots, h_N\}$ which were sequenced at L markers. The hidden states in HMM is denoted as $Z_l = \{Z_l^{(1)}, Z_l^{(2)}\}, l = 1, 2, \ldots L$, representing the copying states for the two haplotypes for the target sample at the lth marker, where $Z_l^{(i)} \in \{1, 2, \ldots, N\}, i = 1, 2$. The transition probabilities switch from state Z_l to state Z_{l+1} is defined as a function of recombination parameter θ_l as displayed in Equation 1.1. MaCH updates the value of θ_l iteratively through the estimation process, whereas IMPUTE sets $\theta_l = 1 - e^{-\frac{4N_e r_l}{N}}$, where N_e stands for the effective population size and r_l stands for the genetic distance between the lth marker and the $(l+1)$th marker.

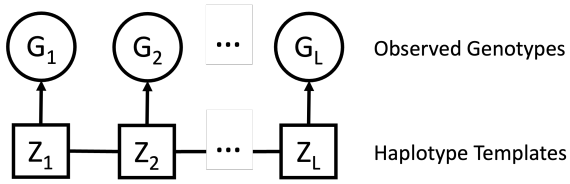


Figure 1.1: The Hidden Markov Model for Genotype Imputation

$$P(Z_{l+1}|Z_l) = \begin{cases} (\frac{\theta_l}{N})^2, & Z_{l+1}^{(1)} \neq Z_l^{(1)} \text{ and } Z_{l+1}^{(2)} \neq Z_l^{(2)} \\ \frac{\theta_l}{N}(1-\theta_l+\frac{\theta_l}{N}), & \text{either } Z_{l+1}^{(1)} \neq Z_l^{(1)} \text{ or } Z_{l+1}^{(2)} \neq Z_l^{(2)} \\ (1-\theta_l+\frac{\theta_l}{N})^2, & Z_{l+1}^{(1)} = Z_l^{(1)} \text{ and } Z_{l+1}^{(2)} = Z_l^{(2)} \end{cases} \tag{1.1}$$

Let $T(Z_l)$ denote the genotype copied exactly from the templates indicated in state Z_l, then the emission probability $P(G_l|Z_l)$ is listed in Table 1.1, where the error rate parameter ϵ_l reflects the combined effects of gene conversion, mutations, and genotyping error. MaCH updates the value of ϵ_l iteratively through the estimation process, whereas IMPUTE sets $\epsilon_l = \frac{1}{2(1+N\sum_{i=1}^{N-1}\frac{1}{i})}$.

		G_l		
		0	1	2
$T(Z_l)$	0	$(1-\epsilon_l)^2$	$2\epsilon_l(1-\epsilon_l)$	ϵ_l^2
	1	$\epsilon_l(1-\epsilon_l)$	$(1-\epsilon_l)^2+\epsilon_l^2$	$\epsilon_l(1-\epsilon_l)$
	2	ϵ_l^2	$2\epsilon_l(1-\epsilon_l)$	$(1-\epsilon_l)^2$

Table 1.1: Emission probabilities in diploid HMM. The probability of observing genotype G_l given hidden state Z_l and error parameter ϵ_l at the lth marker.

It is noteworthy that the forward-backward algorithm used for computing the diploid HMM model described above exhibits an $O(N^4L)$ time complexity, where N represents the number of reference samples and L denotes the number of markers covered by the reference panel. As WGS studies increase in size, encompassing larger sample sizes and an expanding number of rare variants, these methods become computationally challenging.

One of the milestone strategy to reduce the computational complexity of genotype imputation was introduced by Howie et al. in 2012 [39], which proposed separating the phasing and imputation processes. By pre-phasing the study samples, the authors transformed the problem into a

haploid HMM. Instead of searching for a pair of matching haplotypes, the new method imputes each haplotype independently, and thus the state space was reduced from $O(N^2)$ to $O(N)$, significantly decreasing the computational burden to $O(N^2L)$ compared to the previous diploid model. In the simplified model, the hidden state Z_l takes values from $\{1, 2, \ldots, N\}$ and the emission state G_l takes value from $\{0, 1\}$, with the transition probabilities presented in Equation 1.2 and the emission probabilities presented in Equation 1.3.

$$P(Z_{l+1}|Z_l) = \begin{cases} \frac{\theta_l}{N}, & Z_{l+1} \neq Z_l \\ 1 - \theta_l + \frac{\theta_l}{N}, & Z_{l+1} = Z_l \end{cases} \tag{1.2}$$

$$P(G_l|Z_l) = \begin{cases} \epsilon_l, & G_l \neq T(Z_l) \\ 1 - \epsilon_l, & G_l = T(Z_l) \end{cases} \tag{1.3}$$

The pre-phasing imputation was initially implemented in IMPUTE2 [39, 40] and minimac (a successor of MaCH) [31, 52] with their speed and memory usage being enhanced in their subsequent improved versions. Common strategies to improve computational efficiency include parallelization, linear interpolation (computing HMM on genotyped sites only and linearly interpolating untyped sites), using compact data structures for reference panels and adaptive precision for imputed probabilities.

The primary distinction between imputation tools lies in the strategies employed to reduce the state space in the HMM. The IMPUTE series conditions the imputation on a target-specific selection of reference haplotypes instead of including all the reference haplotypes in the state space. IMPUTE2 [40] selects reference haplotypes for each target sample based on Hamming distance; IMPUTE4 [6] utilizes an improved approximation algorithm that selects reference haplotypes according to local (rather than region-wide) sharing; IMPUTE5 [68] incorporates a positional Burrows-Wheeler transform (PBWT) data structure [23], enabling even faster searching of locally matching haplotypes. Beagle4 [4] combines genotyped markers within 0.005 cM windows into a single aggregate marker with more than two possible alleles; Beagle5 [5] introduces a novel

method to reduce the full reference panel into a small number of composite reference haplotypes (a mosaic of reference haplotypes) and adopts the IMPUTE2 idea of conditioning on a target-specific selection of composite reference haplotypes. Minimac3 [19] splits the chromosome into blocks and restricts the state space within each block to unique reference haplotypes, while minimac4 further reduces the number of unique templates by aggregating haplotypes with identical alleles on genotyped markers.

A comparison of computational time and memory usage among IMPUTE5 (v1.1.5), Beagle5 (22Jul22.46e), and minimac4 (v4.1.2) is presented in Table 1 are listed in Table 1.2. We imputed samples from UK Biobank array data [81] using the TOPMed r2 reference panel [78] on chromosome 20. The experiment was conducted on a single core of Intel® Xeon® Platinum 8268 CPU @ 2.90GHz.

Target Sample Size	Time ([h:]mm:ss)			Memory (Gb)		
	IMPUTE5	Beagle5	minimac4	IMPUTE5	Beagle5	minimac4
1	0:42	4:32	2:19	1.88	11.71	12.48
1,000	34:06	37:08	2:47:28	12.25	17.17	15.92
10,000	8:24:19	–	26:44:55	107.73	–	16.11

Table 1.2: A comparison of imputation tools based on time and memory usage for different target sample sizes using the TOPMed r2 reference panel on chromosome 20. Beagle5 failed for 10,000 samples due to excess of maximum Java heap size.

Development of Imputation Servers

Despite the continuous improvements in the computational efficiency of imputation tools, the increasing size of both reference panels and target samples imposes a substantial computational burden. Large-scale studies typically require access to high-performance computing clusters with ample memory and multi-core systems for processing. Additionally, researchers must possess fundamental knowledge of command-line tools and cluster job management, as well as familiarity with the pipeline or software required for quality control and phasing prior to the imputation step.

To make genotype imputation more accessible to the scientific community, cloud-based impu-

tation servers have been developed. Examples include Michigan Imputation Server (MIS) [20], Sanger Imputation Service (SIS) [61] and EagleImp-Web [91]. MIS uses minimac4 [20], whereas the other two rely on the PBWT algorithm [23]. These services provide user-friendly web interfaces, enabling users to securely upload their (pre-phased or unphased) genotype array data and specify imputation options, including the choice of reference panel. The imputation job, including quality control and phasing (if applicable), is performed remotely on the server, with imputed results returned to the user for downstream analyses.

Another challenge in genotype imputation is data sharing restrictions. Imputation requires individual-level genotype data from reference samples, but some reference panels like HRC [61] and TOPMed [78] cannot be made public due to data privacy and consent issues. Imputation servers facilitate the use of controlled-access reference panels. For instance, the TOPMed Imputation Server, powered by the same engine as MIS, was developed specifically to enable researchers to use the TOPMed reference panel for genotype imputation [78]. MIS, SIS, and EagleImp-Web all offer imputation with the 1000G and HRC reference panels. Additionally, MIS supports Genome Asia (GAsP) [13], the multi-ethnic HLA panel [57], CAAPA African American Panel [46], while SIS supports the UK10K reference panel [41] and African Genome Resources.

In summary, imputation servers streamline the process of imputing genotypes using high-quality reference panels by offering standardized workflow pipelines. This reduces the need for users to have extensive computational infrastructure and expertise in command-line tools. These servers play a crucial role in making genotype imputation more accessible to a wider range of researchers, facilitating powerful and high-resolution downstream analyses.

Benefits of Genotype Imputation

By predicting untyped genetic variants based on the reference panel, imputation increases the density of variants available for association tests. This process facilitates fine-mapping to more accurately localize association signals by considering all genetic variants in a given region, which in turn increases the chance of identifying a causal variant.

Genotype imputation also facilitates meta-analysis by enabling merging data or GWAS summary statistics from multiple studies. Different studies often use different genotyping arrays. By imputing array data from different genotyping platforms to the same reference panel, researchers can obtain a consensus set of variants, permitting meta-analysis across all available variants with a larger sample size than any individual study. This becomes particularly beneficial when the individual studies are small and may lack the statistical power needed to detect an effect. Therefore, genotype imputation not only provides a cost-effective alternative to WGS but also helps boost the power of GWAS.

1.3 Development of Statistical Methods in Association Tests

With the dramatic increase in genetic studies and the abundance of genome-wide genotyping and sequencing data, the development of methodologies to maximize the utility of data and facilitate interpretability of results from these studies has been an extremely productive research area. New methods have enabled researchers to gain insights into the biological mechanisms underlying complex traits and diseases. For instance, gene-based burden analyses have improved power by aggregating information from multiple genetic variants, allowing the identification of rare variants even with small sample sizes [18, 49]. Additionally, BOLT-LMM has introduced orders-of-magnitude improvements in the computational efficiency of mixed model methods, making it possible to analyze biobank-scale datasets while accounting for sample relatedness and population substructure [54, 55]. In this section, we will explore the primary challenges faced by genetic association tests and examine how statistical methods have evolved in response to the growing availability of data.

Population Stratification and Relatedness

Population stratification, characterized by allele frequency differences between cases and controls due to systematic ancestry differences, along with sample relatedness are major confounders in genetic association studies. These factors can result in the systematic inflation of test statistics

when methods assume independence between samples, leading to biased or spurious associations. Before the era of GWAS, researchers favored family-based studies for their robustness against population substructure. The transmission disequilibrium test [73], an application of McNemar's test [62], was used to detect genetic linkage by examining the difference in frequency between alleles transmitted from heterozygous parents to affected offspring and those not transmitted.

Devlin and Roeder proposed genomic control (GC) [22], a population-based method intended to be as robust yet less expensive than family-based design, making it more suitable for analyzing complex traits. GC estimates the inflation factor of test statistics by assessing null alleles, then adjusts the test statistics of the candidate gene accordingly. Since null alleles are assumed to be unassociated with the trait, the inflation should arise solely from population stratification. However, a global adjustment may be inappropriate as the extent of inflation can differ among genetic markers [67]. Consequently, the GC inflation factor λ_{GC} is now used as a quality measure to evaluate whether confounding persists after correction in association tests, rather than being directly used for correction. In practice, λ_{GC} is defined as the median of the observed chi-squared test statistics (with 1 degree of freedom) of all tested markers divided by the expected median value under the null hypothesis. $\lambda_{GC} > 1$ indicates stratification or the presence of other confounders.

As sample sizes and admixture in study samples have grown, estimating and matching individual ancestry has become increasingly sophisticated, with principal component analysis (PCA) emerging as a popular option for inferring population substructure [65, 66]. PCA offers advantages such as being assumption-free, parameter-free, and, more importantly, computationally tractable on a genome-wide scale. The top principal components, which may represent broad differences across individuals (although interpretations can be unclear), are widely used as covariates in GWAS. However, PCA does not explicitly account for family structure or cryptic relatedness. Consequently, mixed models have emerged as the preferred method due to their proven ability to account for relatedness among samples while controlling for population stratification and other confounding factors [94]. In practice, association studies often employ a combination of these strategies, correcting for broad sample structure using principal components and then modeling

the association using a mixed model.

Mixed-Model Association Tests in Large-Scale Studies

The standard linear mixed model for genetic association tests can be represented as

$$y = W\alpha + X\beta + b + e \tag{1.4}$$

where y is an n-vector of observed phenotypes, and X denotes the genotype of the variant of interest. W represents a matrix of covariates such as the mean (the intercept), gender and age, α represent the corresponding fixed effects. Note that we may project the covariates out from both the phenotype and genotype, which is equivalent to including them as fixed effects. b and e represent the random effects (the polygenic component) and residuals (the non-genetic component), respectively, which are both normally distributed with mean 0 and covariance $\sigma_g^2 K$ and σ_e^2. Here, K is an $n \times n$ matrix called the genomic relationship matrix (GRM) or empirical kinship matrix which models the genetic similarity between samples. β denotes the fixed effect of the variant of interest, and our goal is to test the null hypothesis $H_0 : \beta = 0$.

One of the challenges of using mixed models for GWAS is the substantial computational cost, as the optimization procedure for the likelihood function or the restricted maximum-likelihood (REML) function requires iterative updates of $\hat{\Sigma} = \hat{\sigma}_g^2 K + \hat{\sigma}_e^2 I$ and $\beta = (X^T \hat{\Sigma}^{-1} X)^{-1} X^T \hat{\Sigma}^{-1} y$. EMMA [45] employs singular value decomposition to avoid redundant matrix inverses and multiplications in the computation of likelihood, and thus reduces the time complexity from cubic to quadratic. EMMAX [44] further reduces the computational cost via a two-stage approach – first, the variance parameters are estimated under the null hypothesis (which avoids repetitive variance component estimation procedure for each variant); second, test the null hypothesis for each variant based on the variance estimates from the first step. It is worth noting that this approach relies on the assumption that each variant has a small effect and may lead to underestimation of p-value and decrease in test power when the assumption is invalid. Methods such as GEMMA [96] and

FaST-LMM [53] provide exact p-values, but here we focus the discussion on two-stage approaches.

BOLT-LMM [55] further improves the efficiency by using the conjugate gradient method to circumvent spectral decomposition, achieving a computational complexity of $O(mn)$, where m is the number of variants. It saves memory usages by operating directly on raw genotypes and calculating the elements GRM as needed instead of pre-computing and storing the entire GRM. The estimation of variance parameters is achieved through a stochastic approximation algorithm, and retrospective mixed-model association statistics are computed for hypothesis testing. In addition to the standard mixed model, BOLT-LMM models non-infinitesimal genetic architecture by placing a Gaussian mixture prior on the effect sizes and calibrating the test statistics using the LD Score regression technique. It also employs the leave-one-chromosome-out (LOCO) scheme to prevent proximal contamination.

Another challenge is the inflated type I error in case-control studies. Since the homoscedasticity assumption no longer holds for binary traits in the presence of covariates, linear mixed models may fail to control type I errors and yield incorrect p-value estimates. Chen et al. proposed GMMAT [8], which applies logistic mixed models and score tests for genome-wide analysis of binary traits. However, its implementation requires $O(mn^2)$ computation and $O(n^2)$ memory. SAIGE [95] has adapted BOLT-LMM's optimization strategies into the logistic mixed model framework, making it scalable for large sample sizes. It also incorporates the saddlepoint approximation to the score test statistics to accommodate unbalanced case-control ratios. REGENIE [60] enables parallel analysis of multiple quantitative or binary traits and further reduces the memory usage by loading only local segments of the genotype matrix. Enhanced computational efficiency in the first step is achieved by partitioning the variants into consecutive blocks and generating a small set of predictors using ridge regressions within each block. LOCO predictors are generated using a second round of ridge regressions with cross-validation and used for association tests in the second step. REGENIE has also proposed an approximate Firth regression approach for the analysis of binary traits.

Improved computational efficiency of GWAS methods offer substantial benefits to genetic research in terms of scalability and capacity. Efficient computational algorithms accommodate anal-

yses of large-scale studies which are growing both in sample size and the number of genetic variants (via WGS or genotype imputation), enhancing the power of GWAS and paving the way for more intricate studies such as multi-trait GWAS or meta-analyses, ultimately broadening our understanding of the complex traits and diseases.

The advancements in the computational efficiency of GWAS methods benefits genetic research in terms of scalability and capacity. These improved algorithms enabled large-scale studies that are growing both in terms of sample size, due to more individuals being genotyped, and in terms of the number of genetic variants (via WGS or genotype imputation), enhancing the power of GWAS and broadening our understanding of the complex genetic traits and diseases.

Statistical Fine-Mapping

GWAS have successfully identified thousands of genetic associations for diseases and complex traits [30, 71, 75, 82]. However, GWAS signals often point to broad regions of the genome which harbor hundreds of genetic variants, among which some are potentially causal while most are implicated due to LD with the true causal variant. With the presence of complex LD structure, it is often challenging to pinpoint the true causal variant, and therefore additional fine-mapping analyses are required to prioritize the candidate causal variants for follow-up functional studies.

One intuitive way to prioritize variants is based on p-values. While it is tempting to assume that the lead variant with the lowest p-value is most likely to be cause, it is not always true – a non-causal variant could have the lowest p-value due to LD with the actual causal variant or due to statistical fluctuations [10]. The limitations of using p-values in this context become apparent when we consider that p-values cannot quantify the uncertainty of a variant being causal [76]. Also, p-values are not comparable across variants or across different studies, given that they are influenced by minor allele frequency and sample size. Therefore, the Bayes factor has been increasingly recognized as a viable alternative to the p-value for summarizing the evidence of associations [84]. The earliest Bayesian fine-mapping approach ranks the associations by posterior probability which is proportional to the Bayes factor of each variant, while assuming that exact one of the variants is

causal in each region [58].

Several methods have been proposed to lift the restriction of one single causal variant in one region by jointly analyzing all the variants in the region. Conditional approach uses a stepwise selection procedure, sequentially selecting variants based on conditional p-values, which are re-calculated at each step [93]. However, this approach carries the drawbacks of using p-values: the necessity to set a significance threshold and the inability of quantifying uncertainty in the selection process, which is sub-optimal in terms of power and precision with the presence of complex LD structure. Consider two SNPs in perfect LD, only one being causal, the conditional approach will randomly choose one of them in the selection process and thus miss the true causal variant in half cases. One of the major methods that overcomes this limitation was presented in Hormozdiari et al. (CAVIAR) [38], in which they took the approach of jointly modeling multiple causal variants rather than sequentially. They framed the issue as a Bayesian Variable Selection (BVS) problem, taking into account all possible combinations of variants and calculating the posterior probability for each variant in a genomic region to be causal. Subsequently, more scalable BVS implementations were proposed, including FINEMAP [3], DAP-G [87, 48], and SuSiE [85], to avoid the need for exhaustive enumeration of all causal configurations.

1.4 Transcriptome-Wide Association Studies

Recent technological advancements have also permitted high-throughput measurement of other omics data. RNA sequencing has facilitated gene expression profiling [86] while mass spectrometry-based techniques have revolutionized proteomics [1] and metabolomics [21] by identifying and quantifying proteins and small molecules. These technologies, combined with advanced statistical methods, have enabled integrative analysis of multi-omics data, enhancing functional annotation of genetic variants, prioritizing candidate genes in association studies, and providing a comprehensive understanding of complex biological processes.

Transcriptome-Wide Association Study (TWAS) is a powerful research approach that integrates

gene expression measurements with GWAS to identify expression-trait associations. TWAS begins by leveraging a panel with both gene expression data and genotype data to build a prediction model for genetically regulated gene expression levels based on genotype [32]. The prediction model is then applied to GWAS data. Instead of testing individual genetic variants, the predicted expression levels of genes are evaluated for their association with the disease or trait in question. This process enables TWAS to identify genes where predicted expression correlates with disease risk, thereby offering a mechanistic hypothesis on how genetic variants may influence the disease.

1.5 Challenges and Purpose

This dissertation is dedicated to addressing urgent challenges and advancing the methods used in genetic association studies, specifically focusing on genotype imputation, fine-mapping, and gene expression imputation.

In Chapter 2, we confront the task of augmenting genotype imputation accuracy by leveraging multiple reference panels. Existing genotype imputation methods typically use one reference panel at a time. Nevertheless, with the increasing availability of large-scale sequencing projects, it is desirable to use multiple reference panels to boost imputation accuracy. But data-sharing restrictions and computational cost create obstacles in directly combining these reference panels. Thus, we introduce a meta-imputation framework, which circumvents the need for accessing individual-level genotype data by imputing target samples using each reference panel separately and then combining the imputed results.

In Chapter 3, we investigated the limitations of statistical fine-mapping methods that employ summary statistics and LD data compared with using individual-level data. We further examine the challenge of non-identifiability arising from complex LD structures within the BVS framework and its implications on fine-mapping outcomes. Additionally, we scrutinize the limitations of BVS implementations utilizing greedy algorithms, in contrast to exact calculation of posterior probabilities via enumeration of all possible model configurations, focusing particularly on their

handling of non-identifiable cases.

In Chapter 4, we pivot our focus to the fundamental part of TWAS – the prediction of genetically regulated expression levels. By employing TOPMed data, we undertook a comparative study of various gene expression imputation methodologies. Our objective is to spotlight the strengths and weaknesses of these methods, assess factors influencing the imputation accuracy of gene expression levels, and provide prediction models trained from the TOPMed data, which researchers can utilize in their TWAS analyses.

Through these projects, our method offers researchers valuable tools and insights for genetic research. We have integrated the meta-imputation feature in our imputation server so that researchers can conveniently improve imputation accuracy with multiple reference panels. Our investigation of fine-mapping methods provides a deeper understanding of non-identifiability issues, sheds light on the limitations of statistical approaches and the utilization of summary statistics. This knowledge empowers researchers to refine their fine-mapping analyses and improve their interpretation of results. Additionally, by providing prediction models trained from the TOPMed data, we offer researchers a practical resource to seamlessly integrate imputed gene expression levels into their analyses. These contributions collectively propel researchers forward in advancing their studies to unravel the mechanisms underlying complex traits and diseases.

1.6 References

[1] Ruedi Aebersold and Matthias Mann. Mass spectrometry-based proteomics. *Nature*, 422(6928):198–207, 2003.

[2] Joshua D Backman, Alexander H Li, Anthony Marcketta, Dylan Sun, Joelle Mbatchou, Michael D Kessler, Christian Benner, Daren Liu, Adam E Locke, Suganthi Balasubramanian, et al. Exome sequencing and analysis of 454,787 uk biobank participants. *Nature*, 599(7886):628–634, 2021.

[3] Christian Benner, Chris CA Spencer, Aki S Havulinna, Veikko Salomaa, Samuli Ripatti, and Matti Pirinen. Finemap: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics*, 32(10):1493–1501, 2016.

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

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

[6] Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O'Connell, et al. The uk biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203–209, 2018.

[7] Brian E Cade, Jiwon Lee, Tamar Sofer, Heming Wang, Man Zhang, Han Chen, Sina A Gharib, Daniel J Gottlieb, Xiuqing Guo, Jacqueline M Lane, et al. Whole-genome association analyses of sleep-disordered breathing phenotypes in the nhlbi topmed program. *Genome medicine*, 13:1–17, 2021.

[8] Han Chen, Chaolong Wang, Matthew P Conomos, Adrienne M Stilp, Zilin Li, Tamar Sofer, Adam A Szpiro, Wei Chen, John M Brehm, Juan C Celedón, et al. Control for population structure and relatedness for binary traits in genetic association studies via logistic mixed models. *The American Journal of Human Genetics*, 98(4):653–666, 2016.

[9] Lei Chen, Haley J Abel, Indraniel Das, David E Larson, Liron Ganel, Krishna L Kanchi, Allison A Regier, Erica P Young, Chul Joo Kang, Alexandra J Scott, et al. Association of structural variation with cardiometabolic traits in finns. *The American Journal of Human Genetics*, 108(4):583–596, 2021.

[10] Wenan Chen, Beth R Larrabee, Inna G Ovsyannikova, Richard B Kennedy, Iana H Haralambieva, Gregory A Poland, and Daniel J Schaid. Fine mapping causal variants with an approximate bayesian method using marginal test statistics. *Genetics*, 200(3):719–736, 2015.

[11] 1000 Genomes Project Consortium et al. A map of human genome variation from population scale sequencing. *Nature*, 467(7319):1061, 2010.

[12] 1000 Genomes Project Consortium et al. A global reference for human genetic variation. *Nature*, 526(7571):68, 2015.

[13] GenomeAsia100K Consortium. The genomeasia 100k project enables genetic discoveries across asia. *Nature*, 576(7785):106–111, 2019.

[14] The International HapMap Consortium. The international hapmap project. *Nature*, 426:789–796, 2003.

[15] The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–678, 2007.

[16] Adrian Cortes and Matthew A Brown. Promise and pitfalls of the immunochip. *Arthritis research & therapy*, 13:1–3, 2011.

[17] Isidro Cortés-Ciriano, Jake June-Koo Lee, Ruibin Xi, Dhawal Jain, Youngsook L Jung, Lixing Yang, Dmitry Gordenin, Leszek J Klimczak, Cheng-Zhong Zhang, David S Pellman, et al. Comprehensive analysis of chromothripsis in 2,658 human cancers using whole-genome sequencing. *Nature genetics*, 52(3):331–341, 2020.

[18] Carlos Cruchaga, Celeste M Karch, Sheng Chih Jin, Bruno A Benitez, Yefei Cai, Rita Guerreiro, Oscar Harari, Joanne Norton, John Budde, Sarah Bertelsen, et al. Rare coding variants in the phospholipase d3 gene confer risk for alzheimer's disease. *Nature*, 505(7484):550–554, 2014.

[19] Sayantan Das, Gonçalo R Abecasis, and Brian L Browning. Genotype imputation from large reference panels. *Annual review of genomics and human genetics*, 19:73–96, 2018.

[20] Sayantan Das, Lukas Forer, Sebastian Schönherr, Carlo Sidore, Adam E Locke, Alan Kwong, Scott I Vrieze, Emily Y Chew, Shawn Levy, Matt McGue, et al. Next-generation genotype imputation service and methods. *Nature genetics*, 48(10):1284–1287, 2016.

[21] Katja Dettmer, Pavel A Aronov, and Bruce D Hammock. Mass spectrometry-based metabolomics. *Mass spectrometry reviews*, 26(1):51–78, 2007.

[22] Bernie Devlin and Kathryn Roeder. Genomic control for association studies. *Biometrics*, 55(4):997–1004, 1999.

[23] Richard Durbin. Efficient haplotype matching and storage using the positional burrows–wheeler transform (pbwt). *Bioinformatics*, 30(9):1266–1272, 2014.

[24] Tal Engel and Uri Kopylov. Ustekinumab in crohn's disease: evidence to date and place in therapy. *Therapeutic advances in chronic disease*, 7(4):208–214, 2016.

[25] Daniel Eriksson, Ellen Christine Røyrvik, Maribel Aranda-Guillén, Amund Holte Berger, Nils Landegren, Haydee Artaza, Åsa Hallgren, Marianne Aardal Grytaas, Sara Ström, Eirik Bratland, et al. Gwas for autoimmune addison's disease identifies multiple risk loci and highlights aire in disease susceptibility. *Nature communications*, 12(1):959, 2021.

[26] Erica B Esrick, Leslie E Lehmann, Alessandra Biffi, Maureen Achebe, Christian Brendel, Marioara F Ciuculescu, Heather Daley, Brenda MacKinnon, Emily Morris, Amy Federico, et al. Post-transcriptional genetic silencing of bcl11a to treat sickle cell disease. *New England Journal of Medicine*, 384(3):205–215, 2021.

[27] Brian G Feagan, William J Sandborn, Geert D'Haens, Julián Panés, Arthur Kaser, Marc Ferrante, Edouard Louis, Denis Franchimont, Olivier Dewit, Ursula Seidler, et al. Induction therapy with the selective interleukin-23 inhibitor risankizumab in patients with moderate-to-severe crohn's disease: a randomised, double-blind, placebo-controlled phase 2 study. *The Lancet*, 389(10080):1699–1709, 2017.

[28] Haydar Frangoul, David Altshuler, M Domenica Cappellini, Yi-Shan Chen, Jennifer Domm, Brenda K Eustace, Juergen Foell, Josu de la Fuente, Stephan Grupp, Rupert Handgretinger, et al. Crispr-cas9 gene editing for sickle cell disease and β-thalassemia. *New England Journal of Medicine*, 384(3):252–260, 2021.

[29] Timothy M Frayling, Nicholas J Timpson, Michael N Weedon, Eleftheria Zeggini, Rachel M Freathy, Cecilia M Lindgren, John RB Perry, Katherine S Elliott, Hana Lango, Nigel W Rayner, et al. A common variant in the fto gene is associated with body mass index and predisposes to childhood and adult obesity. *Science*, 316(5826):889–894, 2007.

[30] Lars G Fritsche, Wilmar Igl, Jessica N Cooke Bailey, Felix Grassmann, Sebanti Sengupta, Jennifer L Bragg-Gresham, Kathryn P Burdon, Scott J Hebbring, Cindy Wen, Mathias Gorski, et al. A large genome-wide association study of age-related macular degeneration highlights contributions of rare and common variants. *Nature genetics*, 48(2):134–143, 2016.

[31] Christian Fuchsberger, Gonçalo R Abecasis, and David A Hinds. minimac2: faster genotype imputation. *Bioinformatics*, 31(5):782–784, 2014.

[32] Eric R Gamazon, Heather E Wheeler, Kaanan P Shah, Sahar V Mozaffari, Keston Aquino-Michaels, Robert J Carroll, Anne E Eyler, Joshua C Denny, GTEx Consortium, Dan L Nicolae, et al. A gene-based association method for mapping traits using reference transcriptome data. *Nature genetics*, 47(9):1091–1098, 2015.

[33] Sara Goodwin, John D McPherson, and W Richard McCombie. Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, 17(6):333–351, 2016.

[34] Bjarni V Halldorsson, Hannes P Eggertsson, Kristjan HS Moore, Hannes Hauswedell, Ogmundur Eiriksson, Magnus O Ulfarsson, Gunnar Palsson, Marteinn T Hardarson, Asmundur Oddsson, Brynjar O Jensson, et al. The sequences of 150,119 genomes in the uk biobank. *Nature*, 607(7920):732–740, 2022.

[35] Isabel M Haugh, Allie K Preston, Dario N Kivelevitch, and Alan M Menter. Risankizumab: an anti-il-23 antibody for the treatment of psoriasis. *Drug Design, Development and Therapy*, pages 3879–3883, 2018.

[36] Martin E Hess, Simon Hess, Kate D Meyer, Linda AW Verhagen, Linda Koch, Hella S Brönneke, Marcelo O Dietrich, Sabine D Jordan, Yogesh Saletore, Olivier Elemento, et al.

The fat mass and obesity associated gene (fto) regulates activity of the dopaminergic midbrain circuitry. *Nature neuroscience*, 16(8):1042–1048, 2013.

[37] George Hindy, Peter Dornbos, Mark D Chaffin, Dajiang J Liu, Minxian Wang, Margaret Sunitha Selvaraj, David Zhang, Joseph Park, Carlos A Aguilar-Salinas, Lucinda Antonacci-Fulton, et al. Rare coding variants in 35 genes associate with circulating lipid levels—a multi-ancestry analysis of 170,000 exomes. *The American Journal of Human Genetics*, 109(1):81–96, 2022.

[38] Farhad Hormozdiari, Emrah Kostem, Eun Yong Kang, Bogdan Pasaniuc, and Eleazar Eskin. Identifying causal variants at loci with multiple signals of association. In *Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, pages 610–611, 2014.

[39] Bryan Howie, Christian Fuchsberger, Matthew Stephens, Jonathan Marchini, and Gonçalo R Abecasis. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nature genetics*, 44(8):955–959, 2012.

[40] Bryan N Howie, Peter Donnelly, and Jonathan Marchini. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS genetics*, 5(6):e1000529, 2009.

[41] Jie Huang, Bryan Howie, Shane McCarthy, Yasin Memari, Klaudia Walter, Josine L Min, Petr Danecek, Giovanni Malerba, Elisabetta Trabetti, Hou-Feng Zheng, et al. Improved imputation of low-frequency and rare variants using the uk10k haplotype reference panel. *Nature communications*, 6(1):1–9, 2015.

[42] Jeroen R Huyghe, Stephanie A Bien, Tabitha A Harrison, Hyun Min Kang, Sai Chen, Stephanie L Schmit, David V Conti, Conghui Qu, Jihyoun Jeon, Christopher K Edlund, et al. Discovery of common and rare genetic risk variants for colorectal cancer. *Nature genetics*, 51(1):76–87, 2019.

