Statistical and Computational Methods for Genome-Wide Association Analysis

by

Corbin T. Quick

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in the University of Michigan
2018

Doctoral Committee:

Professor Michael Boehnke, Co-Chair
Associate Professor Hyun Min Kang, Co-Chair
Professor Gonçalo Abecasis
Associate Professor Xiaoquan (William) Wen
Associate Professor Cristen Willer

Corbin T. Quick

corbinq@umich.edu

ORCID iD: 0000-0001-7199-2930

DEDICATION

To my family and friends.

ACKNOWLEDGMENTS

I'm grateful to my family for their continual support and encouragement over the years; my advisers, committee members, and other faculty and staff for their guidance, patience, encouragement, and help; and my office-mates and other colleagues for their friendship and advice. It has been a great privilege and pleasure to have been a part of the University of Michigan Department of Biostatistics and Center for Statistical Genetics.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Technological and scientific advances in recent years have revolutionized genomics. For example, decreases in whole genome sequencing (WGS) costs have enabled larger WGS studies as well as larger imputation reference panels, which in turn provide more comprehensive genomic coverage from lower-cost genotyping methods. In addition, new technologies and large collaborative efforts such as ENCODE and GTEx have shed new light on regulatory genomics and the function of non-coding variation, and produced expansive publicly available data sets. These advances have introduced data of unprecedented size and dimension, unique statistical and computational challenges, and numerous opportunities for innovation. In this dissertation, we develop methods to leverage functional genomics data in post-GWAS analysis, to expedite routine computations with increasingly large genetic data sets, and to address limitations of current imputation reference panels for understudied populations.

In Chapter 2, we propose strategies to improve imputation and increase power in GWAS of understudied populations. Genotype imputation is instrumental in GWAS, providing increased genomic coverage from low-cost genotyping arrays. Imputation quality depends crucially on reference panel size and the genetic distance between reference and target haplotypes. Current reference panels provide excellent imputation quality in many European populations, but lower quality in non-European, admixed, and isolate populations. We consider a GWAS strategy in which a subset of participants is sequenced and the rest are imputed using a reference panel that

comprises the sequenced participants together with individuals from an external reference panel. Using empirical data from the HRC and TOPMed WGS Project, simulations, and asymptotic analysis, we identify powerful and cost-effective study designs for GWAS of non-European, admixed, and isolated populations.

In Chapter 3, we develop efficient methods to estimate linkage disequilibrium (LD) with large data sets. Motivated by practical and logistical constraints, a variety of statistical methods and tools have been developed for analysis of GWAS summary statistics rather than individual-level data. These methods often rely on LD estimates from an external reference panel, which are ideally calculated on-the-fly rather than precomputed and stored. We develop efficient algorithms to estimate LD exploiting sparsity and haplotype structure and implement our methods in an open-source C++ tool, emeraLD. We benchmark performance using genotype data from the 1KGP, HRC, and UK Biobank, and find that emeraLD is up to two orders of magnitude faster than existing tools while using comparable or less memory.

In Chapter 4, we develop methods to identify causative genes and biological mechanisms underlying associations in post-GWAS analysis by leveraging regulatory and functional genomics databases. Many gene-based association tests can be viewed as instrumental variable methods in which intermediate phenotypes, e.g. tissue-specific expression or protein alteration, are hypothesized to mediate the association between genotype and GWAS trait. However, LD and pleiotropy can confound these statistics, which complicates their mechanistic interpretation. We develop a hierarchical Bayesian model that accounts for multiple potential mechanisms underlying associations using functional genomic annotations derived from GTEx, Roadmap/ENCODE, and other sources. We apply our method to analyze twenty-five complex traits using GWAS summary statistics from UK Biobank, and provide an open-source implementation of our methods.

In Chapter 5, we review our work, discuss its relevance and prospects as new resources emerge,

and suggest directions for future research.

Chapter 1

Introduction

1.1 Background

The past two decades have seen an explosion of technological and scientific advances in genomics. As new sequencing technologies have emerged, the cost of whole-genome sequencing (WGS) has fallen from over $100M USD to under $1,000 USD per genome (KA 2018), enabling larger WGS studies as well as larger haplotype reference panels (e.g., McCarthy et al. 2016), which in turn provide more comprehensive genomic coverage from lower-cost genotyping methods through genotype imputation. New technologies and large collaborative efforts such as the Encyclopedia of DNA Elements (ENCODE) project, NIH Roadmap Epigenomics Mapping Consortium, and Genotype-Tissue Expression (GTEx) project have produced a wealth of data and shed new light on regulatory genomics and the functional non-coding genome (ENCODE Project Consortium 2012; GTEx Consortium 2015; Karczewski and Snyder 2018). Genome-wide association studies (GWAS) have identified thousands of genetic variants associated with hundreds of complex traits (MacArthur et al. 2016), and while the biological mechanisms underlying these associations are often poorly understood, there have been notable breakthroughs in genomic medicine and gene therapy

(e.g., Ribeil et al. 2017; Rangarajan et al. 2017). Altogether, these advances have produced data of various new types and unprecedented volume, presenting new challenges as well as opportunities for statistical innovation.

Challenges & Opportunities from Advances in Genomics

Large sequencing studies and imputation reference panels, made possible by advances in sequencing technologies, require enormous computational resources to process, store, and analyze, presenting a need for more efficient data formats and analysis methods (McCarthy et al. 2016; Das et al. 2016). In addition, large reference panels have enabled far more accurate and comprehensive genotype imputation for many populations, prompting a need to re-assess the relative cost-effectiveness of sequencing and imputation-based genotyping in GWAS, and to develop more powerful and cost-effective genotyping strategies for understudied populations (McCarthy et al. 2016; B. Li and Leal 2008). Large omics studies, empowered by new experimental techniques as well as sequencing technologies, have constructed expansive databases characterizing transcriptomic, proteomic, and metabolomic interactions as well as the functional effects of coding and non-coding variation (ENCODE Project Consortium 2012; GTEx Consortium 2015; De Rie et al. 2017). These data, together with more comprehensive genomic coverage in GWAS, can provide more refined insights into the biological pathways that underlie genetic effects on complex traits, prompting a need for statistical methods to integrate GWAS with diverse types of functional genomic data.

Purpose

In this dissertation, we develop statistical methods to leverage functional genomics databases in post-GWAS analysis, to expedite routine computations with increasingly large genetic data sets, and to address limitations of current imputation reference panels for understudied populations.

1.2 Genotype Imputation: Challenges for Understudied Populations

Genotype imputation has been instrumental in GWAS, enabling more complete meta-analysis of results from multiple studies, and providing increased genomic coverage from array-based genotype callsets (Y. Li et al. 2009). Imputation algorithms typically employ a Hidden Markov Model (HMM) in which partially observed haplotypes in the study sample are modeled as mosaics of complete haplotypes in a reference panel (N. Li and Stephens 2003; e.g., International HapMap 3 Consortium 2010; 1KGP Consortium 2015). Genotyping arrays, e.g. the Illumina OmniExpress, are often used to construct a scaffold for imputation via array-based genotype calls over a sparse set of directly typed marker variants. This array-and-imputation genotyping strategy provides an inexpensive *in silico* alternative to whole genome sequencing, the gold-standard method to capture genetic variation comprehensively across the allele frequency spectrum.

Imputation coverage and accuracy depend crucially on the genetic similarity between reference and target populations (Roshyara and Scholz 2015) and the number of reference haplotypes available (Das et al. 2016). The earliest imputation reference panels, e.g. from the 1000 Genomes Project and International HapMap Consortium, included individuals from diverse worldwide human populations (1KGP Consortium 2015; International HapMap 3 Consortium 2010). These projects provided new insights into haplotype structure and demographic history across human populations, as well as new resources for genotype imputation. By contrast, the largest current imputation reference panels, e.g. from the Haplotype Reference Consortium (HRC; McCarthy et al. 2016) and UK10K Consortium (UK10K Consortium 2015), are largely European. These reference panels have enabled far more accurate and comprehensive imputation for many European populations, but are less effective for non-European, admixed, and isolate populations (UK10K Consortium 2015). In

populations that are underrepresented in current imputation reference panels, population-matched or multi-ethnic reference panels can be constructed to provide improved imputation quality (Deelen et al. 2014; Lencz et al. 2017; Ahmad et al. 2017).

In Chapter 2, we propose strategies to improve genotype imputation and increase power in GWAS of diverse human populations. We consider a strategy in which a subset of participants is sequenced and the rest are imputed using a reference panel that comprises the sequenced participants together with individuals from an external reference panel. Using empirical data from the Haplotype Reference Consortium (HRC) and NHLBI Trans-Omics for Precision Medicine (TOPMed) Whole Genome Sequencing (WGS) Project, simulations, and asymptotic analysis, we identify powerful and cost-effective sequencing-and-imputation study designs for GWAS of non-European, admixed, and isolated populations.

1.3 Leveraging Natural Features of Genetic Data to Expedite Computation

Demographic history and the driving processes of mutation and recombination impart distinctive features to human genetic datasets. For example, the mutation process coupled with explosive population growth produces an abundance of rare variant alleles, resulting in natural sparsity (Kimura 1983; Takahata 1996; Keinan and A. G. Clark 2012); and the sharing of short genomic segments between unrelated individuals produces a high degree of redundancy (Wall and Pritchard 2003; International HapMap 3 Consortium 2010). In Chapter 3, we leverage these properties to expedite linkage disequilibrium estimation with large data sets.

Linkage disequilibrium (LD) refers to association between alleles at different genetic variants, which generally decays with increasing distance between variants on a given chromosome due to

genetic recombination. Accounting for LD is critical for many multi-variant genetic association methods, e.g. conditional analysis and fine-mapping (Benner et al. 2016; Wen et al. 2016; Lee et al. 2018), gene-based association (Lamparter et al. 2016; Bakshi et al. 2016; Feng et al. 2014), and functional enrichment analysis (Finucane et al. 2015; Lamparter et al. 2016). These methods are often applied to GWAS summary association statistics (single-variant test statistics, or effect size estimates and standard errors), and rely on LD estimates that are pre-computed from the GWAS sample (e.g., Feng et al. 2014) or estimated from an external population-matched reference panel (e.g., Lamparter et al. 2016). Existing tools to estimate LD often scale linearly with sample size, prompting a need for more efficient methods for increasingly large genetic data sets. In Chapter 3, we develop efficient algorithms to estimate LD exploiting sparsity and haplotype structure. We implement our methods in an open-source C++ tool, emeraLD, which is up to two orders of magnitude faster than existing tools while using comparable or less memory.

1.4 Leveraging Prior Knowledge from Functional Genomics Studies for Informative, Comprehensive Gene-Based Analysis

New technologies and large collaborative projects in recent years have produced extensive datasets characterizing functional elements throughout the human genome and regulatory effects of non-coding genetic variation. For example, the NIH Roadmap Epigenomics Mapping Consortium has used next-generation sequencing technologies to produce epigenomic datasets across a number of human tissues and cell types (Kundaje et al. 2015); the Encyclopedia of DNA Elements (ENCODE) project has characterized functional elements throughout the human genome across a variety of

tissues and cell types using ChiP-seq, Methyl-seq, RNA-seq and other techniques (Kellis et al. 2014); the Genotype-Tissue Expression (GTEx) project has produced transcriptomic and eQTL mapping datasets across 49 human tissues (GTEx Consortium 2015); and the FANTOM (Functional ANnoTation Of the Mammalian genome) consortium has used RNA sequencing and Cap Analysis of Gene Expression (CAGE) to characterize active enhancers and promoters across a wide range of human and mouse cell types (De Rie et al. 2017). These projects have greatly enhanced our understanding of regulatory genomics and functional non-coding variation, and have enabled new insights into the mechanisms underlying non-coding GWAS associations.

Integrating functional genomic annotations with GWAS data has been an active area of methodological research (e.g., Gusev et al. 2016; Hao et al. 2018; Wu et al. 2018). For example, stratified LD score regression (S-LDSC) and SMART have been applied to partition complex trait heritability across functional elements and detect functional enrichment using annotations from ENCODE, Roadmap, and other sources (Finucane et al. 2015; Hao et al. 2018). Methods such as TWAS and PrediXcan use predictive weights estimated from eQTL mapping datasets (e.g., from GTEx) to assess associations between complex traits and the genetic component of gene expression levels (Gamazon et al. 2015; Gusev et al. 2016). Similarly, the SMR (summary-data-based Mendelian randomization) and HEIDI (heterogeneity in dependent instruments) methods have been applied to assess gene regulatory perturbations underlying GWAS associations using eQTL and mQTL datasets (Wu et al. 2018). Finally, Bayesian finemapping methods have been developed incorporating functional genomic annotations to help prioritize causal variants and assess mechanisms underlying associations (Kichaev, Yang, et al. 2014; Kichaev and Pasaniuc 2015; Wen et al. 2016).

In Chapter 4, we present GaMBIT, a unified statistical framework to infer causative genes, pathways, and biological mechanisms underlying GWAS associations leveraging diverse functional annotations. Our approach accounts for multiple potential mechanisms underlying GWAS

associations to avoid spurious inferences caused by pleiotropy and LD, and leverages trait-specific patterns of functional enrichment to improve prioritization of causal genes and mechanisms. We discuss relationships between the proposed model and existing gene-based tests and fine-mapping methods, and demonstrate that GaMBIT improves prioritization of causal genes and mechanisms through simulation studies. Finally, we apply our method to analyze twenty-five complex traits using GWAS summary statistics from the UK Biobank resource, and provide an open-source implementation of our methods.

1.5 Looking Forward

Human genomics is a rapidly evolving field, and the emergence of new technologies, studies, and data resources present new statistical and computational challenges. In Chapter 5, we review our work and discuss future prospects and opportunities as new resources emerge.

1.6 References

1KGP Consortium (2015). "A global reference for human genetic variation". In: *Nature* 526.7571, pp. 68–74.

Ahmad, Meraj et al. (2017). "Inclusion of Population-specific Reference Panel from India to the 1000 Genomes Phase 3 Panel Improves Imputation Accuracy". In: *Scientific reports* 7, p. 6733.

Bakshi, Andrew et al. (2016). "Fast set-based association analysis using summary data from GWAS identifies novel gene loci for human complex traits". In: *Scientific reports* 6, p. 32894.

Benner, Christian et al. (2016). "FINEMAP: efficient variable selection using summary data from genome-wide association studies". In: *Bioinformatics* 32.10, pp. 1493–1501.

Das, Sayantan et al. (2016). "Next-generation genotype imputation service and methods". In: *Nature genetics* 48.10, pp. 1284–1287.

De Rie, Derek et al. (2017). "An integrated expression atlas of miRNAs and their promoters in human and mouse". In: *Nature biotechnology* 35.9, p. 872.

Deelen, Patrick et al. (2014). "Improved imputation quality of low-frequency and rare variants in European samples using the 'Genome of The Netherlands'". In: *European Journal of Human Genetics* 22.11, pp. 1321–1326.

ENCODE Project Consortium (2012). "An integrated encyclopedia of DNA elements in the human genome". In: *Nature* 489.7414, pp. 57–74.

Feng, Shuang et al. (2014). "RAREMETAL: fast and powerful meta-analysis for rare variants". In: *Bioinformatics* 30.19, pp. 2828–2829.

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

Gamazon, Eric R et al. (2015). "PrediXcan: Trait Mapping Using Human Transcriptome Regulation". In: *bioRxiv*, p. 020164.

GTEx Consortium (2015). "The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans". In: *Science* 348.6235, pp. 648–660.

Gusev, Alexander et al. (2016). "Integrative approaches for large-scale transcriptome-wide association studies". In: *Nature genetics* 48.3, pp. 245–252.

Hao, Xingjie et al. (2018). "Identifying and exploiting trait-relevant tissues with multiple functional annotations in genome-wide association studies". In: *PLoS genetics* 14.1, e1007186.

International HapMap 3 Consortium (2010). "Integrating common and rare genetic variation in diverse human populations". In: *Nature* 467.7311, pp. 52–58.

KA, Wetterstrand (2018). *DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP)*. [Online; accessed 1-August-2018]. URL: https://www.genome.gov/sequencingcostsdata.

Karczewski, Konrad J and Michael P Snyder (2018). "Integrative omics for health and disease". In: *Nature Reviews Genetics* 19.5, p. 299.

Keinan, Alon and Andrew G Clark (2012). "Recent explosive human population growth has resulted in an excess of rare genetic variants". In: *science* 336.6082, pp. 740–743.

Kellis, Manolis et al. (2014). "Defining functional DNA elements in the human genome". In: *Proceedings of the National Academy of Sciences* 111.17, pp. 6131–6138.

Kichaev, Gleb and Bogdan Pasaniuc (2015). "Leveraging functional-annotation data in trans-ethnic fine-mapping studies". In: *The American Journal of Human Genetics* 97.2, pp. 260–271.

Kichaev, Gleb, Wen-Yun Yang, et al. (2014). "Integrating functional data to prioritize causal variants in statistical fine-mapping studies". In: *PLoS genetics* 10.10, e1004722.

Kimura, Motoo (1983). *The neutral theory of molecular evolution*. Cambridge University Press.

Kundaje, Anshul et al. (2015). "Integrative analysis of 111 reference human epigenomes". In: *Nature* 518.7539, p. 317.

Lamparter, David et al. (2016). "Fast and rigorous computation of gene and pathway scores from SNP-based summary statistics". In: *PLoS computational biology* 12.1, e1004714.

Lee, Yeji et al. (2018). "Bayesian Multi-SNP Genetic Association Analysis: Control of FDR and Use of Summary Statistics". In: *bioRxiv*, p. 316471.

Lencz, Todd et al. (2017). "High-depth whole genome sequencing of a large population-specific reference panel: Enhancing sensitivity, accuracy, and imputation". In: *bioRxiv*, p. 167924.

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

Li, Na and Matthew Stephens (2003). "Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data". In: *Genetics* 165.4, pp. 2213–2233.

Li, Yun et al. (2009). "Genotype imputation". In: *Annual review of genomics and human genetics* 10, pp. 387–406.

MacArthur, Jacqueline et al. (2016). "The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog)". In: *Nucleic acids research* 45.D1, pp. D896–D901.

McCarthy, Shane et al. (2016). "A reference panel of 64,976 haplotypes for genotype imputation". In: *Nature genetics* 48.10, p. 1279.

Rangarajan, Savita et al. (2017). "AAV5–factor VIII gene transfer in severe hemophilia A". In: *New England Journal of Medicine* 377.26, pp. 2519–2530.

Ribeil, Jean-Antoine et al. (2017). "Gene therapy in a patient with sickle cell disease". In: *New England Journal of Medicine* 376.9, pp. 848–855.

Roshyara, Nab Raj and Markus Scholz (2015). "Impact of genetic similarity on imputation accuracy". In: *BMC genetics* 16.1, p. 90.

Takahata, Naoyuki (1996). "Neutral theory of molecular evolution". In: *Current opinion in genetics & development* 6.6, pp. 767–772.

UK10K Consortium (2015). "The UK10K project identifies rare variants in health and disease". In: *Nature* 526.7571, pp. 82–90.

Wall, Jeffrey D and Jonathan K Pritchard (2003). "Haplotype blocks and linkage disequilibrium in the human genome". In: *Nature Reviews Genetics* 4.8, pp. 587–597.

Wen, Xiaoquan et al. (2016). "Efficient integrative multi-SNP association analysis via deterministic approximation of posteriors". In: *The American Journal of Human Genetics* 98.6, pp. 1114–1129.

Wu, Yang et al. (2018). "Integrative analysis of omics summary data reveals putative mechanisms underlying complex traits". In: *Nature communications* 9.1, p. 918.

Chapter 2

Sequencing and Imputation in GWAS: Cost-Effective Strategies to Increase Power and Genomic Coverage Across Populations

2.1 Introduction

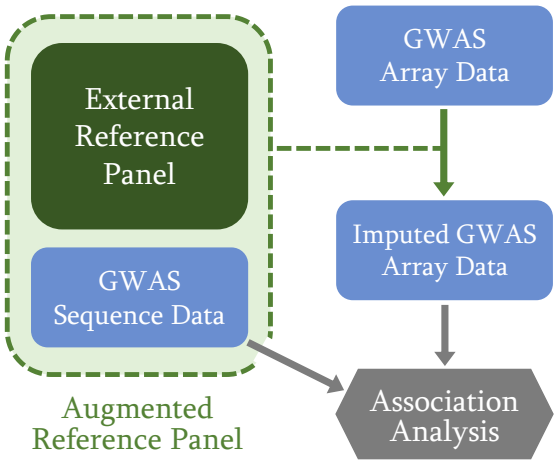
Genome-wide association studies (GWAS) have detected thousands of common genetic variants associated with hundreds of complex diseases and traits (MacArthur et al. 2016). A key aim for the next wave of GWAS is to interrogate the full spectrum of genetic variation underlying human genetic traits, including rare (minor allele frequency [MAF] $\leq$ 0.5%) variants. Detecting association at rare variants requires both more comprehensive genomic coverage and sufficient sample size. Deep whole genome sequencing (WGS) is the gold standard method for capturing rare variation; however, even in the era of the $1,000 genome, large WGS association studies remain prohibitively expensive. Genotype imputation has been a mainstay of GWAS, providing increased genomic coverage from inexpensive array-based genotype call sets. While initial imputation studies only surveyed common variants (e.g., Scott et al. 2007), larger and more diverse reference panels now enable more accurate and comprehensive imputation of rare and low-frequency variants across a wide range of populations (e.g., Mahajan et al. 2018).

Imputation algorithms model haplotypes in the study sample as mosaics of haplotypes in a reference panel (e.g. from the International HapMap Project [International HapMap 3 Consortium 2010] or 1000 Genomes Project [1KGP Consortium 2015]) to predict genotypes at untyped variants (Li et al. 2009). By increasing genomic coverage and accuracy, imputation increases statistical power to detect association, enables more complete meta-analysis of results from multiple studies, and facilitates the identification of causal variants through fine-mapping (Li et al. 2009; Das et al. 2016). Imputation coverage and accuracy depend crucially on the size of the reference panel and the genetic distance between reference and target populations (Li et al. 2009; Roshyara and Scholz 2015). The largest current broadly available reference panels, e.g. from the Haplotype Reference Consortium (S. McCarthy et al. 2016) (HRC) and UK10K project (UK10K Consortium 2015), include tens of thousands of predominantly European individuals. These panels provide near complete imputation of genetic variation down to MAF=0.1% for many European populations, but lower imputation quality for non-European and admixed populations and population isolates, particularly for rare and low-frequency (0.5%< MAF < 5%) variants (Deelen et al. 2014; Pistis et al. 2015; Zhou et al. 2017). The 1000 Genomes Project and HapMap panels include individuals from diverse worldwide populations, but provide more limited imputation coverage and accuracy due to their smaller sample sizes.

Capturing rare variation across diverse populations is crucial to detect population differences in genetic risk factors, accurately predict genetic risk, and identify causal variants and biological mechanisms through trans-ethnic fine-mapping (Kichaev and Pasaniuc 2015; Popejoy and Fullerton 2016). Population-matched or multi-ethnic reference panels can improve imputation quality and coverage for rare variants in GWAS of diverse populations (Deelen et al. 2014; Pistis et al. 2015; Zhou et al. 2017; Ahmad et al. 2017; Lencz et al. 2017; Van Leeuwen et al. 2015); this approach has enabled discovery of novel loci and refinement of association signals for multiple populations and complex traits (Pistis et al. 2015; Auer and Lettre 2015; Holm et al. 2011).

Here, we consider an approach in which a subset of study participants is whole genome sequenced and the rest are array-genotyped and imputed using an augmented reference panel that comprises the sequenced participants and individuals from an external reference panel (Hu et al. 2015; Zeggini 2011). This hybrid sequencing-and-imputation strategy provides more comprehensive coverage than only array genotyping, and is less costly than whole genome sequencing the entire sample. We and others have used this strategy (Van Leeuwen et al. 2015; Fuchsberger et al. 2016; Sidore et al. 2015; Steinthorsdottir et al. 2014), but no analysis of coverage, power, and cost-effectiveness has been carried out to date. Here, we assess how imputation coverage and power to detect association vary across genotyping arrays and as a functions of the number of population-matched individuals sequenced and included in the reference panel for two admixed populations (African Americans and Latino Americans) and two European population isolates (Sardinians and Finns) to identify powerful and cost-effective strategies for GWAS in these populations. We also describe an interactive web-based tool to assist researchers in the design and planning of their own GWAS.

Figure 2.1.1: Sequencing-and-Imputation GWAS Flowchart



Flowchart of sequencing and imputation GWAS strategy. An augmented reference panel that comprises sequenced GWAS participants and an external panel is used to impute array-genotyped GWAS samples. Both sequenced and

imputed GWAS participants are included in association analysis.

2.2 Materials and Methods

We first describe WGS data sources used in our analysis. Next, we describe imputation strategies, and outline procedures and imputation quality metrics to compare these strategies. Finally, we present a novel method to estimate power for the sequencing-only, imputation-only, and sequencing-and-imputation strategies. For ease of presentation, we assume a dichotomous trait and a multiplicative disease model, although our findings generalize easily to continuous traits and other genetic models.

2.2.1 Data Resources

We used WGS data on 3,412 African Americans (participants from the Jackson Heart Study) and 2,068 Latino Americans (participants of Puerto Rican and Mexican descent from the GALA II study and Costa Rican descent from the Genetic Epidemiology of Asthma in Costa Rica and CAMP studies) in the National Heart, Lung, and Blood Institute (NHLBI) Trans-Omics for Precision Medicine (TOPMed) WGS program, and on 2,995 Finns (participants of the GoT2D, 1KGP, SISu, and Kuusamo studies) and 3,445 Sardinians (participants of the SardiNIA study) in the HRC to compare imputation quality between reference panel configurations and genotyping arrays.

2.2.2 Procedures to Evaluate Imputation Coverage and Accuracy

We considered three imputation strategies: (1) using sequenced study participants as a study-specific reference panel, (2) using an external reference panel alone (for this comparison, the HRC or HRC subset excluding individuals from the target population), and (3) using an augmented panel that

comprises sequenced study participants and an external panel.

For African Americans, which are underrepresented in the current version 1.1 of the HRC, we constructed population-specific and HRC-augmented reference panels with 0 to 2,000 African Americans. For Latino Americans, we used the same approach but restricted the study-specific panel size to <1,500 due to the more limited available sample of sequenced Latino American individuals. For Finns and Sardinians, which are present in the HRC, we constructed augmented reference panels that comprised the 29,470 non-Finnish or 29,020 non-Sardinian individuals in the HRC together with 0 to 2,000 Finns or Sardinians from the HRC.

For each population, each imputation strategy, and each of three commonly-used genotyping arrays (Table 2.1), we used sequence-based genotype calls at marker variants present on the array as a scaffold for imputation using Minimac3, masking the remaining sequence-based genotype calls (Das et al. 2016). We then compared the imputed genotype dosages to the true (masked) genotypes to estimate (a) imputation r^2, the squared Pearson correlation between true genotype and imputed dosage, and (b) imputation coverage, the proportion of variants with imputation $r^2 \geq 0.3$ and minor allele count (MAC) ≥ 5 (the MAC threshold used by the HRC panel [S. McCarthy et al. 2016]) in the reference panel.

Table 2.1: Genotyping Arrays Used for Comparisons

Array	No. Marker Variants	List Cost per Sample (Illumina Inc. 2018)
Illumina Infinium Core	307K	$49
Illumina Infinium OmniExpress	710K	$94
Illumina Infinium Omni2.5	2.5M	$172

2.2.3 Estimating Power to Detect Association using Empirical Imputation Quality Data

When sequenced individuals are included in the reference panel, power calculations should account for the interdependence between imputation r^2 and the number of participants sequenced n, and for the possibility that the variant is not imputable (absent in the reference panel or not imputed due to insufficient MAC, or filtered prior to association analysis due to imputation r^2 falling below a given threshold). While common variant associations are likely to be captured by LD proxy SNPs even when the causal variant is not directly genotyped or imputed, rare variant associations are much less likely to be captured by proxy SNPs (Montpetit et al. 2006). Here, we assume that power to detect association for variants that are not imputable is zero. This assumption affects power calculations almost exclusively for rare variants, since common variants are almost uniformly imputable with large reference panels (Das et al. 2016; S. McCarthy et al. 2016).

We assume that the n participants who are sequenced are randomly subsampled from the overall sample of $n+m$ study participants, and that test statistics are calculated separately for the sequenced and imputed subsamples and combined using the effective sample size weighted meta-analysis test statistic $Z_{nm} = c_{nm}^{1/2} Z_n^{\text{seq}} + (1-c_{nm})^{1/2} Z_m^{\text{imp}}$, where $c_{nm} = n/(n+r^2 m)$. The asymptotic distribution of $Z_{nm} - \eta\sqrt{n + r^2 m}$ is normal with mean 0 and variance 1, where r^2 is the squared correlation between imputed dosages and true genotypes, and η is an effect size parameter which is equal to 0 under the null hypothesis of no association. The form of η depends on the association model (additive, dominant, multiplicative), relative risk or odds ratio, MAF, and population prevalence and case-control ratio for binary traits. Under an arbitrary association model for binary traits, we can write

$$\eta = \frac{2(p_{\text{case}} - p_{\text{control}})}{\sqrt{(1+s)\left(v_{\text{case}} + \frac{1}{s}v_{\text{control}}\right) + 4\left(p_{\text{case}} - p_{\text{control}}\right)^2}}$$

where p_{case} and p_{control} are the alternate allele frequencies in the disease-positive and disease-negative populations, v_{case} and v_{control} are the variances of genotypes in the disease-positive and disease-negative populations, and s is the GWAS case-control ratio.

To estimate power while accounting for variability in imputation r^2 and the possibility that a variant is not imputable, we average empirical imputation r^2 values and MACs across variants from experiments with real data described in the previous section. Specifically, we estimate power to detect association when n individuals are sequenced and m are genotyped and imputed as

$$\hat{\text{Power}}(m,n) = \frac{1}{\sum_j w_j} \sum_j w_j C_{nj} \int_{-z_{1-\alpha/2}}^{z_{1-\alpha/2}} \phi\left(u - \eta\sqrt{n + r_{nj}^2 m}\right) du$$

where $\phi(u) = e^{-\frac{u^2}{2}}/\sqrt{2\pi}$ is the standard normal density function, $z_{1-\alpha/2}$ is the α-level significance threshold, r_{nj}^2 is the imputation r^2 value for the j^{th} variant, $C_{nj} = I(MAC_{nj}^{\text{panel}} \geq 5, r_{nj}^2 \geq 0.3)$ is an indicator equal to 1 if the j^{th} variant was imputable and 0 otherwise, and MAC_{nj}^{panel} is the reference panel MAC for the j^{th} variant when the n sequenced individuals from the target population were included in the reference panel.

We define the weights $w_j = P_N^{\text{GWAS}}(\hat{p}_j)/\hat{P}_N(\hat{p}_j)$, where N is the total number of samples used in our analysis for the given population (e.g. N =3,412 for African Americans), $\hat{p}_j$ is the sample MAF for the j^{th} variant in the total sample, $\hat{P}_N(x)$ is the proportion of variants with MAF = x, and $P_N^{\text{GWAS}}(x)$ is the probability of observing sample MAF = x in a sample of size N given the specified association model. For example, in a GWAS with sample size N and case-control ratio s, the sample MAC (which is equal to $2N\hat{p}$, where $\hat{p}$ is the sample MAF) is approximately Poisson distributed with mean $2N(sp_{\text{case}} + p_{\text{control}})/(s+1)$, where $p_{\text{case}} = p\gamma/(1+p(\gamma-1))$ and $p_{\text{control}} = (p - Kp_{\text{case}})/(1-K)$ for

a variant with population MAF p and relative risk γ for a disease with prevalence K. This weighting approach adjusts for differences between the empirical distribution of MACs across variants in real data, and the theoretical MAC distribution for a variant with the specified MAF, effect size, prevalence in a GWAS with sample size N and case-control ratio s.

2.3 Results

First, we compare strategies to improve imputation using study-specific WGS data for African Americans, Latino Americans, Sardinians, and Finns. Next, we assess the effects of genotyping array on imputation quality and coverage for each population and reference panel. We then use these results to estimate statistical power to detect association as a function of study-specific panel size, number of participants imputed, external reference panel, and genotyping array. Finally, we identify cost-effective study designs by comparing statistical power and total experimental (sequencing and genotyping) costs for sequencing-only, imputation-only, and sequencing-and-imputation GWAS designs for each population and genotyping array.

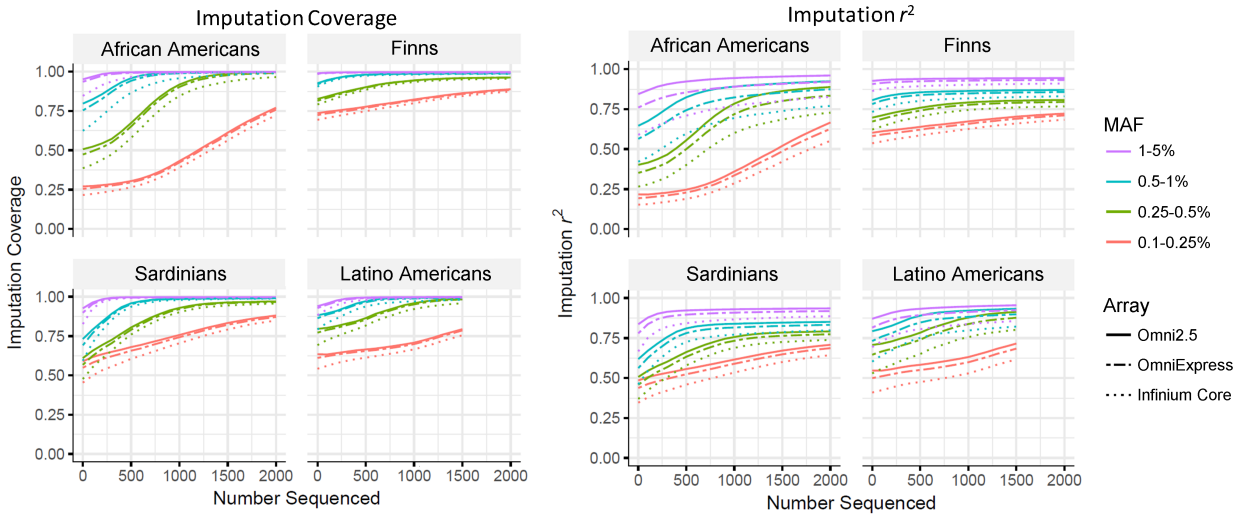
2.3.1 Strategies to Improve Imputation using Study-Specific WGS Data

We compared imputation r^2 and coverage (proportion of variants with imputation r^2>0.3 and reference MAC $\geq$ 5) for three imputation strategies: (1) using an external panel (the HRC or HRC subset) alone, (2) using an augmented panel that combines the study-specific and external panels, and (3) using a study-specific panel alone.

The external panel alone (HRC for Latino Americans and African Americans, and HRC subset that excludes individuals from the target population for Finns and Sardinians) provided 96% imputation coverage for MAF $\geq$ 0.25% variants (where MAF is calculated separately within each population)

for Finns, 84% coverage for Sardinians, 86% coverage for Latino Americans, and 77% coverage for African Americans (Figure 2.3.1A). The relatively lower coverage for African Americans is expected since the HRC consists primarily of Central and Northern Europeans, who are genetically closer to Finns and Sardinians, and includes relatively few Africans or African Americans. Despite the small number of Latino or Native Americans included in the HRC, imputation coverage was slightly higher for Latino Americans than for Sardinians. This may reflect the high degree of European admixture in many Latino American populations (Bryc et al. 2010), and the abundance of population-specific rare and low-frequency variants in the Sardinian population (Sidore et al. 2015).

Figure 2.3.1: Imputation Quality by Population and Genotyping Array.



Imputation and coverage as a function of number of population-matched individuals included in augmented reference panels (Number Sequenced). Here and elsewhere, MAF is calculated separately within each population.

Augmenting an external reference panel with even a relatively small number of sequenced individuals substantially increased coverage, particularly for African Americans and Sardinians and for variants with lower MAF. For example, augmenting the external panel with 500 sequenced individuals from the study population improved overall imputation coverage for MAF=0.25-0.5%

variants by 4% for Finns, 9% for Latino Americans, 16% for African Americans, and 23% for Sardinians genotyped using the OmniExpress relative to the external panel alone (Figure 2.3.1A). Similarly, augmenting the external reference panel with even 200 individuals increased imputation coverage for =0.1-0.25% variants by 3%, 4%, 6%, 10% relative to the external panel alone for Finns, Latino Americans, African Americans, and Sardinians, respectively.