[43] International IBD Genetics Consortium (IIBDGC), Cristina Agliardi, Lars Alfredsson, Mehdi Alizadeh, Carl Anderson, Robert Andrews, Helle Bach Søndergaard, Amie Baker, Gavin Band, et al. Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis. *Nature genetics*, 45(11):1353–1360, 2013.

[44] Hyun Min Kang, Jae Hoon Sul, Susan K Service, Noah A Zaitlen, Sit-yee Kong, Nelson B Freimer, Chiara Sabatti, and Eleazar Eskin. Variance component model to account for sample structure in genome-wide association studies. *Nature genetics*, 42(4):348–354, 2010.

[45] Hyun Min Kang, Noah A Zaitlen, Claire M Wade, Andrew Kirby, David Heckerman, Mark J Daly, and Eleazar Eskin. Efficient control of population structure in model organism association mapping. *Genetics*, 178(3):1709–1723, 2008.

[46] Michael D Kessler, Laura Yerges-Armstrong, Margaret A Taub, Amol C Shetty, Kristin Maloney, Linda Jo Bone Jeng, Ingo Ruczinski, Albert M Levin, L Keoki Williams, Terri H Beaty, et al. Challenges and disparities in the application of personalized genomic medicine to populations with african ancestry. *Nature communications*, 7(1):12521, 2016.

[47] Robert J Klein, Caroline Zeiss, Emily Y Chew, Jen-Yue Tsai, Richard S Sackler, Chad Haynes, Alice K Henning, John Paul SanGiovanni, Shrikant M Mane, Susan T Mayne, et al. Complement factor h polymorphism in age-related macular degeneration. *Science*, 308(5720):385–389, 2005.

[48] Yeji Lee, Francesca Luca, Roger Pique-Regi, and Xiaoquan Wen. Bayesian multi-snp genetic association analysis: control of fdr and use of summary statistics. *BioRxiv*, page 316471, 2018.

[49] Bingshan Li and Suzanne M Leal. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *The American Journal of Human Genetics*, 83(3):311–321, 2008.

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

[51] Yilong Li, Nicola D Roberts, Jeremiah A Wala, Ofer Shapira, Steven E Schumacher, Kiran Kumar, Ekta Khurana, Sebastian Waszak, Jan O Korbel, James E Haber, et al. Patterns of somatic structural variation in human cancer genomes. *Nature*, 578(7793):112–121, 2020.

[52] Yun Li, Cristen J Willer, Jun Ding, Paul Scheet, and Gonçalo R Abecasis. Mach: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genetic epidemiology*, 34(8):816–834, 2010.

[53] Christoph Lippert, Jennifer Listgarten, Ying Liu, Carl M Kadie, Robert I Davidson, and David Heckerman. Fast linear mixed models for genome-wide association studies. *Nature methods*, 8(10):833–835, 2011.

[54] Po-Ru Loh, Gleb Kichaev, Steven Gazal, Armin P Schoech, and Alkes L Price. Mixed-model association for biobank-scale datasets. *Nature genetics*, 50(7):906–908, 2018.

[55] Po-Ru Loh, George Tucker, Brendan K Bulik-Sullivan, Bjarni J Vilhjalmsson, Hilary K Finucane, Rany M Salem, Daniel I Chasman, Paul M Ridker, Benjamin M Neale, Bonnie Berger, et al. Efficient bayesian mixed-model analysis increases association power in large cohorts. *Nature genetics*, 47(3):284–290, 2015.

[56] Xiangfeng Lu, Gina M Peloso, Dajiang J Liu, Ying Wu, He Zhang, Wei Zhou, Jun Li, Clara Sze-man Tang, Rajkumar Dorajoo, Huaixing Li, et al. Exome chip meta-analysis identifies novel loci and east asian–specific coding variants that contribute to lipid levels and coronary artery disease. *Nature genetics*, 49(12):1722–1730, 2017.

[57] Yang Luo, Masahiro Kanai, Wanson Choi, Xinyi Li, Saori Sakaue, Kenichi Yamamoto, Kotaro Ogawa, Maria Gutierrez-Arcelus, Peter K Gregersen, Philip E Stuart, et al. A high-resolution hla reference panel capturing global population diversity enables multi-ancestry fine-mapping in hiv host response. *Nature genetics*, 53(10):1504–1516, 2021.

[58] Julian B Maller, Gilean McVean, Jake Byrnes, Damjan Vukcevic, Kimmo Palin, Zhan Su, Joanna MM Howson, Adam Auton, Simon Myers, Andrew Morris, et al. Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nature genetics*, 44(12):1294–1301, 2012.

[59] Jonathan Marchini, Bryan Howie, Simon Myers, Gil McVean, and Peter Donnelly. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature genetics*, 39(7):906–913, 2007.

[60] Joelle Mbatchou, Leland Barnard, Joshua Backman, Anthony Marcketta, Jack A Kosmicki, Andrey Ziyatdinov, Christian Benner, Colm O'Dushlaine, Mathew Barber, Boris Boutkov, et al. Computationally efficient whole-genome regression for quantitative and binary traits. *Nature genetics*, 53(7):1097–1103, 2021.

[61] Shane McCarthy, Sayantan Das, Warren Kretzschmar, Olivier Delaneau, Andrew R Wood, Alexander Teumer, Hyun Min Kang, Christian Fuchsberger, Petr Danecek, Kevin Sharp, et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nature genetics*, 48(10):1279, 2016.

[62] Quinn McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947.

[63] Kyriaki Michailidou, Sara Lindström, Joe Dennis, Jonathan Beesley, Shirley Hui, Siddhartha Kar, Audrey Lemaçon, Penny Soucy, Dylan Glubb, Asha Rostamianfar, et al. Association analysis identifies 65 new breast cancer risk loci. *Nature*, 551(7678):92–94, 2017.

[64] Christos Pantelis, George N Papadimitriou, Sergi Papiol, Elena Parkhomenko, Michele T Pato, Tiina Paunio, Milica Pejovic-Milovancevic, Diana O Perkins, Olli Pietiläinen, et al. Biological insights from 108 schizophrenia-associated genetic loci. *Nature*, 511(7510):421–427, 2014.

[65] Nick Patterson, Alkes L Price, and David Reich. Population structure and eigenanalysis. *PLoS genetics*, 2(12):e190, 2006.

[66] Alkes L Price, Nick J Patterson, Robert M Plenge, Michael E Weinblatt, Nancy A Shadick, and David Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, 38(8):904–909, 2006.

[67] Alkes L Price, Noah A Zaitlen, David Reich, and Nick Patterson. New approaches to population stratification in genome-wide association studies. *Nature reviews genetics*, 11(7):459–463, 2010.

[68] Simone Rubinacci, Olivier Delaneau, and Jonathan Marchini. Genotype imputation using the positional burrows wheeler transform. *PLoS genetics*, 16(11):e1009049, 2020.

[69] Fredrick R Schumacher, Ali Amin Al Olama, Sonja I Berndt, Sara Benlloch, Mahbubl Ahmed, Edward J Saunders, Tokhir Dadaev, Daniel Leongamornlert, Ezequiel Anokian, Clara Cieza-Borrella, et al. Association analyses of more than 140,000 men identify 63 new prostate cancer susceptibility loci. *Nature genetics*, 50(7):928–936, 2018.

[70] Robert A Scott, Laura J Scott, Reedik Mägi, Letizia Marullo, Kyle J Gaulton, Marika Kaakinen, Natalia Pervjakova, Tune H Pers, Andrew D Johnson, John D Eicher, et al. An expanded genome-wide association study of type 2 diabetes in europeans. *Diabetes*, 66(11):2888–2902, 2017.

[71] Angelo Scuteri, Serena Sanna, Wei-Min Chen, Manuela Uda, Giuseppe Albai, James Strait, Samer Najjar, Ramaiah Nagaraja, Marco Orrú, Gianluca Usala, et al. Genome-wide association scan shows genetic variants in the fto gene are associated with obesity-related traits. *PLoS genetics*, 3(7):e115, 2007.

[72] Elliot Sollis, Abayomi Mosaku, Ala Abid, Annalisa Buniello, Maria Cerezo, Laurent Gil, Tudor Groza, Osman Güneş, Peggy Hall, James Hayhurst, et al. The nhgri-ebi gwas catalog: knowledgebase and deposition resource. *Nucleic Acids Research*, 51(D1):D977–D985, 2023.

[73] Richard S Spielman, Ralph E McGinnis, and Warren J Ewens. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (iddm). *American journal of human genetics*, 52(3):506, 1993.

[74] Cassandra N Spracklen, Momoko Horikoshi, Young Jin Kim, Kuang Lin, Fiona Bragg, Sanghoon Moon, Ken Suzuki, Claudia HT Tam, Yasuharu Tabara, Soo-Heon Kwak, et al. Identification of type 2 diabetes loci in 433,540 east asian individuals. *Nature*, 582(7811):240–245, 2020.

[75] Eli A Stahl, Gerome Breen, Andreas J Forstner, Andrew McQuillin, Stephan Ripke, Vassily Trubetskoy, Manuel Mattheisen, Yunpeng Wang, Jonathan RI Coleman, Héléna A Gaspar, et al. Genome-wide association study identifies 30 loci associated with bipolar disorder. *Nature genetics*, 51(5):793–803, 2019.

[76] Matthew Stephens and David J Balding. Bayesian statistical methods for genetic association studies. *Nature Reviews Genetics*, 10(10):681–690, 2009.

[77] Matthew Stephens and Paul Scheet. Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *The American Journal of Human Genetics*, 76(3):449–462, 2005.

[78] Daniel Taliun, Daniel N Harris, Michael D Kessler, Jedidiah Carlson, Zachary A Szpiech, Raul Torres, Sarah A Gagliano Taliun, André Corvelo, Stephanie M Gogarten, Hyun Min Kang, et al. Sequencing of 53,831 diverse genomes from the nhlbi topmed program. *Nature*, 590(7845):290–299, 2021.

[79] Lam C Tsoi, Sarah L Spain, Jo Knight, Eva Ellinghaus, Philip E Stuart, Francesca Capon, Jun Ding, Yanming Li, Trilokraj Tejasvi, Johann E Gudjonsson, et al. Identification of 15 new psoriasis susceptibility loci highlights the role of innate immunity. *Nature genetics*, 44(12):1341–1348, 2012.

[80] Pim Van Der Harst and Niek Verweij. Identification of 64 novel genetic loci provides an expanded view on the genetic architecture of coronary artery disease. *Circulation research*, 122(3):433–443, 2018.

[81] Cristopher V Van Hout, Ioanna Tachmazidou, Joshua D Backman, Joshua D Hoffman, Daren Liu, Ashutosh K Pandey, Claudia Gonzaga-Jauregui, Shareef Khalid, Bin Ye, Nilanjana Banerjee, et al. Exome sequencing and characterization of 49,960 individuals in the uk biobank. *Nature*, 586(7831):749–756, 2020.

[82] Peter M Visscher, Naomi R Wray, Qian Zhang, Pamela Sklar, Mark I McCarthy, Matthew A Brown, and Jian Yang. 10 years of gwas discovery: biology, function, and translation. *The American Journal of Human Genetics*, 101(1):5–22, 2017.

[83] Benjamin F. Voight, Hyun Min Kang, Jun Ding, Cameron D. Palmer, Carlo Sidore, Peter S. Chines, Noël P. Burtt, Christian Fuchsberger, Yanming Li, Jeanette Erdmann, et al. The metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. *PLOS Genetics*, 8(8):1–12, 08 2012.

[84] Jon Wakefield. Bayes factors for genome-wide association studies: comparison with p-values. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society*, 33(1):79–86, 2009.

[85] Gao Wang, Abhishek Sarkar, Peter Carbonetto, and Matthew Stephens. A simple new approach to variable selection in regression, with application to genetic fine mapping. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(5):1273–1300, 2020.

[86] Zhong Wang, Mark Gerstein, and Michael Snyder. Rna-seq: a revolutionary tool for transcriptomics. *Nature reviews genetics*, 10(1):57–63, 2009.

[87] Xiaoquan Wen, Yeji Lee, Francesca Luca, and Roger Pique-Regi. Efficient integrative multi-snp association analysis via deterministic approximation of posteriors. *The American Journal of Human Genetics*, 98(6):1114–1129, 2016.

[88] Jennifer Wessel, Audrey Y Chu, Sara M Willems, Shuai Wang, Hanieh Yaghootkar, Jennifer A Brody, Marco Dauriz, Marie-France Hivert, Sridharan Raghavan, Leonard Lipovich, et al. Low-frequency and rare exome chip variants associate with fasting glucose and type 2 diabetes susceptibility. *Nature communications*, 6(1):5897, 2015.

[89] Kris A. Wetterstrand. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP). Available at: www.genome.gov/sequencingcostsdata. Accessed: 2023-03-01.

[90] Marsha M Wheeler, Adrienne M Stilp, Shuquan Rao, Bjarni V Halldórsson, Doruk Beyter, Jia Wen, Anna V Mihkaylova, Caitlin P McHugh, John Lane, Min-Zhi Jiang, et al. Whole genome sequencing identifies structural variants contributing to hematologic traits in the nhlbi topmed program. *Nature communications*, 13(1):7592, 2022.

[91] Lars Wienbrandt, Christoph Prieß, Jan Christian Kässens, Andre Franke, Franziska Uhing, and David Ellinghaus. Eagleimp-web: A fast and secure genotype phasing and imputation web service using field-programmable gate arrays. *bioRxiv*, pages 2022–02, 2022.

[92] Naomi R Wray, Stephan Ripke, Manuel Mattheisen, Maciej Trzaskowski, Enda M Byrne, Abdel Abdellaoui, Mark J Adams, Esben Agerbo, Tracy M Air, Till MF Andlauer, et al. Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nature genetics*, 50(5):668–681, 2018.

[93] Jian Yang, Teresa Ferreira, Andrew P Morris, Sarah E Medland, Genetic Investigation of ANthropometric Traits (GIANT) Consortium, DIAbetes Genetics Replication, Meta analysis (DIAGRAM) Consortium, Pamela AF Madden, Andrew C Heath, Nicholas G Martin, Grant W Montgomery, et al. Conditional and joint multiple-snp analysis of gwas summary statistics identifies additional variants influencing complex traits. *Nature genetics*, 44(4):369–375, 2012.

[94] Jianming Yu, Gael Pressoir, William H Briggs, Irie Vroh Bi, Masanori Yamasaki, John F Doebley, Michael D McMullen, Brandon S Gaut, Dahlia M Nielsen, James B Holland, et al. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature genetics*, 38(2):203–208, 2006.

[95] Wei Zhou, Jonas B Nielsen, Lars G Fritsche, Rounak Dey, Maiken E Gabrielsen, Brooke N Wolford, Jonathon LeFaive, Peter VandeHaar, Sarah A Gagliano, Aliya Gifford, et al. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nature genetics*, 50(9):1335–1341, 2018.

[96] Xiang Zhou and Matthew Stephens. Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nature methods*, 11(4):407–409, 2014.

CHAPTER 2

Meta-Imputation: An Efficient Method to Combine Genotype Data after Imputation with Multiple Reference Panels

2.1 Introduction

Genotype imputation, which uses a reference panel of sequenced genomes to estimate unobserved genotypes for samples with sparse microarray data, has been widely used to infer genotypes in genome-wide association studies (GWAS) [11, 17, 26].Genotype imputation helps improve power for detecting association signals, facilitates meta-analyses and enables fine-mapping [20, 7].

Over the last decade, large-scale whole-genome sequencing projects such as 1000 Genomes (1000G) [6], Haplotype Reference Consortium (HRC) [21] and Trans-Omics for Precision Medicine (TOPMed) Program [27] have produced reference panels that include progressively larger numbers of samples. The successive increase in reference sample size captures more rare variants and provides higher resolution mapping in association studies. While these widely used panels have been steadily increasing in resolution and accuracy, particularly in European ancestry samples, the optimal choice of panel is often challenging for other ancestries (for example, the smaller 1000G reference panel sometimes outperforms the larger HRC panel in samples of South Asian ancestry[7, 27]). Furthermore, when imputing samples within a specific study population, smaller customized reference panels exist as alternatives to these widely used public pan-

els and might yield even better imputation quality [9, 22]. Examples where using these customized reference panels can often provide higher accuracy include ongoing studies in Sardinia [25], Finland[16], Norway [15] and Iceland [14], among many others. Unfortunately, these customized reference panels may miss rare variants and haplotypes that could be covered by larger panels and may perform poorly for individuals with unique ancestry. Therefore, it is desirable to utilize genetic information from both customized panels and large-scale panels [32].

An ideal solution is to construct a combined reference panel. However, different studies tend to use different variant calling and filtering strategies, which can make it challenging to merge sequencing data [27, 24]. It is desirable to consider the union set of variants across studies to use as much of the available information as possible. The gold standard method to address such discrepancies between multiple data sets is to jointly call variants from all samples using their original sequence alignment files, which is a highly computationally intensive task. A relatively simple substitute for joint variant calling is cross-imputation, where data sets are used as reference panels for each other and reciprocally imputed up to the union set of variants [13]. Furthermore, another important concern is data-sharing restrictions. For example, individual-level genotype data in many reference panels are not publicly available, it may thus be impossible to directly merge them with other sequencing data sets.

In this paper, we introduce the idea of meta-imputation. Instead of combining the reference panels before imputation, we first impute using different reference panels separately and then combine the imputed results into a consensus data set. By doing so, we can avoid accessing individual-level genotype data of the reference panel samples and achieve the goal of improving imputation accuracy by incorporating genetic information from multiple sources.

2.2 Materials and Methods

Meta-imputation consists of two separate steps (Figure B.1). First, we impute our target samples against two or more different reference panels. Then, we combine the imputation results using

weights that are guided by the empirical performance of each of the panels in stretches of each individual genome. The meta-imputed result at each marker is then a weighted average of the estimated allele counts from imputation against each panel. The weights are individual and region specific and reflect that the optimal choice of reference panel varies along the genome. Weights for each region and individual are estimated through a hidden Markov Model (HMM).

2.2.1 Leave-one-out Imputation

In order to determine the optimal weights for each reference panel along the genome, we need to evaluate the performance of each panel along each imputed haplotype. Theoretically, if we knew the true genotype of the target haplotype at a marker, we could quantify the imputation accuracy at that marker by comparing the true genotype with the imputed haploid dosage. In practice, we mimic this approach by leave-one-out (LOO) imputation, in which each genotyped marker is masked and imputed in turn. Our innovation is to use the genotyped markers in each genome to estimate these local weights for each individual. We do this by masking each observed genotype in turn and then trying to impute it based on information at flanking markers. We call the imputed results from this procedure LOO dosages. We evaluate local imputation performance for each reference panel by comparing the LOO dosages and the original genotypes at the masked sites, and assign local weights accordingly.

Figure 2.1A and Figure 2.1B illustrate a simplified version of the LOO imputation algorithm using two reference panels. For easier understanding, we simplified HMM to estimation based on exact matching haplotypes. The target haplotype is genotyped at three markers (marker 1,3,6). First, we masked the observed allele at marker 1, and searched for matching haplotypes based on marker 3 and marker 6. According to the matching reference haplotypes (shaded in blue), the probability of observing "A" was 0.8 from Panel #1 and 0.3 from Panel #2. Similarly, we could obtain the LOO results at other genotyped markers. Figure 2.1C compares the LOO results from the two reference panels along the genome, Panel #1 was more accurate at the beginning and Panel #2 was more accurate at the end of the chunk, so in the weight estimation process, we would assign

Panel #1 high weights at the beginning and low weights at the end of the chunk. Our expectation is such weights will improve imputation at ungenotyped markers.

In practice, the LOO imputation utilizes the same Markov chain as the regular imputation, and the only difference in the model lies in the genotype emission probability at the masked marker. Let A_m denote the observed allele at marker m in the target haplotype, H_m denote the reference haplotype template at marker m, and M denote the number of markers. In the HMM for the regular imputation, the probability of the underlying template at marker m is given in equation (2.1).

$$P(H_m|A_1,\ldots,A_M) \propto P(A_m|H_m)L_m(H_m)R_m(H_m) \tag{2.1}$$

where $L_m(\cdot)$ and $R_m(\cdot)$ denote the left probability and right probability for the haplotype template at marker m, as defined in equation (2.2) and (2.3) respectively.

$$L_m(H_m) = \begin{cases} 1 & m = 1 \\ \sum_{H_{m-1}} L_{m-1}(H_{m-1})P(A_{m-1}|H_{m-1})P(H_{m-1}|H_m) & 1 < m \leq M \end{cases} \tag{2.2}$$

$$R_m(H_m) = \begin{cases} \sum_{H_{m+1}} R_{m+1}(H_{m+1})P(A_{m+1}|H_{m+1})P(H_{m+1}|H_m) & 1 \leq m < M \\ 1 & m = M \end{cases} \tag{2.3}$$

Assume that the genotype at marker m is observed. When calculating the LOO dosage for marker m, the observed genotype is masked and handled as if it were unknown. Hence, the corresponding genotype emission probability $P(A_m|H_m)$ is set to 1, while other components in equation (2.1) remain the same as in the regular imputation, which yields the LOO posterior probability $\tilde{P}(H_m|A_1,\ldots,A_M) \propto L_m(H_m)R_m(H_m)$. Let $Y_1,\ldots,Y_N$ denote all the haplotypes in the reference panel, and $Y_{n,m}$ denote the alternative allele count at marker m of reference haplotype Y_n, then the LOO dosage at marker m is represented as:

$$d_m = \sum_{n=1}^{N} Y_{n,m} \times \tilde{P}(H_m|A_1,\ldots,A_M) \tag{2.4}$$

The LOO imputation is a built-in feature in the latest version of our minimac4 imputation software [8], which runs at the same time of regular imputation with minimal additional computational cost. The time costs of imputation using minimac4 with and without the LOO imputation feature are displayed in Table B.2. The LOO imputation is computationally inexpensive because it does not require rerunning the forward and backward chains of the HMM that underlie genotype imputation and because it requires limited extra calculations at genotyped markers only.

2.2.2 Model Description

We assume that the target genotypes are pre-phased prior to imputation, so that imputation is conducted on the same set of haplotypes using each reference panel in turn. The key meta-imputation problem is thus combining haploid allele dosages estimated using each of the available panels. Assume that we have K reference panels, containing a union set of M markers, labeled in a chromosome order with indices $1, 2, \ldots, M$. For a target haplotype, we denote the imputed haploid dosages at marker m from panel k as $X_{k,m}$, and the meta-imputed haploid dosage at that marker is represented as their weighted average:

$$X_m = \sum_{k=1}^{K} w_{k,m} X_{k,m} \; m = 1, 2, \ldots, M \tag{2.5}$$

where $w_{k,m}$ represents the weight on panel k at marker m, satisfying $0 \leq w_{k,m} \leq 1$ and $\sum_{k=1}^{K} w_{k,m} = 1$. For each target haplotype, weights are estimated through an HMM that we will describe next. The weights are tailored to each haplotype and vary along the genome. This integration step is implemented in the C++ package MetaMinimac2.

2.2.3 Weight Estimation

As inspired by Li and Stephens model [18], we use an HMM to estimate reference panel weights using the LOO dosages and the observed alleles to guide our decisions about which panel is preferred along the genome. In this HMM the hidden state S_m represents the underlying choice of

reference panel at marker m and the emission state A_m represents the observed allele (0 – reference allele, 1 – alternate allele).

The emission probability $P(A_m|S_m)$ is defined in Equation (2.6) where $d_{k,m}$ denote the LOO dosage from panel k at marker m. An ideal choice of reference panel will maximize the probability of the genotypes that are actually observed.

$$\begin{aligned} P(A_m = 1|S_m = k) &= d_{k,m} \\ P(A_m = 0|S_m = k) &= 1 - d_{k,m} \end{aligned} \tag{2.6}$$

The transition probability $P(S_m|S_{m-1})$ is defined in Equation (2.7) where λ_m represents the probability of a change in optimal reference panel between markers $m-1$ and m. We have found that our model is not very sensitive to reasonable choices of λ_m, and we typically set $\lambda_m = 1 - e^{-c \cdot \text{dist}_m}$, where dist_m is the base pair distance between the two markers and $c = 2 \times 10^{-7}$.

$$P(S_m|S_{m-1}) = \begin{cases} \frac{\lambda_m}{K} & S_m \neq S_{m-1} \\ 1 - \lambda_m + \frac{\lambda_m}{K} & S_m = S_{m-1} \end{cases} \tag{2.7}$$

Finally, these quantities allow us to define the weight for panel k at marker m as the posterior probability $w_{k,m} = P(S_m = k|A_1, \ldots, A_M)$ using the forward and backward algorithm [2].

After obtaining the weights at genotyped markers, weights at intervening markers are interpolated from flanking genotyped markers. When calculating the meta-imputed dosage at a specific marker (equation 2.5), only reference panels including that marker are considered and their weights are scaled so they sum to 1.0. An alternative strategy would be to assume a dosage of 0.0 where the marker is absent, avoiding rescaling. The optimal choice of strategy depends on whether markers are generally absent from a panel due to differences in allele frequency between populations and samples or, instead, due to differences in variant calling and filtering protocols.

2.2.4 Empirical Assessment #1: African American Samples from 1000 Genomes

To evaluate the ability of our method to accurately impute the genomes of admixed individuals, we selected a set of 1000G samples with admixed ancestry and created two panels for imputation – one with individuals with mostly European ancestry, and the other with individuals with mostly African ancestry. This setting is challenging because the optimal choice of panel will vary between individuals (depending on their degree of admixture) and also along the genome of each individual (depending on the ancestral origin of each chromosome segment). Then, we focused on 61 individuals of African Ancestry in Southwest US (ASW) and extracted their genotypes for the Illumina Human1M-Duo Beadchip (19,883 out of 1,803,869 variants on chromosome 20) to mimic a typical GWAS data set. Two reference panels were constructed, one of 503 European (from the 1000G CEU, FIN, GBR, IBS and TSI samples) individuals, and the other including 600 African (from the 1000G ACB, ESN, GWD, LWD, MSL and YRI samples) individuals. The detailed distribution of reference populations is listed in Table S2. All genotype data and ancestry information are from 1000G phase 3 release[6].

We conducted meta-imputation on ASW samples using the European panel and African panel and evaluated the imputation accuracy by calculating aggregated r^2 between the imputed results and the masked genotype data. In order to obtain the aggregated r^2, we grouped the markers by the minor allele frequency (MAF) in the entire 1000G data set. The aggregated r^2 for each group is calculated as the squared Pearson correlation between the imputed dosages and the true minor allele counts across the markers in the group.

2.2.5 Empirical Assessment #2: Evaluation in South Asian Samples from UK Biobank

To illustrate the capability of our method to improve imputation when used together with large reference panels, we tested it on South Asian ancestry individuals in UK Biobank[28]. Genomes

for these individuals are hard to impute using reference panels such as HRC[21] and TOPMed[27] that include relatively few Asian ancestry individuals despite their size. HRC[21] and TOPMed[27] are typically outstanding at imputing missing genotypes in the bulk of the UK Biobank samples, which are of European origin.

The 2019 release of UK Biobank includes approximately 50,000 individuals with both array data and whole exome sequencing data [28]. We imputed the array data across the autosomes and used the exome data as a truth set to evaluate the accuracy of imputed variants. We assigned ancestry to UK Biobank participants by running a supervised ADMIXTURE [1] analysis with the Human Genome Diversity Project (HGDP) data[5] as a reference. Using a threshold of 70% genome content to classify an individual into a population, we identified 762 individuals as South Asian.

We meta-imputed genotypes for these 762 South Asian samples (pre-phased using Eagle v2.3.5 [19] without a reference panel) across the autosomes using the second release of TOPMed panel which includes 97,256 individuals[27] and the 1000G phase 3 (GRCh38) panel which includes 2504 individuals [6]. We evaluated the imputation accuracy by comparing the imputed results with the exome sequencing data. For comparison, we repeated the experiment using several individuals with other ancestries and also after adding half of the exome variants to the array dataset, enabling us to evaluate whether the inclusion of rare and low frequency variants in the scaffold used for imputation might improve results.

Finally, we constructed a combined panel for chromosome 20 by jointly calling the variants in 2504 1000G samples and 86,594 TOPMed samples from their sequence alignment files, split it into two subpanels with singletons excluded, and compared the performance of meta-imputation and imputation using the combined panel.

2.3 Results

2.3.1 Meta-Imputation in African American Samples

We first evaluated our method in the context of the 1000G ASW samples (African Americans from the Southwest US), using reference panels consisting of other 1000G samples of mainly African ancestry (the AFR panel), mainly European ancestry (the EUR panel), or the combination of all these individuals (the AFR+EUR panel). As shown in Figure 2.2, meta-imputation achieved the same accuracy as imputation using the combined AFR+EUR panel, which suggests that meta-imputation can serve as an efficient alternative when a combined reference panel is unavailable or impractical. Importantly, our results also show that the accuracy from meta-imputation was substantially greater than that from imputation using a single reference panel. For variants with minor allele frequency of 0.05%$\tilde{0}$.1%, meta-imputation achieved higher accuracy ($r^2 = 0.427$ between imputed dosages and actual genotypes) than imputation using the AFR panel alone ($r^2 = 0.313$) or using the EUR panel alone ($r^2 = 0.009$), and the accuracy of meta-imputation was comparable to that using the AFR+EUR panel ($r^2 = 0.425$). Overall, we observed the largest advantages of meta-imputation, compared to using one of the smaller panels, for rare variants.

2.3.2 Meta-Imputation in South Asian Samples

Next, we examined whether the benefits of meta-imputation would extend to settings where very large reference panels are available. Generally, these larger reference panels yield better imputation quality, but there are some exceptions. For example, it has been pointed out that the TOPMed panel sometimes underperforms the much smaller 1000G panel, particularly for ancestries (such as South Asian) that are poorly represented in TOPMed [27]. For this assessment, we used UK Biobank samples that have been exome sequenced and compared the results of imputation and meta-imputation with those of exome sequencing.

Figure 2.3 shows that the 1000G panel generally exhibited slightly better accuracy for imputing South Asian genomes than the TOPMed panel for variants with MAF $> 0.2\%$. Our results also

suggested that meta-imputation was able to improve the accuracy even further. For example, the imputation quality for variants with MAF of 0.05%$\tilde{0}$.1% increased from $r^2 = 0.231$ (using the 1000G panel alone) and $r^2 = 0.260$ (using the TOPMed panel alone) to $r^2 = 0.311$ (using meta-imputation with the 1000G panel and TOPMed panel imputation results as input). Also, the number of well-imputed (imputation $r^2 > 0.3$ reported by imputation software) variants on autosomes increased from 16,480,094 (imputation using 1000G panel) to 25,713,394 (meta-imputation), which suggests that 56% more variants would be available for downstream analyses.

We also evaluated a hypothetical combined panel including 1000 Genomes and TOPMed samples. For this analysis, we constructed a combined panel including the 1000G samples and most TOPMed samples, and repeated the experiment on chromosome 20. The result (Figure B.2) shows that meta-imputation achieves comparable accuracy to imputation using the combined panel even in this challenging setting where the reference panels differ greatly in size.

As part of meta-imputation, weights for each reference panel were estimated along each chromosome for each haplotype, reflecting the optimal choice of reference panel at each marker. Figure 2.4A illustrates the pattern of weights along the genome for a typical South Asian ancestry sample, where red indicates a preference for TOPMed and blue indicates a preference for 1000G. In the example, both the 1000G panel and the TOPMed panel are favored in substantial portions of the genome. By contrast, meta-imputation generally places a much heavier weight on the TOPMed panel when tackling a European ancestry sample, as shown in Figure 2.4B.

2.3.3 Computational Time

In principle, meta-imputation is relatively inexpensive (computationally) but there are challenging details in implementation, particularly because input and output file sizes can be extremely large. To achieve computational efficiency, in terms of both memory and CPU usage, we first calculate meta-imputation weights for each haplotype at genotyped markers only. The resulting weight matrices can then be used to scan through imputation results one marker at a time, reading panel specific imputation results, interpolating weights, and outputting weighted meta-imputation

dosages. Since meta-imputation combines imputed dosages, the cost of the meta-imputation step depends on the number of genotyped and imputed markers and on the number of individuals being processed, but not on the reference panel sample sizes.

We tested meta-imputation performance on different numbers of individuals (Table 2.1). For this analysis, we used the 1000G Phase 3 and TOPMed release 2 reference panels, which include 6,771,422 markers on chromosome 20 (1000G contains 1,052,215 markers, TOPMed contains 6,631,674 markers, 912,467 markers overlap). The single-core computational times of meta-imputation for 1000, 2000, 5000 and 10,000 target samples are reported in Table . Generally, the computational requirements for our implementation of meta-imputation are linear with respect to the number samples being imputed (earlier implementations performed in quadratic time because of less efficient memory and input/output usage). The per sample time for the imputation step with the 1000G and TOPMed reference panels using Minimac4 was about 20 seconds and for the meta-imputation using MetaMinimac2 was about 2 seconds. Since chromosome 20 accounts for about 2% of the genome, these estimates translate into about 17 minutes per genome for imputation and 2 minutes for meta-imputation.

2.4 Discussion

We have presented a convenient and efficient meta-imputation framework that enables researchers to merge imputed data generated using multiple reference panels. The meta-imputation procedure consists of two separate steps, imputation and integration, allowing investigators to incrementally consider new reference panels without repeating imputation steps using prior panels. As each panel is added, investigators need only impute the target samples against the new panel and can then combine the results with previously computed imputed result data sets. Our method does not require access to individual-level data from the reference panels and should perform gracefully even when the optimal choice of reference panel varies between individuals or along the genome of each individual. In principle, we expect our method to perform well even when reference panels

have partial overlap.

We first illustrated the performance of our method for meta-imputation in African American ancestry samples using reference panels consisting mainly of European haplotypes, mainly of African haplotypes, or their combination – a challenging situation for meta-imputation. Since the proportion of African ancestry will vary between individuals and along the genome of each individual, achieving accurate meta-imputation requires weights that are highly customizable – varying between individuals and along the genome of each individual. We also evaluated our methods in South Asian samples using reference panels with a large disparity in size. In these scenarios, meta-imputation not only outperformed imputation using either panel alone, but also compared well with imputation against the merged panel in terms of accuracy. Therefore, we propose that it will be safe to use our method even when the reference panels used for the initial imputation step are both sub-optimal, since our MetaMinimac2 algorithm is able to incorporate the best information from the different imputation results to yield much improved genotype dosages.

Improved imputation accuracy brings greater statistical power in GWAS. In the scenarios we examined (see Appendix B.1.1 and Figure B.4), the power of GWAS using meta-imputed dosages is comparable to the power of a hypothetical GWAS using imputed dosages from a merged panel. A previous recommendation for conducting GWAS when multiple reference panels are available was to conduct multiple GWAS (one for each set of imputation results) and to use the smallest p-value at each marker after imputation, carrying out simulations to estimate an appropriate multiple testing correction[32]. This approach also approximates the power of analysis with a combined panel. One of the reasons is that it may captures some of the features of multiple imputation[12], and we speculate that the power of GWAS using our approach might be further improved in the multiple imputation framework. Although the best p-value also performs well, our approach provides important advantages. First, because it produces a single consensus set of imputed dosages, the computational effort required to analyze additional phenotypes is more modest. Additionally, this consensus set of imputed dosages can serve as input to a variety of additional analyses – including trait co-localization [23, 29] and fine-mapping [3, 31, 30].

In the scenarios we examined, meta-imputation consistently produced better accuracy than imputation using only one of the available reference panels. However, it is not necessarily the case that every variant would gain in imputation quality. A challenging question concerns handling of variants that are present in only a subset of the reference panels. If a variant is present in one reference panel only, we opted to preserve the original imputed results for that variant. This is appropriate if we expect the presence or absence of a variant to be due to technical reasons, such as arbitrary differences in filtering criteria or accessibility of different parts of the genome using different sequencing technologies. An alternative would be to score variants that are absent from one panel as if they always match the reference genome in haplotypes from that panel, assigning them a dosage of zero. The optimal choice between these two alternatives will depend on the details of how panels were generated and whether panel specific variants reflect patterns of natural variation or technical artifacts due to variant calling and filtering. Since meta-imputation works on a per haplotype basis, its performance relies on the quality of pre-phasing. Switch errors in phasing may result in decreased imputation accuracy and misleading weights, so meta-imputation should directly benefit from evolving phasing algorithms [19, 10, 4]. The accuracy of meta-imputation could also be affected by factors including the density of genotype array and choice of variants. We would expect that a denser genotype array may bring improved accuracy as it could provide more information and better reflect the local performance of each reference panel. In our experiment (see Appendix B.1.2 and Figure B.3), supplementing the common variant array genotypes with the exome variants did not make a substantial difference in the imputation accuracy. This is because the weights estimated using common variants are also close to the ideal weights for imputation of rare variants.

In the current era, where imputation reference panels are often shared through convenient imputation servers [21, 27, 8], which increase user convenience and protect genetic information in the panel, our approach allows results from different servers to be combined and also allows studies who create their own panels to combine results generated using these panels with results generated from one or more imputation servers. We hope that these meta-imputation strategies will

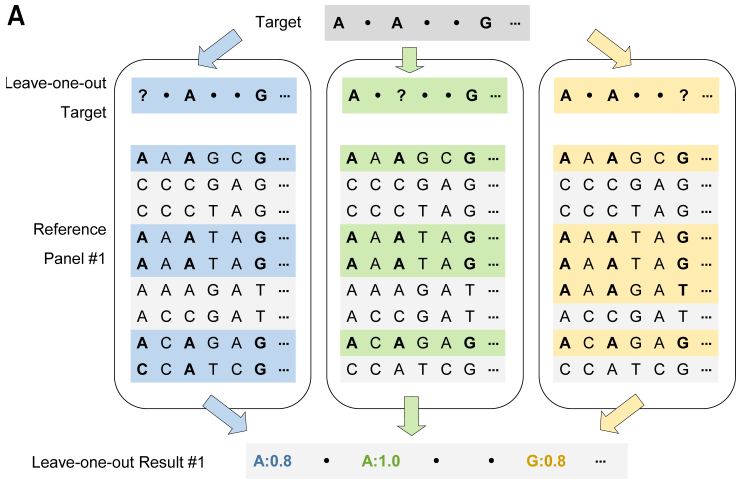
continue to extend the reach of imputation towards rarer and rarer variants and facilitate studies in diverse populations, where supplementing publicly available reference panels with complementary targeted panels is likely to be especially useful.

Table 2.1: Computational time of meta-imputation for UK Biobank samples

Number of Samples	Time ([hh]:mm:ss)			
	Step 1: Minimac4		Step 2: MetaMinimac2	Total
	1000G	TOPMed		
1,000	21:57	5:42:37	38:45	6:43:19
2,000	43:34	11:06:08	1:16:57	13:06:39
5,000	1:45:44	26:40:53	3:12:12	31:38:49
10,000	3:34:10	53:15:35	6:14:16	63:04:01

The analysis was conducted on chromosome 20, which involved 17,388 genotyped markers in the target haplotypes and 6,771,422 markers in reference panels. 1000G phase 3 (GRCh38) panel contains 1,052,215 markers; TOPMed release 2 panel contains 6,631,674 markers; 912,467 markers overlap. All the tests were conducted on Intel Xeon Platinum 8268 CPU @ 2.90GHz using one core at a time.

Figure 2.1: An illustration of leave-one-out imputation

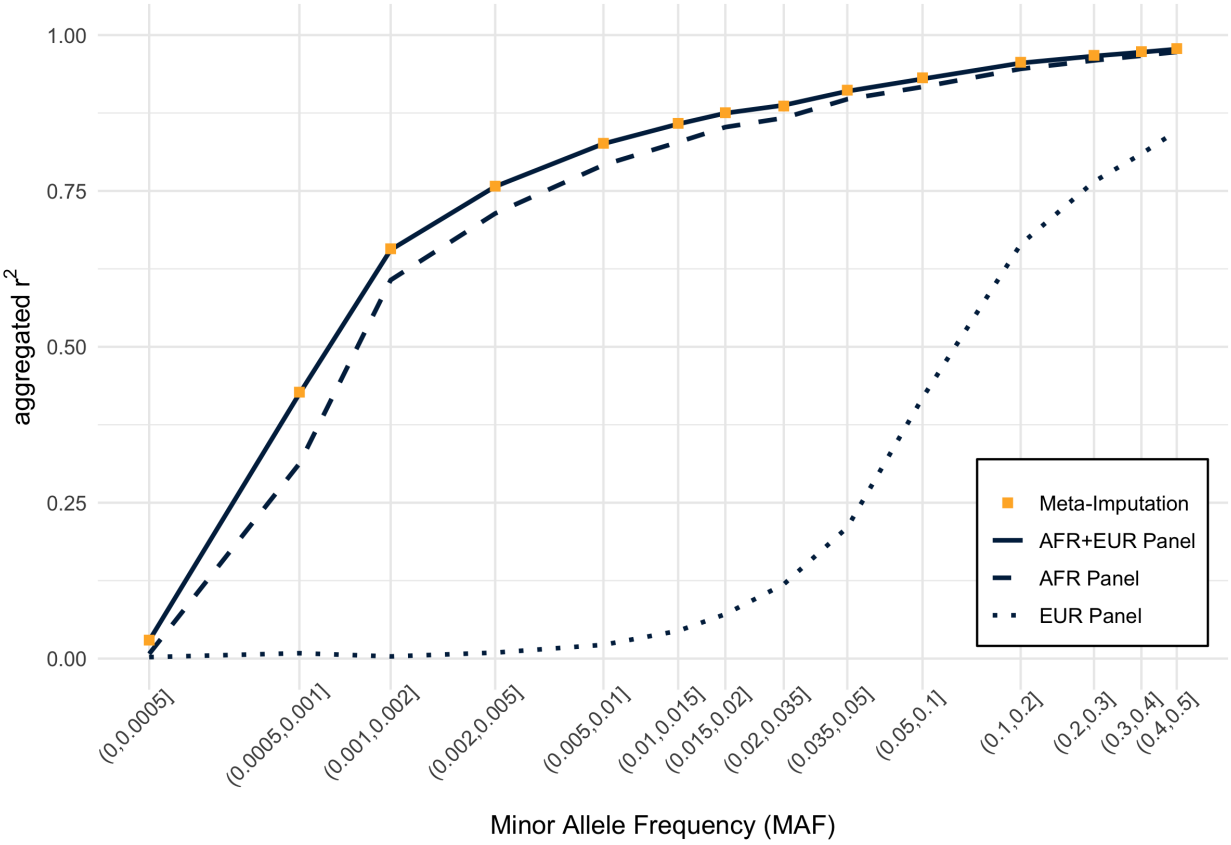


C

Target	A	A	G	A	C	G	A	C	G
Leave-one-out Result #1	**0.8**	**1.0**	**0.8**	**0.8**	0.2	0.3	0.3	0.5	0.3
Leave-one-out Result #2	0.3	0.2	0.3	**0.8**	**1.0**	**0.8**	**0.8**	**0.8**	**1.0**
Weight on Panel #1	High	High	High	Med	Low	Low	Low	Low	Low

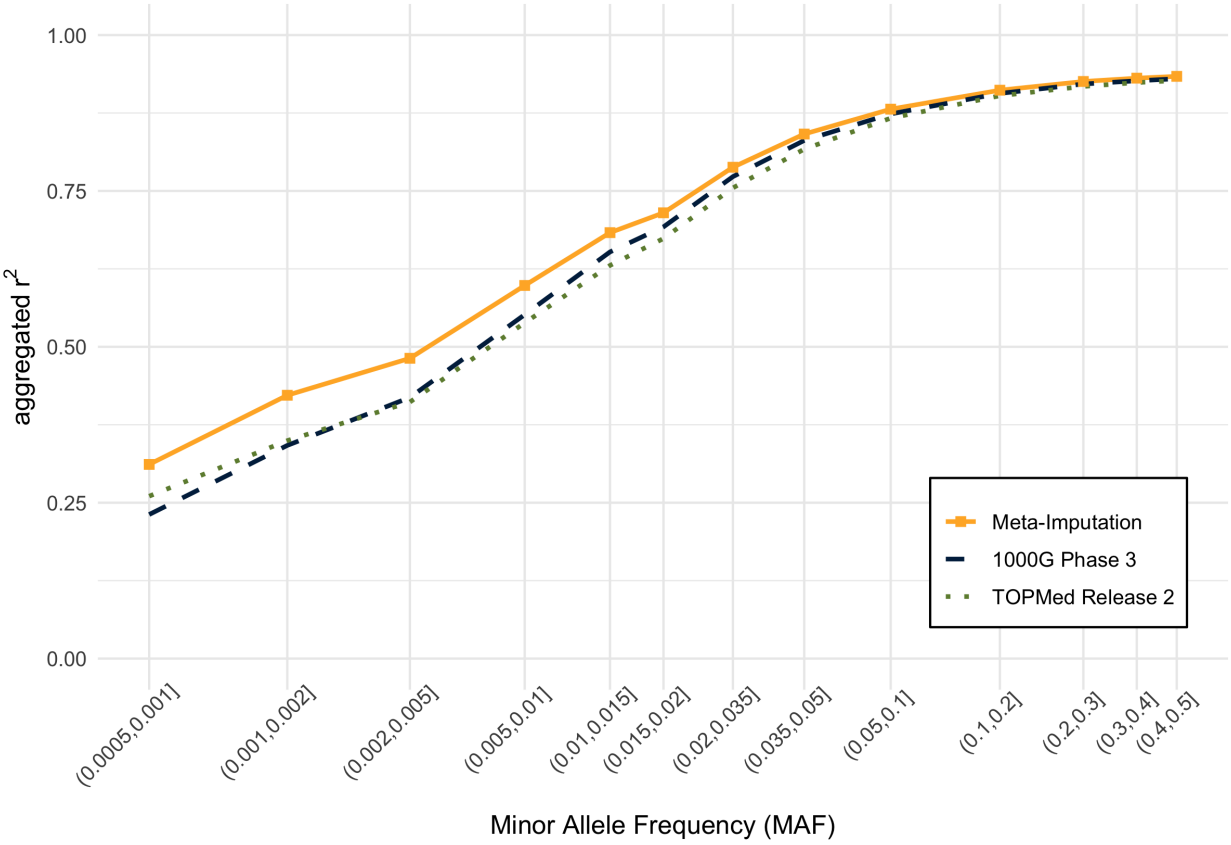
Figure A and B illustrate LOO imputation on a small chunk of 6 genotype markers using two reference panels, respectively. The target haplotype is genotyped at three markers (1,3,6). During the LOO imputation procedure, one marker was masked at a time, denoted as '?'. The figure simplifies the HMM procedure to estimating LOO results based on exact matching according to the unmasked markers (an HMM is used in the actual algorithm). For example, when performing LOO imputation using reference panel #1, we first masked the observed allele "A" at marker 1 and found five haplotype matches (shaded in blue) based on marker 3 and marker 6. The alleles from the five matches at maker 1 were AAAAC, which suggested a result of "A" with probability 0.8. Thus we obtained that the probabilities of observing the true allele at marker (1,3,6) were (0.8, 1.0, 0.8) from panel #1 and (0.3, 0.2, 0.3) from panel #2, which were compared in Figure 1C along with LOO results at other genotyped markers. Panel #1 was more accurate than panel #2 at the beginning but less accurate at the end, so ideally the weight on panel #1 should be high at the beginning and low at the end.

Figure 2.2: Comparison of imputation accuracy in African American samples



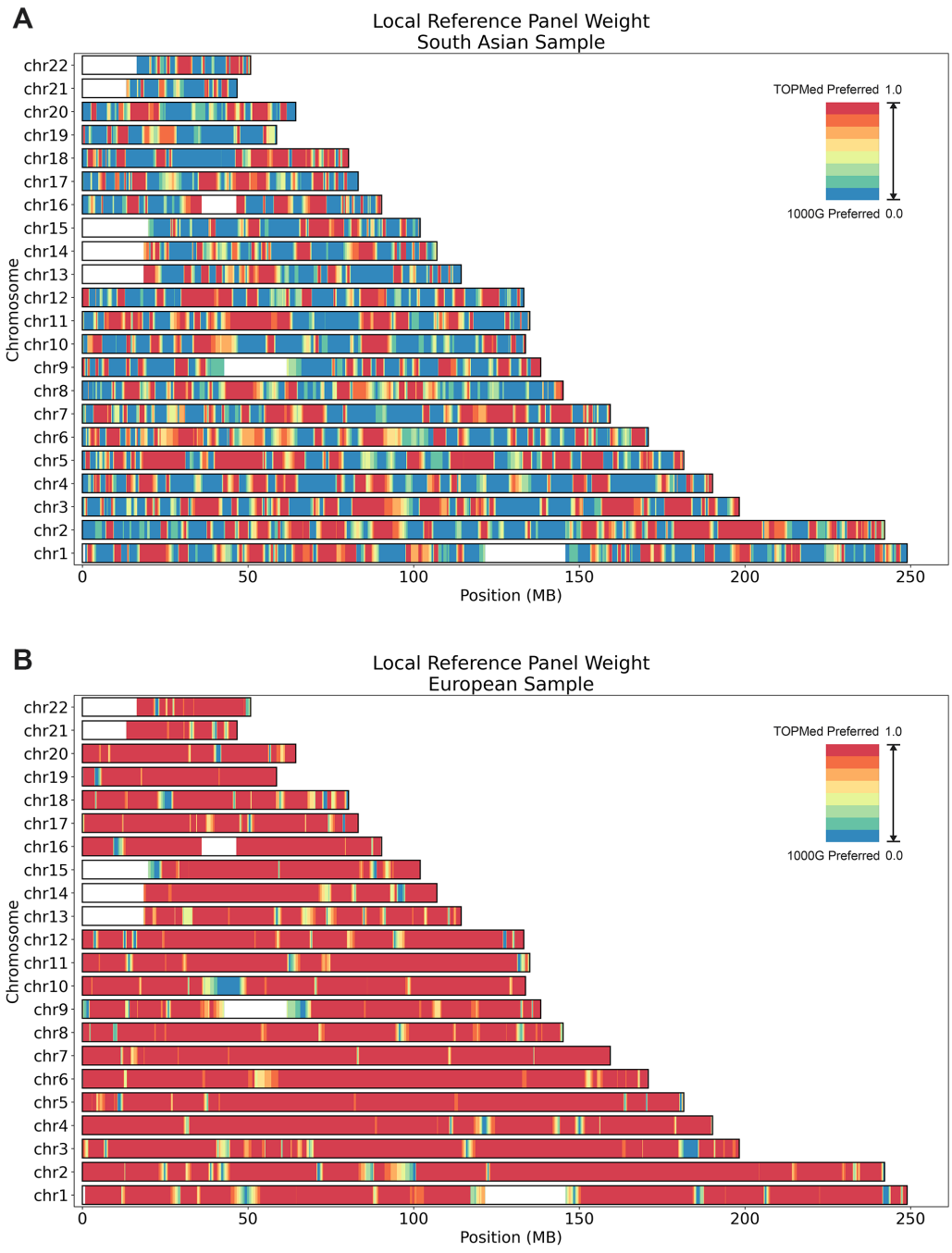
Imputation accuracy for the pseudo-GWAS ASW data set was compared among: 1) meta-imputation; 2) imputation using the combined AFR+EUR panel including both African and European ancestry genomes; 3) imputation using the homogeneous African (AFR) panel; 4) imputation using the homogeneous European (EUR) panel. Variants were grouped according to minor allele frequency, which was estimated from the genotype data of 2504 samples in the 1000 Genomes Project. Aggregated r^2 were calculated for each variant group.

Figure 2.3: Comparison of imputation accuracy in South Asian sample



Imputation accuracy for 762 South Asian samples in UK Biobank data was compared among 1) meta-imputation; 2) imputation using 1000G Phase 3 (GRCh38) panel; 3) imputation using TOPMed Release 2 panel. Aggregated r^2 was computed based on 918,144 variants shared by the 1000G panel, the TOPMed panel and UK Biobank whole exome sequencing data. Variants were binned according to minor allele frequency, which was estimated from exome sequencing data for the 762 samples.

Figure 2.4: Genome-wide summary of weights used in meta-imputation



UK Biobank samples were meta-imputed against the 1000G phase 3 panel and the TOPMed release 2 panel. The figures display the local weights on the TOPMed panel from the weight estimation step, where red indicates a preference for TOPMed and blue indicates a preference for 1000G. Figure A corresponds to the analysis of a sample haplotype with South Asian ancestry, where both the 1000G panel and the TOPMed panel were favored in substantial portions of the genome. Figure B corresponds to the analysis of a sample haplotype with European ancestry, where the TOPMed panel was nearly always favored.

2.5 References

[1] David H Alexander, John Novembre, and Kenneth Lange. Fast model-based estimation of ancestry in unrelated individuals. *Genome research*, 19(9):1655–1664, 2009.

[2] Leonard E Baum, Ted Petrie, George Soules, and Norman Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. *The annals of mathematical statistics*, 41(1):164–171, 1970.

[3] Christian Benner, Chris CA Spencer, Aki S Havulinna, Veikko Salomaa, Samuli Ripatti, and Matti Pirinen. Finemap: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics*, 32(10):1493–1501, 2016.

[4] Brian L Browning, Xiaowen Tian, Ying Zhou, and Sharon R Browning. Fast two-stage phasing of large-scale sequence data. *The American Journal of Human Genetics*, 108(10):1880–1890, 2021.

[5] Howard M Cann, Claudia De Toma, Lucien Cazes, Marie-Fernande Legrand, Valerie Morel, Laurence Piouffre, Julia Bodmer, Walter F Bodmer, Batsheva Bonne-Tamir, Anne Cambon-Thomsen, et al. A human genome diversity cell line panel. *Science*, 296(5566):261–262, 2002.

[6] 1000 Genomes Project Consortium et al. A global reference for human genetic variation. *Nature*, 526(7571):68, 2015.

[7] Sayantan Das, Gonçalo R Abecasis, and Brian L Browning. Genotype imputation from large reference panels. *Annual review of genomics and human genetics*, 19:73–96, 2018.

[8] Sayantan Das, Lukas Forer, Sebastian Schönherr, Carlo Sidore, Adam E Locke, Alan Kwong, Scott I Vrieze, Emily Y Chew, Shawn Levy, Matt McGue, et al. Next-generation genotype imputation service and methods. *Nature genetics*, 48(10):1284–1287, 2016.

[9] Patrick Deelen, Androniki Menelaou, Elisabeth M Van Leeuwen, Alexandros Kanterakis, Freerk Van Dijk, Carolina Medina-Gomez, Laurent C Francioli, Jouke Jan Hottenga, Lennart C Karssen, Karol Estrada, et al. Improved imputation quality of low-frequency and rare variants in european samples using the 'genome of the netherlands'. *European Journal of Human Genetics*, 22(11):1321–1326, 2014.

[10] Olivier Delaneau, Jean-François Zagury, Matthew R Robinson, Jonathan L Marchini, and Emmanouil T Dermitzakis. Accurate, scalable and integrative haplotype estimation. *Nature communications*, 10(1):1–10, 2019.

[11] Lars G Fritsche, Wilmar Igl, Jessica N Cooke Bailey, Felix Grassmann, Sebanti Sengupta, Jennifer L Bragg-Gresham, Kathryn P Burdon, Scott J Hebbring, Cindy Wen, Mathias Gorski, et al. A large genome-wide association study of age-related macular degeneration highlights contributions of rare and common variants. *Nature genetics*, 48(2):134–143, 2016.

[12] Bryan N Howie, Peter Donnelly, and Jonathan Marchini. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS genetics*, 5(6):e1000529, 2009.

[13] Jie Huang, Bryan Howie, Shane McCarthy, Yasin Memari, Klaudia Walter, Josine L Min, Petr Danecek, Giovanni Malerba, Elisabetta Trabetti, Hou-Feng Zheng, et al. Improved imputation of low-frequency and rare variants using the uk10k haplotype reference panel. *Nature communications*, 6(1):1–9, 2015.

[14] Hákon Jónsson, Patrick Sulem, Birte Kehr, Snaedis Kristmundsdottir, Florian Zink, Eirikur Hjartarson, Marteinn T Hardarson, Kristjan E Hjorleifsson, Hannes P Eggertsson, Sigurjon Axel Gudjonsson, et al. Whole genome characterization of sequence diversity of 15,220 icelanders. *Scientific data*, 4(1):1–9, 2017.

[15] S Krokstad, A Langhammer, K Hveem, TL Holmen, K Midthjell, TR Stene, G Bratberg, J Heggland, and J Holmen. Cohort profile: the hunt study, norway. *International journal of epidemiology*, 42(4):968–977, 2013.

[16] Markku Laakso, Johanna Kuusisto, Alena Stančáková, Teemu Kuulasmaa, Päivi Pajukanta, Aldons J Lusis, Francis S Collins, Karen L Mohlke, and Michael Boehnke. The metabolic syndrome in men study: a resource for studies of metabolic and cardiovascular diseases. *Journal of lipid research*, 58(3):481–493, 2017.

[17] James J Lee, Robbee Wedow, Aysu Okbay, Edward Kong, Omeed Maghzian, Meghan Zacher, Tuan Anh Nguyen-Viet, Peter Bowers, Julia Sidorenko, Richard Karlsson Linnér, et al. Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nature genetics*, 50(8):1112–1121, 2018.

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

[19] Po-Ru Loh, Pier Francesco Palamara, and Alkes L Price. Fast and accurate long-range phasing in a uk biobank cohort. *Nature genetics*, 48(7):811–816, 2016.

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

[21] Shane McCarthy, Sayantan Das, Warren Kretzschmar, Olivier Delaneau, Andrew R Wood, Alexander Teumer, Hyun Min Kang, Christian Fuchsberger, Petr Danecek, Kevin Sharp, et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nature genetics*, 48(10):1279, 2016.

[22] Giorgio Pistis, Eleonora Porcu, Scott I Vrieze, Carlo Sidore, Maristella Steri, Fabrice Danjou, Fabio Busonero, Antonella Mulas, Magdalena Zoledziewska, Andrea Maschio, et al. Rare variant genotype imputation with thousands of study-specific whole-genome sequences: implications for cost-effective study designs. *European Journal of Human Genetics*, 23(7):975–983, 2015.

[23] Vincent Plagnol, Deborah J Smyth, John A Todd, and David G Clayton. Statistical independence of the colocalized association signals for type 1 diabetes and rps26 gene expression on chromosome 12q13. *Biostatistics*, 10(2):327–334, 2009.

[24] Allison A Regier, Yossi Farjoun, David E Larson, Olga Krasheninina, Hyun Min Kang, Daniel P Howrigan, Bo-Juen Chen, Manisha Kher, Eric Banks, Darren C Ames, et al. Functional equivalence of genome sequencing analysis pipelines enables harmonized variant calling across human genetics projects. *Nature communications*, 9(1):1–8, 2018.

[25] Angelo Scuteri, Serena Sanna, Wei-Min Chen, Manuela Uda, Giuseppe Albai, James Strait, Samer Najjar, Ramaiah Nagaraja, Marco Orrú, Gianluca Usala, et al. Genome-wide association scan shows genetic variants in the fto gene are associated with obesity-related traits. *PLoS genetics*, 3(7):e115, 2007.

[26] Eli A Stahl, Gerome Breen, Andreas J Forstner, Andrew McQuillin, Stephan Ripke, Vassily Trubetskoy, Manuel Mattheisen, Yunpeng Wang, Jonathan RI Coleman, Héléna A Gaspar, et al. Genome-wide association study identifies 30 loci associated with bipolar disorder. *Nature genetics*, 51(5):793–803, 2019.

[27] Daniel Taliun, Daniel N Harris, Michael D Kessler, Jedidiah Carlson, Zachary A Szpiech, Raul Torres, Sarah A Gagliano Taliun, André Corvelo, Stephanie M Gogarten, Hyun Min Kang, et al. Sequencing of 53,831 diverse genomes from the nhlbi topmed program. *Nature*, 590(7845):290–299, 2021.

[28] Cristopher V Van Hout, Ioanna Tachmazidou, Joshua D Backman, Joshua D Hoffman, Daren Liu, Ashutosh K Pandey, Claudia Gonzaga-Jauregui, Shareef Khalid, Bin Ye, Nilanjana Banerjee, et al. Exome sequencing and characterization of 49,960 individuals in the uk biobank. *Nature*, 586(7831):749–756, 2020.

[29] Chris Wallace, Maxime Rotival, Jason D Cooper, Catherine M Rice, Jennie HM Yang, Mhairi McNeill, Deborah J Smyth, David Niblett, François Cambien, Cardiogenics Consortium, et al. Statistical colocalization of monocyte gene expression and genetic risk variants for type 1 diabetes. *Human molecular genetics*, 21(12):2815–2824, 2012.

[30] Gao Wang, Abhishek Sarkar, Peter Carbonetto, and Matthew Stephens. A simple new approach to variable selection in regression, with application to genetic fine mapping. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(5):1273–1300, 2020.

[31] Xiaoquan Wen, Yeji Lee, Francesca Luca, and Roger Pique-Regi. Efficient integrative multi-snp association analysis via deterministic approximation of posteriors. *The American Journal of Human Genetics*, 98(6):1114–1129, 2016.

[32] Wei Zhou, Lars G Fritsche, Sayantan Das, He Zhang, Jonas B Nielsen, Oddgeir L Holmen, Jin Chen, Maoxuan Lin, Maiken B Elvestad, Kristian Hveem, et al. Improving power of association tests using multiple sets of imputed genotypes from distributed reference panels. *Genetic epidemiology*, 41(8):744–755, 2017.

CHAPTER 3

Exploring the Limitations of Statistical Fine-Mapping Analysis of Genetic Association Signals

3.1 Introduction

3.1.1 Background and Motivations

Genome-wide association studies (GWAS) have successfully identified thousands of genetic associations for diseases and complex traits [7, 15, 17, 21]. However, GWAS signals often point to broad regions of the genome which harbor hundreds of genetic variants, among which some are potentially causal while most are implicated due to linkage disequilibrium (LD) with the true causal variant. With the presence of complex LD structure, it is often challenging to pinpoint the true causal variant, and therefore additional fine-mapping analyses are required to prioritize the candidate causal variants for follow-up functional studies.

Prioritizing variants intuitively often hinges on p-values. While it might seem logical to infer that the variant with the smallest p-value is most likely to be causal, this isn't always the case – a non-causal variant could have the lowest p-value due to LD with the actual causal variant or due to statistical fluctuations [3]. The limitations of using p-values in this context become apparent when we consider that p-values cannot quantify the uncertainty of a variant being causal [18]. Also, p-

values are not comparable across variants or across different studies, given that they are influenced by minor allele frequency and sample size. Therefore, the Bayes factor has been increasingly recognized as a viable alternative to the p-value for summarizing the evidence of associations [22]. The earliest Bayesian fine-mapping approach ranks the associations by posterior probability which is proportional to the Bayes factor of each variant, while assuming that exact one of the variants is causal in each region [13]. Several methods have been proposed to lift this restriction by jointly analyzing all the variants in the region. Conditional approach uses a stepwise selection procedure, sequentially selecting variants based on conditional p-values, which are recalculated at each step [26]. However, this approach carries the drawbacks of using p-values: the necessity to set a significance threshold and the inability of quantifying uncertainty in the selection process, which is sub-optimal in terms of power and precision with the presence of complex LD structure. Consider two SNPs in perfect LD, only one being causal, the conditional approach will randomly choose one of them in the selection process and thus miss the true causal variant in half cases.

One of the major methods that overcomes this limitation was presented in Hormozdiari et al. (CAVIAR) [10], in which they took the approach of jointly modeling multiple causal variants rather than sequentially. They framed the issue as a Bayesian Variable Selection (BVS) problem, taking into account all possible combinations of variants and calculating the posterior probability for each variant in a genomic region to be causal. Subsequently, more scalable BVS implementations were proposed to avoid the need for exhaustive enumeration of all causal configurations. FINEMAP optimizes computational efficiency in fine-mapping by using a stochastic shotgun search [2], a technique that emphasizes the most probable subset of causal configurations. DAP-G [11, 24]leverages an efficient deterministic search strategy to discern plausible models and approximates the normalizing constants using the well-established statistical principle of Sure Independence Screening [6]. SuSiE introduces the ”Sum of Single Effects” model [23] , implementing an iterative Bayesian stepwise selection algorithm that enables effective computation in fine-mapping by associating each credible set with variants sharing similar marginal impacts on traits.

While BVS offers a robust alternative to traditional methods, it’s not exempt from complica-

tions, especially when dealing with complex LD structures. One prominent issue arises when a non-causal variant is linked with multiple causal variants. In such scenarios, the non-causal variant may manifest with higher significance than the actual causal variants. For example, the sole 95% credible set variant located in SKIV2L intron pinpointed by fine-mapping was proved not truly associated with age-related macular degeneration according to haplotype analysis. Instead, it tags with two CFB missense variants and shows stronger association than either of these true causal variants [7].

Another big challenge for fine-mapping is data-sharing restrictions. Privacy concerns have grown in prominence in the era of big data, especially when dealing with genomic data as individual genetic data can be personally identifiable and potentially misused in ways that could discriminate or harm individuals [16]. To address these concerns, data sharing policies and best practices have been instituted across research communities, limiting the availability of individual-level data to a broader audience. This limitation has necessitated the development of methods tailored to work with summary statistics.

In this chapter, we investigated non-identifiability issues caused by complex LD structure within the BVS framework and the resulting false discoveries in fine-mapping. We evaluated the limitations of BVS implementations which employ greedy algorithms, especially when they handle the non-identifiable cases. Also, we evaluated the limitations of fine-mapping analyses that utilize summary statistics and compared their power and coverage against those using individual-level data.

3.1.2 Overview of Bayesian Variable Selecion Framework

Assume that we have observed genotype and phenotype data for n samples. Let $y = (y_1, y_2, ..., y_n)^T$ represent the trait of interest of n samples and $X_{n\times p} = (x_1, x_2, ..., x_p)$ represents the genotype matrix, where x_j is an n-vector of genotype at the jth variant, $j = 1, 2, ..., p$. The

relationship between genotype and the trait of interest is modeled as a multiple regression.

$$y = X\beta + \epsilon, \epsilon\ N(0, \sigma^2 I_n) \tag{3.1}$$

where β denote the effect sizes of the $pvariants$,ϵ is an n-vector of error terms, and σ^2 is the residual variance. For simplicity, we assume that both X and y have been mean-centered before analysis so that we do not have to consider the intercept term when modeling the association.