With 2,000 individuals from the target population (or 1,500 for Latino Americans), population-specific panels provided roughly equivalent imputation r^2 compared to augmented panels (Supplementary Figure 2.1A); however, augmented panels provided higher imputation coverage overall for low MAF variants (Supplementary Figure 2.1B). For example, augmented panels with 2,000 individuals from the target population (or 1,500 for Latino Americans) provided 86%, 80%, 79%, and 86% coverage for 0.1-0.25% MAF variants for Finns, Latino Americans, African Americans, and Sardinians respectively, whereas population-specific panels alone provided 72%, 51%, 78%, and 72% coverage using the Omni Express array. However, imputation coverage for variants with MAF>0.25% differed by <1% between augmented and population-specific panels with 2,000 individuals from the target population (or 1,500 for Latino Americans) for all populations and genotyping arrays. When only a small number (less than 500) of individuals from the target population are sequenced, augmented reference panels provided substantially higher imputation coverage and r^2 than population-specific panels alone. For example, augmented panels with 500 individuals from the target population provided 90%, 85%, 65%, and 85% coverage for 0.25-0.5% MAF variants for Finns, Latino Americans, African Americans, and Sardinians respectively, whereas population-specific panels of 500 individuals provided <30% coverage using the Omni Express array.

Even very rare variants (MAF=0.1-0.25%) attained high coverage across all populations given a sufficient number of population-matched individuals in the reference panel. For example, attaining >70% imputation coverage for MAF=0.1-0.25% variants required a study-specific panel of >1,800

individuals for African Americans, 1,000 for Latino Americans, 700 for Sardinians, and 0 for Finns using the OmniExpress. These increases in imputation coverage primarily reflect increasing numbers of population-specific variants captured in the reference panel, which are absent from or present in low copy number in the external panel.

We also assessed potential drawbacks of augmented reference panels relative to population-specific panels. Using a reference panel that includes individuals outside of the target population can result in false positive imputed variants: variants that are monomorphic in the target population but have imputed MAC $\gg 0$ and high imputation quality metrics (see also Zeggini 2011). We found that many false positive variants can be identified and filtered by comparing allele frequencies between the sequenced and imputed samples from the target population (Supplementary Figure 2.2).

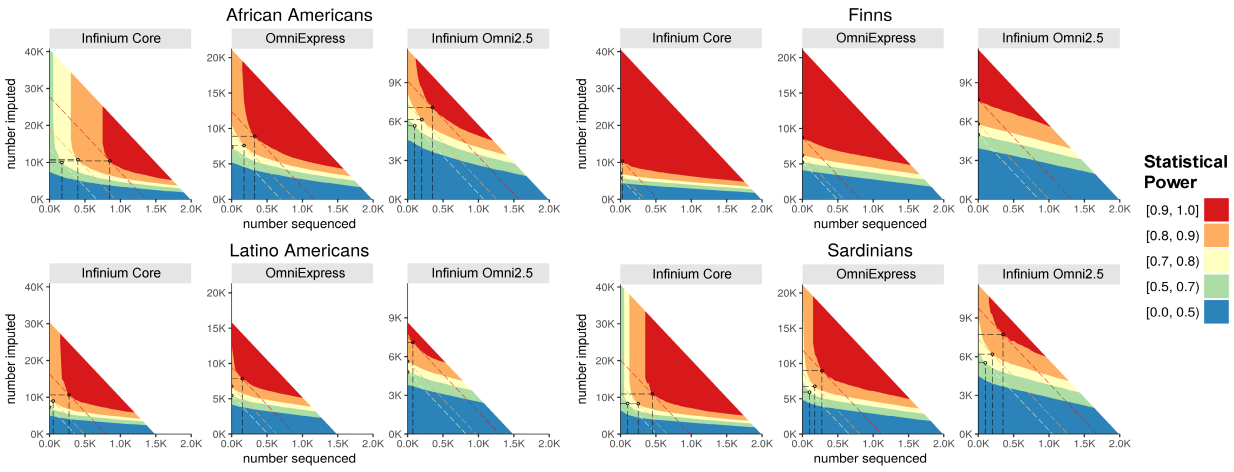
2.3.2 Imputation Coverage and Quality across Genotyping Arrays

Imputation coverage was generally similar for the OmniExpress and Omni2.5 arrays, but consistently lower for the less dense Core array. Coverage differed by <7% between the OmniExpress and Omni2.5 across all MAF bins, populations, and reference panels, whereas the Core provided up to 24% lower coverage than the Omni2.5 (Figure 2.3.1A). Imputation coverage was more heterogeneous across arrays for populations with greater genetic distance from the external reference panel (e.g., African Americans and the HRC panel), particularly with smaller (or absent) study-specific panels (Figure 2.3.1A). Because we used the same reference panels for each genotyping array, differences in imputation coverage between arrays are solely due to differences in the proportion of variants that attained imputation $r^2 \geq 0.3$. Imputation r^2 varied more across genotyping arrays than did imputation coverage (Figure 2.3.1B versus 2.3.1A); however, the magnitude of differences in imputation r^2 between arrays was still generally modest, particularly for the Finns and Sardinians.

2.3.3 Powerful and Cost-Effective Strategies for GWAS across Populations

We compared the cost-effectiveness of sequencing-only, imputation-only, and sequencing-and-imputation strategies by analyzing statistical power to detect association as a function of numbers of study participants sequenced and imputed, genotyping array, and reference panel across a range of genetic models. Here, we define the most *cost-effective* strategy as either (1) minimizing total experimental (sequencing and genotyping) cost while attaining power at or above a given threshold, or equivalently (2) maximizing power while maintaining cost no greater than a specified constraint.

Figure 2.3.2: Power and Optimal Design by Population and Genotyping Array.

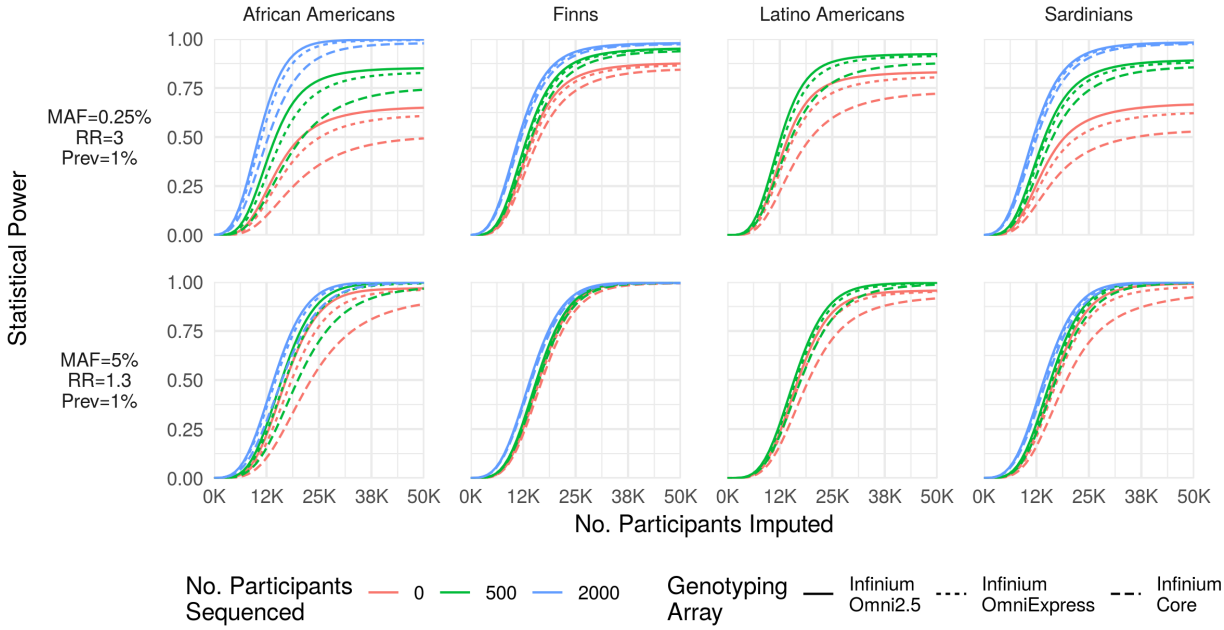


Power to detect association for case-control studies with equal numbers of cases and controls as a function of sequenced subsample size (x-axis) and imputed subsample size (y-axis) for a variant with MAF 0.5% and relative risk 4 for a disease with prevalence 1%. Axes are scaled to reflect costs of genotyping arrays (Table 2.1) and sequencing ($1K per sample). Power shown only for designs with total genotyping cost $2M ($1.5M for Latino Americans). Dashed diagonal lines indicate study designs with the same total cost, given by $y = a - bx$ where $a = (Total\ Cost)/(Array\ Cost)$ and $b = (Sequencing\ Cost)/(Array\ Cost)$. Circled points indicate optimal study designs, which attain the indicated power level at minimum total experimental cost (or, maximize power at the indicated total experimental cost).

The cost-effectiveness of sequencing a subset of study participants varied greatly across populations. For Finns, imputation-only designs were most powerful to detect association and adding

sequenced individuals increased power only minimally, even for low-frequency and rare variants. For Sardinians, Latino Americans, and African Americans, sequencing a subset of study participants was optimal, and often achieved substantially greater power than imputation-only or sequencing-only studies. For example, a GWAS of African Americans with equal numbers of cases and controls in which 800 participants are sequenced and 10,500 are imputed using the Illumina Infinium Core array has 90% power to detect a risk variant with MAF = 0.5% and RR = 4 for a disease with prevalence 1%, whereas an imputation-only GWAS with the same total cost (27,000 participants) has power <70% (Figure 2.3.2). Even for populations in which optimal sequencing-and-imputation designs had substantially greater power than imputation-only, the optimal number to sequence was often modest. For example, only 175 participants are sequenced under the optimal design using the Illumina OmniExpress to attain 80% power in the previous example (Figure 2.3.3). This is expected because even a relatively small study-specific panel can substantially increase imputation coverage (Figure 2.3.1A).

Figure 2.3.3: Power as a Function of Minor Allele Frequency and Effect Size.



Statistical power (y-axis) to detect a rare large-effect variant (MAF=0.25%, RR=3; top row) and common modest-effect variant (MAF=5%, RR=1.3; bottom row) for a disease with prevalence 1% as a function of the number of participants array-genotyped and imputed (x-axis) when 0, 500, or 2,000 participants are sequenced and included in an augmented reference panel. The number of participants sequenced has a far greater impact on statistical power for the rare variant association. Importantly, statistical power is bounded above by the probability that the variant is imputable ($r^2 > 0.3$ and reference $MAC \geq 5$), causing power to asymptote below 1 as a function of the number of imputed participants (e.g., upper-left panel).

2.3.4 Denser Genotyping Arrays vs. Sequencing: Which is More Cost-Effective to Increase Power?

Imputation coverage and power to detect association can be increased by using denser genotyping arrays, which provide a more informative framework for imputation, or by sequencing population-matched individuals and adding them to the reference panel. We assessed the cost-effectiveness of these two strategies by comparing power to detect association across genotyping

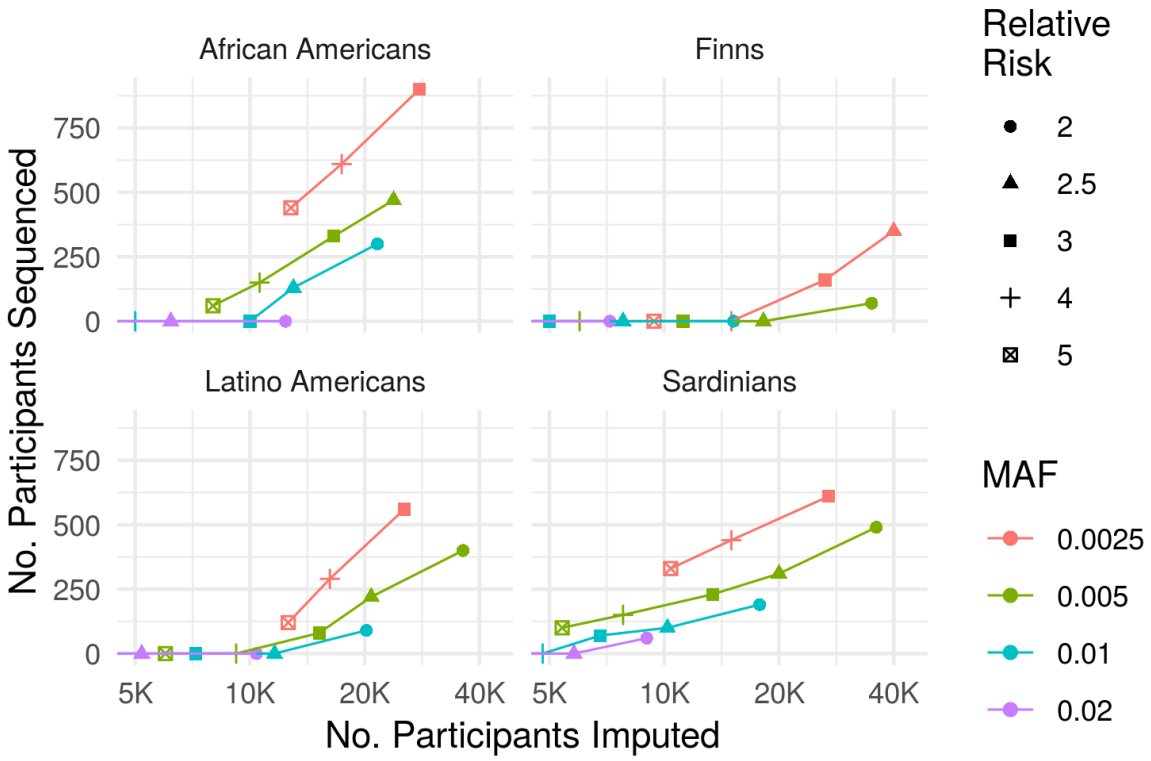
arrays for study designs that have the same total cost assuming $1000 for WGS and current list prices for genotyping arrays (Table 2.1). As expected, the optimal number of participants sequenced to maximize power given fixed total cost generally decreased with increasing array density. For example, the optimal number sequenced to maximize power to detect association was 625, 275, and 100 for the Infinium Core, OmniExpress, and Omni2.5 respectively for Sardinians given a risk variant with RR = 2, MAF = 1%, and disease prevalence 1%. Power to detect association under the optimal design given a fixed total cost was generally greater for sparser arrays; in the previous example, power under the optimal design was 97%, 88%, and 54% for the Infinium Core, OmniExpress, Omni2.5.

We also compared optimal designs to attain power above a given threshold at minimum total cost across genotyping arrays based on the per-sample genotyping costs reported in Table 2.1. Generally, sparser arrays were more cost-effective (reached the power threshold with lower total cost) than dense arrays. In fact, the sparsest genotyping array in our analysis, the Infinium Core, was most cost-effective across all disease models and populations apart from African Americans, for whom the Infinium OmniExpress was often most cost-effective. This last result is unsurprising given the substantial difference in imputation coverage between the Infinium Core and Omni arrays for African Americans (Figure 2.3.1B). Importantly, our analysis assumes 1) a direct trade-off between the GWAS sample size and sequencing/genotyping costs, and 2) no additional costs per GWAS sample other than sequencing/genotyping. Under these assumptions, we found that denser arrays are generally less cost-effective than sparser arrays; however, denser arrays provide higher imputation coverage given a fixed GWAS sample size.

2.3.5 Optimal Study Design as a Function of Minor Allele Frequency and Effect Size

Power to detect association under a given study design depends on MAF, effect size (relative risk or odds ratio), and population prevalence (Sham and Purcell 2014). These parameters also influence the relative cost-effectiveness of sequencing and imputation. While common variants can be accurately imputed with small reference panels, large population-matched reference panels are needed to capture rare (population-specific) variants. In Figure 2.3.4, we illustrate the impact of sequencing on statistical power for two combinations of MAF and effect size in each of the four study populations.

Figure 2.3.4: Optimal Design as a Function of Minor Allele Frequency and Effect Size.



Optimal numbers of participants sequenced (y-axis) and imputed (x-axis; using the Illumina OmniExpress array) to

attain statistical power 80%.

The optimal number of study participants sequenced to attain $\geq$80% power to detect association at minimum total cost increases with decreasing MAF (Figure 2.3.4). This is expected, since larger reference panels are needed to capture variants with lower frequency. In addition, the total cost required to attain $\geq$80% power increases with decreasing MAF, which is expected given that power to detect association for a given sample size decreases with decreasing MAF. The optimal number to sequence to attain $\geq$80% power decreases with increasing effect size magnitude. This is expected, since the expected number of risk alleles captured in the reference panel increases with effect size magnitude.

2.4 Discussion

While the cost of genome sequencing has fallen dramatically ((KA, 2018)), large genome sequencing studies remain prohibitively expensive. Large reference panels are now enabling accurate imputation of even very rare variants (S. McCarthy et al. 2016; Zhou et al. 2017; Mahajan et al. 2018), making imputation-based GWAS viable and cost-effective for detecting associations across the allele frequency spectrum. For populations with limited reference panel data, we have shown that sequencing a subset of study participants can substantially increase imputation coverage and accuracy, particularly for rare and population-specific variants, at a fraction of the cost of sequencing the entire study cohort. Our results also suggest that it is almost always advantageous to augment existing reference panels, except when the study-specific panel is large or the target population has high genetic distance from the external panel.

Complementary sequencing-and-imputation GWAS strategies have been applied to refine association signals and discover novel associations for several populations and complex traits (Pistis

et al. 2015; Auer and Lettre 2015; Holm et al. 2011). While most sequencing-and-imputation studies to date have been carried out in European isolated populations, our results suggest that this strategy can also be powerful and cost-effective for admixed and non-European populations. In addition to increasing genomic coverage and power to detect association for the study itself, sequencing a subset of study participants provides a data resource that can be used to enhance imputation in future studies of the same or related populations.

Directly augmenting an existing reference panel with study-specific sequence data is not always feasible due to technical, logistical, and privacy constraints. However, we and others have found that the distributed reference panel approach (separately imputing with two or more reference panels and combining the results) provides nearly equivalent imputation quality (Supplementary Figure 2.3). Thus, study-specific WGS data can be used to improve imputation even when directly augmenting an external panel is not feasible.

While large reference panels enable accurate imputation across a wide range of the allele frequency spectrum (S. McCarthy et al. 2016; Zhou et al. 2017), the extent of genetic variation that can be captured through imputation is limited relative to WGS. For example, *de novo* mutations cannot be imputed regardless of reference panel size. This is particularly salient for monogenic disorders; for example, over 80% of achondroplasia cases occur from recurrent *de novo* mutations in *FGFR3*(Bellus et al. 1995). Thus, imputation may be unable to detect causative alleles for traits with extreme genetic architectures, even with very large reference panels.