Statistical fine-mapping approaches primarily aim to identify causal variants rather than determining their effect sizes, thus transforming the issue into a variable selection problem. The causal status of the jth variant is denoted as $\gamma_j = I(\beta_j \neq 0)$, so that it is referred to as a causal variant if $\gamma_j = 1$. In order to quantify the uncertainty in variants selected, most existing fine-mapping methods employ the BVS framework (rather than non-Bayesian methods including LASSO [20] and COJO [26]), in which we focus on the inference of the posterior distribution of the indicator vector $\gamma = (\gamma_1, \gamma_2, ..., \gamma_p)^T$:

$$P(\gamma|X, y) = \frac{\pi(\gamma)BF(\gamma)}{\sum_{\gamma' \in \Gamma} \pi(\gamma')BF(\gamma')} \tag{3.2}$$

where Γ denotes the model space of 2^p possible configurations of γ, $\pi(\cdot)$ denotes the prior probability and $BF(\gamma) = (P(y|X, \gamma))/(P(y|X, \gamma = 0))$ denotes the Bayes factor for γ. The prior probability is a function of prior inclusion probabilities of variants $\overrightarrow{\pi} = (\pi_1, \pi_2, \cdots, \pi_p)'$ where π_j denotes the prior probability that variant j is causal, and we set $\pi_j = \frac{1}{p}$ by default:

$$\pi(\gamma) = \prod_{j=1}^{p} \pi_j^{\gamma_j} (1 - \pi_j)^{1-\gamma_j} \tag{3.3}$$

Subsequently, we can obtain the posterior inclusion probability (PIP) of each variant by marginalizing the posterior model probability:

$$PIP_j := P(\gamma_j = 1|X, y) = \sum_{\gamma:\gamma_j=1} P(\gamma|X, y) \tag{3.4}$$

3.1.3 Using Summary Statistics for Fine Mapping

It is challenging to obtain in practice due to data privacy concerns and sharing restrictions. Such obstacles have motivated methodological development of fine-mapping frameworks that require only summary statistics.

Studies have shown that given the summary-level data $(R, \hat{b}, \hat{s}, y^T y, n)$, we could achieve inference results identical to those obtained through fine-mapping using individual-level data. Here, R is the correlation matrix of genotypes between genetic variants, $\hat{b} = (\hat{b}_1, \hat{b}_2, \cdots, \hat{b}_p)^T$ and $\hat{s} = (\hat{s}_1, \hat{s}_2, \cdots, \hat{s}_p)^T$ represent the effect size and the corresponding standard errors from a simple linear regression for each variant, $y^T y$ denotes the sum of squares of the centered phenotype, and n denotes the sample size of the GWAS study. Proofs can be found in the supplementary materials.

However, the in-sample LD matrix R may not be available in practice and is typically substituted with an out-of-sample LD estimate $\hat{R}$ – the sample correlation matrix of the same set of variants from a reference panel of the same or a genetically similar population as the study samples.

3.1.4 Non-Identifiability Issues caused by LD

A model is non-identifiable when theoretically it is not possible to learn about the true values of the underlying model parameters, even given an infinite number of observations [14]. This is common in genetic research due to the presence of complex LD structure. Within the fine-mapping framework specifically, Bayesian non-identifiability occurs when two configurations γ and γ' yield identical posteriors, i.e., $P(\gamma|X, y) = P(\gamma'|X, y)$. Contrastingly, the definition of non-identifiability from a classical perspective refers to the cases with identical likelihoods $P(X, y|\gamma) = P(X, y|\gamma')$. It is recommended to consider both the identifiability of the likelihood and the identifiability of the posterior in analyses [4].

Consider an intuitive example where one non-causal variant is completely correlated with a causal variant, denoted as $x_1 = x_2$. The two configurations $\gamma = (1, 0)^T$ and $\gamma' = (0, 1)^T$ generate identical Bayes factors. Given identical priors, these configurations will consequently yield the same posteriors. In this case, the true causal variant cannot be distinguished from the non-

causal one based on the posterior or likelihood, regardless of the number of observations available, meaning that the model is non-identifiable. DAP-G [11] introduced the concept of a signal cluster. Variants within the same cluster exhibit modest to high LD with each other, and the posterior of a configuration that includes one variant resembles the posterior of a configuration that includes another variant within the same cluster. Similarly, SuSiE [23] introduced the concept of a credible set, which is defined as the smallest subset of variants that contains at least one causal variant, with a probability exceeding a predetermined threshold. As a result, we could make causal inference at the cluster level, rather than relying on a randomly selected variant, which could potentially overlook the true causal variant.

However, these methods may not work well under more complex LD structures, for example, when a non-causal variant is correlated with multiple causal variants. Let x_1 and x_2 denote two causal variants in weak correlation and let x_3 denote a non-causal variant correlated with both x_1 and x_2. Consider the following two models, corresponding to two configurations $\gamma_(M_1) = (1,1,0)^T$ and $\gamma_(M_2) = (1,1,0)^T$, respectively.

$$y = x_1\beta_1 + x_2\beta_2 + \epsilon_{M_1} \tag{M1}$$

$$y = x_3\beta_3 + \epsilon_{M_2} \tag{M2}$$

where y is the trait of interest, $\beta_1, \beta_2, \beta_3$ represent the effect sizes of x_1, x_2, x_3, respectively, and ϵ_{M_1} and ϵ_{M_2} represent the error terms. In an edge case where $x_3 \propto x_1\beta_1 + x_2\beta_2$, they two configurations will yield identical likelihood as x_3 provides as much information as the combination of x_1 and x_2. Therefore, an ideal inference should conclude that either both x_1 and x_2 are causal or x_3 is causal. The concepts of signal clusters or credible sets might not be effective in accurately capturing such an inference. The variant x_3 could be incorporated into either the x_1 cluster or the x_2 cluster, or be included in a standalone cluster. The latter case may lead to a false-positive discovery.

The multimodality posteriors will induce computational difficulties in practice, as it is almost intractable to explore the neighborhoods of all modes. Most fine-mapping methods including

SuSiE and DAP-G implement greedy algorithms for exploration of the model space and will get stuck at one mode while ignoring the other(s). For example, in the toy example, a greedy algorithm may explore the configurations that include x_3 while disregarding the configuration that contains x_1 and x_2 only, which will overestimate x_3's PIP and lead to a false discovery.

3.2 Results

3.2.1 Overdispersion of Sample Correlation Matrix

Fine-mapping using summary statistics requires the LD information between the variants. Ideally, this LD should be calculated directly from the GWAS samples. However, in many cases, the original GWAS data are not available, prompting the use of LD information derived from publicly accessible reference genotype panels as a substitute in the analysis. This substitution operates on the underlying assumption that when the sample size is suitably large, the sample correlation matrix can serve as an accurate approximation of the population correlation matrix. To explore the validity of this assumption, we carried out simulations under different settings of the ratio of number of variants to sample size.

Figure 3.1 demonstrates that when the ratio stands is 10:1 (with the number of variants $p = 500$ and the sample size $n = 50$), the largest eigenvalue of the sample correlation matrix is much higher than that of the population correlation matrix. Note that the eigenvalues of the correlation matrix R, denoted as $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$, are related to the spectral radius $\|R\|_2 = \lambda_1$. This allows us to evaluate the difference between the sample correlation matrix and population correlation matrix by comparing their eigenvalues since $\|\hat{R} - R\|_2 \geq \left| \|\hat{R}\|_2 - \|R\|_2 \right| = |\lambda_1(\hat{R}) - \lambda_1(R)|$. The large disparity in the largest eigenvalue suggests that the sample correlation may not provide a reliable estimate for the population correlation in such circumstances. An increase in the sample size leads to a decrease in this divergence. When the ratio reaches 1:10 (with the number of variants $p = 500$ and the sample size $n = 5000$) – a situation that is relatively rare in actual datasets – the sample correlation approximates the population correlation quite closely.

Our findings illustrates that when the ratio of the number of variants to sample size is high, a sample LD matrix could deviate substantially from the corresponding population LD matrix. This highlights the potential inaccuracies that can emerge in such situations and the need for careful consideration when using out-of-sample LD for fine-mapping.

3.2.2 Comparing Fine-Mapping using Summary Statistics and Individual-Level Data

The above analysis demonstrated that the application of summary statistics and LD data drawn from a reference population may be suboptimal in comparison to fine-mapping conducted directly with individual-level genotype and phenotype data. To assess the potential impacts of using summary statistics on the power and coverage of fine-mapping in practical scenarios, we conducted a series of comparative analyses utilizing genotype data from the TOPMed dataset [19].

We randomly selected 500 unrelated European samples to form a GWAS panel, and constructed two LD panels from the remaining European samples – one comprising 500 samples and the other 2500 samples. For fine-mapping with summary statistics, we examined three LD matrices, which were derived from the GWAS panel and the two LD panels, respectively. The power (the proportion of causal variants covered by signal clusters with SPIP$>$95%) and the coverage (the proportion of signal clusters with SPIP$>$95% that contain at least one causal variant) were compared between fine-mapping using individual-level data and using summary statistics with LD data from different panels.

The results indicate that both the power (Table 3.1) and coverage (Table 3.2) of fine-mapping when utilizing individual-level genotype and phenotype data are superior to those achieved when using summary statistics. When comparing the results derived from different LD matrices, we found that the choice of LD panel does not significantly impact the power of fine-mapping. However, the difference in coverage is striking, and this disparity widens with an increase in PVE. The coverage of fine-mapping using an in-sample LD consistently outperforms that from using an out-of-sample LD, even when the sample size of the LD panel is considerably larger than that of the

GWAS panel.

Under ideal circumstances – when the method is well-calibrated – we would expect that the coverage of signal clusters with SPIP $> 95\%$ should exceed 95%. However, when summary statistics are used, the coverage is way off the expectation as PVE increases, especially with the usage of an out-of-sample LD matrix. Notably, when using the LD matrix derived from an LD panel of 500 samples, the coverage dips below 95% as soon as the PVE reaches 0.10. At a PVE of 0.40, the coverage drops to 68.3%, in contrast to the 94.8% coverage achieved with the use of individual-level data.

3.2.3 Non-Identifiability Issues in Simulation Studies

To investigate the conditions under which non-identifiability due to complex LD structure can occur in fine-mapping, we conducted a simulation study involving 1000 individuals and 6 variants. We assumed that the phenotype is associated with five variants as formulated in Equation 3.5:

$$y = \beta_1 x_1 + \beta_2 x_2 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \epsilon \tag{3.5}$$

Among these, two causal variants, x_1 and x_2, are correlated, whereas the other three causal variants x_4, x_5, x_6 are independently and identically distributed following the standard normal distribution. A non-causal variant, x_3, exhibits a correlation with both x_1 and x_2, with correlation coefficients given by $cor(x_1, x_3) = cor(x_2, x_3) = 0.8$. The variance of the error term, ϵ, was adjusted such that the proportion of variance in the phenotype explained by x_1 and x_2 matches a pre-determined value (PVE = 0.05, 0.1, 0.15, 0.2).

We enumerated all 2^6 model configurations to determine the exact posterior probabilities, assuming a prior probability of 0.05 for the inclusion of each variant. We then compared the posterior probabilities for the true model configuration, denoted by $\gamma = (1, 1, 0, 1, 1, 1)^T$, and an alternative model configuration, denoted by $\gamma' = (0, 0, 1, 1, 1, 1)^T$. We examined the incidence of multimodality, defined as $\left|\log \frac{P(\gamma'|X,y)}{P(\gamma|X,y)}\right| < 1$, and inevitable false discoveries, defined as $\log \frac{P(\gamma'|X,y)}{P(\gamma|X,y)} \geq 1$.

Figure 3.2 illustrates that with a relatively high PVE, the posterior probability of the true model outweighs that of the alternative model in a majority of the cases. However, as the PVE decreases, the proportion of non-identifiable cases increases. Consequently, the data may increasingly favor the alternative model, resulting in a greater likelihood of false discoveries.

The presence of non-identifiability may trap the greedy algorithm at a local optimum. For example, the algorithm may select x_3 as the optimal starting point and consequently explore configurations that include x_3, while it may disregard configurations that include x_1, x_2 and other variants exclusive of x_3. Such a scenario could lead to an overestimation of x_3's PIP, resulting in a false discovery. A possible solution to this predicament is to refine the greedy algorithm by launching it from multiple starting points rather than solely the optimal one, to incorporate configurations that include either x_1 or x_2. By integrating this multi-starting-point strategy into DAP-G, we developed a new implementation known as DAP-MS.

To validate this idea, we expanded our simulated datasets under the condition of PVE=0.05 and $cor(x_1, x_2) = 0.5$ with an additional 994 independent non-variants. Under these conditions, the models were non-identifiable in 42.6% of the 500 simulated cases according to the posteriors obtained from exact calculations. We then performed fine-mapping among the 1000 variants using three methods: SuSiE, DAP-G, and our new implementation, DAP-MS. The PIPs generated by each method were then compared with those derived from exact calculations.

As depicted in Figure 3.3, both SuSiE and DAP-G demonstrated a tendency to underestimate the PIPs for x_1 and x_2, and to overestimate the PIP for x_3 when compared to the exact calculations. In contrast, the estimates provided by DAP-MS proved closer to those from exact calculation. Notably, in four cases where the exact calculation favored the true model, the PIP for x_3 was nearly 1 by DAP-G and SuSiE, suggesting potential false discoveries. However, these cases were better addressed with the DAP-MS implementation.

3.2.4 Numerical Comparisons with Real Genotype Data

To demonstrate the potential inflated false discoveries introduced by the usage of greedy algorithms with the presence of complex LD structures, we first evaluated the performance of SuSiE and DAP-G in fine-mapping using individual-level genotype and phenotype data. To ensure our analysis closely mirrored the LD structure in real-world situations, we simulated phenotypes based on genotype data from the Genotype-Tissue Expression (GTEx) project [5]. Details of the simulation study are available in the Methods section.

As illustrated in Figure C.1, the coverage and power of both methods tend to increase with rising PVE and decrease with an increasing number of causal variants. Generally, SuSiE demonstrates higher power compared to DAP-G, albeit with lower coverage. It is noteworthy that when the number of causal variants is three or more, the coverage begins to fall below 95%, which indicates more false discoveries than expected.

We then conducted a comparison of the cross-entropy measures among DAP-G, SuSiE, and DAP-MS. Cross-entropy is a measure of the difference between two probability distributions, and in this context, it is used to assess the calibration of the variant-wise PIPs. Lower cross-entropy signifies a closer match between the predicted and actual association status. The results from simulations under the parameter settings of PVE=0.1 and 3 causal variants are displayed in Table 3.3 and Figure 3.4. DAP-MS was found to exhibit slightly lower cross-entropy, which aligns with our hypothesis that the implementation of DAP-MS with multiple starting points may enhance the calibration. However, the improvement introduced by DAP-MS over SuSiE and DAP-G is rather marginal.

In addition to performance measures, it is also crucial to consider the computational costs associated with each of these methodologies. The computational costs, displayed in Table 3.4, highlight an interesting trade-off. With an increase in the number of causal variants, the computational time required for DAP-MS escalates significantly. This increase is anticipated as the number of model configurations inspected with the multiple-starting-point strategy grows faster than the greedy algorithm. Therefore, although DAP-MS can potentially enhance the calibration of the posterior

probabilities, the substantial computational overhead it imposes could limit its applicability in large-scale studies.

3.3 Discussion

In this chapter, we presented a comprehensive exploration of the trade-offs associated with different fine-mapping strategies, particularly highlighting the impacts of the choice of data type (summary statistics versus individual-level data) and the choice of algorithmic approach (greedy versus multiple starting-point strategy).

First, our evaluations underscored that using summary statistics, as opposed to individual-level genotype and phenotype data, typically leads to decreased power and coverage in fine-mapping. While this issue can be solved by providing sufficient statistics, such information is often available only with access to the individual data. An additional challenge we found pertains to the choice of LD panel when working with summary statistics. The most accurate results can be achieved when the LD matrix is derived directly from the GWAS panel. However, due to practical limitations, researchers often resort to using an LD matrix derived from a publicly available reference genotype panel. The out-of-sample LD matrix may not accurately replicate the sample correlation between genetic markers in the GWAS data. This discrepancy is particularly pronounced when the sample size of the LD panel is relatively small compared to the number of variants being considered, which may lead to an inflated rate of false discoveries.

To enhance accuracy in fine-mapping studies using summary statistics, researchers have suggested the use of shrinkage estimation or regularization for the sample LD matrix [25, 28]. This technique involves shrinking the off-diagonal entries towards zero, resulting in a sparse matrix. Notably, the sparsity of the matrix also contributes to faster computation, making it particularly advantageous for large genomic regions. While the shrinkage-based estimate may improve the reliability of fine-mapping results compared to using the sample LD from a reference panel (evaluations showed that SuSiE and DAP-G performed similarly at all levels of regularization though

[28]), it is essential to acknowledge that it is still less accurate than utilizing the in-sample LD matrix derived from the study samples [27].

Based on these findings, we recommend that researchers prudently evaluate their choice of data input and LD panel, particularly considering the available sample size and the number of variants under investigation. Also, as a forward-looking suggestion, we strongly advocate for researchers to publish their fine-mapping results in conjunction with GWAS results. This practice would ensure the provision of comprehensive data, enhancing the utility of these results for future research and facilitating more effective meta-analyses.

Furthermore, we investigated the challenges posed by non-identifiability in the presence of complex LD structures. Notably, even with an exact calculation of posteriors by enumerating all conceivable model configurations, it may still prove impossible to unambiguously identify the true association model based on the observed data. We observed that with the presence of multimodal posteriors, a greedy search strategy, such as the one employed by DAP-G, could be trapped at local optima, potentially leading to an overconfidence in the identified signals while overlooking alternate model configurations. Consequently, this could contribute to an increase in false discoveries.

In response to this challenge, we proposed an enhancement to the greedy algorithm – DAP-MS – incorporating a multiple starting-point strategy. This alternative approach demonstrated potential in improving the calibration of posterior probabilities and reducing false discoveries in multimodality cases. However, it is important to note that this refinement primarily emphasizes inference at the model level and lacks features like signal clusters. As a result, we were able to compare variant-level PIP in terms of probability calibration and cross entropy but were unable to directly compare signal-level coverage and power with other methods, such as SuSiE and DAP-G.

Furthermore, this refinement is accompanied by an increase in computational demands. In conventional greedy algorithm, the exploration ends up with one local optima, while as the multiple starting-point strategy aims to visit the neighborhood of all the modes, the number of model configurations it explore increases dramatically, especially as the number of causal variants escalates, making it less suitable for large-scale studies.

Consequently, the choice of algorithm ultimately becomes a trade-off between computational efficiency and accuracy. Note that DAP-MS is a better solution than greedy algorithms in terms of lower false discovery rate only with the presence of multimodal posteriors, while in practice there is no straightforward way to examine whether multimodality exists without running and comparing results from both algorithms. Considering the growing size of the data sets being investigated, methods like SuSiE or DAP-G may still be the optimal choices for fine-mapping tasks. Nevertheless, it's important for researchers to stay aware of the potential limitations of these methods.

3.4 Methods

Owing to the restrictions on sharing individual-level data, statistical fine-mapping methods have been adapted to utilize only GWAS summary data in conjunction with an LD estimate. In theory, given sufficient statistics — including LD estimates derived directly from GWAS samples — we could achieve fine-mapping results identical to those using individual-level data. However, this is often impractical because LD information typically isn't shared alongside GWAS summary statistics, necessitating estimation from a reference panel. In the first part of this study, we embarked on simulation studies using TOPMed genotype data to assess the influence of LD panel choice on the power and coverage of fine-mapping.

In the second part of our study, we examined the limitations of statistical fine-mapping, even when individual-level data is accessible. While prior studies, such as [28], have noted that the greedy algorithms used in most fine-mapping methods can be trapped in local optima, leading to false positives, this issue have not been extensively explored. We delved into scenarios where a non-causal variant correlates with multiple causal variants and assessed the occurence and influence of multimodal posterior distribution in fine-mapping. Through simulations using GTEx data, we evaluated the improvements brought by a multi-starting-point strategy which is expected to reduce the false signals typically attributed to the greedy algorithm.

3.4.1 Simulations on TOPMed data

TOPMed [19] is a research program aiming to advance precision medicine for heart, lung and blood traits through the integration of whole-genome sequencing (WGS) and other omics data with high-quality epidemiological data in ongoing studies of these traits.

In our study, we identified 3,115 individuals of European descent (EUR) within the TOPMed whole blood samples. This was accomplished through ADMIXTURE, using a threshold of 80% genomic content as a criterion for inclusion [1]. From this pool, we randomly selected 500 individuals to form a GWAS panel. The remaining individuals were utilized to construct two LD panels of different sizes – one comprising 500 individuals and the other, more comprehensive panel, consisting of 2,500 individuals.

For this analysis, we randomly selected 1,000 genes. For each gene, we chose 5,000 variants from its cis-region (within a 1Mb window of the transcription start site) that had a minor allele frequency (MAF) greater than 1%. Among these variants, we arbitrarily designated three as causal variants, and simulated phenotypes under varying conditions, specifically different proportions of variance explained by the genotype (PVE = 0.05, 0.1, 0.2, 0.4).

We carried out fine-mapping on each of the simulated datasets using DAP-G [11]. It's important to note that when fine-mapping with summary statistics, the input data for DAP-G includes only effect sizes and the corresponding standard deviations from single-variant association tests, as well as LD data, which do not constitute sufficient statistics. This remains the case even when the LD data is derived in-sample from the GWAS panel. The GWAS summary statistics were derived from simple linear regression analyses using individual-level genotype data and the simulated phenotype data in the GWAS panel.

3.4.2 Simulations on GTEx V8 data

The GTEx V8 dataset [9, 12] presents an extensive collection of WGS genotype data and RNA-seq expression data from 838 donors covering 54 human tissue types, including 670 donors with whole blood samples. Both RNA-seq and genotype data underwent pre-processing in accordance with the

protocols of GTEx data processing. We adjusted normalized gene expression levels using the same set of covariates as those applied in GTEx v8 single-SNP Expression Quantitative Loci (eQTL) mapping [8].

In our study, we randomly selected 1000 genomic regions from the autosomes within the GTEx dataset, each containing 1000 variants. Using a similar strategy to the one detailed in the SuSiE paper [23], we simulated synthetic expression levels for each genomic region under varying settings of the number of causal variants (denoted by S) and the proportion of variance explained by genotypes (denoted by ϕ). Initially, for each genomic region, we sampled S causal variants from the 1000 variants under analysis. Subsequently, we assigned an effect size drawn from the normal distribution $N(0, 0.6^2)$ to each causal variant and set the effect sizes for the remaining variants to zero. The expression levels y were simulated from $N(Xb, \sigma^2)$, where X represents the $n \times p$ matrix of genotypes, b is a p-vector of effect sizes, and $\sigma^2 = \frac{1-\phi}{\phi} Var(Xb)$. This approach resulted in a total of $5 \times 4 \times 1000 = 20,000$ datasets, from all pairwise combinations of $S \in \{1, 2, \cdots, 5\}$ and $\phi \in \{0.05, 0.1, 0.2, 0.4\}$.

We conducted a fine-mapping analysis on each simulated dataset using existing methods such as DAP-G, SuSiE, and our new implementation, DAP-MS (Section C.2), which employs the multiple starting-point strategy. We evaluated the performance of these methods based on: 1) the calibration of variant-wise posterior inclusion probabilities; 2) cluster-wise coverage and power; and 3) computational efficiency.

3.5 Tables and Figures

Table 3.1: Power of Fine-Mapping using Individual-level Data and Summary Statistics

PVE	Individual-level data	Summary statistics		
		In-sample LD	Out-of-sample LD	Out-of-sample LD
	(n=500)	(n=500)	(n=2500)	(n=500)
0.05	10.8%	11.5%	11.6%	11.7%
0.10	32.1%	31.9%	32.4%	32.5%
0.20	49.4%	47.2%	46.9%	47.1%
0.40	63.1%	56.0%	56.5%	56.6%

Table 3.2: Coverage of Fine-Mapping using Individual-level Data and Summary Statistics

PVE	Individual-level data	Summary statistics		
		In-sample LD	Out-of-sample LD	Out-of-sample LD
	(n=500)	(n=500)	(n=2500)	(n=500)
0.05	97.1%	96.7%	96.4%	96.7%
0.10	95.4%	95.2%	95.1%	93.9%
0.20	94.9%	94.6%	92.2%	89.7%
0.40	94.8%	89.1%	81.6%	68.3%

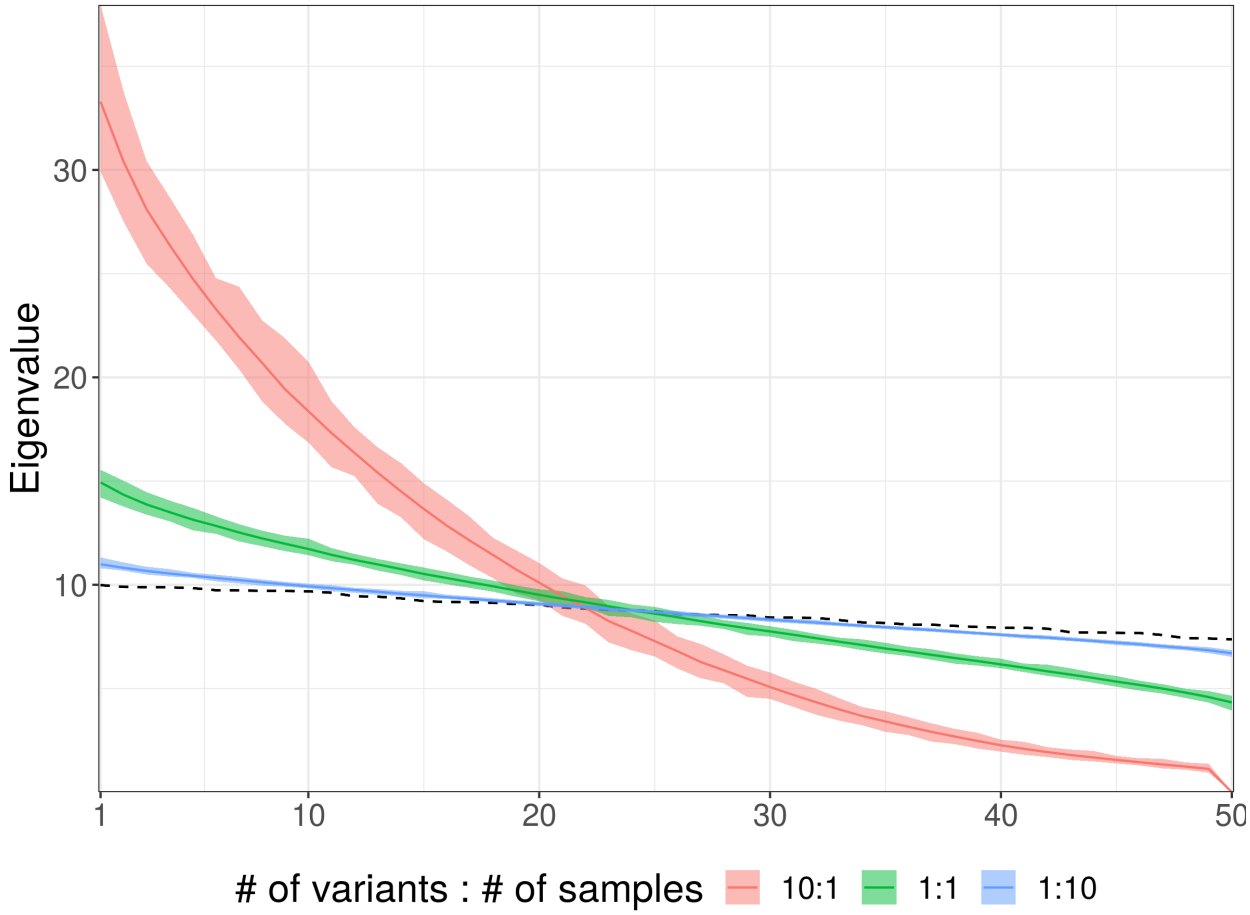
Table 3.3: Comparisons of Cross Entropy across different Methods

# of Causal Variants	SuSiE	DAP-G	DAP-MS
1	0.0043	0.0024	0.0025
2	0.0119	0.0094	0.0088
3	0.0197	0.0178	0.0165
4	0.0271	0.0269	0.0249
5	0.0343	0.0361	0.0335

Table 3.4: Comparisons of Computational Cost across different Methods

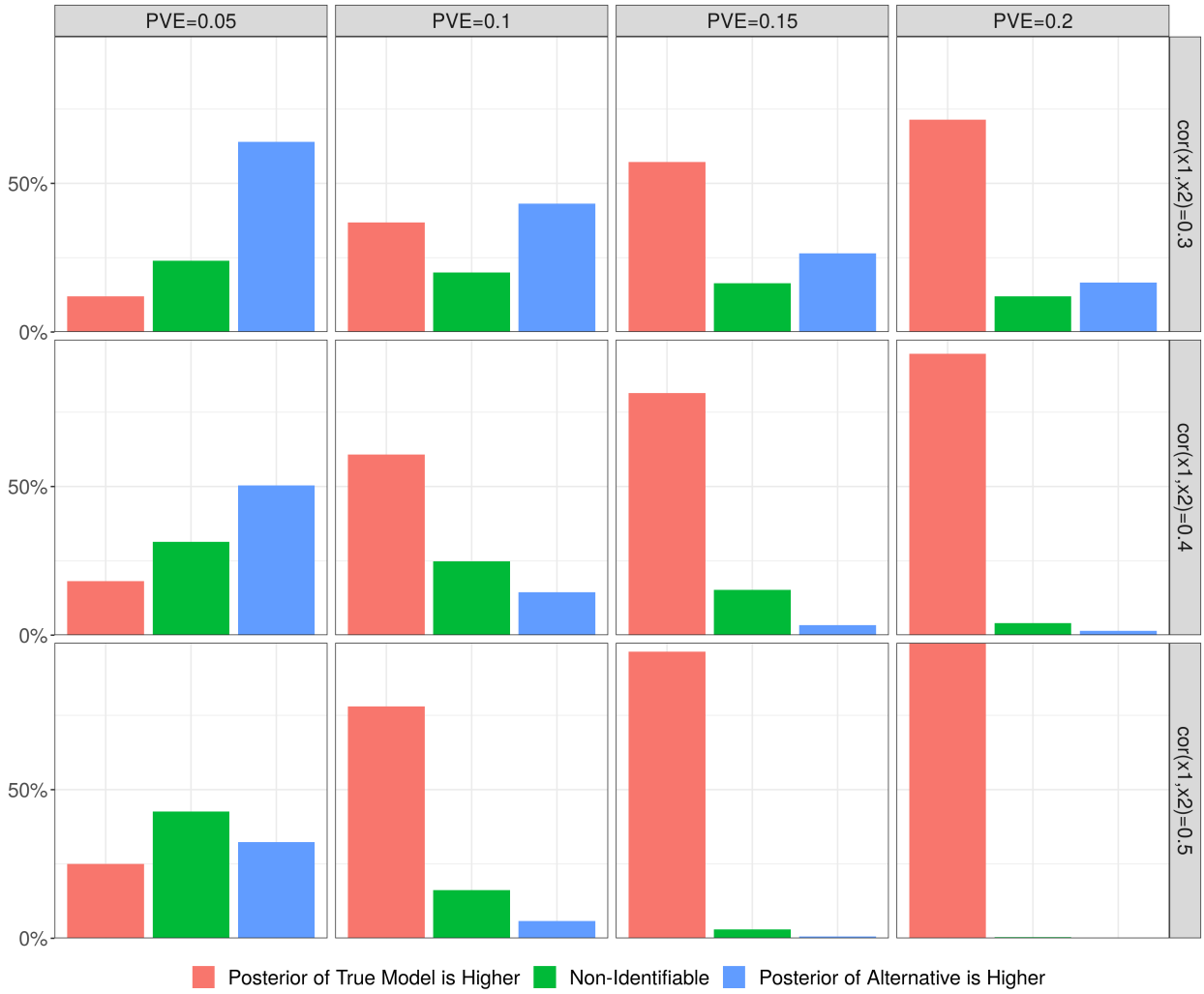
# of Causal Variants	SuSiE	DAP-G	DAP-MS
1	26:42	5:23	23:33
2	25:54	12:32	45:29
3	28:17	43:04.10	3:06:20
4	23:16	1:00:12	3:37:54
5	29:33	1:12:12	3:47:38

Figure 3.1: Comparison of Eigenvalues of sample LD and population LD



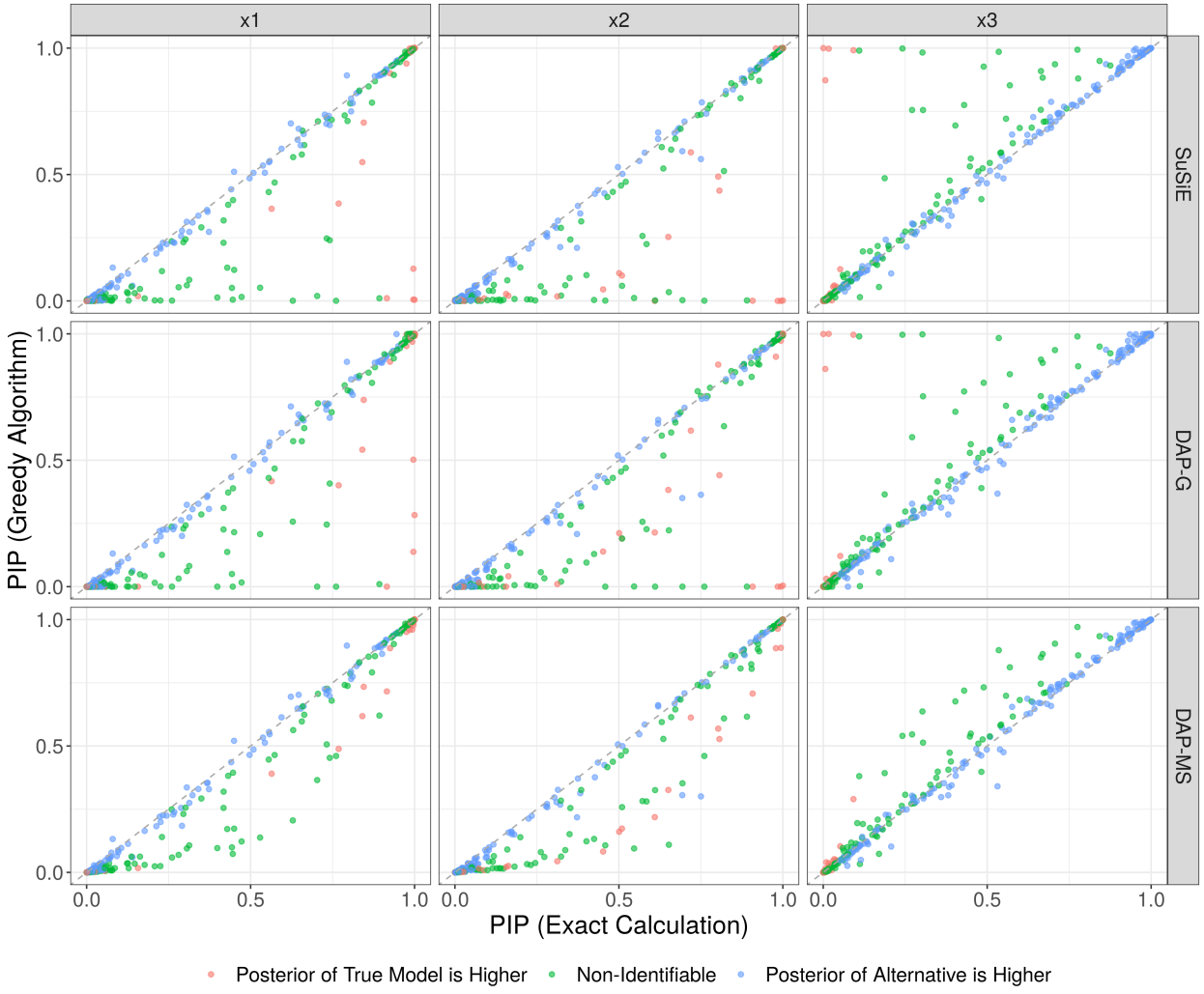
We simulated a blockwise LD structure for 500 variants, each block containing 10 variants with pairwise correlation drawn from uniform distribution from 0.7 to 1, and generated 100 genotype datasets based on the simulated LD structure with sample size $n = 50, 500, 5000$, respectively. We compared the first 50 eigenvalues from the population LD matrix (dashed line) and sample LD matrix. The solid line represents the mean of the eigenvalues of LD matrix calculated from the 100 datasets, the upper bound represents the maximum value of the eigenvalues and the lower bound represents the minimum value.