As increasingly large and diverse sequencing projects are conducted, larger and more diverse reference panels will become available. In the design and planning of GWAS, it may be prudent to consider resources under development and pending release in addition to resources that are currently available. More broadly, our analysis highlights the utility of collaboration and coordination across institutions for effective study design and resource allocation. For example, the optimal design to

maximize power in an individual study does not necessarily maximize meta-analysis power across multiple studies of the same trait and population.

Our analysis of cost-effectiveness and optimal design depends crucially on the relative per-sample costs of sequencing and array genotyping. Both sequencing and genotyping costs have fallen markedly in recent years, and are likely to continue to do so. Depending on the relative rates of change, cost-effectiveness and optimal design also may change. In addition, the cost of participant recruitment and DNA sample collection may alter the relative cost-effectiveness of sequencing and genotyping. Finally, our cost-effectiveness analysis assumes that sample size is unconstrained, and may not apply for small populations or rare diseases. While our results are illustrative, investigators may wish to explore questions of the relative cost-effectiveness of sequencing and array genotyping strategies in the context of their own study and relevant assumptions about population, reference panels, and sequencing and array genotyping costs. To enable this exploration, we have developed a flexible, easy-to-use software tool, APSIS (Analysis of Power for Sequencing and Imputation Studies), to analyze power and identify optimal study designs while accounting for imperfect imputation coverage and accuracy.

2.4.1 Conclusions

Here, we assessed the genomic coverage, statistical power, and cost-effectiveness of sequencing and imputation-based designs for GWAS in a variety of populations and a range of genetic models. We developed a novel method to account for available reference haplotype data in power calculations using empirical data, which can be applied to inform GWAS planning and design. For European populations that are well-represented in current reference panels, our results suggest that imputation-based GWAS is cost-effective and well-powered to detect both common- and rare-variant associations. For populations with limited representation in current reference panels, we found that

sequencing a subset of study participants can substantially increase genomic coverage and power to detect association, particularly for rare and population-specific variants. Our results also suggest that larger and more diverse reference panels will be important to facilitate GWAS in global populations.

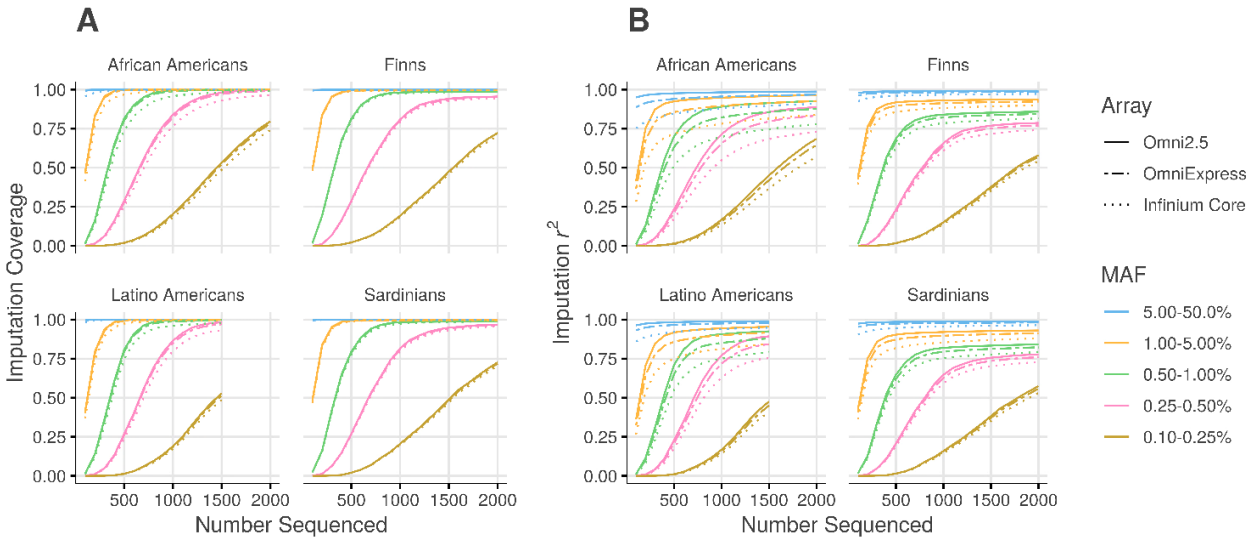
2.5 Acknowledgments

I thank our collaborators Pramod Anugu,[2] Solomon Musani,[2] Scott T. Weiss,[3,4,5] Esteban G. Burchard,[6,7] Marquitta J. White,[6] and Kevin L. Keys[6] for the opportunity to analyze WGS data from the NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium[†]. I thank Francesco Cucca[8,9] and Carlo Sidore[8] for the opportunity to analyze WGS data from the Sardinian cohorts of the Haplotype Reference Consortium (HRC). I also thank the staffs and participants of the Jackson Heart Study, the Genetics of Asthma in Costa Rica project, the Genes-Environments and Admixture in Latino Americans (GALA II) Study, TOPMed WGS Project, and HRC. Finally, I thank Michael Boehnke[1] and Christian Fuchsberger[1,10,11], who jointly supervised this project.

[1] Department of Biostatistics and Center for Statistical Genetics, University of Michigan School of Public Health, Ann Arbor, MI
[2] University of Mississippi Medical Center, Jackson, MS
[3] Harvard Medical School, Boston, MA
[4] Channing Department of Network Medicine, Brigham and Women's Hospital, Boston, MA
[5] Partners HealthCare Personalized Medicine, Boston, MA
[6] Department of Medicine, University of California San Francisco, San Francisco, California
[7] Department of Bioengineering and Therapeutic Sciences, University of California San Francisco, San Francisco, California
[8] Istituto di Ricerca Genetica e Biomedica (IRGB), CNR, Monserrato, Italy
[9] Dipartimento di Scienze Biomediche, Università di Sassari, Sassari, Italy
[10] Division of Genetic Epidemiology, Department of Medical Genetics, Molecular and Clinical Pharmacology, Medical University of Innsbruck, Innsbruck, Austria
[11] Institute for Biomedicine, Eurac Research, Affiliated Institute of the University of Lübeck, Bolzano, Italy
† A complete list of NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium co-investigators is available online.

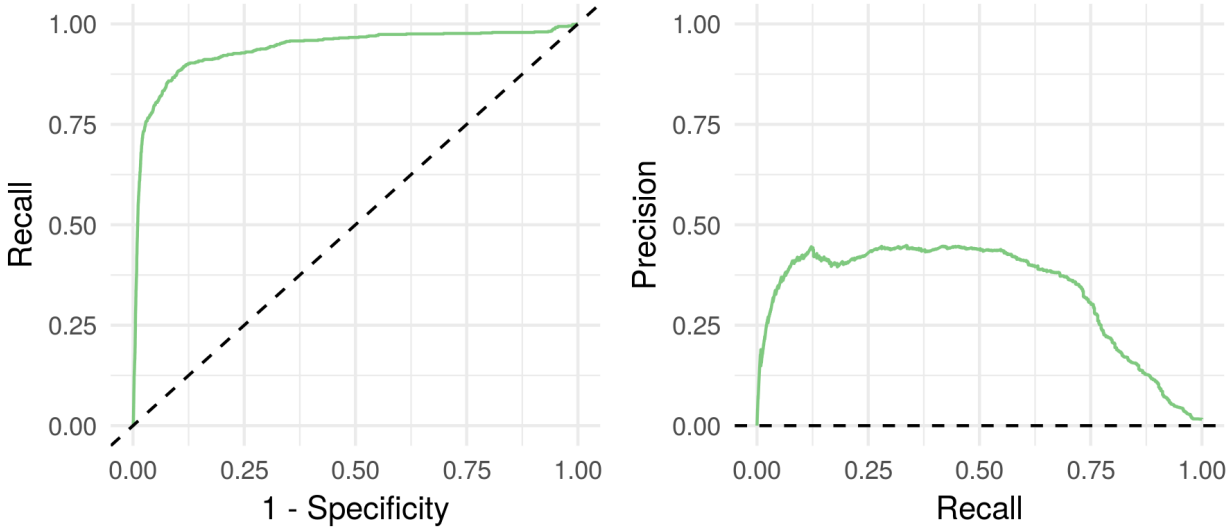
2.6 Appendix: Supplementary Figures

Supplementary Figure 2.1: Imputation Coverage and r^2 as Functions of Population-Specific Reference Panel Size.



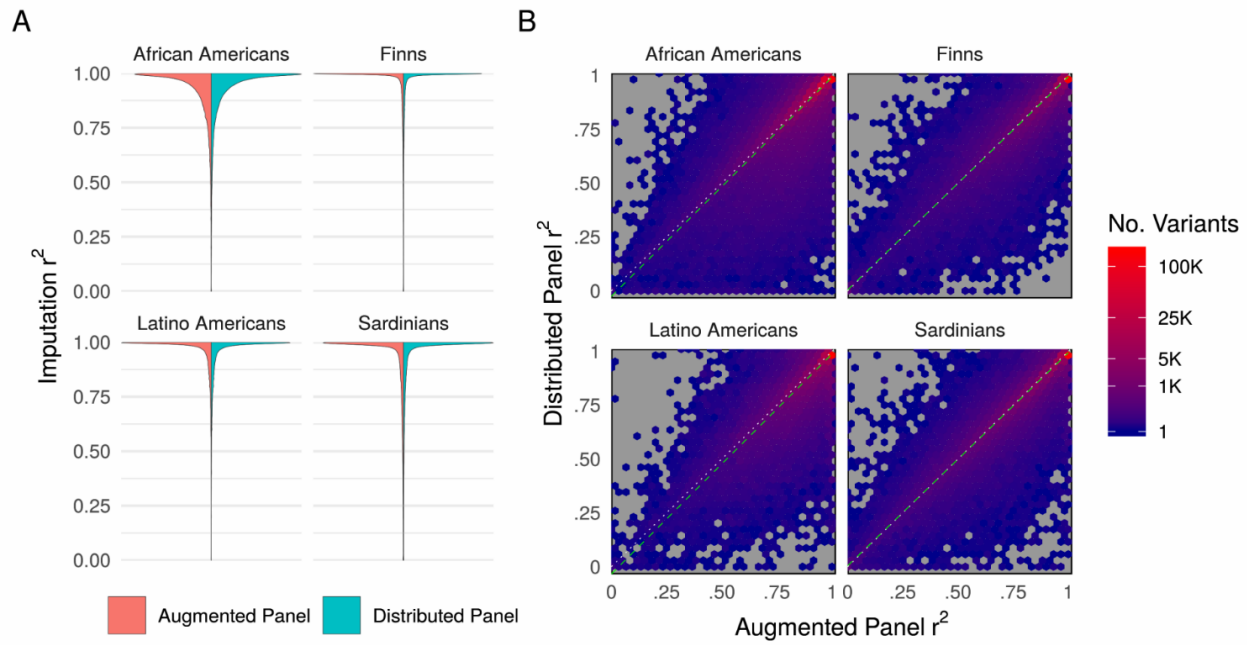
(**A**) Imputation coverage, defined as the proportion of variants with imputation $r^2 \geq 0.3$ and minor allele count (MAC) ≥ 5 in the reference panel, and (**B**) imputation r^2, defined as the squared Pearson correlation between true genotype and imputed dosage, as a function of study-specific reference panel size (Number Sequenced).

Supplementary Figure 2.2: Filtering False-Positive Imputed Variants.



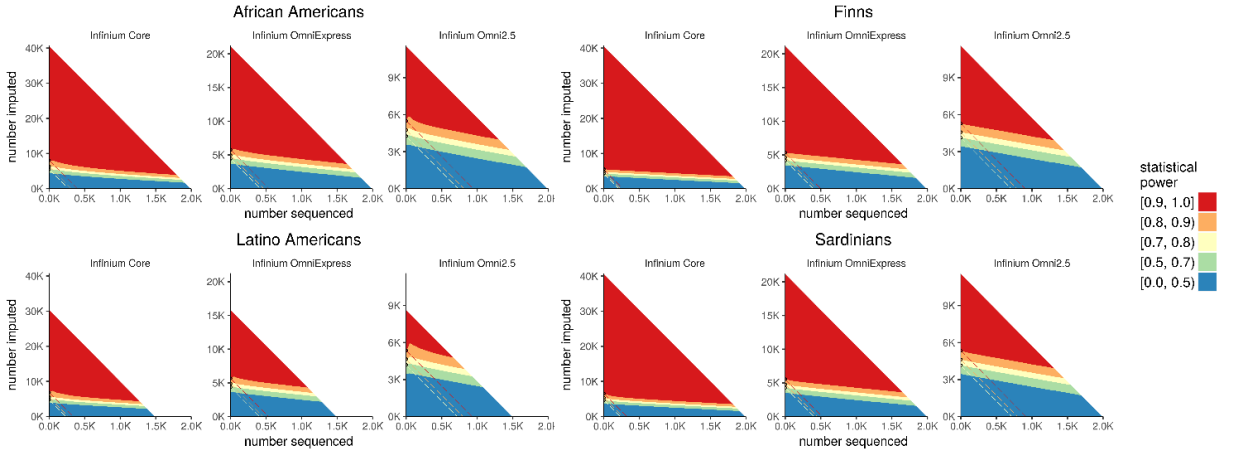
ROC and Precision-Recall curves for detecting false-positive imputed variants (variants with true $MAC = 0$, but $MAC_{imputed} \geq 5$ and imputation $r^2 \geq 0.3$) using the two-sample t-test statistic of allele frequencies between the imputed and sequenced subsamples. Results are shown for African Americans variants based on a sequenced subsample size of 1,400 and imputed subsample size of 1,412. Out of 656K variants with imputed MAF$\geq 0.1\%$ and imputation $r^2 \geq 0.3$, 11K variants (1.7%) have true MAC=0. Area under the ROC curve = 0.94, and area under Precision-Recall curve = 0.34.

Supplementary Figure 2.3: Imputation r^2 for Augmented versus Distributed Reference Panels.



Imputation r^2 for augmented reference panels (directly augmenting the HRC or HRC subset with study-specific sequence data) versus a distributed reference panel approach, in which participants are separately imputed with the HRC (or HRC subset) and the study-specific reference panel merged by selecting the MaCH- $\hat{r}^2$. Results are shown for the Omni Express array, and study specific reference panel sizes of 2,000 for African Americans, Finns, and Sardinians and 1,500 for Latino Americans. he mean pairwise difference in imputation r^2 across variants between augmented and distributed panels was < 0.005 in magnitude for each population.

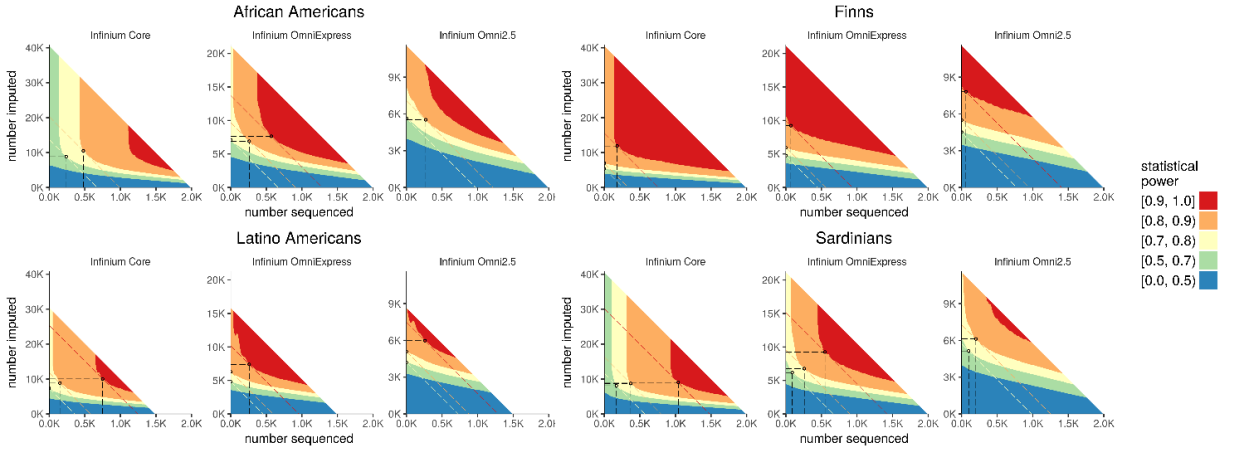
Supplementary Figure 2.4: Optimal Designs for a Common Variant with Moderate Effect.



Power to detect association for a variant with RR=1.5, MAF=10%, and disease prevalence 1%

Power to detect association for case-control studies with equal numbers of cases and controls as a function of sequenced subsample size (x-axis) and imputed subsample size (y-axis) for a variant with MAF=10% and relative risk 1.5 for a disease with prevalence 1%. Axes scaled to reflect costs of genotyping arrays (Table 2.1) and sequencing ($1K per sample). Power shown only for designs with total genotyping cost $\leq$ $2M ($1.5M for Latino Americans). Dashed diagonal lines indicate study designs with the same total cost, given by $y = a - bx$ where $a = (Total\ Cost)/(Array\ Cost)$ and $b = (Sequencing\ Cost)/(Array\ Cost)$. Circled points indicate optimal study designs, which attain the indicated power level at minimum total experimental cost (or, maximize power at the indicated total experimental cost). In this example, exclusively array-based genotyping and imputing from the external HRC panel is optimal for all populations considered.

Supplementary Figure 2.5: Optimal Designs for a Rare Variant with Large Effect.



Power to detect association for a variant with RR=6, MAF=0.3%, and disease prevalence 1%

Power to detect association for case-control studies with equal numbers of cases and controls as a function of sequenced subsample size (x-axis) and imputed subsample size (y-axis) for a variant with MAF=0.3% and relative risk 6 for a disease with prevalence 1%. Axes scaled to reflect costs of genotyping arrays (Table 2.1) and sequencing ($1K per sample). Power shown only for designs with total genotyping cost $\leq$ $2M ($1.5M for Latino Americans). Dashed diagonal lines indicate study designs with the same total cost, given by $y = a - bx$ where $a = (Total\ Cost)/(Array\ Cost)$ and $b = (Sequencing\ Cost)/(Array\ Cost)$. Circled points indicate optimal study designs, which attain the indicated power level at minimum total experimental cost (or, maximize power at the indicated total experimental cost). In this example, sequencing a subset of participants is almost uniformly optimal across populations, with the exception of Finns.

2.7 References

1KGP Consortium (2015). "A global reference for human genetic variation". In: *Nature* 526.7571, pp. 68–74.

Ahmad, Meraj et al. (2017). "Inclusion of Population-specific Reference Panel from India to the 1000 Genomes Phase 3 Panel Improves Imputation Accuracy". In: *Scientific reports* 7, p. 6733.