Figure 3.2: Proportion of Multimodal Cases in Simulation Studies



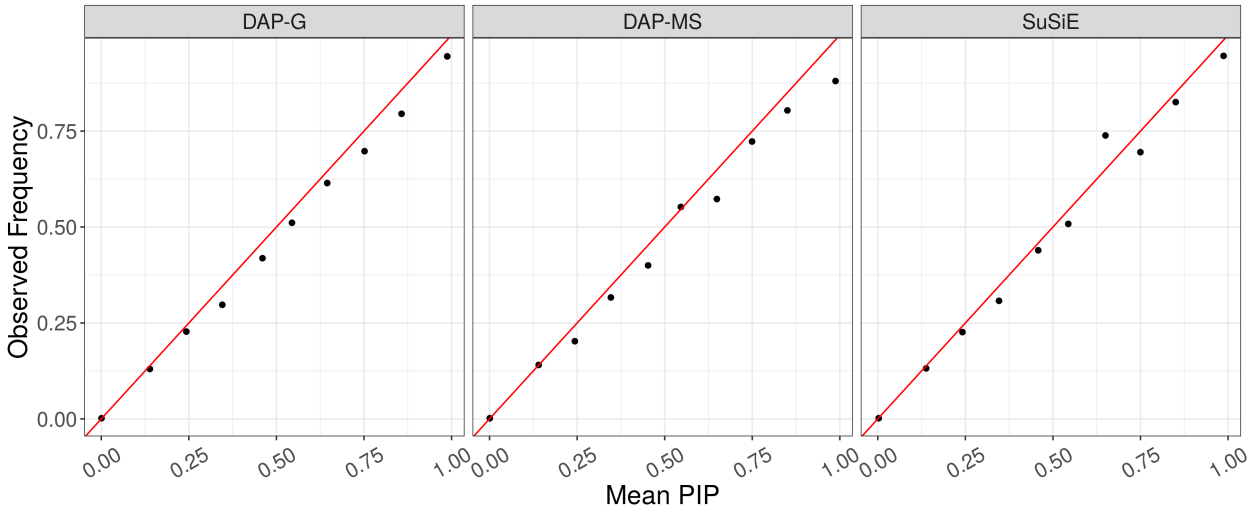
The phenotype in this analysis is associated with five variants. Among these, two causal variants, x_1 and x_2, are correlated with each other, and a non-causal variant, x_3, shares a correlation with both. The remaining three causal variants are statistically independent from x_1, x_2, and x_3. Under each combination of settings – PVE values of 0.05, 0.1, 0.15, 0.2, and correlation coefficients for x_1 and x_2 of 0.3, 0.4, and 0.5 – we simulated 500 datasets. Subsequently, we enumerated all 2^6 model configurations to yield the exact posterior probabilities. We compared the posterior probabilities of the true model $\gamma = (1, 1, 0, 1, 1, 1)^T$, with those of the alternative model $\gamma' = (0, 0, 1, 1, 1, 1)^T$. For each setting, we summarized the proportion of cases deemed multimodal, defined as $\left|\log \frac{P(\gamma'|X,y)}{P(\gamma|X,y)}\right| < 1$.

Figure 3.3: Comparisons between PIPs from Greedy Algorithms and Exact Calculations



Two causal variants, x_1 and x_2, are correlated with a non-causal variant x_3. We determined the exact PIPs by enumerating all possible model configurations and then compared these exact PIPs with those derived from SuSiE, DAP-G, and DAP-MS, respectively. A data point beneath the diagonal line denotes an underestimated PIP in comparison to the exact calculation, whereas a point above the diagonal line signals an overestimated PIP. Large deviations from the exact calculation indicate that that the greedy algorithm may have overlooked some model configurations that carry unignorable posteriors during the search process.

Figure 3.4: Assessment of PIP calibration



Variants across all simulations were classified into bins based on their reported PIP, using ten bins of equal spacing, ranging from 0 to 1. The graphs display the average PIP for each bin against the proportion of causal variants within that respective bin. A method demonstrating good calibration should yield data points in close proximity to the diagonal red line.

3.6 References

[1] David H Alexander, John Novembre, and Kenneth Lange. Fast model-based estimation of ancestry in unrelated individuals. *Genome research*, 19(9):1655–1664, 2009.

[2] Christian Benner, Chris CA Spencer, Aki S Havulinna, Veikko Salomaa, Samuli Ripatti, and Matti Pirinen. Finemap: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics*, 32(10):1493–1501, 2016.

[3] Wenan Chen, Beth R Larrabee, Inna G Ovsyannikova, Richard B Kennedy, Iana H Haralambieva, Gregory A Poland, and Daniel J Schaid. Fine mapping causal variants with an approximate bayesian method using marginal test statistics. *Genetics*, 200(3):719–736, 2015.

[4] Diana Cole. *Parameter redundancy and identifiability*. CRC Press, 2020.

[5] GTEx Consortium, Kristin G Ardlie, David S Deluca, Ayellet V Segrè, Timothy J Sullivan, Taylor R Young, Ellen T Gelfand, Casandra A Trowbridge, Julian B Maller, Taru Tukiainen, et al. The genotype-tissue expression (gtex) pilot analysis: multitissue gene regulation in humans. *Science*, 348(6235):648–660, 2015.

[6] Jianqing Fan and Jinchi Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911, 2008.

[7] Lars G Fritsche, Wilmar Igl, Jessica N Cooke Bailey, Felix Grassmann, Sebanti Sengupta, Jennifer L Bragg-Gresham, Kathryn P Burdon, Scott J Hebbring, Cindy Wen, Mathias Gorski, et al. A large genome-wide association study of age-related macular degeneration highlights contributions of rare and common variants. *Nature genetics*, 48(2):134–143, 2016.

[8] Eric R Gamazon, Ayellet V Segrè, Martijn Van De Bunt, Xiaoquan Wen, Hualin S Xi, Farhad Hormozdiari, Halit Ongen, Anuar Konkashbaev, Eske M Derks, François Aguet, et al. Using an atlas of gene regulation across 44 human tissues to inform complex disease-and trait-associated variation. *Nature genetics*, 50(7):956–967, 2018.

[9] Eric R Gamazon, Heather E Wheeler, Kaanan P Shah, Sahar V Mozaffari, Keston Aquino-Michaels, Robert J Carroll, Anne E Eyler, Joshua C Denny, GTEx Consortium, Dan L Nicolae, et al. A gene-based association method for mapping traits using reference transcriptome data. *Nature genetics*, 47(9):1091–1098, 2015.

[10] Farhad Hormozdiari, Emrah Kostem, Eun Yong Kang, Bogdan Pasaniuc, and Eleazar Eskin. Identifying causal variants at loci with multiple signals of association. In *Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, pages 610–611, 2014.

[11] Yeji Lee, Francesca Luca, Roger Pique-Regi, and Xiaoquan Wen. Bayesian multi-snp genetic association analysis: control of fdr and use of summary statistics. *BioRxiv*, page 316471, 2018.

[12] John Lonsdale, Jeffrey Thomas, Mike Salvatore, Rebecca Phillips, Edmund Lo, Saboor Shad, Richard Hasz, Gary Walters, Fernando Garcia, Nancy Young, et al. The genotype-tissue expression (gtex) project. *Nature genetics*, 45(6):580–585, 2013.

[13] Julian B Maller, Gilean McVean, Jake Byrnes, Damjan Vukcevic, Kimmo Palin, Zhan Su, Joanna MM Howson, Adam Auton, Simon Myers, Andrew Morris, et al. Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nature genetics*, 44(12):1294–1301, 2012.

[14] Andreas Raue, Clemens Kreutz, Thomas Maiwald, Julie Bachmann, Marcel Schilling, Ursula Klingmüller, and Jens Timmer. Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics*, 25(15):1923–1929, 2009.

[15] Angelo Scuteri, Serena Sanna, Wei-Min Chen, Manuela Uda, Giuseppe Albai, James Strait, Samer Najjar, Ramaiah Nagaraja, Marco Orrú, Gianluca Usala, et al. Genome-wide association scan shows genetic variants in the fto gene are associated with obesity-related traits. *PLoS genetics*, 3(7):e115, 2007.

[16] Suyash S Shringarpure and Carlos D Bustamante. Privacy risks from genomic data-sharing beacons. *The American Journal of Human Genetics*, 97(5):631–646, 2015.

[17] Eli A Stahl, Gerome Breen, Andreas J Forstner, Andrew McQuillin, Stephan Ripke, Vassily Trubetskoy, Manuel Mattheisen, Yunpeng Wang, Jonathan RI Coleman, Héléna A Gaspar, et al. Genome-wide association study identifies 30 loci associated with bipolar disorder. *Nature genetics*, 51(5):793–803, 2019.

[18] Matthew Stephens and David J Balding. Bayesian statistical methods for genetic association studies. *Nature Reviews Genetics*, 10(10):681–690, 2009.

[19] Daniel Taliun, Daniel N Harris, Michael D Kessler, Jedidiah Carlson, Zachary A Szpiech, Raul Torres, Sarah A Gagliano Taliun, André Corvelo, Stephanie M Gogarten, Hyun Min Kang, et al. Sequencing of 53,831 diverse genomes from the nhlbi topmed program. *Nature*, 590(7845):290–299, 2021.

[20] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[21] Peter M Visscher, Naomi R Wray, Qian Zhang, Pamela Sklar, Mark I McCarthy, Matthew A Brown, and Jian Yang. 10 years of gwas discovery: biology, function, and translation. *The American Journal of Human Genetics*, 101(1):5–22, 2017.

[22] Jon Wakefield. Bayes factors for genome-wide association studies: comparison with p-values. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society*, 33(1):79–86, 2009.

[23] Gao Wang, Abhishek Sarkar, Peter Carbonetto, and Matthew Stephens. A simple new approach to variable selection in regression, with application to genetic fine mapping. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(5):1273–1300, 2020.

[24] Xiaoquan Wen, Yeji Lee, Francesca Luca, and Roger Pique-Regi. Efficient integrative multi-snp association analysis via deterministic approximation of posteriors. *The American Journal of Human Genetics*, 98(6):1114–1129, 2016.

[25] Xiaoquan Wen and Matthew Stephens. Using linear predictors to impute allele frequencies from summary or pooled genotype data. *The annals of applied statistics*, 4(3):1158, 2010.

[26] Jian Yang, Teresa Ferreira, Andrew P Morris, Sarah E Medland, Genetic Investigation of ANthropometric Traits (GIANT) Consortium, DIAbetes Genetics Replication, Meta analysis (DIAGRAM) Consortium, Pamela AF Madden, Andrew C Heath, Nicholas G Martin, Grant W Montgomery, et al. Conditional and joint multiple-snp analysis of gwas summary statistics identifies additional variants influencing complex traits. *Nature genetics*, 44(4):369–375, 2012.

[27] Xiang Zhu and Matthew Stephens. Bayesian large-scale multiple regression with summary statistics from genome-wide association studies. *The annals of applied statistics*, 11(3):1561, 2017.

[28] Yuxin Zou, Peter Carbonetto, Gao Wang, and Matthew Stephens. Fine-mapping from summary data with the "sum of single effects" model. *PLoS Genetics*, 18(7):e1010299, 2022.

CHAPTER 4

Gene Expression Imputation Analysis on TOPMed RNAseq data

4.1 Introduction

Transcriptome-wide association studies (TWAS) have emerged as a powerful tool for causal-gene prioritization [37], a crucial yet challenging process that seeks to unravel the mechanisms through which the genetic variants, identified via genome-wide association studies (GWAS), contribute to the trait. By integrating transcriptomic datasets with GWAS, TWAS prioritize candidate genes at GWAS loci by testing the mediating effects of gene expression levels on the trait of interest and thus provide mechanistic insights of gene regulation underlying complex traits and diseases. TWAS have detected new susceptibility genes and highlighted potential regulatory targets for schizophrenia [16], autism spectrum disorder [33] and Crohn's disease [7] among many others [25].

A conventional TWAS framework consists of two stages – first, the genetically regulated expression (GReX) is estimated for each individual in the GWAS cohort; second, the association tests are performed between GReX and the trait of interest. The first step, often referred to as gene expression imputation, involves constructing prediction models for gene expression based on a reference dataset that contains both genotype and transcriptomic data. These prediction models are then utilized to estimate the GReX. Therefore, the reliability of TWAS results heavily depends on the quality of gene expression imputation and the reference datasets that enable it.

The estimation of GReX shares common ground with the polygenic modeling of complex dis-

eases and traits, or Polygenic Risk Score (PRS) calculation [20], as they both involve phenotype prediction by jointly modeling genotypes across multiple genetic variants, potentially spanning the entire genome. Therefore, many PRS techniques might be suitable for gene expression imputation. For instance, the Bayesian Sparse Linear Mixed Model (BSLMM) [45] is incorporated in the efficient TWAS tool FUSION [15], and the Latent Dirichlet Process Regression (DPR) [41] is utilized in TIGAR [28, 29]. However, these methods require access to individual-level data and lack scalability to data with millions of genetic variants. Hence, when applied for gene expression imputation, they include only genotypes in the cis-region of the target gene. In contrast, PRS methods using summary statistics including LDpred [36, 30], SBayesR [21], and MegaPRS [42], consider genetic variants genome-wide, offering the potential for a more comprehensive prediction of gene expression. In this study, we assess the performance of PRS methods using individual-level data in the cis-region and methods using summary statistics across the genome within the context of gene expression imputation.

Most gene expression imputation models have been trained using data from the Genotype-Tissue Expression (GTEx) project [5, 23], which predominantly consists of subjects of European descent. However, many genetic studies include samples from multi-ethnic populations. Here we explore the potential of utilizing the Trans-Omics for Precision Medicine (TOPMed) [34] dataset, distinguished by its larger sample size and greater population diversity, as a reference panel for gene expression imputation. We conducted comparative analyses on both European and African samples, using models trained from varying reference sample sizes and degrees of ancestry alignment. Furthermore, we assessed the performance of whole-blood expression models, trained using TOPMed data, when applied to GTEx samples.

4.2 Methods

4.2.1 TOPMed RNA-sequencing Data

TOPMed [34] is a research program aiming to advance precision medicine for heart, lung and blood traits through the integration of whole-genome sequencing (WGS) and other omics data with high-quality epidemiological data in ongoing studies of these traits .

This analysis included 9264 unrelated participants from 8 TOPMed studies with both WGS data and transcriptomics data (Table D.1): Genetic Epidemiology of COPD Study (COPDGene, n=382) [31], Framingham Heart Study (FHS, n=793) [2], Genes-Environments and Admixture in Latino Americans (GALA II, n=1897) [11], Lung Tissue Research Consortium (LTRC, n=1360) [35], Multi-Ethnic Study of Atherosclerosis (MESA, n=1271) [3], Study of African Americans, Asthma, Genes and Environments (SAGE, n=705) [4], SubPopulations and InteRmediate Outcome Measures In COPD Study (SPIROMICS, n=1578) [6], Women's Health Intiative (WHI, n=1279) [17].

RNA libraries were prepared using the Illumina TruSeqTM stranded mRNA kit, sequenced for a target depth of 75 million reads per 2×101 bp paired-end reads in whole blood samples, and 40 million reads per 2×101 bp paired-end reads in other tissues. Alignment and quality control were conducted via the TOPMed RNA-seq pipeline [39]. Briefly, RNA-seq reads were aligned to GRCh38 reference genome with STAR [8] and gene-level expression was quantified with RNA-SeQC 2 [14] based on GENCODE 34 annotations.

WGS was generated to an average depth of 38x by seven sequencing centers (Broad Institute of MIT and Harvard, Northwest Genomics Center, New York Genome Center, Illumina Genomic Services, Macrogen, Baylor College of Medicine Human Genome Sequencing Center, and McDonnell Genome Institute at Washington University). TOPMed freeze 9b [24] includes WGS data with joint genotype calling and variant-level quality control for about 158,000 samples across more than 80 studies. We matched the RNA samples with WGS samples by comparing genotypes at variants with MAF$\geq$5% in coding exons and excluded samples without a match. Also, we esti-

mate subject based on autosomal SNPs with MAF$\geq$1% relatedness using KING v.2.2.7 [26], and included only unrelated samples in the study. SNPs and short indels (<50 bp) that passed quality control and exhibited MAF$\geq$1% in the selected unrelated participants were used for analysis.

Gene counts were filtered to include only autosomal and chromosome X genes, and TMM-normalized using edgeR [32], followed by inverse normal transformation. Genes with low expression levels were excluded from the analysis. To control for potential confounders, we adjusted the normalized gene expression data for covariates including cohort, inferred sex, 15 genotype principal components (PCs), and a varying number of gene expression PCs depending on the tissue: 30 for peripheral blood mononuclear cells (PBMCs), nasal epithelial, T cells, and monocytes; 75 for lung; and 100 for whole blood.

4.2.2 Building Gene Expression Imputation Models

We evaluated four gene expression imputation models, each being a key component of well-established TWAS tools: the elastic net model [47] in PrediXcan [13], BSLMM [45] in FUSION [15], DPR [41] in TIGAR [28, 29], and DAP-G [38, 19] in the probabilistic TWAS framework [43].

These models estimate the effect sizes of cis-SNPs by modeling the relationship between expression level of genes and genotypes in the cis-region based on a linear model (Equation 4.1) in a transcriptomic reference data set.

$$y = X\beta + \epsilon \tag{4.1}$$

where y is an n-vector of normalized gene expression level for n samples, X is an $n \times p$ matrix of genotypes for p cis-SNPs, β is a p-vector of effect sizes of the cis-SNPs, ϵ is an n-vector of error terms following a multivariate normal distribution. The effect sizes β are estimated based on different assumptions depending on the method (Table 4.1), and are subsequently used as weights to calculate GReX, which is derived as a weighted average of the genotypes across the cis-SNPs.

The elastic net is a regularized linear regression that combines L1 penalty and L2 penalty, which

Method	Priori Assumption
Elastic Net	$\pi(\beta) \propto \exp\{\alpha\|\beta\|_1 + (1-\alpha)\|\beta\|_2\}$
BSLMM	$\beta \sim N(0, \sigma_a^2 + \sigma_b^2) + (1-\pi)N(0, \sigma_b^2)$
DPR	$\beta \sim \sum_{i=1}^{\infty} \pi_i N(0, \sigma_i^2)$
DAP-G	$\beta \sim \pi N(0, \sigma^2) + (1-\pi)\delta_0$

Table 4.1: Comparison of priori assumptions on the effect size in different methods.

is equivalent to a mixture of Gaussian and Laplace prior on the effect size [47]. The cv.glmnet function from the R package glmnet [9] was used to fit the elastic net models, with the mixture parameter fixed at $\alpha = 0.5$ and the penalty parameter chosen by 5-fold cross validation.

BSLMM is a combination of Bayesian variable selection (BVS) model and linear mixed model [45]. It assumes that all SNPs have non-zero effects, while a small proportion of them have additional effects and follows a mixture of two normal distributions. GEMMA [46] was used to fit BSLMM using Markov chain Monte Carlo (option '-bslmm 1') with default settings.

DPR is a more generalized model that includes BLSMM as a special case [41]. It assumes a non-parametric prior on the effect size, which is equivalent to a mixture of infinitely many normal distributions. The model was fitted using the variational Bayesian algorithm (option '-dpr 1') with default settings.

DAP-G is a BVS method specially designed for fine-mapping [19, 38] and later extended for robust effect size estimation through Bayesian model averaging [43]. It uses an indicator variable $\gamma_j = I(\beta_j \neq 0)$ to indicate the association status of the jth SNP. We refer to the indicator vector $\gamma = (\gamma_1, \gamma_2, \cdots, \gamma_p)'$ as model configuration. DAP-G provides a cost-efficient implementation for estimating the posterior probability of all plausible model configurations. SNP-level posterior inclusion probability can be obtained for fine-mapping purposes by marginalizing the posterior model probabilities. The effect sizes can be obtained as a weighted average of model-specific effect estimates. Given posterior model probabilities, model-specific effect estimates are weighted by the probability of the corresponding model.

The models mentioned above employ individual-level genotype data at cis-SNPs (within a 1Mb window of the transcription start site) to impute the expression level of the target gene. To assess

if incorporating information from trans-SNPs could enhance imputation accuracy, we also include megaPRS [42] in comparison. The megaPRS software implements methods with four different priors (LASSO [44], Ridge [18], BOLT-LMM [22], BayesR [27]) for genetic prediction of complex traits using genome-wide summary statistics. It performs cross-validation to determine optimal model parameters of the prior distribution. The input for megaPRS comprised GWAS summary statistics and the LD matrix, calculated from the individual-level data within the reference dataset. Subsequently, the effect size estimates derived from this optimal model served as weights for imputing gene expression.

4.2.3 Evaluation on GTEx data

The GTEx V8 dataset [13, 23] presents an extensive collection of WGS genotype data and RNA-seq expression data from 838 donors covering 54 human tissue types. In this study, we focused on 670 whole blood samples to assess the accuracy of gene expression imputation models. Both RNA-seq and genotype data underwent pre-processing in accordance with the protocols of GTEx data processing. We adjusted normalized gene expression levels using the same set of covariates as those applied in GTEx v8 single-SNP eQTL mapping [12]. Additionally, we imputed the genotype data with the TOPMed release 2 reference panel [34].

To evaluate the prediction accuracy among different gene expression imputation models (ElasticNet, BSLMM, DPR, DAP-G and megaPRS), we randomly selected 1000 autosomal genes with whole-blood expression levels measured in both TOPMed and GTEx data. Models were trained on 3000 European samples in TOPMed and tested on 670 samples in GTEx. The imputation accuracy of the models was quantified by the squared Spearman's Rank correlation coefficient (r^2) between the imputed expression level and the measured expression level. We compared the mean r^2, median r^2 and the proportion of well-imputed genes among methods, and performed paired t-tests to see if there is a significant difference in mean r^2 between methods.

4.2.4 Evaluation of Factors Affecting Imputation Accuracy

Two main factors that we aim to investigate regarding gene expression imputation accuracy include 1) the sample size of the reference dataset, and 2) the degree of ancestry matching between the reference and target samples.

Using a threshold of 80% genome content in ADMIXTURE analysis [1], we identified 3,115 European (EUR) individuals and 709 African (AFR) individuals within the TOPMed whole blood samples. The ancestral composition of all 6602 whole blood samples are displayed in Figure D.1. We designated 115 EUR samples and 109 AFR samples as test samples, while the remaining 3000 EUR samples and 600 AFR samples were used to construct reference panels of varying sizes and ancestral compositions. The DAP-G method was employed to analyze each reference dataset, with the derived effect sizes serving as weights for gene expression imputation.

For assessing the value of reference sample size, we maintained a fixed AFR to EUR ratio in the reference panel at 1:5 and then compared the imputation accuracy across models trained on reference panels of different sample sizes (n=600, 1200, 1800, 3600). To evaluate the effect of the ancestral composition of the reference panel, we kept the sample size constant at n=600, and compared the imputation accuracy among models trained on the following reference panels: 600 AFR, 300 AFR + 300 EUR, 200 AFR + 400 EUR, 100 AFR + 500 EUR, and 600 EUR.

4.3 Results

4.3.1 Comparisons of Imputation Models

To evaluate different strategies for building gene expression imputation models using the TOPMed data, we first compared the accuracy of various models, including DAP-G, Elastic Net, DPR, BSLMM, and megaPRS, by comparing their predictions when appied to GTEx with actual gene expression levels. These models were trained on 3,000 TOPMed European individuals with whole-blood gene expression data and tested on 670 GTEx samples. The accuracy of gene expression

imputation was evaluated using the squared Spearman's correlation (r^2) between the measured normalized expression levels and the imputed gene expression levels.

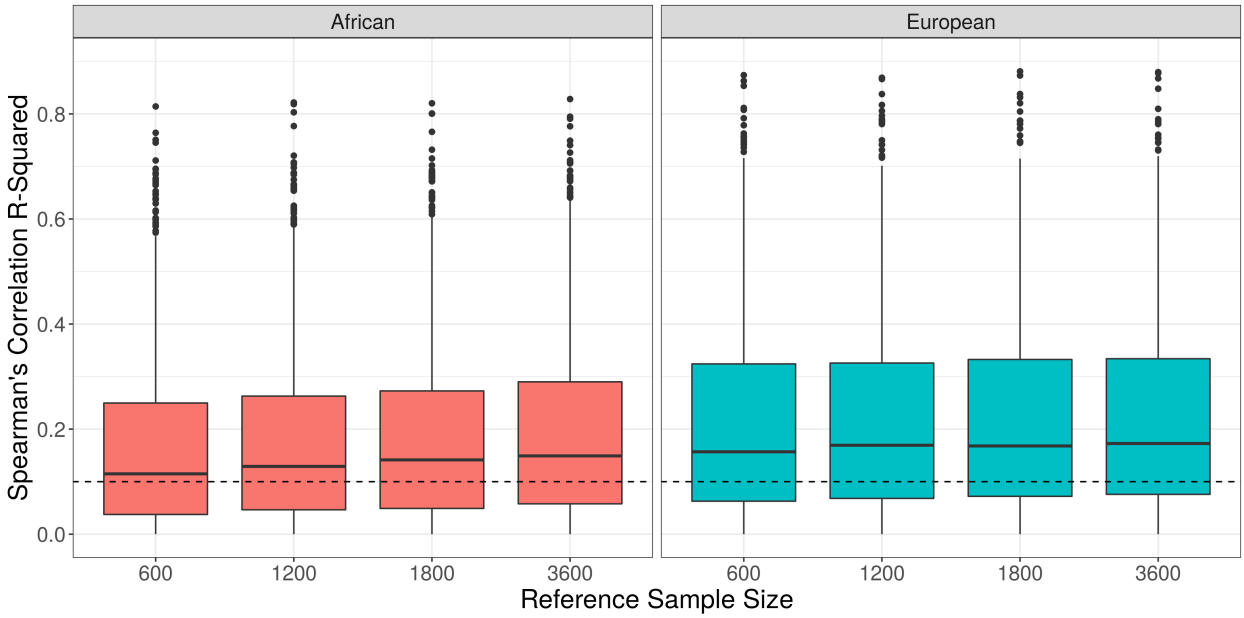
Our evaluations were performed on 1,000 autosomal genes. However, some methods yielded small effect sizes (weights for gene expression imputation) across all variants from certain genes, with absolute values less than 10^{-4}, and thus were unable to generate effective imputation models for those genes. The number of effective imputation models generated by DAP-G, Elastic Net, DPR, BSLMM, and MegaPRS were 970, 909, 981, 987, and 920, respectively, with 815 genes shared across all methods. We excluded one gene (DND1P1) for which the Spearman's correlation was less than -0.5 for all methods (Figure D.2). Our comparison of the performance of the five methods was thus conducted across the remaining 814 genes.

Our comparison revealed that DAP-G was the top model (with highest r) for 41.6% genes, compared with 24.9% for Elastic Net, 18.1% for DPR, 8.0% for BSLMM and 7.3% for megaPRS (Figure D.2). The comparison is detailed in Table 4.2. DAP-G outperformed the other four methods, demonstrating the highest mean r^2 of 0.133 and the highest median r^2 of 0.088. Furthermore, DAP-G achieved the highest percentage of well-imputed genes with 44.3% of tested genes achieving $r^2 > 0.1$. As we increased the r^2 threshold, the number of well imputed genes decreased for all methods, reaching 4.67% when we set the threshold at 0.4. Elastic Net performed similarly, although it presented slightly lower r^2 values (paired t-test $p = 1.3 \times 10^{-9}$).

Method	Mean r^2	Median r^2	% $r^2 > 0.1$	> 0.2	> 0.3	> 0.4
DAP-G	0.133	0.088	44.3%	21.6%	11.8%	4.67%
Elastic Net	0.130	0.085	43.4%	20.8%	10.9%	4.55%
DPR	0.127	0.078	41.5%	20.4%	10.8%	4.42%
BSLMM	0.120	0.073	38.9%	19.2%	9.58%	3.93%
megaPRS	0.113	0.069	37.3%	18.1%	8.23%	2.95%

Table 4.2: Comparison of Spearman's r^2 across different methods.

Figure 4.1: Comparison of Spearman's r^2 using Reference Panels of Different Sizes



4.3.2 Impacts of Sample Size on Imputation Accuracy

To assess the influence of sample size on gene expression imputation accuracy, we performed a comparison using weights derived from DAP-G based on reference panels of varying sizes ($n = 600, 1200, 1800, 3600$), maintaining a fixed AFR to EUR ratio of 1:5.

As depicted in Figure 4.1, imputation accuracy increases with the reference sample size. Notably, the increase in accuracy for AFR target samples (median r^2 increases from 0.115 to 0.149) outstrips that of EUR target samples (median r^2 increases from 0.157 to 0.173). This finding suggests that populations underrepresented in the reference panel will benefit more from an increase in sample size.