Auer, Paul L and Guillaume Lettre (2015). "Rare variant association studies: considerations, challenges and opportunities". In: *Genome medicine* 7.1, p. 16.

Bellus, Gary A et al. (1995). "Achondroplasia is defined by recurrent G380R mutations of FGFR3." In: *American journal of human genetics* 56.2, p. 368.

Bryc, Katarzyna et al. (2010). "Genome-wide patterns of population structure and admixture among Hispanic/Latino populations". In: *Proceedings of the National Academy of Sciences*, p. 200914618.

Das, Sayantan et al. (2016). "Next-generation genotype imputation service and methods". In: *Nature genetics* 48.10, pp. 1284–1287.

Deelen, Patrick et al. (2014). "Improved imputation quality of low-frequency and rare variants in European samples using the 'Genome of The Netherlands'". In: *European Journal of Human Genetics* 22.11, pp. 1321–1326.

Fuchsberger, Christian et al. (2016). "The genetic architecture of type 2 diabetes". In: *Nature* 536.7614, p. 41.

Holm, Hilma et al. (2011). "A rare variant in MYH6 is associated with high risk of sick sinus syndrome". In: *Nature genetics* 43.4, p. 316.

Hu, Yi-Juan et al. (2015). "Integrative analysis of sequencing and array genotype data for discovering disease associations with rare mutations". In: *Proceedings of the National Academy of Sciences* 112.4, pp. 1019–1024.

Illumina Inc. (2018). *Microarray kits for genotyping and epigenetic analysis*. [Online; accessed 1-August-2018]. URL: `https://www.illumina.com/products/by-type/microarray-kits.html`.

International HapMap 3 Consortium (2010). "Integrating common and rare genetic variation in diverse human populations". In: *Nature* 467.7311, pp. 52–58.

KA, Wetterstrand (2018). *DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP)*. [Online; accessed 1-August-2018]. URL: `https://www.genome.gov/sequencingcostsdata`.

Kichaev, Gleb and Bogdan Pasaniuc (2015). "Leveraging functional-annotation data in trans-ethnic fine-mapping studies". In: *The American Journal of Human Genetics* 97.2, pp. 260–271.

Lencz, Todd et al. (2017). "High-depth whole genome sequencing of a large population-specific reference panel: Enhancing sensitivity, accuracy, and imputation". In: *bioRxiv*, p. 167924.

Li, Yun et al. (2009). "Genotype imputation". In: *Annual review of genomics and human genetics* 10, pp. 387–406.

MacArthur, Jacqueline et al. (2016). "The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog)". In: *Nucleic acids research* 45.D1, pp. D896–D901.

Mahajan, Anubha et al. (2018). "Refining the accuracy of validated target identification through coding variant fine-mapping in type 2 diabetes". In: *Nature genetics* 50.4, p. 559.

McCarthy, Shane et al. (2016). "A reference panel of 64,976 haplotypes for genotype imputation". In: *Nature genetics* 48.10, p. 1279.

Montpetit, Alexandre et al. (2006). "An evaluation of the performance of tag SNPs derived from HapMap in a Caucasian population". In: *PLoS genetics* 2.3, e27.

Pistis, Giorgio et al. (2015). "Rare variant genotype imputation with thousands of study-specific whole-genome sequences: implications for cost-effective study designs". In: *European Journal of Human Genetics* 23.7, p. 975.

Popejoy, Alice B and Stephanie M Fullerton (2016). "Genomics is failing on diversity". In: *Nature* 538.7624, p. 161.

Roshyara, Nab Raj and Markus Scholz (2015). "Impact of genetic similarity on imputation accuracy". In: *BMC genetics* 16.1, p. 90.

Scott, Laura J et al. (2007). "A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants". In: *science*.

Sham, Pak C and Shaun M Purcell (2014). "Statistical power and significance testing in large-scale genetic studies". In: *Nature Reviews Genetics* 15.5, pp. 335–346.

Sidore, Carlo et al. (2015). "Genome sequencing elucidates Sardinian genetic architecture and augments association analyses for lipid and blood inflammatory markers". In: *Nature genetics* 47.11, p. 1272.

Steinthorsdottir, Valgerdur et al. (2014). "Identification of low-frequency and rare sequence variants associated with elevated or reduced risk of type 2 diabetes". In: *Nature genetics* 46.3, p. 294.

UK10K Consortium (2015). "The UK10K project identifies rare variants in health and disease". In: *Nature* 526.7571, pp. 82–90.

Van Leeuwen, Elisabeth M et al. (2015). "Population-specific genotype imputations using minimac or IMPUTE2". In: *Nature protocols* 10.9, p. 1285.

Zeggini, Eleftheria (2011). "Next-generation association studies for complex traits". In: *Nature genetics* 43.4, p. 287.

Zhou, Wei et al. (2017). "Improving power of association tests using multiple sets of imputed genotypes from distributed reference panels". In: *Genetic epidemiology* 41.8, pp. 744–755.

Chapter 3

emeraLD: Rapid Linkage Disequilibrium Estimation with Massive Data Sets

[1]

3.1 Introduction

Linkage disequilibrium (LD) – pairwise association between alleles at different genetic variants – is of fundamental interest in population genetics as a vestige of natural selection and demographic history, and is essential for a wide range of analyses from summary statistics in genome-wide association studies (GWAS). Motivated by restrictive data sharing policies and logistical constraints, a variety of methods have been developed for analysis of GWAS summary statistics (single-variant association statistics) rather than individual-level data. For example, summary statistics-based methods have been developed for fine-mapping (Benner, Spencer, et al. 2016), conditional association (Yang et al. 2012), gene-based association (Bakshi et al. 2016; Barbeira et al. 2016; Lamparter et al. 2016), heritability estimation (Bakshi et al. 2016), and functional enrichment analysis (Finucane et al. 2015;

[1] This work has been published (Quick et al. 2018)

Lamparter et al. 2016). These methods generally rely on LD estimates from an external data set, which are ideally calculated on-the-fly rather than precomputed and stored due to prohibitive storage costs. For example, the 1000 Genomes Project Phase 3 panel includes over 35M shared variants (1KGP Consortium 2015), which corresponds to over 4×10^{11} pairwise LD coefficients within 1 Mbp windows genome-wide.

3.1.1 Existing Tools to Estimate LD

Existing tools to estimate LD generally scale linearly with sample size, prompting a need for more efficient methods for large data sets. PLINK is a widely used software toolkit for analyzing genetic data, and is among the most computationally efficient tools for estimating LD (Shaun Purcell et al. 2007; Purcell and Chang 2016). PLINK's BED genotype data format allows efficient querying and data processing, but demands prohibitive storage space for large sample sizes and large numbers of markers (e.g., 7.6TB for the TOPMed Whole Genome Sequencing Project, which includes >60K individuals). VCFtools is another widely used software toolkit for manipulating and analyzing genetic data in the Variant Call Format (VCF) (Danecek et al. 2011). Compressed VCF files (VCF.gz) require far less storage space than BED files (e.g., >30$\times$ less storage space for the TOPMed WGS Project), and permit random access of genomic regions through block-compression and Tabix indexing (Danecek et al. 2011; Li 2011). VCFtools provides utilities to estimate LD from VCF files, but is computationally burdensome for large data sets. M3VCF format uses a compact haplotype representation that requires far less storage than genotype formats (Das et al. 2016). m3vcftools provides efficient utilities for estimating LD with M3VCF format, but is substantially slower than PLINK with BED file input.

3.2 Methods

3.2.1 LD Statistics

Three common measures of LD are the LD coefficient D (the covariance of genotypes), the standardized LD coefficient D' (D divided by its maximum value given allele frequencies), and the Pearson correlation r or its square (Gabriel et al. 2002). Each of these statistics can be written as a function of allele frequency estimates, sample size, and dot product of genotype vectors. Importantly, only the dot product must be calculated for each pair of variants to calculate LD, since allele frequencies and haplotype counts can be precomputed when processing genotype data.

3.2.2 Computational Approach

We tailored our computational approach to exploit the structure of each supported input data format. For genotype formats (e.g., VCF Danecek et al. 2011), we calculate the dot product using sparse-by-dense and sparse-by-sparse vector products. Using haplotype block format (M3VCF Das et al. 2016), we can calculate the dot product using within-block and between-block haplotype intersections.

Sparse Representation of Phased Genotypes For each variant, we keep a $\{0,1\}^{2n}$ vector of genotypes (where 1 indicates the minor allele) and sparse vector containing the indexes of non-zero entries. If the major allele is non-reference in the input file (allele count greater than n), we reverse the sign of its LD coefficients for consistency. Letting $C_j = \{i|G_{ij} = 1\}$ denote the set indexing minor-allele carriers of variant j, the dot product $m_{jk} \coloneqq \boldsymbol{G}_j \cdot \boldsymbol{G}_k$ between variants j and k can be calculated in $\min(m_j, m_k)$ operations, where m_j is the minor allele count (MAC) for variant j, by using the sparse-by-dense product formula $m_{jk} = \sum_{i \in C_j} G_{ik}$.

Sparse Representation of Unphased Genotypes For unphased genotypes, we store a $\{0, 1, 2\}^n$ vector of genotypes and sparse vectors indexing heterozygotes and minor-allele homozygotes for each variant. In this case, LD between two variants can be calculated in $\min(N_{j1} + N_{j2}, N_{k1} + N_{k2})$ operations, where N_{ji} is the number of individuals with genotype i at variant j.

Haplotype Block Representation A haplotype is a sequence of contiguous alleles along a chromosome within a genomic region, or haplotype block. Due to the limited diversity of human haplotypes (Wall and Pritchard 2003), the number of distinct haplotypes in a block with J biallelic variants is typically small relative to the sample size n or number of possible haplotypes 2^J (whichever is smaller). M3VCF format maps each sample to a haplotype within each block, and maps each variant in a block to the set of haplotypes that contain the non-reference allele (Das et al. 2016). Given M3VCF input, we precompute the number of observations N_h^b of each haplotype h for each block b, and index the set of haplotypes H_j^b containing the minor allele at each variant j in block b. For two variants j and k in the same block, the dot product can then be calculated in at most $\min(c_j^b, c_k^b)$ operations, where $c_k^b = \#H_k^b$ is the number of distinct haplotypes that carry the minor allele at variant k, using the sparse-by-dense product formula $m_{jk} = \sum_{h \in H_j^b} 1_{H_k^b}(h) N_h^b$. To calculate LD for variants in different blocks, we can compute a between-block count matrix $N_{hh'}^{ab}$, the number of samples with haplotype h in block a and haplotype h' in block b. The dot product between variants j and k can then be calculated in $c_j^a \times c_k^b$ operations using the formula $m_{jk} = \sum_{h \in H_j^a} \sum_{h' \in H_k^b} N_{hh'}^{ab}$. In practice, sparse-by-dense genotype products are typically more efficient for between-block calculations.

Informed Subsampling to Estimate LD with Large Sample Sizes When both variants j and k have large MAC (e.g., common variants and/or large sample sizes), calculating sparse-by-dense products to estimate LD becomes expensive. In this case, we use an informed subsampling approach to efficiently estimate LD while maintaining a user-specified bound on the precision of LD estimates.

We treat the sample correlation $r = (p_{jk} - p_j p_k)/s_j s_k$ as a parameter to be estimated by informed subsampling. Here, p_j, p_k, s_j and s_k can be calculated efficiently and stored; because p_{jk} must be calculated for each pair of variants, we subsample from the carriers of the rarest allele to increase computational efficiency. In Supplementary Materials, we show that the approximate estimator $\tilde{r}_\ell$ can be calculated in at most ℓ operations for any pair of variants, and increases the mean squared error (MSE) by no more than $1/\ell$ relative to exact LD estimates (or $2/\ell$ for unphased genotypes), where ℓ is a user-specified parameter. In very large data sets ($n > 50$K), subsampling with $\ell = 250$ decreased computation time for common variants (MAF > 5%) by an order of magnitude or more.

3.3 Results

3.3.1 Implementation and Usage

We implemented our algorithms as an open-source C++ tool, emeraLD (efficient methods for estimation and random access of LD), which can be used via command line or through an R interface included with source files. emeraLD accepts block compressed VCF.gz and M3VCF.gz input, and leverages Tabix (Li 2011) and the C library HTSlib to support rapid querying and random access of genotype data over genomic regions. emeraLD implements several options to customize output fields (variant information and LD statistics) and formats (long tables or square symmetric matrices). We also provide tools to facilitate estimating LD from a reference panel for analysis of GWAS summary statistics.

3.3.2 Performance

Table 3.1: Benchmarking: CPU Time and Memory Usage

Tool:	m3vcftools	PLINK 1.9	LDstore	emeraLD*	Absolute*
Format:	M3VCF.gz	BED	BGEN	M3VCF.gz	
CPU Time Relative to emeraLD					
1KGP	18.8	1.3	4.4	1.0	8.5 m
HRC	44.7	6.8	16.8	1.0	2.6 m
UKB	473.7	128.4	250.6	1.0	19.9 m
Memory Usage Relative to emeraLD					
1KGP	0.7	137.6	372.4	1.0	43.8 MiB
HRC	0.6	10.7	26.1	1.0	156.9 MiB
UKB	0.4	4.7	4.8	1.0	4.8 GiB

Time and memory to calculate LD in a 1Mbp region of chr20 (28,126 variants in 1KGP; 13,174 in HRC; and 32,783 in UKB). All experiments were run on a 2.8GHz Intel Xeon CPU. emeraLD
*Absolute time or memory for emeraLD as reference

We used WGS genotype data from the 1000 Genomes Project Phase 3 (1KGP; n = 2,504) (1KGP Consortium 2015), Haplotype Reference Consortium (HRC; n = 32,470) (McCarthy et al. 2016), and imputed genotype data from the UK Biobank (UKBB; n = 487,409) to compare performance between emeraLD and PLINK v1.9 (Purcell and Chang 2016), LDstore (Benner, Havulinna, et al. 2017), VCFtools (Danecek et al. 2011), and m3vcftools (Das et al. 2016). For UKB, emeraLD from M3VCF.gz file input is >100$\times$ faster than PLINK from BED files (Table 3.1), which are >10$\times$ larger than VCF.gz and >30$\times$ larger than M3VCF.gz. For HRC, which includes 32K individuals and only variants with MAC $\geq$5, emeraLD calculates LD from M3VCF.gz files >6$\times$ faster than PLINK from BED files, which are >4$\times$ larger than VCF.gz and >20$\times$ larger than M3VCF.gz. Times reported for emeraLD used $\ell = 1,000$ (MSE of approximation ≤ 0.001); this has a negligible effect for 1KGP, but reduced overall computation time by ~50% for UKB and HRC. Using M3VCF.gz files reduced computation time for emeraLD by ~30-50% relative to VCF.gz.

3.3.3 Applications

Our approach will be implemented in a forthcoming web-based service capable of providing LD information from large panels with >60K samples, such as the TOPMed WGS project, in real time. This enables use of improved LD information by rapidly emerging and gaining in popularity web-based interactive analysis and visualization tools such as LocusZoom (Pruim et al. 2010).

We have also used emeraLD to estimate LD for gene-based association and functional enrichment analysis of GWAS summary statistics. This approach avoids precomputing and storing LD without compromising speed – for example, we developed an implementation of the MetaXcan gene-based association method (Barbeira et al. 2016) using emeraLD to estimate LD on-the-fly, which is ~5× faster than the original implementation using precomputed LD estimates. To enable simple integration with R scripts or libraries, we include an R interface to emeraLD with source files.

3.4 Conclusions

Here we described computational and statistical methods to efficiently estimate LD with large data sets. Our methods exploit two natural features of genetic data: sparsity that arises from the abundance of rare variation, and high redundancy that arises from haplotype structure. We also developed an informed subsampling approach to further improve computational efficiency while maintaining a user-specified bound on precision relative to exact LD estimates. Finally, we described an open-source software implementation that can be used to facilitate analysis of GWAS summary statistics.

3.5 Acknowledgments

I thank Ryan Welch, Daniel Taliun, Sayanatan Das, and Christian Fuchsberger for helpful suggestions and assistance with the algorithms and software implementation. I also thank the cohorts and staffs of the Haplotype Reference Consortium, 1000 Genomes Project Consortium, and the UK Biobank Resource. This research has been conducted using the UK Biobank Resource under Application Number 24460.

3.6 Appendix: Supplementary Methods & Figures

Here we describe subsampling techniques to approximate linkage disequilibrium (LD) between biallelic variants. We begin with the case where haplotype phase is known (genotypes take values 0 or 1), followed by the case where phase is unknown (genotypes take values 0, 1, or 2).

We treat the sample correlation $r = (p_{jk} - p_j p_k)/s_j s_k$ as a parameter to be estimated by subsampling. Here, minor allele frequencies p_j and p_k (and standard deviations s_j and s_k) can be calculated efficiently and stored; because p_{jk} must be calculated for each pair of variants, we approximate to increase computational efficiency. For convenience, we treat allele frequencies as known constants.

Informed Subsampling with Phased Genotypes

Here, we describe a subsampling approach to approximate the sample correlation $r = (p_{jk} - p_j p_k)/s_j s_k$ using phased genotypes. Consider the estimator $\tilde{r}(\ell, \Delta) = [\tilde{p}_{jk}(\ell, \Delta) - p_j p_k]/s_j s_k$, where

$$\tilde{p}_{jk}(\ell, \Delta) = \begin{cases} \frac{p_j}{\ell} \sum_{i=1}^{\ell} \tilde{G}_{ik}^{(j)} & \Delta = 1 \\ \frac{p_k}{\ell} \sum_{i=1}^{\ell} \tilde{G}_{ij}^{(k)} & \Delta = 0 \end{cases}$$

for $\Delta \in \{0, 1\}$ and where each $\tilde{G}_{ik}^{(j)}$ (or $\tilde{G}_{ij}^{(k)}$) is independently sampled from the subset of haplotypes with $G_{ij} = 1$ (or $G_{ik} = 1$).

Clearly $\tilde{r}(\ell, \Delta)$ is an unbiased estimator for r, and has empirical variance

$$\mathrm{var}_n[\tilde{r}(\ell, \Delta)] = \frac{p_{jk}}{\ell s_j^2 s_k^2} \left[\Delta p_j^2 (p_j - p_{jk}) + (1 - \Delta) p_k^2 (p_k - p_{jk})\right].$$

Therefore, given that we sample ℓ minor allele carriers of either variant j or variant k, the optimal

estimator $\tilde{r}_\ell$ is given by taking $\Delta = I(p_j \leq p_k)$. Intuitively, carriers of the rarer allele are more informative for estimating the size of the intersection.