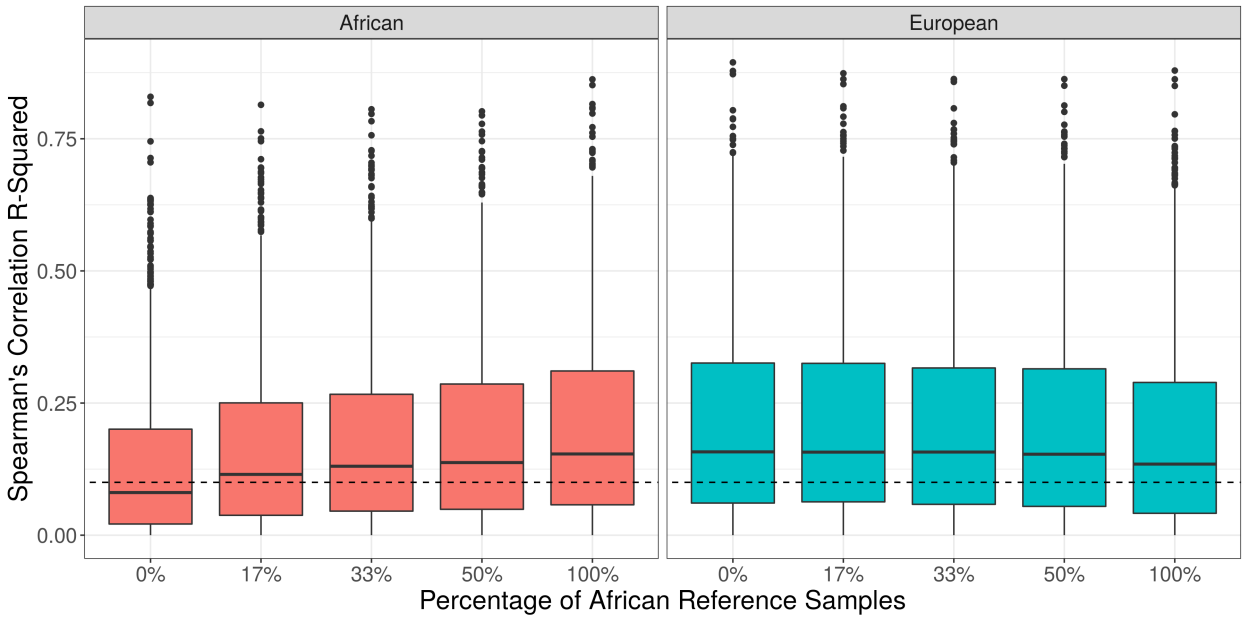
4.3.3 Impacts of Ancestry Matching on Imputation Accuracy

To evaluate the impact of ancestry matching between the reference panel and target samples on gene expression imputation accuracy, we performed a comparison using DAP-G-derived weights, based on reference panels with varying AFR-to-EUR ratios, with a fixed sample size at 600.

Figure 4.2 demonstrates that imputation accuracy escalates as the proportion of matching ancestry in the reference samples rises. Interestingly, the influence of ancestry matching appears to be more substantial on AFR target samples compared to EUR target samples. Imputing gene expression levels in AFR samples using a pure EUR reference panel yielded a median r^2 of 0.085, while the same task with EUR samples using a pure AFR reference panel resulted in a median r^2 of 0.137. Imputing with a 100% ancestry-matching reference panel yielded a median r^2 of 0.157 for AFR and 0.158 for EUR samples. This data clearly underscores the importance of ancestry matching in gene expression imputation.

Using a combined reference panel including all 600 EUR and 600 AFR ancestry samples (that is, with 1200 samples in total) yielded median imputation accuracy r^2 of 0.159 for AFR and 0.161 for EUR, which are both higher than using the ancestry-specific sub-panel. However, Figure D.3 revealed that a combined reference panel of 600 AFR and more than 600 EUR samples may have a worse performance as the proportion of AFR decreases in the reference panel. Other metrics including mean r^2 and proportion of well-imputed genes indicate that an ancestry-specific reference panel may perform better. The paired t-tests on the Spearman's correlation r showed that the difference is significant for African samples ($\Delta\bar{r} = 0.005, p = 0.003$) and yet insignificant for European samples ($\Delta\bar{r} = 0.0001, p = 0.947$).

Figure 4.2: Comparison of Spearman's r^2 using Reference Panels of Different Ancestral Compositions



4.3.4 Computational Costs

We evaluated the computational costs of expression imputation in UK Biobank samples on chromosome 20, using TOPMed-imputed genotype dosages (6,631,680 variants) and weights (557 genes, 40,541 variants) derived from TOPMed dataset using DAP-G as inputs. As demonstrated in Table 4.3, the computational time displays a linear increase with the sample size, presenting a per-sample cost of approximately 0.43 seconds. As the genotype matrix does not need to be stored while calculating the imputed expression, memory usage remains effectively managed, escalating only with the number of genes and sample size.

Sample Size	Time ([hh]:mm:ss)	Memory (MB)
1000	7:13	15.6
10,000	1:11:44	57.2
100,000	11:51:32	474.56

Table 4.3: Computational Cost of Expression Imputation using TOPMed-Imputed Genotype Data

The most time-intensive segment of the process involves reading through the genotype file; therefore, the per-sample computational time will correspondingly expand with the number of variants. Also, the computational costs of expression imputation using weights derived from alternative methods should be similar to that uing DAP-G.

The experiment was conducted on a single core of Intel® Xeon® Platinum 8268 CPU @ 2.90GHz. The C++ package for expression imputation is available on Github (https://github.com/yukt/PRScal).

4.4 Discussion

Through our systematic comparison, we concluded that DAP-G emerges as the most effective model for gene expression imputation based on the TOPMed data. This conclusion was substantiated by the fact that DAP-G was the top model for a significant proportion of genes and consistently outperformed other models in terms of the mean and median Spearman's r^2.

We also show that the accuracy of gene expression imputation improves with increased reference sample size, highlighting the benefit of a larger reference panel, especially for less represented populations. Although imputation accuracy increases only slowly with reference panel size, it increased steadily in the range we evaluated and we expect it will continue to increase as larger reference panels become available. Increasing reference panel size from 600 to 3000 individuals increased the mean correlation between estimated and actual expression values by 20 - 30%.

Our evaluations also showed that imputation accuracy is positively correlated with the degree of ancestry matching between the reference and target samples. Therefore, using a reference panel that is more diverse or that has a higher proportion of samples with matching ancestry leads to a more accurate imputation. The imputation accuracy was comparable between African and European samples when using matching reference panels. However, a notable decrease in accuracy was observed when imputing African samples using a European reference panel, in contrast to imputing European samples using an African reference panel. This asymmetry in accuracy can be attributed to the differing size of LD blocks between the two populations. Generally, European populations exhibit larger LD blocks compared to African populations, which implies that the causal variant might be correlated with more surrounding variants in European samples [10]. As discussed in the previous chapter, a complex LD structure can complicate the identification of true causal associations, and thus it becomes more challenging to identify the true causal variant using data from European samples compared to African samples. Consequently, when using an expression imputation model inferred from European samples on African samples, the presence of this confounding variant may lead to decreased accuracy. This observed asymmetry highlights the importance of carefully evaluating the cross-ancestry transferability of gene expression models by ancestry group.

Figure D.3 demonstrated that utilizing a fully African panel of 600 samples can achieve comparable accuracy on African target samples as a considerably larger mixed panel comprising 600 African and 3000 European samples. However, this also implies that with a large mixed reference panel, there is no necessity to impute gene expression using ancestry-specific subpanels. Instead,

samples across all populations can be processed together using the same mixed panel, a procedure which is notably more convenient and efficient.

Building upon our findings, we have developed DAP-G weights based on the TOPMed data, specifically designed to facilitate transcriptome-wide association studies (TWAS). We compared the imputation performance with heritability estimated from the previous study [40] (Figure D.4) and the imputation r^2 exceeds the lower bound of heritability estimate for 43.5% genes. In an effort to make these resources readily available and user-friendly, we will integrate this feature into the TOPMed imputation server. This enhancement will allow users to conveniently access both imputed gene expressions and genotypes in a unified platform.

4.5 References

[1] David H Alexander, John Novembre, and Kenneth Lange. Fast model-based estimation of ancestry in unrelated individuals. *Genome research*, 19(9):1655–1664, 2009.

[2] Charlotte Andersson, Andrew D Johnson, Emelia J Benjamin, Daniel Levy, and Ramachandran S Vasan. 70-year legacy of the framingham heart study. *Nature Reviews Cardiology*, 16(11):687–698, 2019.

[3] Diane E Bild, David A Bluemke, Gregory L Burke, Robert Detrano, Ana V Diez Roux, Aaron R Folsom, Philip Greenland, David R JacobsJr, Richard Kronmal, Kiang Liu, et al. Multi-ethnic study of atherosclerosis: objectives and design. *American journal of epidemiology*, 156(9):871–881, 2002.

[4] Luisa N Borrell, Elizabeth A Nguyen, Lindsey A Roth, Sam S Oh, Haig Tcheurekdjian, Saunak Sen, Adam Davis, Harold J Farber, Pedro C Avila, Emerita Brigino-Buenaventura, et al. Childhood obesity and asthma control in the gala ii and sage ii studies. *American journal of respiratory and critical care medicine*, 187(7):697–702, 2013.

[5] GTEx Consortium, Kristin G Ardlie, David S Deluca, Ayellet V Segrè, Timothy J Sullivan, Taylor R Young, Ellen T Gelfand, Casandra A Trowbridge, Julian B Maller, Taru Tukiainen, et al. The genotype-tissue expression (gtex) pilot analysis: multitissue gene regulation in humans. *Science*, 348(6235):648–660, 2015.

[6] David Couper, Lisa M LaVange, MeiLan Han, R Graham Barr, Eugene Bleecker, Eric A Hoffman, Richard Kanner, Eric Kleerup, Fernando J Martinez, Prescott G Woodruff, et al. Design of the subpopulations and intermediate outcomes in copd study (spiromics). *Thorax*, 69(5):492–495, 2014.

[7] Yulin Dai, Guangsheng Pei, Zhongming Zhao, and Peilin Jia. A convergent study of genetic variants associated with crohn's disease: evidence from gwas, gene expression, methylation, eqtl and twas. *Frontiers in genetics*, 10:318, 2019.

[8] Alexander Dobin, Carrie A Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R Gingeras. Star: ultrafast universal rna-seq aligner. *Bioinformatics*, 29(1):15–21, 2013.

[9] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1, 2010.

[10] Stacey B Gabriel, Stephen F Schaffner, Huy Nguyen, Jamie M Moore, Jessica Roy, Brendan Blumenstiel, John Higgins, Matthew DeFelice, Amy Lochner, Maura Faggart, et al. The structure of haplotype blocks in the human genome. *science*, 296(5576):2225–2229, 2002.

[11] Joshua M Galanter, Christopher R Gignoux, Dara G Torgerson, Lindsey A Roth, Celeste Eng, Sam S Oh, Elizabeth A Nguyen, Katherine A Drake, Scott Huntsman, Donglei Hu, et al. Genome-wide association study and admixture mapping identify different asthma-associated loci in latinos: the genes-environments & admixture in latino americans study. *Journal of Allergy and Clinical Immunology*, 134(2):295–305, 2014.

[12] Eric R Gamazon, Ayellet V Segrè, Martijn Van De Bunt, Xiaoquan Wen, Hualin S Xi, Farhad Hormozdiari, Halit Ongen, Anuar Konkashbaev, Eske M Derks, François Aguet, et al. Using an atlas of gene regulation across 44 human tissues to inform complex disease-and trait-associated variation. *Nature genetics*, 50(7):956–967, 2018.

[13] Eric R Gamazon, Heather E Wheeler, Kaanan P Shah, Sahar V Mozaffari, Keston Aquino-Michaels, Robert J Carroll, Anne E Eyler, Joshua C Denny, GTEx Consortium, Dan L Nicolae, et al. A gene-based association method for mapping traits using reference transcriptome data. *Nature genetics*, 47(9):1091–1098, 2015.

[14] Aaron Graubert, François Aguet, Arvind Ravi, Kristin G Ardlie, and Gad Getz. Rna-seqc 2: Efficient rna-seq quality control and quantification for large cohorts. *Bioinformatics*, 37(18):3048–3050, 2021.

[15] Alexander Gusev, Arthur Ko, Huwenbo Shi, Gaurav Bhatia, Wonil Chung, Brenda WJH Penninx, Rick Jansen, Eco JC De Geus, Dorret I Boomsma, Fred A Wright, et al. Integrative approaches for large-scale transcriptome-wide association studies. *Nature genetics*, 48(3):245–252, 2016.

[16] Alexander Gusev, Nicholas Mancuso, Hyejung Won, Maria Kousi, Hilary K Finucane, Yakir Reshef, Lingyun Song, Alexias Safi, Schizophrenia Working Group of the Psychiatric Genomics Consortium, Steven McCarroll, et al. Transcriptome-wide association study of schizophrenia and chromatin activity yields mechanistic disease insights. *Nature genetics*, 50(4):538–548, 2018.

[17] Jennifer Hays, Julie R Hunt, F Allan Hubbell, Garnet L Anderson, Marian Limacher, Catherine Allen, and Jacques E Rossouw. The women's health initiative recruitment methods and results. *Annals of epidemiology*, 13(9):S18–S77, 2003.

[18] Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[19] Yeji Lee, Francesca Luca, Roger Pique-Regi, and Xiaoquan Wen. Bayesian multi-snp genetic association analysis: control of fdr and use of summary statistics. *BioRxiv*, page 316471, 2018.

[20] Cathryn M Lewis and Evangelos Vassos. Polygenic risk scores: from research tools to clinical instruments. *Genome medicine*, 12(1):1–11, 2020.

[21] Luke R Lloyd-Jones, Jian Zeng, Julia Sidorenko, Loïc Yengo, Gerhard Moser, Kathryn E Kemper, Huanwei Wang, Zhili Zheng, Reedik Magi, Tõnu Esko, et al. Improved polygenic prediction by bayesian multiple regression on summary statistics. *Nature communications*, 10(1):5086, 2019.

[22] Po-Ru Loh, George Tucker, Brendan K Bulik-Sullivan, Bjarni J Vilhjalmsson, Hilary K Finucane, Rany M Salem, Daniel I Chasman, Paul M Ridker, Benjamin M Neale, Bonnie Berger, et al. Efficient bayesian mixed-model analysis increases association power in large cohorts. *Nature genetics*, 47(3):284–290, 2015.

[23] John Lonsdale, Jeffrey Thomas, Mike Salvatore, Rebecca Phillips, Edmund Lo, Saboor Shad, Richard Hasz, Gary Walters, Fernando Garcia, Nancy Young, et al. The genotype-tissue expression (gtex) project. *Nature genetics*, 45(6):580–585, 2013.

[24] National Heart Lung and Blood Institute. TOPMed Whole Genome Sequencing Methods: Freeze 9. Available at: https://topmed.nhlbi.nih.gov/topmed-whole-genome-sequencing-methods-freeze-9. Accessed: 2023-03-01.

[25] Nicholas Mancuso, Huwenbo Shi, Pagé Goddard, Gleb Kichaev, Alexander Gusev, and Bogdan Pasaniuc. Integrating gene expression with summary association statistics to identify genes associated with 30 complex traits. *The American Journal of Human Genetics*, 100(3):473–487, 2017.

[26] Ani Manichaikul, Josyf C Mychaleckyj, Stephen S Rich, Kathy Daly, Michèle Sale, and Wei-Min Chen. Robust relationship inference in genome-wide association studies. *Bioinformatics*, 26(22):2867–2873, 2010.

[27] Gerhard Moser, Sang Hong Lee, Ben J Hayes, Michael E Goddard, Naomi R Wray, and Peter M Visscher. Simultaneous discovery, estimation and prediction analysis of complex traits using a bayesian mixture model. *PLoS genetics*, 11(4):e1004969, 2015.

[28] Sini Nagpal, Xiaoran Meng, Michael P Epstein, Lam C Tsoi, Matthew Patrick, Greg Gibson, Philip L De Jager, David A Bennett, Aliza P Wingo, Thomas S Wingo, et al. Tigar: an improved bayesian tool for transcriptomic data imputation enhances gene mapping of complex traits. *The American Journal of Human Genetics*, 105(2):258–266, 2019.

[29] Randy L Parrish, Greg C Gibson, Michael P Epstein, and Jingjing Yang. Tigar-v2: Efficient twas tool with nonparametric bayesian eqtl weights of 49 tissue types from gtex v8. *Human Genetics and Genomics Advances*, 3(1):100068, 2022.

[30] Florian Privé, Julyan Arbel, and Bjarni J Vilhjálmsson. Ldpred2: better, faster, stronger. *Bioinformatics*, 36(22-23):5424–5431, 2020.

[31] Elizabeth A Regan, John E Hokanson, James R Murphy, Barry Make, David A Lynch, Terri H Beaty, Douglas Curran-Everett, Edwin K Silverman, and James D Crapo. Genetic epidemiology of copd (copdgene) study design. *COPD: Journal of Chronic Obstructive Pulmonary Disease*, 7(1):32–43, 2011.

[32] Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. edger: a bioconductor package for differential expression analysis of digital gene expression data. *bioinformatics*, 26(1):139–140, 2010.

[33] Cristina Rodriguez-Fontenla and Angel Carracedo. Utmost, a single and cross-tissue twas (transcriptome wide association study), reveals new asd (autism spectrum disorder) associated genes. *Translational psychiatry*, 11(1):256, 2021.

[34] Daniel Taliun, Daniel N Harris, Michael D Kessler, Jedidiah Carlson, Zachary A Szpiech, Raul Torres, Sarah A Gagliano Taliun, André Corvelo, Stephanie M Gogarten, Hyun Min Kang, et al. Sequencing of 53,831 diverse genomes from the nhlbi topmed program. *Nature*, 590(7845):290–299, 2021.

[35] Robert Vassallo, Paula R Walters, Jeffrey Lamont, Theodore J Kottom, Eunhee S Yi, and Andrew H Limper. Cigarette smoke promotes dendritic cell accumulation in copd; a lung tissue research consortium study. *Respiratory research*, 11:1–13, 2010.

[36] Bjarni J Vilhjálmsson, Jian Yang, Hilary K Finucane, Alexander Gusev, Sara Lindström, Stephan Ripke, Giulio Genovese, Po-Ru Loh, Gaurav Bhatia, Ron Do, et al. Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *The american journal of human genetics*, 97(4):576–592, 2015.

[37] Michael Wainberg, Nasa Sinnott-Armstrong, Nicholas Mancuso, Alvaro N Barbeira, David A Knowles, David Golan, Raili Ermel, Arno Ruusalepp, Thomas Quertermous, Ke Hao, et al. Opportunities and challenges for transcriptome-wide association studies. *Nature genetics*, 51(4):592–599, 2019.

[38] Xiaoquan Wen, Yeji Lee, Francesca Luca, and Roger Pique-Regi. Efficient integrative multi-snp association analysis via deterministic approximation of posteriors. *The American Journal of Human Genetics*, 98(6):1114–1129, 2016.

[39] Kris A. Wetterstrand. TOPMed RNA-seq pipeline. Available at: https://github.com/broadinstitute/gtex-pipeline/blob/master/TOPMed_RNAseq_pipeline.md. Accessed: 2023-03-01.

[40] Heather E Wheeler, Kaanan P Shah, Jonathon Brenner, Tzintzuni Garcia, Keston Aquino-Michaels, GTEx Consortium, Nancy J Cox, Dan L Nicolae, and Hae Kyung Im. Survey of the heritability and sparse architecture of gene expression traits across human tissues. *PLoS genetics*, 12(11):e1006423, 2016.

[41] Ping Zeng and Xiang Zhou. Non-parametric genetic prediction of complex traits with latent dirichlet process regression models. *Nature communications*, 8(1):456, 2017.

[42] Qianqian Zhang, Florian Privé, Bjarni Vilhjálmsson, and Doug Speed. Improved genetic prediction of complex traits from individual-level data or summary statistics. *Nature communications*, 12(1):4192, 2021.

[43] Yuhua Zhang, Corbin Quick, Ketian Yu, Alvaro Barbeira, GTEx Consortium, Francesca Luca, Roger Pique-Regi, Hae Kyung Im, and Xiaoquan Wen. Ptwas: investigating tissue-relevant causal molecular mechanisms of complex traits using probabilistic twas analysis. *Genome biology*, 21:1–26, 2020.

[44] Peng Zhao and Bin Yu. Stagewise lasso. *The Journal of Machine Learning Research*, 8:2701–2726, 2007.

[45] Xiang Zhou, Peter Carbonetto, and Matthew Stephens. Polygenic modeling with bayesian sparse linear mixed models. *PLoS genetics*, 9(2):e1003264, 2013.

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

[47] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320, 2005.

CHAPTER 5

Conclusions and Discussions

In this dissertation, we developed an meta-imputation framework that enables highly-accurate genotype imputation to benefit from multiple distributed reference panels, examined the potential and limitations of statistical fine-mapping analysis for refining genetic association signals, and assessed the performance of various approaches for gene expression imputation to facilitate transcriptome-wide association studies. We believe that our work brings improvements in genetic research particularly through the integration of data from diverse sources. In the following sections, we review the findings from each of these chapters, deliberate their limitations and discuss potential future directions for integrative data analysis in genetic research.

5.1 Genotype Imputation with Multiple Reference Panels

In Chapter 2, we introduced meta-imputation, a convenient and efficient framework that enables genotype imputation using multiple reference panels. Our method involves imputing target samples separately with different reference panels and combining the imputed results into a consensus dataset using weights determined by the empirical performance of each panel in stretches of individual genomes. By adopting this approach, we can improve imputation accuracy by incorporating genetic information from multiple sources without accessing individual-level genotype data of the reference panel samples.

Our research demonstrated that meta-imputation not only outperformed imputation using individual panels alone, leading to greater statistical power in genome-wide association studies

(GWAS), but also achieved comparable accuracy to imputation against a merged panel. This highlights the potential of meta-imputation to enhance imputation accuracy for rarer variants and support studies in diverse populations, where supplementing publicly available reference panels with customized panels for the study population can be particularly beneficial.

However, it is important to acknowledge that not all variants will necessarily experience improvements in imputation quality. In cases where a variant is present in only one reference panel, we chose to retain the original imputed results for that variant, which is appropriate when the absence is due to technical reasons. For variants that are absent from one panel because they always match the reference genome in haplotypes from that panel, assigning them a dosage of zero would be a better choice. It is also worth noting that the accuracy of meta-imputation can be influenced by factors such as the quality of pre-phasing, the density of the genotype array, and the selection of variants.

Meta-Imputation based on Low-Coverage Sequencing Data

An intuitive extension of meta-imputation is for imputation from low-coverage sequencing. The decreasing cost of next-generation sequencing (NGS) has facilitated large-scale, high-coverage whole-genome sequencing (WGS) projects [29, 31], leading to a transition from microarray platforms to NGS. However, the financial burden associated with sequencing large sample sizes remains a challenge at the current stage. To address this, low-coverage WGS followed by genotype imputation has been proposed as a cost-effective alternative [23, 27], potentially offering higher statistical power than standard GWAS designs based on microarrays under the same cost [19].

Similar to microarray-based imputation, the accuracy of imputation from low-coverage sequencing is influenced by the reference sample size and its alignment with the ancestry of the target samples. Therefore, the meta-imputation framework may enhance the accuracy in this context, and still enjoy the advantage that it does not require access to individual-level data of reference samples. Hence, the meta-imputation framework may enhance accuracy in this context while still enjoying the advantage of not requiring access to individual-level data of reference samples.

Notably, the imputation algorithm for low-coverage WGS data differs from that used for microarray data, as discussed in Chapter 1. In low-coverage WGS, observed genotypes are typically provided as genotype likelihoods rather than hard calls, necessitating an additional layer of emission probabilities on top of the Li and Stephens model [16, 23]. Moreover, unlike microarray data where genotypes are available at sparse markers, genotype likelihoods are available for almost all variants in the reference panel. The computational time for imputing low-coverage WGS data is substantially longer than that for imputing microarray data due to the larger number of variants. To address this, PBWT [8] is used to reduce the state space at each variant, making the computation more tractable. However, integrating the leave-one-out feature in this context poses a challenge as it requires considering changes in the state space from PBWT after masking a marker. Overcoming this challenge would not only facilitate the incorporation of the leave-one-out feature into the existing methodology but also unlock meta-imputation on IMPUTE5-derived results [22], thereby benefiting microarray-based imputation as well.

By extending the meta-imputation approach to imputation from low-coverage sequencing data, researchers can leverage the advantages of higher statistical power in GWAS through cost-effective imputation with enhanced accuracy.

Leave-One-Out Feature for Ensemble Learning

Methodology-wise, a notable aspect of the meta-imputation framework is the empirical evaluation of the local performance of each panel. We achieve this by systematically masking one genotyped marker at a time and comparing the leave-one-out dosages and the original genotypes at the masked sites. This approach allows us to assign appropriate weights that may vary along the genome, which realize the improved accuracy through meta-imputation.

The leave-one-out method holds promise for its potential application in ensemble learning on partially observed data, particularly for predictions on network data and spatial data. By masking local observations, we can evaluate the local predictive accuracy of each model and assign appropriate weights to enhance ensemble predictions. One compelling example is the study of

seizure spread patterns. During an epileptic seizure, the electrical activity of the brain can be observed using intracranial electroencephalography, which involves electrodes implanted in specific brain regions. However, due to practical constraints, only selected brain regions can be implanted, creating a need to build prediction models for unobserved areas based on observations in the implanted regions. It is assumed that the seizure spreads along the white-matter structural connections [20, 18], resulting in correlated outcomes in nearby regions. Therefore, we can empirically evaluate the local prediction accuracy of different models by masking observations in one region and comparing the predicted outcome with ground truth. This allows us to obtain patient-specific and region-specific weights to build an ensemble of models for optimal prediction [25].

Similarly, this idea can be extended for different types of spatial data where only partial observations are available, including but not limited to neuroimaging-based biomarker prediction [6, 21] and climate forecasting [9, 11]. Ensemble learning techniques, coupled with the leave-one-out approach, offer promising avenues for enhancing predictions and understanding complex spatial processes.

5.2 Limitations of Statistical Fine Mapping

Fine-mapping is a critical technique in genetic research that refines the initial genetic association signals from GWAS and provides insights into the functional impact of genetic variants. It plays a pivotal role in precision medicine by identifying variants associated with disease susceptibility or treatment response and guiding the prioritization of functional studies [2, 24].

In Chapter 3, we conducted a comprehensive investigation into non-identifiability issues and false discoveries arising from the complex linkage disequilibrium (LD) structure in fine-mapping analysis of genetic association signals. Non-identifiability poses a significant challenge in distinguishing causal variants from non-causal ones, even with a substantial amount of observational data. Our evaluation of existing methodologies revealed that the use of greedy algorithms could exacerbate this problem, resulting in a higher rate of false discoveries. Incorporating multiple starting

points in the implementation can potentially mitigate these false discoveries at the expense of increased computational cost. Therefore, a trade-off between accuracy and computational efficiency needs to be carefully considered.

Furthermore, we examined the effectiveness of fine-mapping approaches using summary statistics in comparison to individual-level data. The growing prominence of summary statistics is primarily driven by data sharing restrictions and the computational challenges associated with large sample sizes. While theoretically, using summary statistics may yield results equivalent to those derived from individual-level data [14, 32], our findings demonstrate that in practical terms, this approach often compromises power and coverage, particularly when LD is obtained from an external reference panel and the sample size is relatively small. Consequently, we advocate for researchers to publish their fine-mapping results alongside their GWAS findings, allowing for a comprehensive assessment of the genetic architecture underlying complex traits and diseases.

Meta-Analysis of Fine-Mapping Results

GWAS meta-analysis has become a popular method for discovering genetic risk variants due to its higher statistical power through the synthesis of information from multiple studies and its ability to utilize summary data, bypassing restrictions on sharing individual-level data. As a result, the majority of genetic risk variants discovered in recent years have emerged from large-scale meta-analyses [4, 15, 17].

However, recent research highlighted an important limitation in fine-mapping GWAS signals derived from meta-analyses. Heterogeneity arising from disparities in sample size, phenotyping processes, genotyping, or imputation can lead to miscalibration and false discoveries [13]. Achieving robust fine-mapping analysis based on summary statistics from GWAS meta-analysis requires a challenging harmonization procedure that necessitates full genotype data from all cohorts, which is often infeasible. Given the potential pitfalls, an alternative strategy is warranted. One promising idea is to conduct a meta-analysis on the fine-mapping results obtained directly from individual studies, further highlighting the significance of researchers sharing their fine-mapping findings.

This approach has the potential to address the challenges posed by heterogeneity, leading to enhanced reliability and robustness in fine-mapping analyses.

Fine-Mapping with Related Samples

As discussed in Chapter 1, addressing population structure and relatedness is crucial in both GWAS and fine-mapping analyses. While popular methods like CAVIAR [3], SuSiE [30], and DAP-G [14] incorporate genotype principal components as covariates to adjust for population structure, they do not fully account for sample relatedness within the Bayesian variable selection framework. This limitation becomes apparent when performing eQTL fine-mapping on TOPMed RNAseq data [28], where the exclusion of related samples results in a smaller effective sample size and compromises statistical power. Although these methods also accommodate summary statistics for fine-mapping, a comprehensive evaluation is needed to assess the performance of summary statistics from linear mixed models for related samples in fine-mapping. To enhance the accuracy and power of association analyses, a future direction for fine-mapping should incorporate the modeling of sample relatedness.

5.3 Gene Expression Imputation

In Chapter 4, we embarked on a detailed examination of various strategies for gene expression imputation based on the TOPMed dataset [28]. Our analysis unveiled that the accuracy of gene expression imputation improves with an increasing reference sample size, underscoring the importance of utilizing a larger reference panel, especially for underrepresented populations. Furthermore, our evaluations showed that imputation accuracy increases with the degree of ancestry matching between the reference and target samples. Therefore, using a reference panel that exhibits more diversity or possesses a higher proportion of samples with matching ancestry results in more accurate imputation outcomes.

However, our findings also revealed that a substantial mixed reference panel can yield compa-

rable results to a smaller ancestry-specific subpanel. This led us to argue that there is no necessity to impute gene expression using ancestry-specific subpanels. Instead, samples from diverse populations can be efficiently and conveniently processed together using the same mixed panel. This eliminates the need for separate processing and allows for increased convenience and efficiency when imputing gene expression for target samples from diverse populations and admixed samples.

Finally, we presented a set of imputation models derived from the TOPMed dataset, and will integrate this feature into the TOPMed imputation server. This strategy aims to offer a seamless user experience where researchers can conveniently access both imputed gene expressions and genotypes within a single, unified platform. We believe that it will greatly facilitate transcriptome-wide association analyses.

Cross-Tissue Gene Expression Imputation

Recent studies have demonstrated that incorporating expression information from multiple tissues can improve imputation accuracy and enhance the reliability of detecting causal genes compared to single-tissue methods [1, 12]. However, in our analysis of the TOPMed dataset, we focused on single-tissue strategies only due to several reasons. Firstly, the TOPMed dataset only includes RNA-seq data for a limited number of distinct tissues, and the correlations among expressions across these tissues are relatively low. Secondly, the sample sizes of these tissues, with the exception of whole blood, are considerably smaller, which could introduce bias if all tissues were to be analyzed collectively. Lastly, the small sample sizes of these tissues may result in large variance in parameter estimation when constructing multi-tissue expression models, potentially compromising the reliability of the resulting imputation models.

Nevertheless, with the continuous advancements in RNA-seq technology, we anticipate that the near future will witness the availability of gene expression data from multiple tissues on a larger scale. This expanded availability can serve as a valuable reference for building accurate and robust imputation models that consider multiple tissues simultaneously, enabling comprehensive and reliable gene expression imputation across tissues.

Imputation of Other Molecular Phenotypes

From a statistical perspective, transcriptome-wide association analysis (TWAS) can be viewed as a specific application of instrumental variable analysis. In TWAS, a prediction model is constructed using a reference panel to impute gene expression levels for samples in a genome-wide association study (GWAS). These imputed gene expression levels act as instrumental variables, enabling the assessment of associations between genetically regulated expression and the trait of interest. This framework can be readily extended to other molecular phenotypes, such as DNA methylation, metabolomics, and proteomics. Recent studies have demonstrated the potential of integrating these phenotypes to refine GWAS signals and uncover novel insights into complex traits and diseases [5, 7, 10, 26].

In line with the development of gene expression imputation, the use of instrumental variables can be extended to other molecular phenotypes through the construction of polygenic risk scores (PRS). However, our evaluation in Chapter D showcased the slightly superior performance of fine-mapping-based prediction models compared to PRS methods. As future directions, it would be valuable to focus on localizing quantitative trait loci for methylation, metabolites, and proteins, and subsequently develop imputation models leveraging these loci and their effect sizes.

5.4 Closing Remarks

With the continuous development of sequencing technology, we are entering an era of larger-scale and more diverse genomic datasets, enabling more comprehensive analysis of genetic variation and its impact on human health. The methods presented in this dissertation, such as meta-imputation and fine-mapping analysis, take advantage of these advancements. Meta-imputation leverages multiple reference panels to improve genotype imputation accuracy, while fine-mapping can better pinpoint causal variants and unravel their functional impacts as sequencing data resolution and coverage improve. Moreover, the progress in technology allows for profiling additional molecular phenotypes, including transcriptomics, metabolomics, and proteomics. Integrating these multi-

omics data in genetic association studies enhances our understanding of the complex genetic architecture underlying traits and diseases.

The ongoing development of sequencing technology brings new possibilities to genetic research. We hope that our work could inspire future research, driving further progress in precision medicine and expanding our knowledge of human genetics.

5.5 References

[1] Alvaro N Barbeira, Owen J Melia, Yanyu Liang, Rodrigo Bonazzola, Gao Wang, Heather E Wheeler, François Aguet, Kristin G Ardlie, Xiaoquan Wen, and Hae K Im. Fine-mapping and qtl tissue-sharing information improves the reliability of causal gene identification. *Genetic Epidemiology*, 44(8):854–867, 2020.

[2] Christian Benner, Chris CA Spencer, Aki S Havulinna, Veikko Salomaa, Samuli Ripatti, and Matti Pirinen. Finemap: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics*, 32(10):1493–1501, 2016.

[3] Wenan Chen, Beth R Larrabee, Inna G Ovsyannikova, Richard B Kennedy, Iana H Haralambieva, Gregory A Poland, and Daniel J Schaid. Fine mapping causal variants with an approximate bayesian method using marginal test statistics. *Genetics*, 200(3):719–736, 2015.