Letting ρ denote the true LD value in the population, the MSE of the approximate estimator is

$$\mathrm{MSE}(\tilde{r}_\ell) \coloneqq \mathbb{E}[(\tilde{r}_\ell - \rho)^2] = \mathbb{E}[(r-\rho)^2] + \mathbb{E}[(\tilde{r}_\ell - r)^2],$$

so for $p_j \leq p_k$ (WLOG) we have $\mathrm{MSE}(\tilde{r}_\ell) - \mathrm{MSE}(r) = (p_j - p_{jk})p_{jk}/\ell s_j^2 s_k^2$.

The variance of the estimator is maximized with respect to p_{jk} when $p_{jk} = p_j/2$, and maximized with respect to p_j when $p_j = 1/2$ (because $1/2 \geq s_k \geq s_j \geq p_j$). It follows that $\mathrm{MSE}(\tilde{r}_\ell) - \mathrm{MSE}(r) \leq 1/\ell$.

Informed Subsampling with Unphased Genotypes

Here, we describe a subsampling approach to approximate the sample correlation $r = c_{jk}/s_j s_k$ using unphased genotypes. We define the sample covariance between variants j and k as $c_{jk} = \frac{1}{n}\sum_{i=1}^n G_{ij}G_{ik} - 4p_j p_k$, and we can write

$$\frac{1}{n}\sum_{i=1}^n G_{ij}G_{ik} = p_{k,1}\hat{\mathbb{E}}(G_j|G_k = 1) + 2p_{k,2}\hat{\mathbb{E}}(G_j|G_k = 2)$$

where $p_{k,m}$ is the proportion of individuals with genotype m at variant k, and $\hat{\mathbb{E}}(G_j|G_k = m)$ is the mean genotype at variant j among individuals with genotype m at variant k in the overall sample of n individuals.

Define the approximate estimator

$$\tilde{c}_{jk}(\ell_1, \ell_2) = p_{k,1}\tilde{\mathbb{E}}_{\ell_1}(G_j|G_k = 1) + 2p_{k,2}\tilde{\mathbb{E}}_{\ell_2}(G_j|G_k = 2) - 4p_j p_k,$$

where $\tilde{\mathbb{E}}_\ell(G_j|G_k = m)$ is estimated by sampling ℓ genotypes from individuals with genotype m at variant k. The approximate estimator is unbiased and has empirical variance

$$\mathrm{var}_n[\tilde{c}_{jk}(\ell_1, \ell_2)] = \frac{p_{k,1}^2}{\ell_1}\mathrm{var}_n(G_j|G_k = 1) + \frac{4p_{k,2}^2}{\ell_2}\mathrm{var}_n(G_j|G_k = 2).$$

Supposing that variants j and k are independent (which maximizes the variability of the estimator),

$$\mathrm{var}_n[\tilde{c}_{jk}(\ell_1, \ell_2)] = \left(\frac{p_{k,1}^2}{\ell_1} + \frac{4p_{k,2}^2}{\ell_2}\right) s_j^2,$$

which is minimized by choosing $\ell_1 : \ell_2$ in proportion to $p_{k,1} : 2p_{k,2}$, or in other words oversampling homozygotes by a factor of 2.

We can now define the optimal approximate estimator $\tilde{c}_{jk}^{\ell} = \tilde{c}_{jk}(\ell_1^*, \ell_2^*)$, where

$$\ell_1^* = \frac{p_{k,1}}{2p_{k,2} + p_{k,1}}\ell \quad \text{and} \quad \ell_2^* = \frac{2p_{k,2}}{2p_{k,2} + p_{k,1}}\ell.$$

Therefore, the optimal approximate estimator has $\mathrm{var}_n(\tilde{c}_{jk}^{\ell}) \le 4p_k^2 s_j^2/\ell$ (note that $2p_k = p_{k,1} + 2p_{k,2}$), and letting $\tilde{r}_\ell = \tilde{c}_{jk}^{\ell}/s_j s_k$, we have

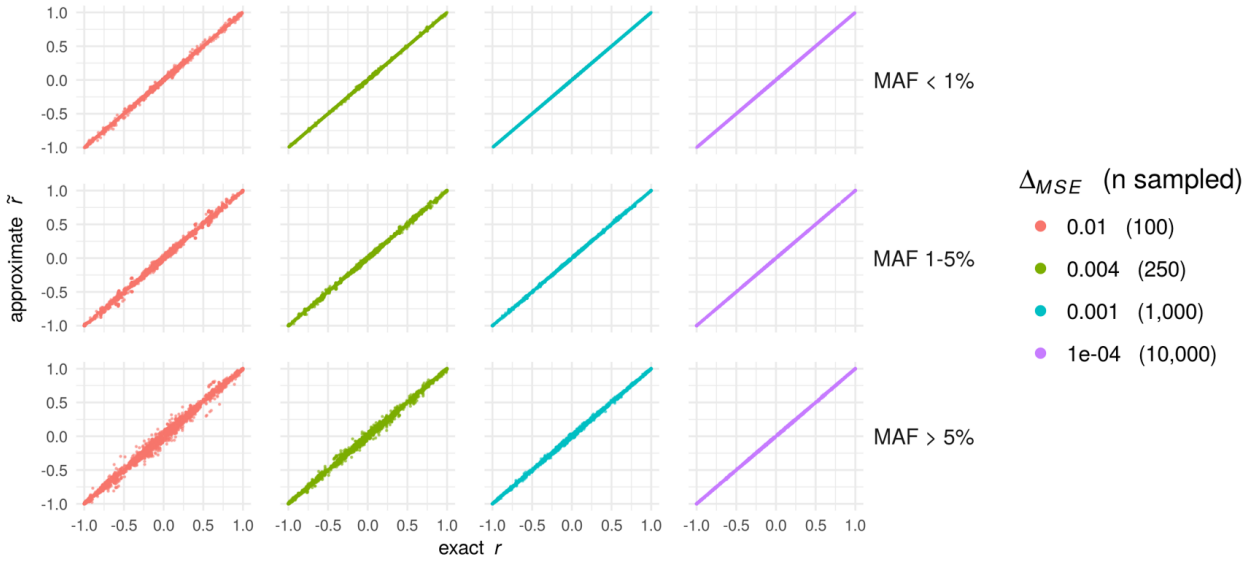
$$\mathrm{MSE}(\tilde{r}_\ell) - \mathrm{MSE}(r) = \mathrm{var}_n(\tilde{r}_\ell) \le \frac{4p_k^2}{\ell s_k^2} \le \frac{2}{\ell}.$$

Here, we have not assumed Hardy-Weinberg Equilibrium (HWE) for either variant. Supposing that both variants are in HWE, we can write $\hat{\mathbb{E}}(G_j G_k) = 2p_{jk}(1+p_j+p_k-p_{jk})+2(p_k-p_{jk})(p_j-p_{jk})$, and because p_{jk} is the only unknown parameter, the most efficient subsampling estimator would use as many minor-allele homozygotes as possible before sampling any heterozygotes. We avoid this assumption to ensure that estimates are robust.

Time Complexity of Approximation by Informed Subsampling

By subsampling ℓ individuals or haplotypes whenever $\min(MAC_j, MAC_k) > \ell$, we are guaranteed at most ℓ operations for each pair of variants. For computational efficiency, we sample subsets of minor-allele carriers once for each variant as genotype data are processed.

Supplementary Figure 3.1: Approximate vs. Exact LD Estimates.



Here, we show approximate vs. exact LD estimates from the Haplotype Reference Consortium. The number of minor-allele carriers sampled ℓ is equal to $1/\Delta_{MSE}$, where Δ_{MSE} is the maximum MSE induced by approximation.

3.7 References

1KGP Consortium (2015). "A global reference for human genetic variation". In: *Nature* 526.7571, pp. 68–74.

Bakshi, Andrew et al. (2016). "Fast set-based association analysis using summary data from GWAS identifies novel gene loci for human complex traits". In: *Scientific reports* 6, p. 32894.

Barbeira, A et al. (2016). "MetaXcan: Summary statistics based gene-level association method infers accurate PrediXcan results. bioRxiv: 045260". In:

Benner, Christian, Aki S Havulinna, et al. (2017). "Prospects of Fine-Mapping Trait-Associated Genomic Regions by Using Summary Statistics from Genome-wide Association Studies". In: *The American Journal of Human Genetics* 101.4, pp. 539–551.

Benner, Christian, Chris CA Spencer, et al. (2016). "FINEMAP: efficient variable selection using summary data from genome-wide association studies". In: *Bioinformatics* 32.10, pp. 1493–1501.

Danecek, Petr et al. (2011). "The variant call format and VCFtools". In: *Bioinformatics* 27.15, pp. 2156–2158.

Das, Sayantan et al. (2016). "Next-generation genotype imputation service and methods". In: *Nature genetics* 48.10, pp. 1284–1287.

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

Gabriel, Stacey B et al. (2002). "The structure of haplotype blocks in the human genome". In: *Science* 296.5576, pp. 2225–2229.

Lamparter, David et al. (2016). "Fast and rigorous computation of gene and pathway scores from SNP-based summary statistics". In: *PLoS computational biology* 12.1, e1004714.

Li, Heng (2011). "Tabix: fast retrieval of sequence features from generic TAB-delimited files". In: *Bioinformatics* 27.5, pp. 718–719.

McCarthy, Shane et al. (2016). "A reference panel of 64,976 haplotypes for genotype imputation". In: *Nature genetics* 48.10, p. 1279.

Pruim, Randall J et al. (2010). "LocusZoom: regional visualization of genome-wide association scan results". In: *Bioinformatics* 26.18, pp. 2336–2337.

Purcell, S and CC Chang (2016). *PLINK 1.9 package*.

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

Quick, Corbin et al. (2018). "emeraLD: Rapid Linkage Disequilibrium Estimation with Massive Data Sets". In: *Bioinformatics*.

Wall, Jeffrey D and Jonathan K Pritchard (2003). "Haplotype blocks and linkage disequilibrium in the human genome". In: *Nature Reviews Genetics* 4.8, pp. 587–597.

Yang, Jian et al. (2012). "Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits". In: *Nature genetics* 44.4, pp. 369–375.

Chapter 4

Leveraging Functional Genomic Annotations to Identify Causative Genes and Biological Mechanisms Underlying GWAS Associations

4.1 Introduction

Genome-wide association studies (GWAS) have identified thousands of genetic loci associated with hundreds of complex traits (Welter et al. 2013). However, the biological mechanisms underlying these associations are often poorly understood. The majority of GWAS associations to date are in non-coding regions of the genome, making it difficult to identify causal genes, let alone dissect genetic etiology in more detail. The Roadmap Epigenomics project (Bernstein et al. 2010), ENCODE (ENCODE Project Consortium 2012), GTEx (GTEx Consortium 2015), FANTOM5 (Lizio et al. 2015), and other consortia have fueled significant advances in regulatory genomics, and have provided valuable public data resources for studying the activity of regulatory elements and potential functional effects of non-coding variation. Integrating these large and complex data to better understand GWAS associations is a highly active area of research, and has presented enticing new

possibilities for GWAS researchers as well as formidable computational and statistical challenges.

Identifying Causal Variants Narrowing down the most likely causal variants underlying GWAS association signals is an important step toward identifying causal mechanisms. While linkage disequilibrium (LD) has been invaluable for identifying causal loci– first exploited in gene mapping studies, and later in genotype imputation algorithms– it complicates the identification of causal variants, forming dense clusters of highly correlated association statistics. To address this challenge, a Bayesian framework is often used to finemap associated loci, which involves calculating Bayes factors to identify sets or configurations of variants with the highest posterior probability of being causal (Y. Lee et al. 2018; Wen et al. 2016; Benner et al. 2016). Variants identified through finemapping can be cross-referenced with functional annotations to assess potential causal mechanisms (e.g., J. Z. Liu et al. 2012; Farh et al. 2015; Huang et al. 2017); however, this approach may or may not be informative for identifying causal genes.

Identifying Causal Genes Gene-based association tests provide a more interpretable framework for association analysis, and can increase power to detect association by aggregating effects across variants and reducing the burden of multiple testing (D. J. Liu et al. 2014; Sham and Purcell 2014). Often, variants are grouped based on their putative functional effects, e.g., rare non-synonymous variants for a given gene (D. J. Liu et al. 2014; Morrison et al. 2013). A more recent class of gene-based association tests have been developed for eQTL variants, e.g. PrediXcan (Gamazon et al. 2015; Barbeira et al. 2016) and TWAS (Gusev et al. 2016), which were proposed as tests of association between the genetic component of gene expression and traits. These expression-based tests employ the logic of Mendelian randomization to estimate causal effects: since genotype precedes phenotype, the direction of causality between phenotype and the genetic component of expression– a function of genotype– is unambiguous (Gusev et al. 2016; Hauberg et al. 2017). More broadly, many gene-based

association tests can be viewed as proxy variable methods in which intermediate phenotypes, e.g. protein dysfunction or tissue-specific expression, are hypothesized to mediate association. However, as in the analysis of single variants, LD is a potential source of confounding for gene-based association, which complicates the mechanistic interpretation of these statistics.

Purpose Here, we develop a statistical framework and computational tool to integrate regulatory and functional genomic datasets with GWAS summary statistics to identify causal genes and biological mechanisms underlying associations. We describe a novel gene-centered Bayesian model that accounts for multiple potential mechanisms underlying GWAS association signals, and use an approximate E-M algorithm to efficiently estimate hyperparameters. A central premise of this approach is to utilize functional enrichment to update prior weights for groups of variants; this approach has been used to improve the accuracy of single-variant finemapping (Y. Lee et al. 2018; Wen et al. 2016), but has not been applied in gene-based analysis. We compiled an integrative annotation dataset that maps functional variants to likely target genes by aggregating data sets from Roadmap/ENCODE (Cao et al. 2017; Bernstein et al. 2010; ENCODE Project Consortium 2012), FANTOM5 (Marbach et al. 2016; Cao et al. 2017; Lizio et al. 2015), and GTEx (GTEx Consortium 2015; Gamazon et al. 2015; Barbeira et al. 2016). We present a software toolkit and implementation of our methods, which leverages functional annotations to simultaneously examine a range of potential genes, mechanisms, and pathways underlying association signals. Finally, we discuss an application to 25 diverse traits using GWAS summary statistics from the UK Biobank.

4.2 Methods

We describe 1) a statistical model that explicitly maps functional variants to genes to identify genes implicated by association signals, 2) methods to aggregate variants for gene-based analysis,

3) algorithms to estimate the model and inference procedures using GWAS association summary statistics and LD estimates, 4) functional genomic data resources we used to annotate functional variants, 5) procedures to simulate GWAS data using real genotype and functional annotation data, and 6) GWAS data from the UK Biobank to which we applied our methods.

4.2.1 Model Definitions and Assumptions

Mediation as a Conceptual Basis for Genetic Association

The molecular mechanisms underlying complex traits are intricate, and involve many complex processes interacting over time at the cellular level and beyond. Here, we describe a simple model for complex traits as a conceptual basis, which is motivated by the intuition that cellular phenotypes (e.g., gene expression levels, protein functionality) mediate the causal associations between genetic variants and traits.

For a phenotype Y_i with mean 0 and unit variance, we consider a linear model

$$Y_i = \sum_g \boldsymbol{\alpha}_g^\top \boldsymbol{M}_{ig} + \epsilon_i$$

where $\boldsymbol{M}_{ig}$ is a vector of cellular phenotypes for gene g and ϵ_i is an environmental component. In turn, we model $\boldsymbol{M}_i = (\boldsymbol{M}_{i1}^\top, \boldsymbol{M}_{i2}^\top, ...)^\top$ as a linear function of genotypes,

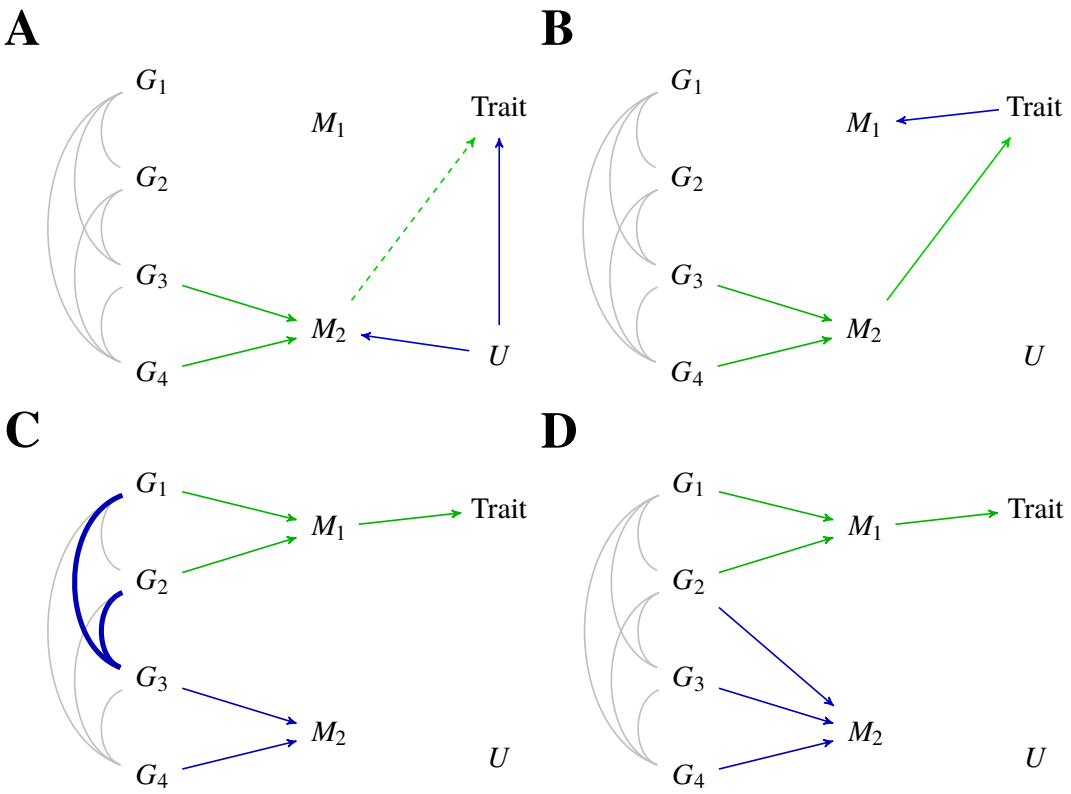
$$\boldsymbol{M}_i = \boldsymbol{\Lambda}\boldsymbol{G}_i + \mathbf{e}_i$$

where each element of the genotype vector $\boldsymbol{G}_i$ has mean 0 and unit variance. In practice, the functional effects of genetic variants (denoted by $\boldsymbol{\Lambda}$) are typically unknown, and cellular phenotypes $\boldsymbol{M}_i$ are typically unobserved in GWAS (unless additional omics data have been collected). We use functional genomic data to construct a proxy matrix $\tilde{\boldsymbol{\Lambda}}$ specifying effects of genetic variants

on gene-based cellular phenotypes. We then aggregate groups of variants based on functional annotations to construct proxy cellular phenotypes, which serve as the unit of association in our approach.

Using external functional genomic data to construct proxies for cellular phenotypes has been widely applied in methods such as TWAS and PrediXcan (Gusev et al. 2016; Gamazon et al. 2015), which can be conceptualized as tests of association between the predicted genetic component of tissue-specific gene expression levels and GWAS trait. However, these methods can be confounded by LD and pleiotropy (illustrated in Figure 4.2.1), rendering a causal interpretation ambiguous. Here, we aggregate diverse and comprehensive functional genomic annotation databases to account for multiple possible mechanisms underlying association. This approach lessens the potential for LD or pleiotropy-induced confounding, although such confounding is still possible due to incomplete and imprecise annotations. In addition, we assume that genetic associations reflect causal genetic effects, and our methods will be affected by confounding caused by un-adjusted technical factors, population structure, or relatedness in GWAS data.

Figure 4.2.1: Causal Diagrams: Mediation as a Conceptual Basis for Genetic Association



Path diagrams for four scenarios. In each subpanel, $\boldsymbol{G} = (G_1, G_2, G_3, G_4)^\top$ represents genotypes at four genetic variants, which are potentially correlated due to LD (gray undirected edges). $\boldsymbol{M} = (M_1, M_2)^\top$ represent intermediate cellular phenotypes (e.g., cell-type-specific mRNA expression levels for a given transcript, protein functionality, protein stability, etc.). The trait of interest is indicated in the upper right corner of each subpanel, and U in the lower right represents an unmeasured confounding variable. Here, we do not use $\boldsymbol{M}$ directly (and in general do not measure $\boldsymbol{M}$ directly); instead, we use genotypes to construct proxies $\widetilde{\boldsymbol{M}}$ to assess association with the trait of interest. Importantly, since $\widetilde{\boldsymbol{M}}$ is constructed solely from genotype, it can be used for valid causal inference even when $\boldsymbol{M}$ is confounded.
In Panel A, the intermediate phenotype M_2 and the trait of interest are both downstream of counfounding variable U. However, because genotypes $\boldsymbol{G}$ are independent of U, we can construct a valid instrumental variable from $\boldsymbol{G}$ to estimate the causal effect of M_2 on trait (dashed line) if it exists.
In Panel B, $\boldsymbol{G}$ affects the trait through M_2. However, because M_1 is downstream of the trait, $\boldsymbol{G}$ also indirectly affects M_1. Thus, a proxy variable $\widetilde{M}_1$ constructed from $\boldsymbol{G}$ may lead to the erroneous conclusion that M_1 affects the trait.
Panel C depicts LD-induced confounding. Here, only M_1 has a causal affect on the trait. However, we may detect a statistical association between $\widetilde{M}_2$ and trait due to LD between G_3 (which affects M_2) with G_1 and G_2 (which affect M_1).
Panel D depicts pleiotropy. Here, only M_1 has a causal affect on the trait. However, we may detect a statistical association between $\widetilde{M}_2$ and trait because G_2 affects both M_1 and M_2.

4.2.2 Aggregating Variants for Gene-Based Analysis

Gene-based Test Statistics

Table 4.1: Gene-Based Association Tests

Gene-based Test	Test Statistic	Comments	Synonyms & Special Cases
Q-form (Quadratic)	$\boldsymbol{Z}^\top \mathbf{W} \boldsymbol{Z}$	$\mathbf{W}$ = diagonal weight matrix	SKAT, SOCS
L-form (Linear)	$\boldsymbol{w}^\top \boldsymbol{Z}$	$\boldsymbol{w}$ = weight vector	Burden test, TWAS & PrediXcan
M-form (Min/Max)	$\max_j \lvert Z_j \rvert^2$	Alternatively, $\min_j$ p-value$_j$	MOCS

Here, we review common gene-based association tests and their interpretation in the proposed framework. The oldest and most widely used gene-based tests are linear combinations of z-scores (B. Li and Leal 2008; S. Lee, Wu, and Lin 2012), here referred to as L-form tests (Table 4.1). Examples of L-form tests include the burden test (B. Li and Leal 2008), which aggregates rare, putatively deleterious mutations; and TWAS/PrediXcan tests (Gamazon et al. 2015; Gusev et al. 2016), which aggregate eQTL variants using prediction weights estimated from external eQTL mapping data, e.g. GTEx. These can be viewed as tests of association between GWAS trait and an explicit proxy variable constructed as a linear combination of genotypes. Importantly, L-form tests rely on prior knowledge regarding the directions of effect across variants. For example, burden tests are appropriate when rare deleterious alleles are hypothesized to increase risk for disease, and TWAS/PrediXcan tests (using predictive weights) are appropriate when gene expression levels are hypothesized to affect trait. Another common form of gene-based test is the sum of squared z-scores across variants, here referred to as Q-form tests. While less tractable than L-form, analytical p-values for Q-form tests can be calculated using a variety of techniques to approximate the tail probabilities of multivariate normal quadratic forms (e.g., Davies 1980; H. Liu, Tang, and H. H. Zhang 2009). Q-form tests are most appropriate when a sizable proportion of variants are hypothesized to have non-zero effects of unknown and inconsistent direction (S. Lee, Wu, and Lin 2012). Finally, perhaps the simplest

gene-based test is the maximum absolute z-score across variants (or equivalently, the minimum p-value), here referred to as M-form tests. Analytical p-values for M-form tests can be calculated by directly integrating the multivariate normal density of z-scores within the hypercube given by $\boldsymbol{x} \in \mathbb{R}^m : \max_k |x_k| \leq \max_j |Z_j|$ where m is the number of variants, or approximated by adjusting the minimum p-value across variants by the effective number of tests (Conneely and Boehnke 2007). M-form tests are most appropriate when only a small proportion of variants are hypothesized to have non-zero effects. Unlike L-form tests, Q-form and M-form do not involve constructing an explicit proxy variable; however, they can be viewed as testing association between GWAS trait and a proxy variable constructed with stochastic weights. For example, a M-form test across variants within a regulatory element could be used to assess evidence that regulatory perturbations of a given gene affects GWAS trait, supposing that only a single unknown variant within the regulatory element perturbs gene regulation.

It is interesting to note the mathematical relationships among these three forms and the variety generalizations and possible extensions. Q-form and M-form can both be viewed as special cases of a more general statistic $S_p = (\sum_j |Z_j|^p)^{1/p}$, which is equivalent to Q-form (with $\mathbf{W} = \mathrm{I}$) when $p = 2$, and equivalent to M-form when $p \to \infty$; the S_p generalized form has been used, for example, in the aSPU gene-based test (Kwak and Pan 2015). Similarly, Q-form and L-form can both be viewed as special cases of a more general statistic $S_\pi = \boldsymbol{Z}^\top(\pi\mathbf{W} + (1-\pi)\boldsymbol{w}\boldsymbol{w}^\top)\boldsymbol{Z}$, which is equivalent to Q-form when $\pi = 1$ and L-form when $\pi = 0$; the S_π generalized form has been used, for example, in the SKAT-O gene-based test (S. Lee, Wu, and Lin 2012). Here, we focus primarily on using prior biological knowledge to inform gene-based analysis, and use only the basic gene-based test forms given in Table 4.1.

Gene-based Bayes Factors

In this section, we outline a Bayesian perspective of gene-based association analysis and provide Bayes factors corresponding to various prior distributions for the genetics effects on GWAS trait. Here and elsewhere, genetic effect sizes $\boldsymbol{\beta}$ are scaled such that β_j^2 is equal to the proportion of trait variance accounted for by the effect of variant j. We consider three distributional forms for the prior distribution of genetic effect sizes $\boldsymbol{\beta}$:

1. **Linear prior** in which weights $\boldsymbol{w}$ are used to construct an explicit proxy phenotype $\tilde{M}_i = \boldsymbol{w}^\top \boldsymbol{G}_i$. We assume that the effect size of the intermediate phenotype on trait is $\alpha \sim \mathcal{N}(0, \tau)$, which implies $\boldsymbol{\beta}$ is multivariate normal with mean $\mathbf{0}$ and the degenerate covariance matrix $\boldsymbol{w}\boldsymbol{w}^\top\tau$. We use this approach to aggregate eQTL variants using precomputed prediction weights $\boldsymbol{w}$ scaled such that $\boldsymbol{w}^\top \mathrm{R} \boldsymbol{w} = 1$.

2. **Dispersed prior** in which all variants have non-zero effect sizes, but magnitude and direction are unknown. Specifically, we assume $\boldsymbol{\beta} \sim \mathcal{N}_m(\mathbf{0}, I_m \frac{\tau}{m})$ where m is the number of variants. We use this approach to aggregate groups of coding variants and variants in regulatory elements.

3. **Sparse prior** in which only a single variant has a non-zero effect. Specifically, the causal variant j^* is uniformly distributed over {1, 2, ..., m} where m is the number of variants, and $\beta_{j*} \sim \mathcal{N}(0, \tau)$. We use this approach to aggregate groups of coding variants and variants in regulatory elements.

We define the Bayes factor BFs for a prior model A as

$$BF = \frac{f_{\hat{\boldsymbol{\beta}}|A}(\hat{\boldsymbol{\beta}})}{f_{\hat{\boldsymbol{\beta}}|\boldsymbol{\beta}\equiv\mathbf{0}}(\hat{\boldsymbol{\beta}})}, \tag{4.2.1}$$

where $\hat{\boldsymbol{\beta}}$ denotes the MLE of effect sizes. In general, we do not explicitly calculate $\hat{\boldsymbol{\beta}}$; rather, we reconstruct the Bayes factor using single-variant z-scores $\boldsymbol{Z}$ and LD matrix R. This approach has been widely applied previously for single-variant Bayesian finemapping (e.g., Y. Lee et al. 2018). The sampling distribution of the MLE $\hat{\boldsymbol{\beta}}$ is approximately

$$n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \overset{a}{\sim} \boldsymbol{\mathcal{N}}_m(\mathbf{0}, \sigma_Y^2 \mathbf{R}^{-1}) \tag{4.2.2}$$

where σ_Y^2 is the trait variance, and $\mathbf{R}$ is the LD matrix. Equation 4.2.2 holds exactly under a linear regression model with i.i.d. normal residuals where genotypes are scaled with unit variance.

For prior distributions of the form $A : \boldsymbol{\beta} \sim \mathcal{N}_m(\mathbf{0}, \boldsymbol{\Lambda})$, where $\boldsymbol{\Lambda}$ is a positive semi-definite matrix, we can use the binomial inverse theorem to find

$$BF = \left|\mathrm{I}_m + \frac{n}{\sigma_Y^2}\mathbf{R}\boldsymbol{\Lambda}\right|^{-1/2} \exp\left\{-\frac{1}{2\sigma_Y^2}\boldsymbol{Z}^\top \boldsymbol{\Lambda}(\frac{\sigma_Y^2}{n}\mathbf{I}_m + \mathbf{R}\boldsymbol{\Lambda})^{-1}\boldsymbol{Z}\right\}.$$

When $\boldsymbol{\Lambda} = \tau^2 \boldsymbol{w}\boldsymbol{w}^\top$, where $\boldsymbol{w}$ is scaled such that $\boldsymbol{w}^\top \mathrm{R}\boldsymbol{w} = 1$, this expression can be simplified by again applying the binomial inverse theorem and Sylvester's determinant theorem. Bayes factors for each form of prior are given in Table 4.2.

Table 4.2: Gene-Based Bayes Factors

Prior	Effect Size Distribution	Log Bayes Factor
Linear	$\boldsymbol{\beta} \sim \mathcal{N}_m(\mathbf{0}, \tau^2 \boldsymbol{w}\boldsymbol{w}^\top)$ where $\boldsymbol{w}^\top \mathbf{R}\boldsymbol{w} = 1$	$\frac{n\tau^2}{2\sigma_Y^2(\sigma_Y^2 + n\tau^2)}(\boldsymbol{Z}^\top \boldsymbol{w})^2 - \frac{1}{2}\log\left(1 + \frac{n\tau^2}{\sigma_Y^2}\right)$
Dispersed	$\boldsymbol{\beta} \sim \mathcal{N}_m(\mathbf{0}, \frac{1}{m}\tau^2 \mathrm{I}_m)$	$\frac{\tau^2}{2m\sigma_Y^2}\boldsymbol{Z}^\top(\frac{\sigma_Y^2}{n}\mathbf{I}_m + \frac{\tau^2}{m}\mathbf{R})^{-1}\boldsymbol{Z} - \frac{1}{2}\log\left\lvert\mathrm{I}_m + \frac{n\tau^2}{m\sigma_Y^2}\mathbf{R}\right\rvert$
Sparse	$\beta_j \sim \mathcal{N}_m(0, \delta_{j^*,j}\tau^2)$ where $j^* \sim \mathrm{Unif}\{1, 2, ..., m\}$	$\log\left(\frac{1}{m}\sum_{j=1}^m e^{\frac{n\tau^2}{2\sigma_Y^2(\sigma_Y^2 + n\tau^2)}Z_j^2}\right) - \frac{1}{2}\log\left(1 + \frac{n\tau^2}{\sigma_Y^2}\right)$

It is interesting to note the close relationships between these gene-based BFs and the gene-based

test statistics described in the previous section. The linear-prior BF is a direct function of the L-form gene-based test. Similarly, the dispersed-prior BF can be expressed as a function of a Q-form gene-based test. Finally, the sparse-prior BF can be written as a function of $\mathrm{LSE}(cZ_1^2, ..., cZ_m^2)$, where the LogSumExp function $\mathrm{LSE}(x_1, ..., x_m) = \log\left(\sum_j e^{x_j}\right)$ is a well-known smooth approximation to $\max(x_1, ..., x_m)$, and is thus closely related to the M-form gene-based test.

4.2.3 Model Fitting Algorithms and Statistical Inference

Hierarchical Bayesian Model for Gene-based Association

Here we describe a hierarchical Bayesian model for gene-based association. For each gene, we stratify variants by annotation class $j = 1, 2, ..., J$ (coding variants, eQTLs, enhancers, and UTR regions), and further stratify variants within each class j by annotation subclass $k = 1, 2, ..., K_j$ (tissue-type for eQTLs and enhancers; and non-synonymous, splice-site, ... for coding variants). Let $\boldsymbol{\beta}_{gjk}$ denote the effect sizes of variants in class j and subclass k with respect to gene g, and let ξ_{gjk} denote an indicator function such that

$$f_{\boldsymbol{\beta}|\xi}(\boldsymbol{\beta}_{gjk}; \xi_{gjk}) = \begin{cases} f_k & \xi_{gjk} = 1 \\ \delta_{\mathbf{0}} & \xi_{gjk} = 0 \end{cases}$$

where the forms of prior densities f_k are given in the previous section (Table 4.2). We assume that each gene has at most one causal class and subclass; this assumption simplifies computation, and is analogous to an adjusted minimum p-value across stratified gene-based tests. We introduce indicator variables θ_g and ξ_{gj} for each gene g and for each annotation class j, and assume

$$P(\theta_g = 1) = \pi_\theta, \qquad P(\xi_{gj} = 1|\theta_g) = \begin{cases} \pi_j & \theta_g = 1 \\ 0 & \theta_g = 0 \end{cases}, \qquad P(\xi_{gjk} = 1|\xi_{gk}) = \begin{cases} \pi_{jk} & \xi_{gj} = 1 \\ 0 & \xi_{gj} = 0 \end{cases}.$$

In other words, θ_g is equal to 1 if any functional variants for gene g affect GWAS trait and 0 otherwise; ξ_{gj} is equal to 1 if any functional variants of annotation class j with respect to gene g affect GWAS trait; and ξ_{gjk} is equal to 1 if functional variants of annotation class j and subclass k affect GWAS trait and 0 otherwise. Under the assumption that each gene has at most one causal annotation class and subclass, we can calculate an overall gene-based Bayes factor as

$$BF_g = \frac{f_{\hat{\boldsymbol{\beta}}|\theta=1}(\hat{\boldsymbol{\beta}}_g)}{f_{\hat{\boldsymbol{\beta}}|\boldsymbol{\beta}\equiv\mathbf{0}}(\hat{\boldsymbol{\beta}}_g)} = \frac{\sum_j P(\xi_{gj} = 1|\theta_g = 1) \sum_k P(\xi_{gjk} = 1|\xi_{gj} = 1) f_{\hat{\boldsymbol{\beta}}|\boldsymbol{\xi}}(\hat{\boldsymbol{\beta}}_g|\xi_{gjk} = 1)}{f_{\hat{\boldsymbol{\beta}}|\boldsymbol{\beta}\equiv\mathbf{0}}(\hat{\boldsymbol{\beta}}_g)}$$

$$= \sum_j \pi_j \sum_k \pi_{jk} \frac{f_{\hat{\boldsymbol{\beta}}|\boldsymbol{\xi}}(\hat{\boldsymbol{\beta}}_g|\xi_{gjk} = 1)}{f_{\hat{\boldsymbol{\beta}}|\boldsymbol{\beta}\equiv\mathbf{0}}(\hat{\boldsymbol{\beta}}_g)}$$

$$= \sum_j \pi_j \sum_k \pi_{jk} BF_{gjk},$$

where $\hat{\boldsymbol{\beta}}_g$ denotes the effect size estimates for functional variants for gene g and $BF_{gjk} \coloneqq f_{\hat{\boldsymbol{\beta}}|\boldsymbol{\xi}}(\hat{\boldsymbol{\beta}}_g|\xi_{gjk} = 1)/f_{\hat{\boldsymbol{\beta}}|\boldsymbol{\beta}\equiv\mathbf{0}}(\hat{\boldsymbol{\beta}}_g)$ has the corresponding form given in Table 4.2. We can similarly calculate the posterior probability for annotation class j and subclass k for gene g as

$$P(\xi_{gjk} = 1|\hat{\boldsymbol{\beta}}_g) = \frac{P(\hat{\boldsymbol{\beta}}_g|\xi_{gjk} = 1)P(\xi_{gjk} = 1)}{P(\hat{\boldsymbol{\beta}}_g)}$$

$$= \frac{P(\hat{\boldsymbol{\beta}}_g|\xi_{gjk}=1)P(\xi_{gjk}=1|\xi_{gj}=1)P(\xi_{gj}=1|\theta_g=1)P(\theta_g=1)}{P(\hat{\boldsymbol{\beta}}_g|\theta_g=0)P(\theta_g=0)+P(\hat{\boldsymbol{\beta}}_g|\theta_g=1)P(\theta_g=1)}$$

$$= \frac{\frac{P(\hat{\boldsymbol{\beta}}_g|\xi_{gjk}=1)}{P(\hat{\boldsymbol{\beta}}_g|\theta_g=0)}\pi_{jk}\pi_j\pi_\theta}{(1-\pi_\theta)+\frac{P(\hat{\boldsymbol{\beta}}_g|\theta_g=1)}{P(\hat{\boldsymbol{\beta}}_g|\theta_g=0)}\pi_\theta}$$

$$= \frac{\pi_j\pi_{jk}BF_{gjk}}{\frac{1-\pi_\theta}{\pi_\theta}+BF_g},$$

which can be viewed as a form of Bayesian model averaging (Hoeting et al. 1999).