[4] David V Conti, Burcu F Darst, Lilit C Moss, Edward J Saunders, Xin Sheng, Alisha Chou, Fredrick R Schumacher, Ali Amin Al Olama, Sara Benlloch, Tokhir Dadaev, et al. Trans-ancestry genome-wide association meta-analysis of prostate cancer identifies new susceptibility loci and informs genetic risk prediction. *Nature genetics*, 53(1):65–75, 2021.

[5] Cajsa Davegårdh, Sonia García-Calzón, Karl Bacos, and Charlotte Ling. Dna methylation in the pathogenesis of type 2 diabetes in humans. *Molecular metabolism*, 14:12–25, 2018.

[6] Nico UF Dosenbach, Binyam Nardos, Alexander L Cohen, Damien A Fair, Jonathan D Power, Jessica A Church, Steven M Nelson, Gagan S Wig, Alecia C Vogel, Christina N Lessov-Schlaggar, et al. Prediction of individual brain maturity using fmri. *Science*, 329(5997):1358–1361, 2010.

[7] Alexandra Dumitriu, Javad Golji, Adam T Labadorf, Benbo Gao, Thomas G Beach, Richard H Myers, Kenneth A Longo, and Jeanne C Latourelle. Integrative analyses of proteomics and rna transcriptomics implicate mitochondrial processes, protein folding pathways and gwas loci in parkinson disease. *BMC medical genomics*, 9(1):1–17, 2015.

[8] Richard Durbin. Efficient haplotype matching and storage using the positional burrows–wheeler transform (pbwt). *Bioinformatics*, 30(9):1266–1272, 2014.

[9] Dapeng Feng, Kathryn Lawson, and Chaopeng Shen. Mitigating prediction error of deep learning streamflow models in large data-sparse regions with ensemble modeling and soft data. *Geophysical Research Letters*, 48(14):e2021GL092999, 2021.

[10] Eilis Hannon, Mike Weedon, Nicholas Bray, Michael O'Donovan, and Jonathan Mill. Pleiotropic effects of trait-associated genetic variation on dna methylation: utility for refining gwas loci. *The American Journal of Human Genetics*, 100(6):954–959, 2017.

[11] Sue Ellen Haupt, Jim Cowie, Seth Linden, Tyler McCandless, Branko Kosovic, and Stefano Alessandrini. Machine learning for applied weather prediction. In *2018 IEEE 14th international conference on e-science (e-Science)*, pages 276–277. IEEE, 2018.

[12] Yiming Hu, Mo Li, Qiongshi Lu, Haoyi Weng, Jiawei Wang, Seyedeh M Zekavat, Zhaolong Yu, Boyang Li, Jianlei Gu, Sydney Muchnik, et al. A statistical framework for cross-tissue transcriptome-wide association analysis. *Nature genetics*, 51(3):568–576, 2019.

[13] Masahiro Kanai, Roy Elzur, Wei Zhou, Kuan-Han H Wu, Humaira Rasheed, Kristin Tsuo, Jibril B Hirbo, Ying Wang, Arjun Bhattacharya, Huiling Zhao, et al. Meta-analysis fine-mapping is often miscalibrated at single-variant resolution. *Cell genomics*, 2(12):100210, 2022.

[14] Yeji Lee, Francesca Luca, Roger Pique-Regi, and Xiaoquan Wen. Bayesian multi-snp genetic association analysis: control of fdr and use of summary statistics. *BioRxiv*, page 316471, 2018.

[15] Daniel F Levey, Murray B Stein, Frank R Wendt, Gita A Pathak, Hang Zhou, Mihaela Aslan, Rachel Quaden, Kelly M Harrington, Yaira Z Nuñez, Cassie Overstreet, et al. Bi-ancestral depression gwas in the million veteran program and meta-analysis in¿ 1.2 million individuals highlight new therapeutic directions. *Nature neuroscience*, 24(7):954–963, 2021.

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

[17] Mike A Nalls, Cornelis Blauwendraat, Costanza L Vallerga, Karl Heilbron, Sara Bandres-Ciga, Diana Chang, Manuela Tan, Demis A Kia, Alastair J Noyce, Angli Xue, et al. Identification of novel risk loci, causal insights, and heritable risk for parkinson's disease: a meta-analysis of genome-wide association studies. *The Lancet Neurology*, 18(12):1091–1102, 2019.

[18] Christopher S Parker, Jonathan D Clayden, M Jorge Cardoso, Roman Rodionov, John S Duncan, Catherine Scott, Beate Diehl, and Sebastien Ourselin. Structural and effective connectivity in focal epilepsy. *NeuroImage: Clinical*, 17:943–952, 2018.

[19] Bogdan Pasaniuc, Nadin Rohland, Paul J McLaren, Kiran Garimella, Noah Zaitlen, Heng Li, Namrata Gupta, Benjamin M Neale, Mark J Daly, Pamela Sklar, et al. Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. *Nature genetics*, 44(6):631–635, 2012.

[20] Timothée Proix, Fabrice Bartolomei, Maxime Guye, and Viktor K Jirsa. Individual brain structure and modelling predict seizure propagation. *Brain*, 140(3):641–654, 2017.

[21] Saima Rathore, Mohamad Habes, Muhammad Aksam Iftikhar, Amanda Shacklett, and Christos Davatzikos. A review on neuroimaging-based classification studies and associated feature extraction methods for alzheimer's disease and its prodromal stages. *NeuroImage*, 155:530–548, 2017.

[22] Simone Rubinacci, Olivier Delaneau, and Jonathan Marchini. Genotype imputation using the positional burrows wheeler transform. *PLoS genetics*, 16(11):e1009049, 2020.

[23] Simone Rubinacci, Diogo M Ribeiro, Robin J Hofmeister, and Olivier Delaneau. Efficient phasing and imputation of low-coverage sequencing data using large reference panels. *Nature Genetics*, 53(1):120–126, 2021.

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

[25] Viktor Sip, Meysam Hashemi, Anirudh N Vattikonda, Marmaduke M Woodman, Huifang Wang, Julia Scholly, Samuel Medina Villalon, Maxime Guye, Fabrice Bartolomei, and Viktor K Jirsa. Data-driven method to infer the seizure propagation patterns in an epileptic brain from intracranial electroencephalography. *PLoS computational biology*, 17(2):e1008689, 2021.

[26] Silvia Sookoian and Carlos J Pirola. Liver enzymes, metabolomics and genome-wide association studies: from systems biology to the personalized medicine. *World Journal of Gastroenterology: WJG*, 21(3):711, 2015.

[27] Athina Spiliopoulou, Marco Colombo, Peter Orchard, Felix Agakov, and Paul McKeigue. Geneimp: fast imputation to large reference panels using genotype likelihoods from ultralow coverage sequencing. *Genetics*, 206(1):91–104, 2017.

[28] Daniel Taliun, Daniel N Harris, Michael D Kessler, Jedidiah Carlson, Zachary A Szpiech, Raul Torres, Sarah A Gagliano Taliun, André Corvelo, Stephanie M Gogarten, Hyun Min Kang, et al. Sequencing of 53,831 diverse genomes from the nhlbi topmed program. *Nature*, 590(7845):290–299, 2021.

[29] Erwin L Van Dijk, Hélène Auger, Yan Jaszczyszyn, and Claude Thermes. Ten years of next-generation sequencing technology. *Trends in genetics*, 30(9):418–426, 2014.

[30] Gao Wang, Abhishek Sarkar, Peter Carbonetto, and Matthew Stephens. A simple new approach to variable selection in regression, with application to genetic fine mapping. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(5):1273–1300, 2020.

[31] Kris A. Wetterstrand. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP). Available at: www.genome.gov/sequencingcostsdata. Accessed: 2023-03-01.

[32] Yuxin Zou, Peter Carbonetto, Gao Wang, and Matthew Stephens. Fine-mapping from summary data with the "sum of single effects" model. *PLoS Genetics*, 18(7):e1010299, 2022.

APPENDIX A

Quality Metrics for Post-Imputation Filtering

Quality control of genotype imputation results is crucial for downstream genome-wide association studies (GWAS) analyses, as poorly-imputed genotype dosages may lead to false-positive associations, ultimately affecting the validity of the findings. Ideally, imputation accuracy should be evaluated by comparing the imputed results to the ground-truth of sequenced genotypes using the correlation R^2. However, ground-truth data are often unavailable in practice. To provide insights into the confidence in the imputed results for each genetic variant, imputation tools report variant-level quality metrics that serve as proxies for the actual correlation R^2 between imputed and true genotypes. Examples of such metrics include Beagle AR2 (Allelic R^2) [1], MaCH Rsq [4] and IMPUTE INFO [5]. These metrics were designed under the original imputation models, and the documentations have not been updated since the introduction of the pre-phasing strategy, which has now become a standard approach in genotype imputation.

Here we present the updated formulae for imputation quality metrics under the pre-phasing imputation model, where each haplotype is imputed independently. It is worth noting that AR2 has been deprecated in Beagle5 [2]; instead, they report DR2 (dosage R^2) as the imputation quality score, which shares the same definition as the Rsq in minimac3 and minimac4 [3] (successors of MaCH). Therefore, our discussion will focus primarily on minimac Rsq and IMPUTE INFO.

Let $X = \{X_1, X_2, \ldots, X_{2N}\}$ denote the genotypes of N target samples at the SNP of interest, and let $x = \{x_1, x_2, \ldots, x_{2N}\}$ denote the imputed dosages. For simplicity, let $n = 2N$ so that n represents the number of pre-phased haplotypes being imputed. $X_i = 0$ indicates that the ith

haplotype carries the reference allele, and $X_i = 1$ indicates the alternate allele, $i = 1, 2, \ldots, n$. Also, let p denote the population allele frequency of the SNP, with the estimated allele frequency from the imputed results given by $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} x_i$. In this section, we focus on the cases where $\hat{p} \in (0, 1)$ and show that minimac Rsq and IMPUTE INFO are equivalent to each other.

For edge cases where $\hat{p} = 0$ (or $\hat{p} = 1$), meaning the imputation predicts all target samples carry the reference (or alternate) alleles with probability 1, the quality score is defined as 0 by minimac3/4 [3] and Beagle5 [2], while it is defined as 1 by IMPUTE5 [6].

A.1 Minimac Rsq

The minimac Rsq measure $\hat{R}^2$ is represented as the ratio of the observed sample variance of the dosage $s_X^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2$ to the expected variance assuming the allele X_i follows a Bernoulli distribution with the mean being the observed allele frequency $\hat{p}$.

$$\hat{R}^2 = \frac{s_X^2}{\hat{p}(1-\hat{p})} = \frac{\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2/n}{\sum_{i=1}^{n} x_i - (\sum_{i=1}^{n} x_i)^2/n}, \hat{p} \in (0, 1) \tag{A.1}$$

Since $0 \leq x_i \leq 1$, it follows that $\sum_{i=1}^{n} x_i^2 \leq \sum_{i=1}^{n} x_i$, ensuring that the numerator is always less than or equal to the denominator. As a result, the values of $\hat{R}^2$ are restricted to the range $[0, 1]$. We emphasize this point in response to criticisms against the MaCH Rsq that its value may exceed 1 [5], which is no longer true for minimac Rsq.

A.2 IMPUTE INFO

The IMPUTE INFO metric, I_A, is derived from Fisher Information and represents the ratio of observed information to complete information. Marchini & Howie provided the derivation of the INFO metric in the supplementary material of their 2010 publication [5], which was designed for the original imputation model before pre-phasing was introduced. In this work, we present a detailed derivation of the INFO metric adapted for the pre-phasing imputation model and demonstrate

that the INFO metric is equivalent to the minimac Rsq metric.

The full data likelihood is given by

$$L(p) = \prod_{i=1}^{n} p^{X_i}(1-p)^{1-X_i} \tag{A.2}$$

We can derive the likelihood score and information function:

$$U(p) = \frac{\partial \log L(p;X)}{\partial p} = \frac{\sum_{i=1}^{n} X_i - np}{p(1-p)} \tag{A.3}$$

$$I(p) = -\frac{\partial U(p)}{\partial p} = \frac{1}{p^2}\sum_{i=1}^{n} X_i + \frac{1}{(1-p)^2}\sum_{i=1}^{n}(1-X_i) \tag{A.4}$$

The complete information given the observed imputed dosages is

$$\mathcal{I}(\hat{p}) = \mathbb{E}[I(p)]|_{X=x} = \frac{n}{\hat{p}(1-\hat{p})} \tag{A.5}$$

Note that the variance for each X_i give the observed data x_i is $V(X_i)|_{X_i=x_i} = x_i(1-x_i)$, then we can derive the variance of the likelihood score given the imputed dosages as follows:

$$V(U)|_{X=x} = \frac{\sum_{i=1}^{n} x_i(1-x_i)}{\hat{p}^2(1-\hat{p})^2} \tag{A.6}$$

The observed information is defined as $\mathcal{I}^*(\hat{p}) = \mathcal{I}(\hat{p}) - V(U)|_{X=x}$. Therefore, the IMPUTE INFO metric which represents the ratio of observed information to complete information is given by:

$$I_A = 1 - \frac{V(U)|_{X=x}}{\mathbb{E}[I(p)]|_{X=x}} = 1 - \frac{\sum_{i=1}^{n} x_i(1-x_i)}{n\hat{p}^2(1-\hat{p})} = \frac{\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2/n}{\sum_{i=1}^{n} x_i - (\sum_{i=1}^{n} x_i)^2/n} \tag{A.7}$$

which is equivalent to the minimac Rsq metric shown in Equation A.1 when $\hat{p} \in (0,1)$.

A.3 References

[1] Brian L Browning and Sharon R Browning. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *The American Journal of Human Genetics*, 84(2):210–223, 2009.

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

[3] Sayantan Das, Lukas Forer, Sebastian Schönherr, Carlo Sidore, Adam E Locke, Alan Kwong, Scott I Vrieze, Emily Y Chew, Shawn Levy, Matt McGue, et al. Next-generation genotype imputation service and methods. *Nature genetics*, 48(10):1284–1287, 2016.

[4] Yun Li, Cristen J Willer, Jun Ding, Paul Scheet, and Gonçalo R Abecasis. Mach: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genetic epidemiology*, 34(8):816–834, 2010.

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

[6] Simone Rubinacci, Olivier Delaneau, and Jonathan Marchini. Genotype imputation using the positional burrows wheeler transform. *PLoS genetics*, 16(11):e1009049, 2020.

APPENDIX B

Supplemental Materials for Chapter 2

B.1 Supplemental Materials and Methods

B.1.1 Power Analysis

To evaluate the improvement brought by meta-imputation to the downstream genome-wide association studies (GWAS), we conducted simulations upon 9936 European samples from the UK Biobank exome sequencing data[4].

We randomly divided the TOPMed samples into two halves and constructed two subpanels for imputation. We imputed the UK Biobank array data on chromosome 1 using the two subpanels and the whole panel separately, and meta-imputed using the two subpanels. We then carried out a series of analyses: first, using the imputation results from the original combined panel; next, using the imputation results from each of the two subpanels; then, using the meta-imputation results; and finally, using two previously suggested approaches for GWAS when multiple imputation reference panels are available[5]: we tested each marker for association after imputation with subpanel 1 and subpanel 2, and retained the most significant result among the two, or retained the one with higher estimated imputation accuracy. The phenotypes for the association tests were simulated based on exome sequencing data. We pruned the exome sequencing data based on linkage disequilibrium (pairwise LD $r^2 < 0.2$), and randomly selected 5,000 variants on chromosome 1 with MAF < 0.0005 and estimated imputation $r^2 > 0.3$ from at least one subpanel.

For each selected variant, the phenotype was generated in the following steps so that the power

of association test using the original exome data could achieve a family-wise type I error rate of 0.05 and a statistical power of 50% with Bonferroni correction.

1. Determine the non-centrality parameter (ncp) of the chi-square distribution under the alternative hypothesis given the desired power and type I error rate.
2. The effect size $\beta = \sqrt{\frac{\text{ncp}}{2nf(1-f)}}$, where n denotes the sample size and f denotes the MAF of the variant.
3. The phenotype $y = G\beta + \epsilon, \epsilon \sim N(0, 1 - \frac{\text{ncp}}{n})$.

We compared the power among four strategies:

- GWAS using meta-imputed dosages.
- GWAS using TOPMed-imputed dosages.
- best r-square strategy – GWAS using imputed dosages from the subpanel with higher estimated imputation r^2 for each variant.
- best p-value strategy – GWAS using imputed dosages from the two subpanels separately and use the most significant p-value, adjusting for the additional variants tested.

The significance threshold for each strategy was determined by permutation tests. The BMI metrics of the target samples were permuted for 1000 times, followed by association tests on each of the 5000 variants using the four strategies separately. For each strategy, the most significant p-value from association test on each permuted trait was recorded, and the significance threshold was chosen as the 50th smallest one among the 1000 recorded p-values.

B.1.2 Meta-Imputation with Denser Array Data

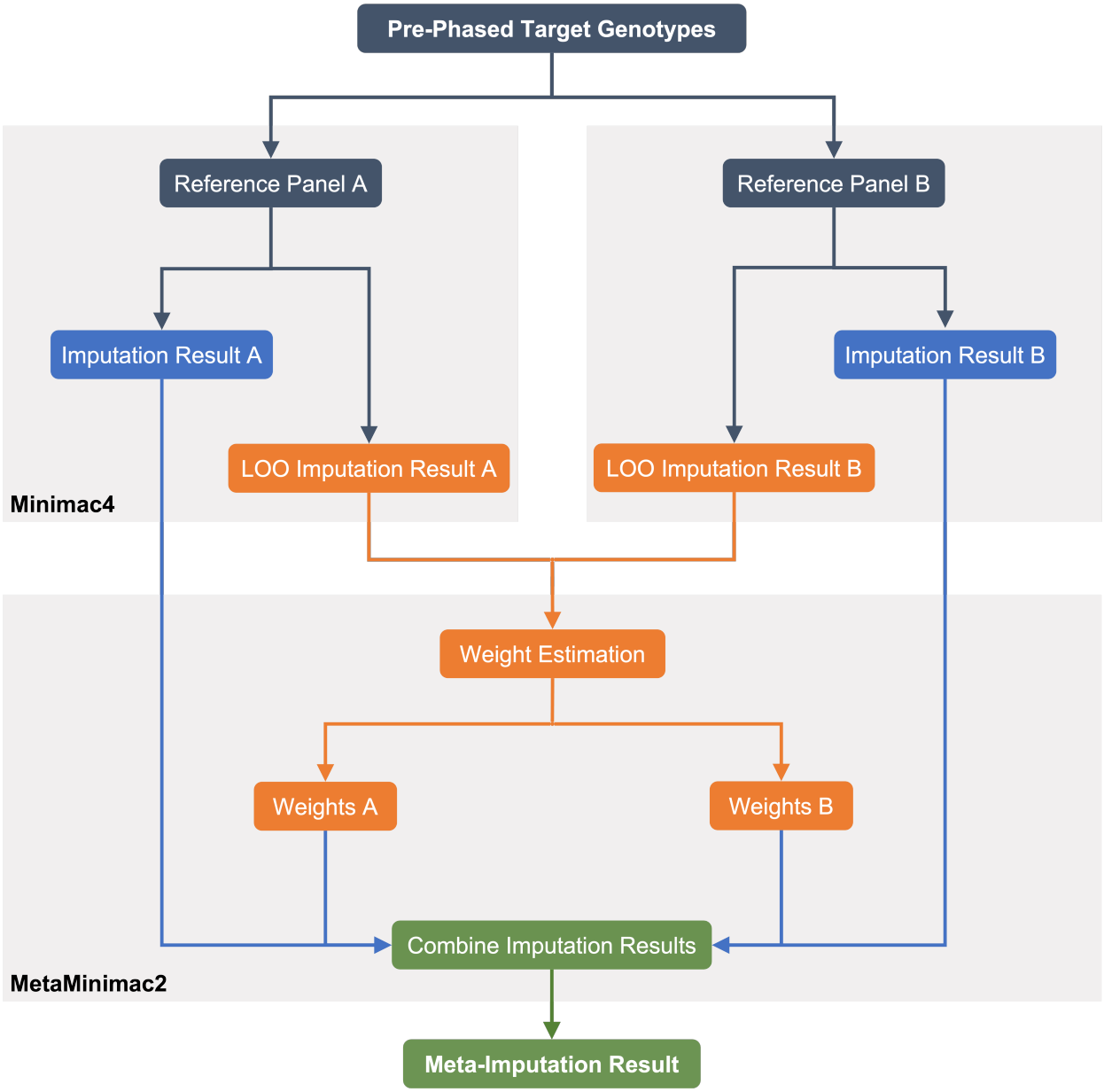
Typically, genotyping arrays mainly focus on common variants that are selected to tag other common variants and haplotypes. It is well established that larger arrays, with larger numbers of carefully selected common variants, provide for improved imputation accuracy. Our meta-imputation

approach should also benefit from increased array density. Our model includes an additional calibration step, where weights for each region of the genome are estimated using a leave-one-out approach where each array genotype is masked and re-imputed in turn. Potentially, the results of this calibration step would be different if rare variants were available in the array. To assess the value of including rare variants in the array datasets used for meta-imputation, we compared the accuracy between meta-imputation using the original UK Biobank[4] array data and meta-imputation using the original array UK Biobank array data together with half of the available UK Biobank exome variants.

We randomly selected half of the exome variants which have complete data for the 762 South Asian samples from UK Biobank and combined them with the original UK Biobank array. We rephased the merged dataset using Eagle v2.4[2] and conducted meta-imputation using 1000G[1] and TOPMed[3] panel across the autosomes. The Imputation accuracy was evaluated by comparing the final meta-imputation results for the remaining exome variants. The results, in Figure B.3, show that supplementing the common variant array genotypes with the exome variants does not make a substantial difference in the quality of the final meta-imputation result. We speculate that this is because the weights estimated using the leave-one-out approach using common variants are also close to the ideal weights for imputation of rare variants.

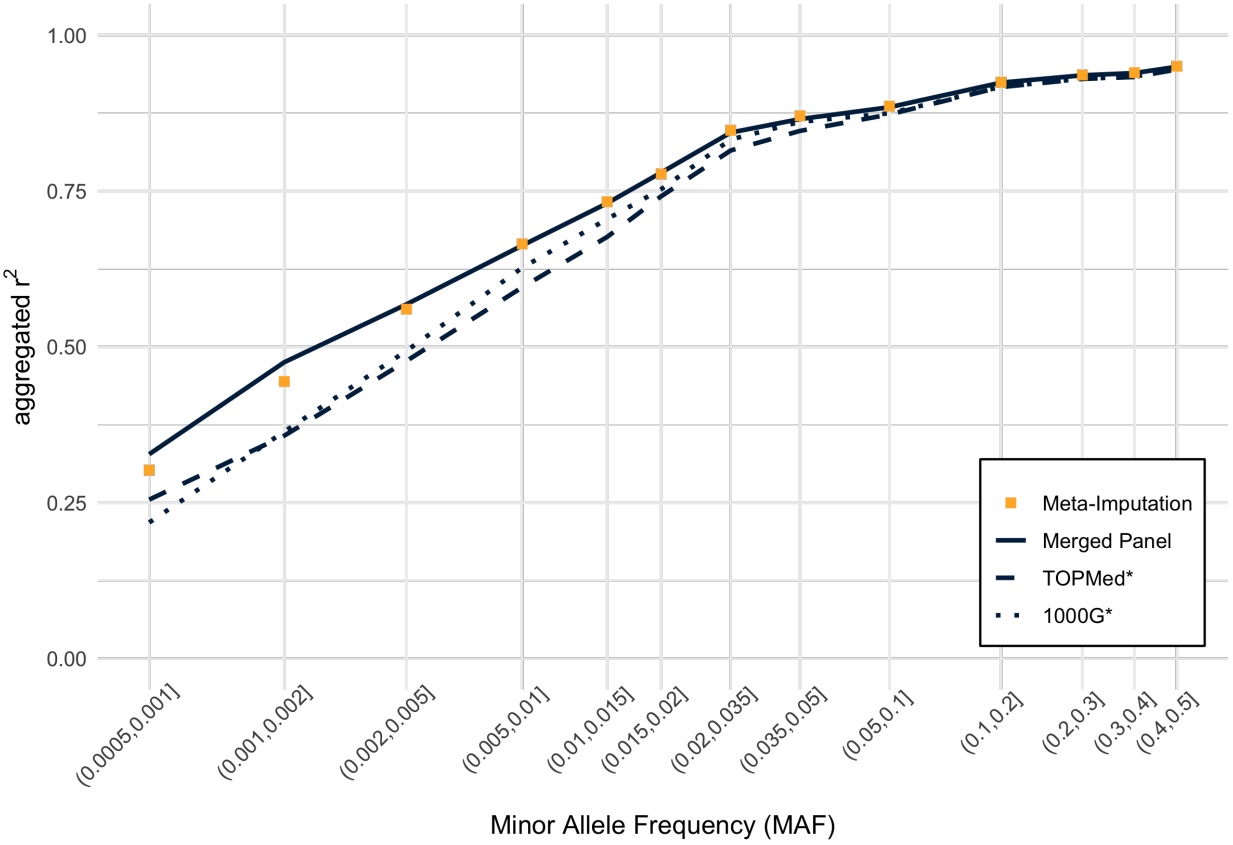
B.2 Supplemental Figures

Figure B.1: Workflow of meta-imputation



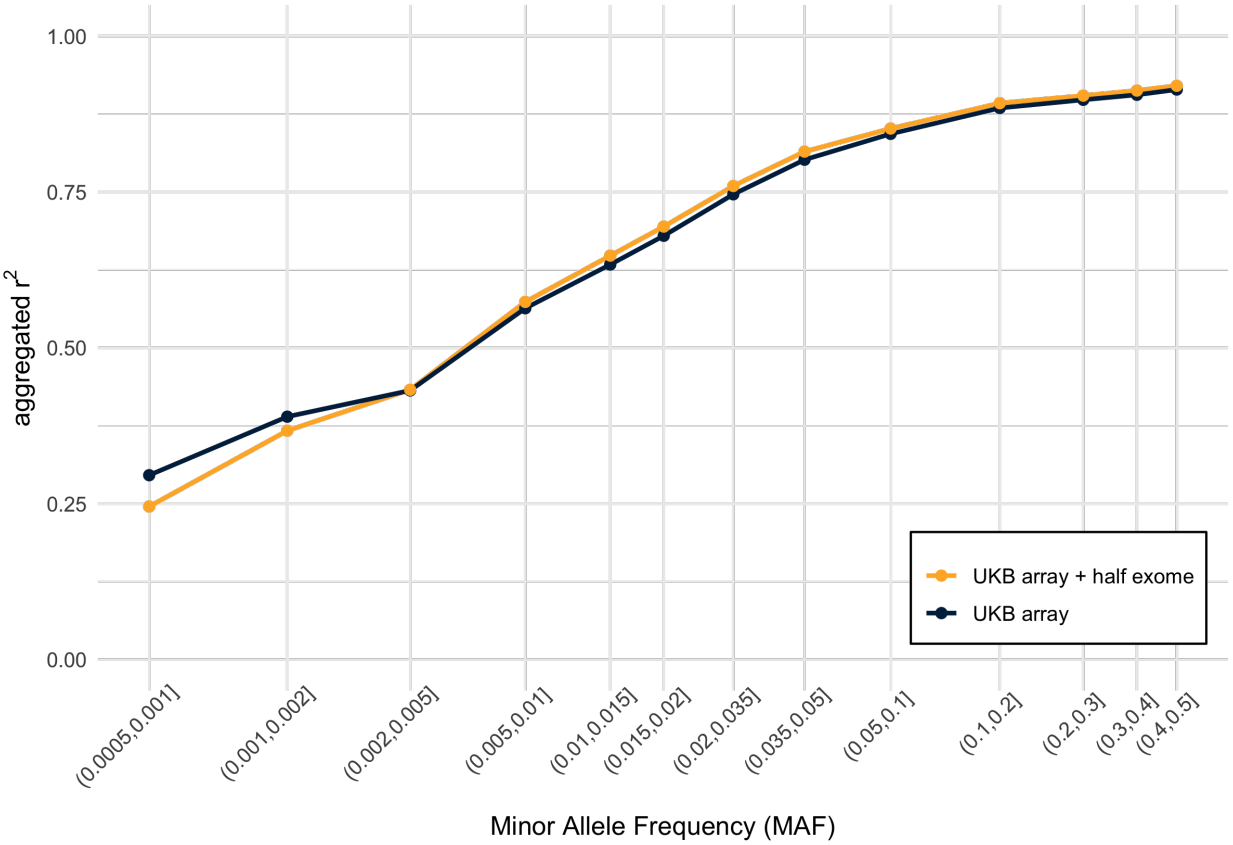
First, minimac4 imputes the target samples against two or more different reference panels. Then, MetaMinimac2 estimates the weights on each of the panels according to the empirical performance in stretches of each individual genome which is measured by leave-one-out (LOO) imputation results from minimac4. The weights are individual and region specific and reflect that the optimal choice of reference panel varies along the genome. The meta-imputation result at each marker is then a weighted average of the estimated allele counts from imputation against each panel.

Figure B.2: Comparison of accuracy between meta-Imputation and imputation using the merged panel for 762 South Asian samples on chromosome 20



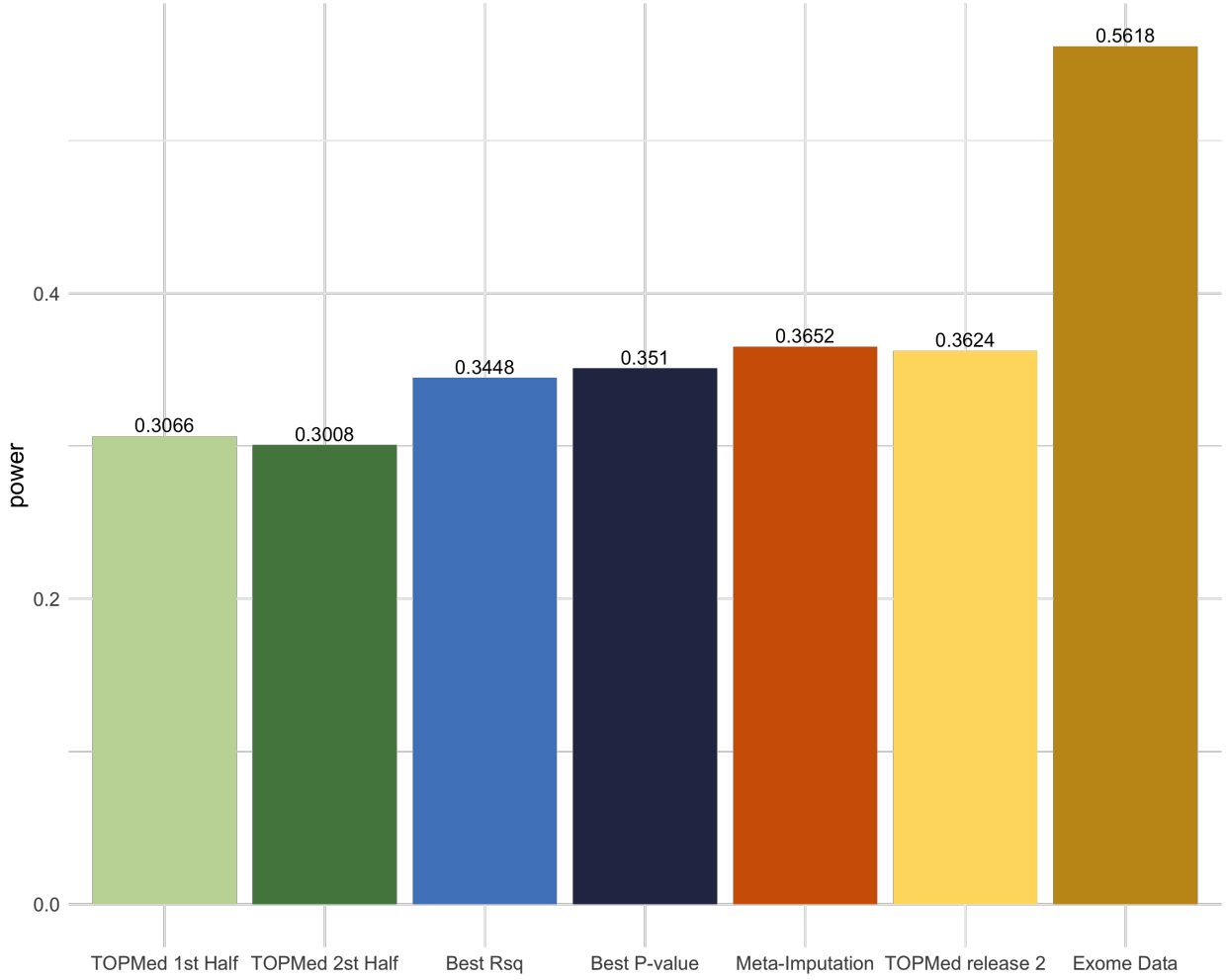
We constructed the merged panel by jointly calling variants in 2,504 1000G samples and 86,594 TOPMed samples, and reconstructed the 1000G* panel and TOPMed* panel accordingly by separating the samples and excluding the singletons. The 1000G* panel contains 2,046,899 variants on chromosome 20, and the TOPMed* panel contains 8,782,465 variants. 1,768,427 variants overlap. The imputation accuracy was evaluated based on 11,268 variants shared by 1000G*, TOPMed* and the exome sequencing data on chromosome 20. MAF was calculated based on the exome sequencing data of the target samples.

Figure B.3: Comparison of imputation accuracy between using the UK Biobank array data and using the array variants plus half of exome variants



We conducted meta-imputation on 762 South Asian samples from UK Biobank 50K exome data set using 1000G panel and TOPMed r2 panel. Imputation accuracy was evaluated by comparing imputed results and the remaining exome sequencing data on 151,719 variants across autosomes. MAF was calculated from the exome sequence data of the study samples.

Figure B.4: Comparison of power of association tests among different strategies



The evaluation was performed on 9936 European samples from UK Biobank 50K exome dataset, and 5,000 LD pruned variants with MAF < 0.0005 on chromosome 1. The significance thresholds from permutation tests are 1.75×10^{-5} (TOPMed 1st half), 1.57×10^{-5} (TOPMed 2nd half), 1.60×10^{-5} (best rsq), 1.09×10^{-5} (best p-value), 1.98×10^{-5} (meta-imputation), 1.73×10^{-5} (the whole TOPMed panel), 2.02×10^{-5} (exome data), respectively.

B.3 Supplemental Tables

Table B.1: Distribution of sample populations of the reference panels used for imputing the African American individuals in the Southwest US

Panel	Population Code	Population Description	Number of Samples
African (AFR)	ACB	African Caribbean in Barbados	96
	ESN	Esan in Nigeria	99
	GWD	Gambian in Western Division, The Gambia - Mandinka	113
	LWD	Luhya in Webuye, Kenya	99
	MSL	Mende in Sierra Leone	85
	YRI	Yoruba in Ibadan, Nigeria	108
European (EUR)	CEU	Utah residents with Northern and Western European ancestry	99
	FIN	Finnish in Finland	99
	GBR	British in England and Scotland	91
	IBS	Iberian populations in Spain	107
	TSI	Toscani in Italy	107

Table B.2: Comparison of computational time between imputation using Minimac4 with and without the meta-imputation option

The meta-imputation option in Minimac4 triggers the leave-one-out imputation (which is carried out with the inner loop used for standard imputation) and writes the leave-one-out imputation dosage file which is required for the downstream meta-imputation analysis in MetaMinimac2. The tests were performed on chromosome 20 for UK Biobank samples using the TOPMed panel. All the tests were conducted on Intel Xeon Platinum 8268 CPU @ 2.90GHz, using one single core at a time.

Number of Samples	Time ([hh]:mm:ss)	
	Minimac4	Minimac4 with Meta-Imputation Option
1,000	5:27:37	5:42:37
2,000	10:39:41	11:06:08
5,000	26:23:14	26:40:53
10,000	51:55:23	53:15:35

B.4 References

[1] 1000 Genomes Project Consortium et al. A global reference for human genetic variation. *Nature*, 526(7571):68, 2015.

[2] Po-Ru Loh, Pier Francesco Palamara, and Alkes L Price. Fast and accurate long-range phasing in a uk biobank cohort. *Nature genetics*, 48(7):811–816, 2016.

[3] Daniel Taliun, Daniel N Harris, Michael D Kessler, Jedidiah Carlson, Zachary A Szpiech, Raul Torres, Sarah A Gagliano Taliun, André Corvelo, Stephanie M Gogarten, Hyun Min Kang, et al. Sequencing of 53,831 diverse genomes from the nhlbi topmed program. *Nature*, 590(7845):290–299, 2021.

[4] Cristopher V Van Hout, Ioanna Tachmazidou, Joshua D Backman, Joshua D Hoffman, Daren Liu, Ashutosh K Pandey, Claudia Gonzaga-Jauregui, Shareef Khalid, Bin Ye, Nilanjana Banerjee, et al. Exome sequencing and characterization of 49,960 individuals in the uk biobank. *Nature*, 586(7831):749–756, 2020.

[5] Wei Zhou, Lars G Fritsche, Sayantan Das, He Zhang, Jonas B Nielsen, Oddgeir L Holmen, Jin Chen, Maoxuan Lin, Maiken B Elvestad, Kristian Hveem, et al. Improving power of association tests using multiple sets of imputed genotypes from distributed reference panels. *Genetic epidemiology*, 41(8):744–755, 2017.

APPENDIX C

Supplemental Materials for Chapter 3

C.1 Sufficient Statistics for Fine-Mapping

Consider the following multivariate regression model for fine-mapping

$$y = X\beta + \epsilon, \epsilon \sim N(0, \sigma^2 I) \tag{C.1}$$

For simplicity, we assume that both the phenotype y and genotype X have been centered to have mean 0. The log likelihood can be written as

$$l(\beta, \sigma^2; X, y) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(y^T y - 2y^T X\beta + \beta^T X^T X\beta) \tag{C.2}$$

Equation C.2 indicates that $(X^T X, X^T y, y^T y, n)$ are sufficient statistics. Here we show that this set of statistics is equivalent to $(R, \hat{b}, \hat{s}, y^T y, n)$, where R denote the sample correlation matrix between variants calculated from the GWAS samples, $\hat{b}$ and $\hat{s}$ denote the estimated effect sizes and corresponding standard deviation obtained from the single-variant association test, $y^T y$ denotes the sum of squared phenotype, and n denotes the sample size.

For the jth variant, we consider the single-variant association test model $y = x_j b_j + e_j, e_j \sim N(0, \sigma_j^2)$, where x_j denote the jth column of the genotype matrix X. We are able to obtain the estimated effect size as $\hat{b}_j = \frac{x_j^T y}{x_j^T x_j}$ with the variance of the estimator as $s_j^2 = \frac{\sigma_j^2}{x_j^T x_j}$. When σ_j^2 is unknown, which is the typical case in practice, we approximate s_j^2 by plugging the esimtated

residual variance $\hat{\sigma}^2 = \frac{1}{n-1}(y - x_j\hat{b}_j)^T(y - x_j\hat{b}_j)$. Therefore, the z-score can be wriiten as $\hat{z}_j = \frac{\hat{b}_j}{\hat{s}_j} = \frac{x_j^T y}{\hat{\sigma}\sqrt{x_j^T x_j}}$ and $\hat{z}_j$ follows a t-distribution with $(n-1)$ degrees of freedom, which is close ot the normal distribution when n is large.

Let $\hat{r}_j$ denote the sample correlation between x_j and y, i.e. $\hat{r}_j = \frac{x_j^T y}{\sqrt{x_j^T x_j}\sqrt{y^T y}}$. Since in a simple linear regression, the test statistic for the null hypothesis $b_j = 0$ is equivalent to that for testing the correlation coefficient $r_j = 0$, the z-score can also be represented as $\hat{z}_j = \sqrt{n-1}\frac{\hat{r}_j}{\sqrt{1-\hat{r}_j^2}}$, which induces that $\hat{r}_j^2 = \frac{\hat{z}_j^2}{\hat{z}_j^2+n-1}$. Note that the residual sum of squares $RSS = (1-\hat{r}_j^2)y^T y = \frac{n-1}{\hat{z}_j^2+n-1}y^T y$, the estimated variance can be written as $\hat{\sigma}^2 = \frac{RSS}{n-1} = \frac{1}{\hat{z}_j^2+n-1}y^T y$. Therefore, we are able to write $x_j^T x_j$ and $x_j^T y$ as a function of $(\hat{b_j}, \hat{s}_j, y^T y, n)$ as follows.

$$x_j^T x_j = \frac{\hat{\sigma}_j^2}{\hat{s}_j^2} = \frac{y^T y}{\hat{b}_j^2 + (n-1)\hat{s}_j^2} \tag{C.3}$$

$$x_j^T y = \hat{b}_j x_j^T x_j = \frac{\hat{b}_j y^T y}{\hat{b}_j^2 + (n-1)\hat{s}_j^2} \tag{C.4}$$

To this point, we can recover $X^T y$ and the diagonal elements of $X^T X$. Let Λ denotes a diagonal matrix with the jth diagnoal element being $\Lambda_{jj} = x_j^T x_j, \forall j = 1, 2, \cdots, p$. Note that $X^T X = \Lambda^{\frac{1}{2}} R \Lambda^{\frac{1}{2}}$, we can able to fully recover $X^T X$ from $(R, \hat{b_j}, \hat{s}_j, y^T y, n)$.

In summary, given the effect sizes along with the standard deviation $\hat{b_j}, \hat{s}_j$ from the single-variant association test, the sample size n, sample variance of the phenotype $(y^T y)/(n-1)$, and the in-sample genotype correlation matrix R, we are able to obtain the same results as fine-mapping using the individual-level data (X, y).

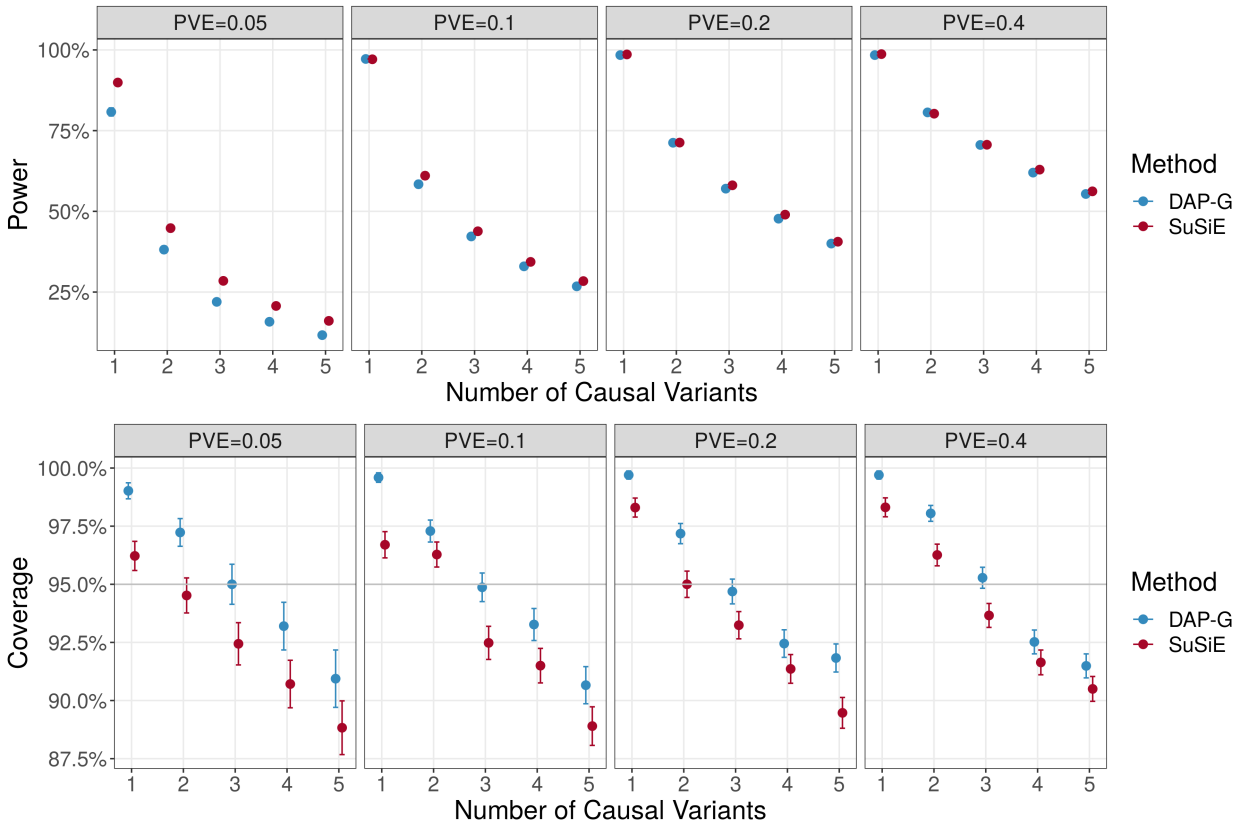
C.2 DAP-MS Algorithm

- **Require:** Data X, y
- **Require:** Priors for variants, $\frac{1}{p}$ by default
- **Require:** Model size limit, $T = p$ by default

- **Step 1:** Scan all single-variant models, calculate posterior score $S(\gamma_i|X, y) = \pi_i BF(\gamma_i)$, where all elements in γ_i are zero except that the ith element is 1.
- **Step 2:** Select models with score$> T$ as starting points (M_1 – candidate models of size 1).
- **While** model size $t \leq T$, do
  - **For** each candidate model $m_{t,i}$ in M_t
    1. Expand the model by including one more variant.
    2. Mark the new model with the highest posterior score as $m_{t+1,i}$.
    3. If $BF(m_{t+1,i}) > BF(m_{t,i})$, insert $m_{t+1,i}$ into M_{t+1}.
  - Calculate the sum of posterior scores of all models of size $t + 1$ scanned.
  - Stop if $|M_{t+1}| = 0$, or the sum of posterior scores of models of size $t + 1$ is smaller than 10% that of size t.
  - Update $t = t + 1$

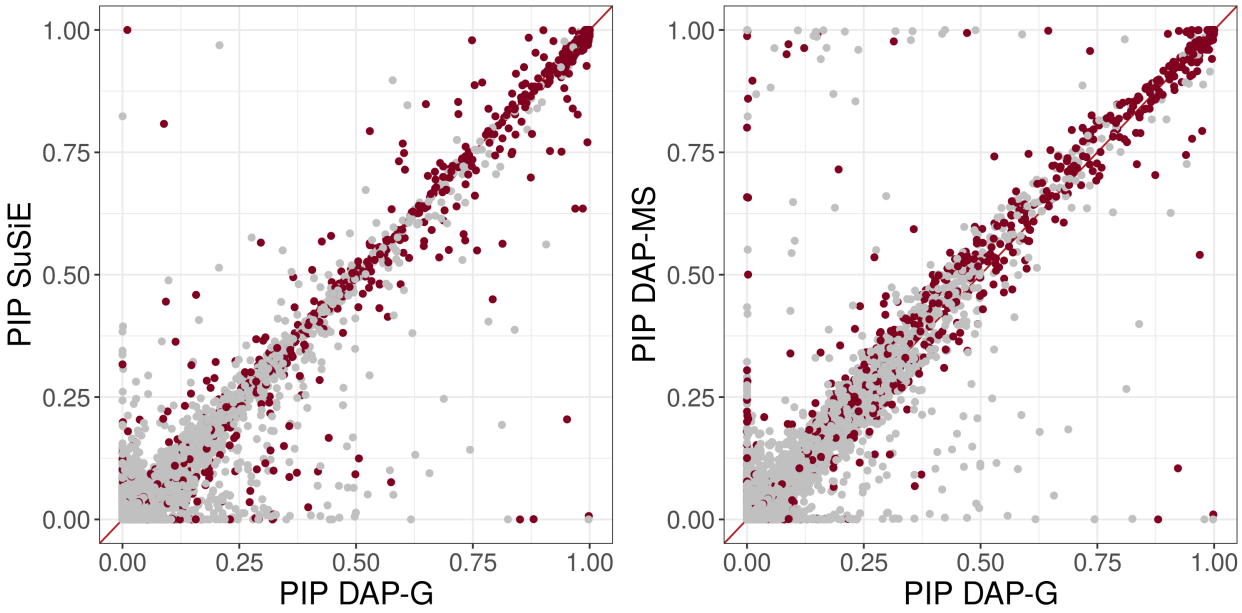
C.3 Supplemental Figures

Figure C.1: Comparisons of Power and Coverage between SuSiE and DAP-G



We randomly selected 1,000 genomic regions from the GTEx data, each encompassing 1,000 variants. For the 670 samples, we simulated phenotypes under varying settings of PVE$\in \{0.05, 0.1, 0.2, 0.4\}$ and number of causal variants ranging from 1 to 5. We then conducted fine-mapping using individual-level data and evaluated the 95% credible sets from SuSiE and signal clusters with SPIP$> 95\%$ from DAP-G. This comparison was performed in two aspects: 1) power, defined as the fraction of simulated causal variants included in a signal, and 2) coverage, denoting the fraction of signals that encapsulate at least one causal variant.

Figure C.2: Comparisons of PIPs from different Methods

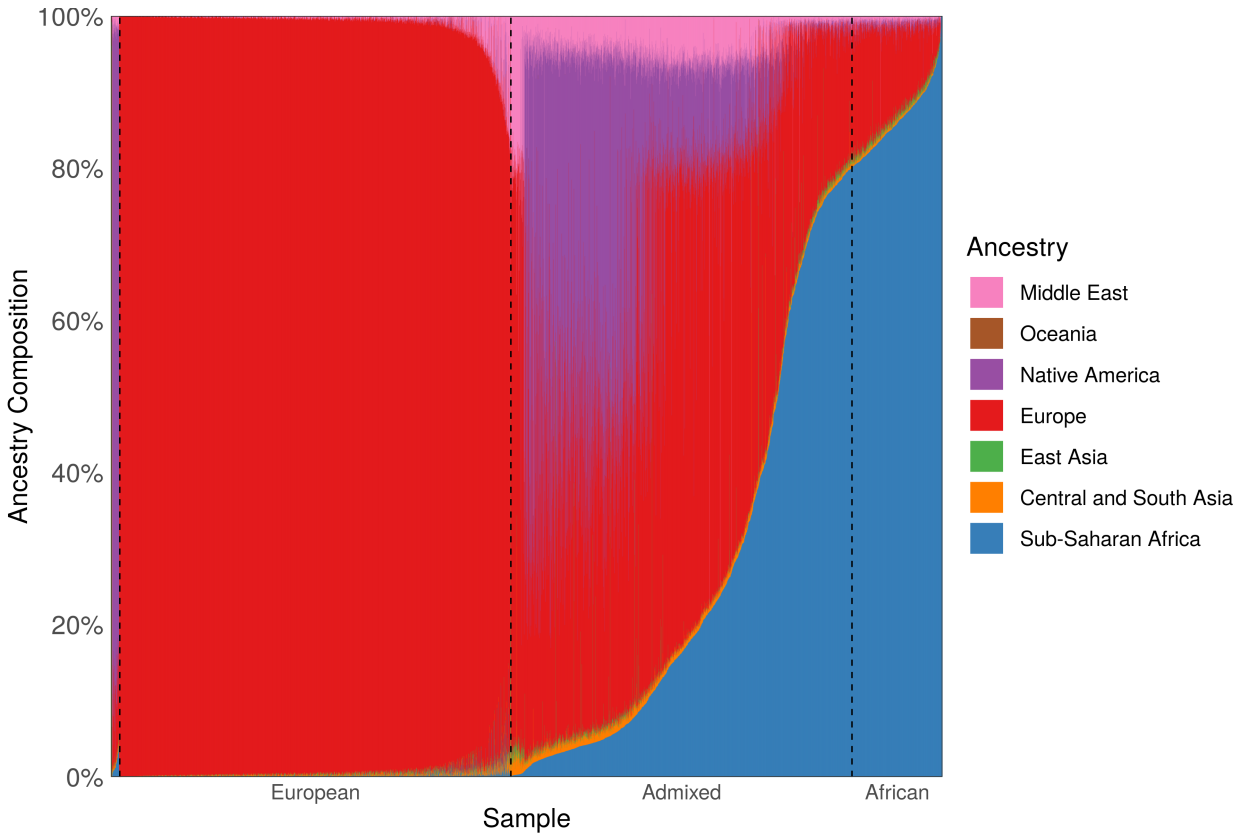


We conducted a comparison of PIPs obtained from DAP-G against those derived from SuSiE and DAP-MS, respectively. Each point within the visualization represents a single variant from one of the simulations, with causal variants indicated in red and non-causal variants depicted in gray. These simulations were performed under conditions of PVE set at 0.2 with 3 causal variants.

APPENDIX D

Supplemental Figures and Tables for Chapter 4

Figure D.1: Ancestral Composition of TOPMed Whole Blood Samples



We conducted ADMIXTURE analysis on 6602 individuals with whole blood RNAseq data from TOPMed. Using a threshold of 80%, we identified (from right to left in the figure) 719 individuals as African, 2708 as admixed, 3105 as European, and 70 as others (East Asia: 12; Central and South Asia: 2; Middle East: 1; Native America: 55).

Figure D.2: Comparison of Spearman's correlation r across different methods.

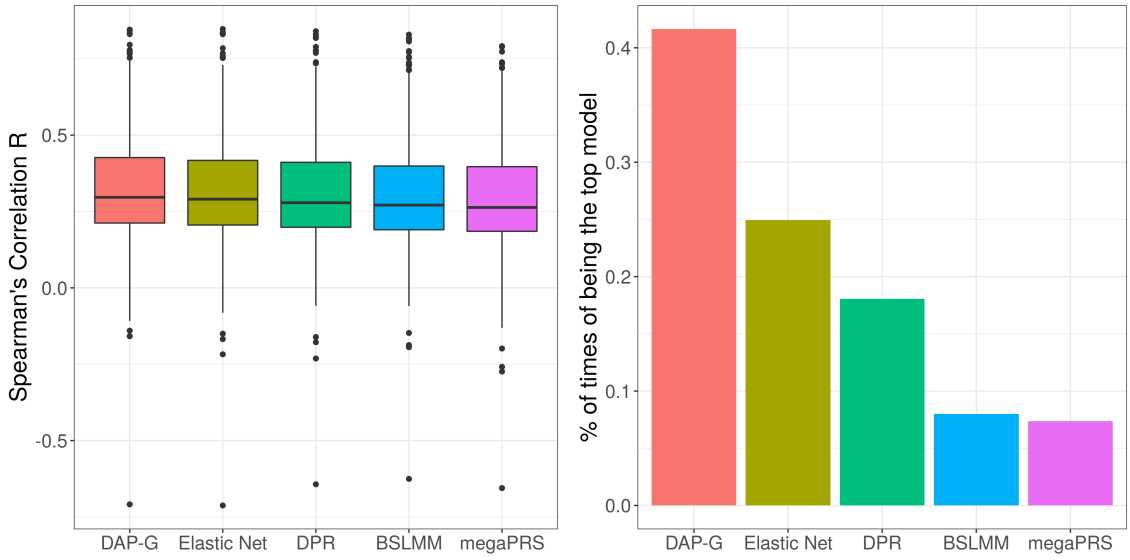


Figure D.3: Performance of Reference Panels of Different Size and Ancestral Compositions

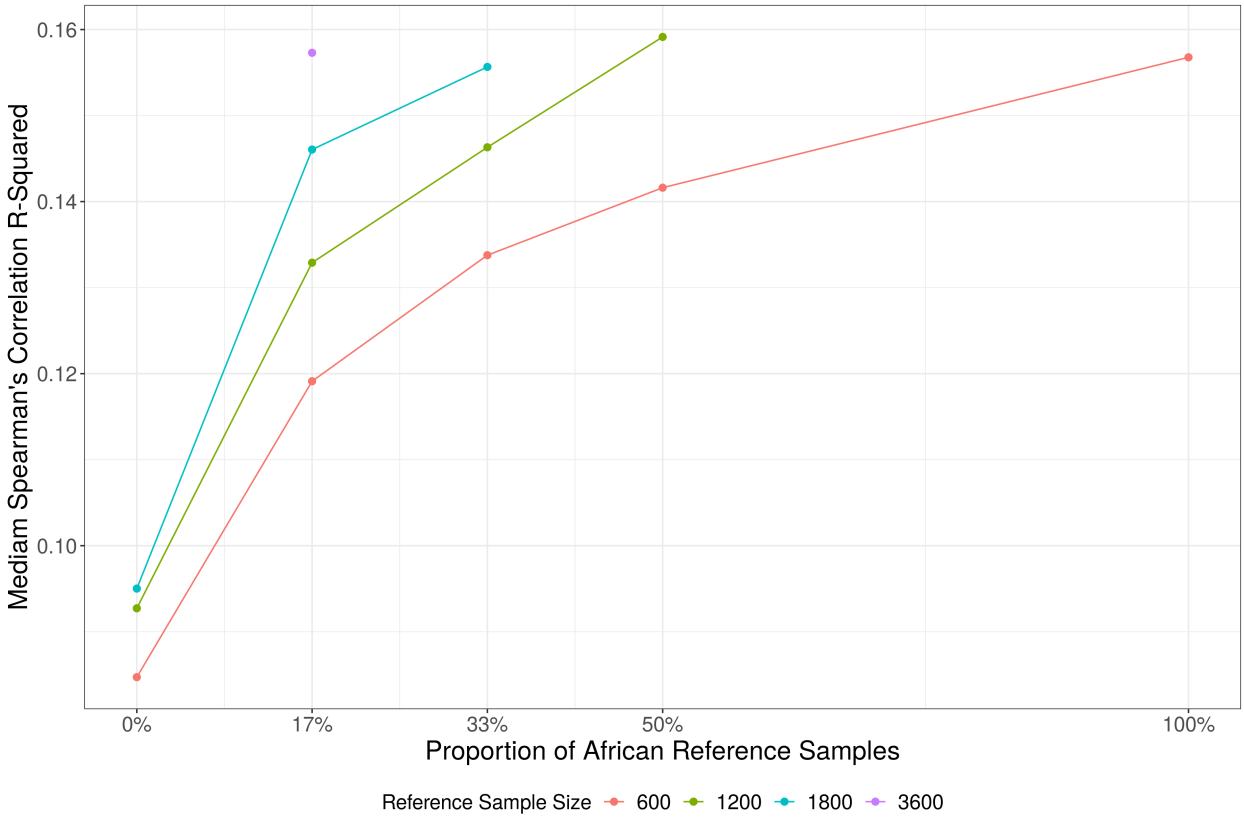
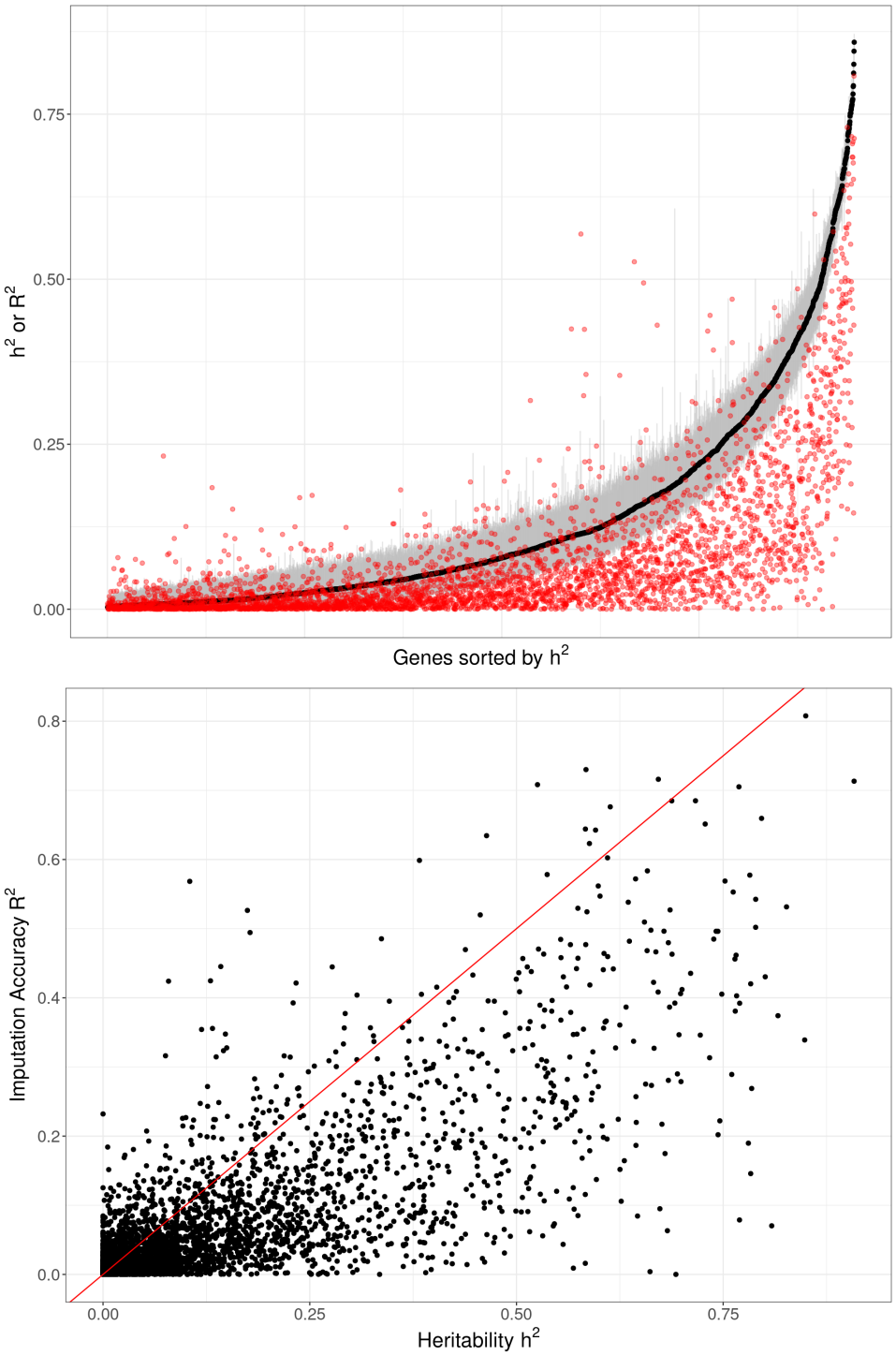


Figure D.4: Prediction Performance versus Heritability



We compared the prediction performance (red) in comparison to gene expression heritability estimates (black; 95% confidence intervals in gray). Performance was assessed in GTEx whole-blood cohort. Accuracy was measured by Pearson's r^2 between the imputed expression levels using DAP-G weights obtained the from TOPMed panel and the observed expression levels. The heritability was estimated by Wheeler et al. [1] using BSLMM in DGN dataset.

Table D.1: Summary of TOPMed RNA-Sequencing Samples included in Analysis

Short Name	Study Name	Sample Size by Tissue					
		Whole Blood	Lung	Monocyte	Nasal epithelial	PBMC	T cell
COPDGene	Genetic Epidemiology of COPD	350	-	-	359	-	-
FHS	Framingham Heart Study	793	-	-	-	-	-
GALA II	Genes-Environments and Admixture in Latino Americans	1897	-	-	-	-	-
LTRC	Lung Tissue Research Consortium	-	1360	-	-	-	-
MESA	Multi-Ethnic Study of Atherosclerosis	-	-	352	-	1265	368
SAGE	Study of African Americans, Asthma, Genes and Environments	705	-	-	-	-	-
SPIROMICS	SubPopulations and InteRmediate Outcome Measures In COPD Study	1578	-	-	-	-	-
WHI	Women's Health Intiative	1279	-	-	-	-	-
Total		6602	1360	352	359	1265	368

Table D.2: Comparison of Gene Expression Imputation using Reference Panels of Different Sample Sizes and Ancestral Compositions

Reference Panel			Mean r^2		Median r^2		% $r^2 > 0.1$		% $r^2 > 0.2$		% $r^2 > 0.3$		% $r^2 > 0.4$	
Size	AFR	EUR	AFR	EUR	AFR	EUR	AFR	EUR	AFR	EUR	AFR	EUR	AFR	EUR
600	0	600	0.142	0.217	0.085	0.158	45.7%	63.8%	25.6%	42.7%	15.5%	28.7%	9.6%	18.4%
600	100	500	0.170	0.217	0.119	0.157	55.3%	64.1%	32.6%	41.8%	19.3%	28.1%	11.3%	18.2%
600	200	400	0.182	0.212	0.134	0.157	59.4%	63.8%	35.7%	41.5%	20.5%	28.0%	12.1%	17.4%
600	300	300	0.191	0.208	0.142	0.152	61.3%	62.3%	39.4%	40.3%	23.2%	26.8%	12.6%	16.9%
600	600	0	**0.209**	0.191	0.157	0.137	**66.0%**	58.5%	**41.6%**	35.7%	**26.9%**	24.8%	**15.8%**	14.6%
1200	0	1200	0.147	0.225	0.093	0.170	48.0%	67.1%	27.6%	43.3%	15.7%	29.9%	9.4%	18.6%
1200	200	1000	0.182	0.224	0.133	0.171	59.4%	66.7%	35.8%	43.5%	20.6%	29.6%	12.1%	18.4%
1200	400	800	0.194	0.220	0.146	0.166	61.3%	66.4%	38.7%	42.8%	23.8%	28.4%	13.5%	18.3%
1200	600	600	0.203	0.216	**0.159**	0.161	64.1%	65.6%	41.0%	41.2%	24.8%	27.8%	14.3%	17.8%
1800	0	1800	0.150	0.227	0.095	0.173	48.1%	67.6%	27.6%	44.0%	16.0%	29.9%	9.7%	**19.3%**
1800	300	1500	0.189	0.227	0.146	0.169	61.4%	66.9%	38.1%	44.0%	21.8%	30.2%	12.4%	19.0%
1800	600	1200	0.200	0.223	0.156	0.167	64.2%	67.5%	41.1%	43.4%	24.2%	29.0%	14.4%	18.5%
3600	600	3000	0.198	0.**229**	0.157	**0.175**	63.8%	**68.7%**	39.6%	**44.2%**	24.5%	**30.3%**	13.0%	**19.3%**

D.1 References

[1] Heather E Wheeler, Kaanan P Shah, Jonathon Brenner, Tzintzuni Garcia, Keston Aquino-Michaels, GTEx Consortium, Nancy J Cox, Dan L Nicolae, and Hae Kyung Im. Survey of the heritability and sparse architecture of gene expression traits across human tissues. *PLoS genetics*, 12(11):e1006423, 2016.