Algorithms for Empirical Bayes Estimation

Here, we describe algorithms to obtain empirical Bayes estimates of prior weights for each annotation class and subclass. In general, we use E-M algorithms in which the latent variables are indicators variables ξ_{gjk} for each gene g, annotation class j, and annotation subclass k. We estimate subclass priors π_{jk} separately within each annotation class j to avoid penalizing genes with association signals across multiple annotation classes.

Within each annotation class j, we use a logistic prior of the form

$$P(\xi_{gjk}=1) = 1/(1+e^{-\boldsymbol{x}_{gjk}^\top\boldsymbol{\gamma}})$$

where $\boldsymbol{x}_{gjk}$ denotes an annotation vector. In the simplest case, $\boldsymbol{x}_{gjk}$ indicates the annotation subclass (e.g., tissue-type), and if annotation subclasses are mutually exclusive and collectively exhaustive, then we simply have

$$P(\xi_{gjk}=1) = 1/(1+e^{-\gamma_{jk}})$$

where the intercept is omitted to avoid rank deficiency. In this case, the logistic MLE has closed form; namely, $\hat{\gamma}_{jk} = \text{logit}(\frac{1}{N_{jk}} \sum_g \xi_{gjk})$ where N_{jk} is the number of genes with one or more variants of annotation class j and subclass k. Otherwise, $\hat{\boldsymbol{\gamma}}$ can be estimated using Fisher scoring or gradient descent. We note that this logistic prior does not explicitly account for interdependence between neighboring groups of variants, and can thus be viewed as a form of composite likelihood estimation (Varin, Reid, and Firth 2011).

Because ξ_{gjk} are unobserved, we use an E-M algorithm to estimate $\hat{\boldsymbol{\gamma}}$. In the t^{th} E step for class j, we update each

$$\tilde{\xi}_{gjk}^{(t+1)} = \hat{\mathbb{E}}^{(t+1)}(\xi_{gjk}|\hat{\boldsymbol{\beta}}_{gj}, \boldsymbol{x}_{gjk}) = \frac{e^{\boldsymbol{x}_{gjk}^\top \hat{\boldsymbol{\gamma}}^{(t)}} BF_{gjk}}{1 + \sum_{k'} e^{\boldsymbol{x}_{gjk'}^\top \hat{\boldsymbol{\gamma}}^{(t)}} BF_{gjk'}}, \qquad k = 1, 2, ..., K_j \text{ and } g = 1, 2, ..., N_{jk}$$

and update $\hat{\boldsymbol{\gamma}}^{(t)}$ in the subsequent M-step using Fisher scoring or closed form (when possible) as described previously. We then calculate subclass priors as $\hat{\pi}_{jk} = \hat{P}(\xi_{gjk} = 1|\xi_{gj} = 1) = e^{\hat{\gamma}_{jk}} / \sum_{k'} e^{\hat{\gamma}_{jk'}}$.

After updating annotation subclass priors π_{jk}, we collapse across subclasses and use the same approach to calculate annotation class priors π_j. For example, the E step becomes

$$\tilde{\xi}_{gj}^{(t+1)} = \hat{\mathbb{E}}^{(t+1)}(\xi_{gj}|\hat{\boldsymbol{\beta}}_{g}, \boldsymbol{x}_{gj}) = \frac{e^{\boldsymbol{x}_{gj}^\top \hat{\boldsymbol{\gamma}}^{(t)}} BF_{gj}}{1 + \sum_{j'} e^{\boldsymbol{x}_{gj'}^\top \hat{\boldsymbol{\gamma}}^{(t)}} BF_{gj'}},$$

where $BF_{gj} = \sum_k \hat{\pi}_{jk} BF_{gjk}$. In practice, we may estimate subclass priors for annotation classes of direct interest and use a flat prior elsewhere for convenience.

Calculating Posterior Probabilities in Hit Regions

The composite likelihood approach described above ignores LD between neighboring genes. To calculate gene-based posterior probabilities accounting for signals at neighboring genes, we finemap regions with one or more marginal Bayes factors above a specified threshold. We define hit regions by forming 1 cM windows around each of the identified genes and merging regions that overlap. We then calculate posterior probabilities under the assumption that each hit region contains at most one causal gene.

4.2.4 Regulatory and Functional Annotation Data

To establish regulatory and functional relationships between variants and genes, we aggregated annotations from four databases. To identify potentially protein-altering variants, we used TabAnno/EPACTS (Kang 2014). We used expression and genotype data from GTEx to estimate eQTL weights using forward selection and elastic net models for variable selection (GTEx Consortium 2015; Gamazon et al. 2015; Barbeira et al. 2016). To capture additional regulatory variation, we used promoter-target pairs and enhancer-target pairs from regulatorycircuits.org inferred using FANTOM5 CAGE data (Marbach et al. 2016), and tissue-specific enhancer-target pairs from JEME (joint effect of multiple enhancers) inferred using FANTOM5 and ENCODE/Roadmap data (Cao et al. 2017).

Table 4.3: Functional Annotation Sources

Annotation Source	Data Source	Annotation Class	Reference
Anno/EPACTS	GENCODE, RefSeq	Coding variation	Kang, 2014
GTEx eQTLs	GTEx	eQTLs	GTEx Consortium, 2015
RegulatoryCircuits	FANTOM5	Enhancers, promoters	Marbach et al., 2016
JEME	ENCODE,FANTOM5	Enhancers	Cao et al., 2017

4.2.5 Simulation Procedures

Here, we describe procedures to simulate GWAS summary statistics using real genotype data or LD estimates. We begin by defining summary statistics and deriving their distribution. We next outline procedures to simulate GWAS summary statistics under the desired distribution. Finally, we describe procedures to simulate configurations of causal genes, causal variants, and effect sizes using real functional genomic annotation data.

GWAS Summary Statistics

In the absence of covariates, we can write the single-variant z-score test Z_j for association between genotype G_j and a continuous trait Y as

$$Z_j = n^{-1/2}\tilde{\boldsymbol{G}}_j^\top \tilde{\boldsymbol{Y}},$$

where $\tilde{\boldsymbol{Y}}$ and $\tilde{\boldsymbol{G}}_j$ are scaled and centered with mean 0 and standard deviation 1. Similarly, we can write the vector of z-scores across variants as

$$\boldsymbol{Z} = n^{-1/2}\tilde{\mathbf{G}}^\top \tilde{\boldsymbol{Y}}.$$

Given the additive association model $Y_i = \alpha_0 + \boldsymbol{\alpha}^\top \boldsymbol{G}_i + \varepsilon_i$, we can write a scaled model $\tilde{\boldsymbol{Y}} = \tilde{\mathbf{G}}\boldsymbol{\beta} + \tilde{\boldsymbol{\varepsilon}}$, where $\beta_j = \mathrm{sd}(G_j)\alpha_j/\mathrm{sd}(Y)$ are heritability-scale effect sizes such that β_j^2 is equal to the proportion of trait variance accounted for by the effect of the j^{th} variant. Substituting this expression for $\tilde{\boldsymbol{Y}}$, we can re-write $\boldsymbol{Z}$ as

$$\boldsymbol{Z} = n^{-1/2}\tilde{\mathbf{G}}^\top \tilde{\mathbf{G}}\boldsymbol{\beta} + n^{-1/2}\tilde{\mathbf{G}}^\top \tilde{\boldsymbol{\varepsilon}}$$

$$= n^{1/2}\mathbf{R}\boldsymbol{\beta} + n^{-1/2}\tilde{\mathbf{G}}^\top\tilde{\boldsymbol{\varepsilon}},$$

where $\mathbf{R} \coloneqq \frac{1}{n}\tilde{\mathbf{G}}^\top\tilde{\mathbf{G}}$ denotes the LD matrix. If the residuals are i.i.d. normal, i.e. $\tilde{\boldsymbol{\varepsilon}} \sim \boldsymbol{\mathcal{N}}_n(\mathbf{0}, I_n)$ where $\boldsymbol{\mathcal{N}}_n$ denotes the n-dimensional multivariate normal distribution, then the z-scores are distributed

$$\boldsymbol{Z} \sim n^{1/2}\mathbf{R}\boldsymbol{\beta} + \boldsymbol{\mathcal{N}}_m(\mathbf{0}, \mathbf{R}) \tag{4.2.3}$$

For convenience, we define $\boldsymbol{Z}_0 \coloneqq n^{-1/2}\tilde{\mathbf{G}}^\top\tilde{\boldsymbol{\varepsilon}} \sim \boldsymbol{\mathcal{N}}_m(\mathbf{0}, \mathbf{R})$ as the *z-score residual*. Under the null hypothesis that $\boldsymbol{\beta} = \mathbf{0}$, the z-scores $\boldsymbol{Z}$ and $\boldsymbol{Z}_0$ are equivalent.

Equation (1) also holds approximately for non-normally distributed traits, which can be shown by asymptotic arguments under a sequence of local alternatives. For binary traits, the z-score corresponding to the score test of association in the absence of covariates can still be written $Z_j = n^{-1/2}\tilde{\boldsymbol{G}}_j^\top\tilde{\boldsymbol{Y}}$, and $\boldsymbol{Z} - n^{1/2}\mathbf{R}_n\boldsymbol{\gamma} \xrightarrow{d} \boldsymbol{\mathcal{N}}_m(\mathbf{0}, \mathbf{R})$, where $\boldsymbol{\gamma}$ is an effect size parameter that depends on the odds-ratio or relative risk, minor allele frequency (MAF), population prevalence of the trait, and the GWAS case-control ratio.

Procedures to Simulate GWAS Summary Statistics

Here, we describe procedures to simulate $\boldsymbol{Z}$ following equation (1) given a specified vector of effect sizes $\boldsymbol{\beta}$, GWAS sample size n, and LD matrix $\mathbf{R}$. While the GWAS sample size n does not explicitly appear in the distribution of the z-score residual, the rank of $\mathbf{R}$ and statistical precision of estimates r_{jk} do depend on n. Specifically, $\operatorname{rank}(\mathbf{R}) \leq \operatorname{rank}(\mathbf{G}) \leq \min(m, n)$, where m is the number of variants. In addition, the number of distinct variants in a sample of size n under a neutral model (constant population size and no selection) is $O(\log n)$; however, we ignore this dependence because

the number of observed variants in array and imputation-based studies is fixed (restricted to variants that are directly typed, or imputable from the reference panel). In GWAS simulations, we condition on the reference LD structure (so that the precision of r_{jk} estimates is irrelevant) to simulate $\boldsymbol{Z}$ with arbitrary n given a fixed number of LD reference samples n_{ref}, and avoid rank deficiency when necessary by LD pruning or Tikhonov regularization.

To simulate summary statistics following equation (1), we can consider two approaches:

1. Compute $\mathbf{R}^{1/2}$, simulate $\boldsymbol{S} \sim \mathcal{N}_m(\mathbf{0}, I_m)$, multiply to obtain $\boldsymbol{Z}_0 = \mathbf{R}^{1/2}\boldsymbol{S} \sim \boldsymbol{\mathcal{N}}_m(\mathbf{0}, \mathbf{R})$, and calculate $\boldsymbol{Z} = n^{1/2}\mathbf{R}\boldsymbol{\beta} + \boldsymbol{Z}_0$.

2. Find $m \times n$ $\mathbf{V}$ such that $\mathbf{V}\mathbf{V}^\top = \mathbf{R}$, simulate $\boldsymbol{T} \sim \mathcal{N}_m(\mathbf{0}, I_m)$, multiply to obtain $\boldsymbol{Z}_0 = \mathbf{V}\boldsymbol{T} \sim \boldsymbol{\mathcal{N}}_m(\mathbf{0}, \mathbf{R})$, and calculate $\boldsymbol{Z} = n^{1/2}\mathbf{R}\boldsymbol{\beta} + \boldsymbol{Z}_0$. When reference genotypes are directly available, we can simply take $\mathbf{V} = \tilde{\mathbf{G}}_{ref}^\top / \sqrt{n_{ref}}$.

To simulate summary statistics for whole chromosomes (with $m \gg n$), we use Approach 2 with reference genotypes from the European subset of the 1000 Genomes Phase 3 panel.

Simulating Configurations of Causal Genes, Variants, and Effect Sizes

To generate variant effect sizes for simulations under the alternative hypothesis, we first select K causal annotation classes using real functional genomic annotation data described above. We then sample M_k causal genes for each causal annotation class $k = 1, 2, ...K$, ensuring each that each causal gene has at least one variant matching its causal annotation. For example, a gene with the causal group "liver-specific eQTL" must have one or more liver-specific eVariant.

Given a configuration of causal genes and annotation groups, we simulate effect sizes for each variant following the effect size prior forms given in Table 4.2. We then rescale β_j's for each causal gene to ensure that $\sum_{j \in C_g} \beta_j^2 = h_L^2$, where C_g is the set of causal SNPs for gene g and h_L^2 is the

specified per-locus heritability.

4.2.6 The UK Biobank Resource

We selected 25 traits from the UK Biobank for primary analysis with effective sample size of at least 7,500, one or more significant GWAS association, and one or more relevant Mendelian gene (Table 4.4).

Table 4.4: UK Biobank: Traits Included for Primary Analysis

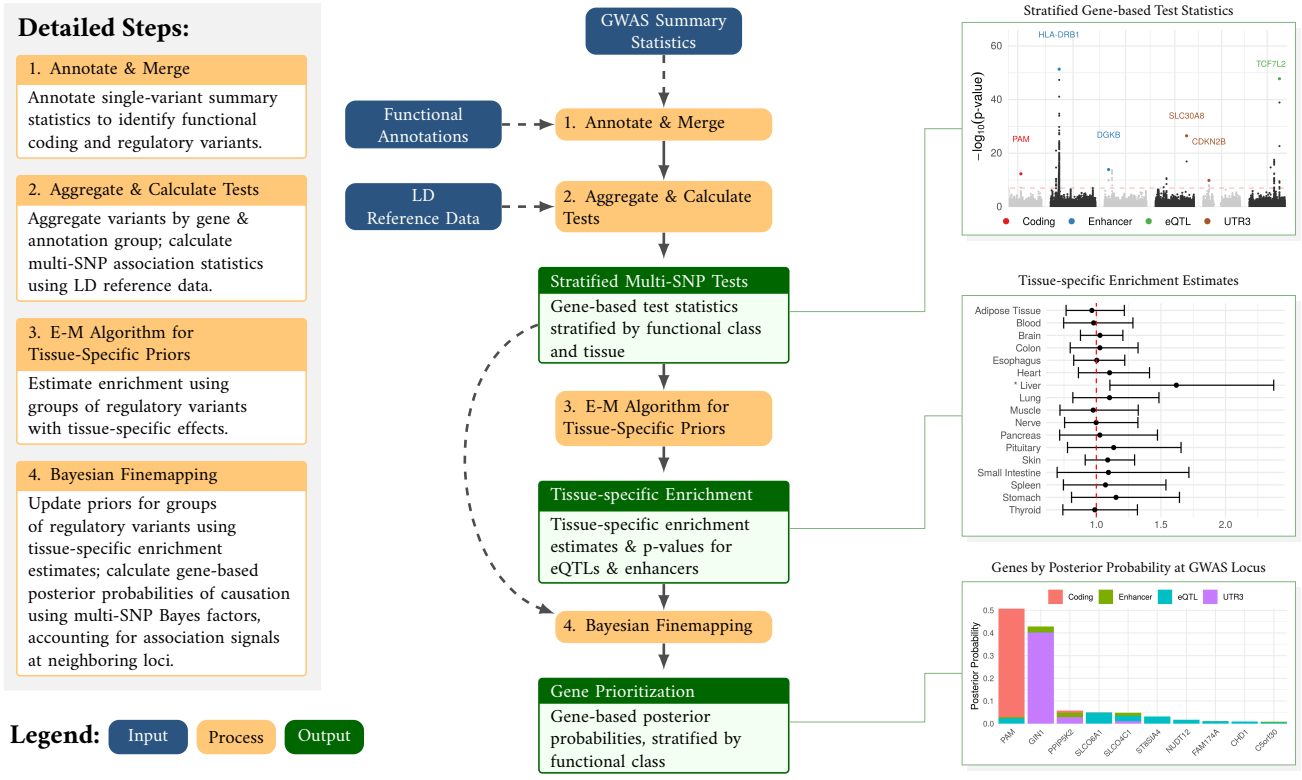
Phenotype	Category	n_{eff}	No. cases	No. controls
Gout	Endocrine/Metabolic	12679.9	3195	404630
Hypercholesterolemia	Endocrine/Metabolic	122016.0	33242	370349
Hyperlipidemia	Endocrine/Metabolic	130804.5	35844	372951
Hypothyroidism NOS	Endocrine/Metabolic	54687.0	14171	388068
Lipoid Metabolism Disorders	Endocrine/Metabolic	131083.3	35927	373034
Mineral Metabolism Disorders	Endocrine/Metabolic	8463.8	2127	406834
Overweight, Obesity & Other Hyperalimentation	Endocrine/Metabolic	42695.4	10968	397993
Type 2 Diabetes	Endocrine/Metabolic	72247.7	18945	387496
Angina Pectoris	Circulatory System	61928.5	16175	361420
Atrial Fibrillation & Flutter	Circulatory System	56983.3	14820	367697
Cardiac Dysrhythmias	Circulatory System	92666.4	24681	377558
Circulatory Disease NEC	Circulatory System	62782.7	16366	383215
Hypertension	Circulatory System	252255.9	77977	329748
Ischemic Heart Disease	Circulatory System	115780.3	31355	376600
Myocardial Infarction	Circulatory System	45325.9	11703	356948
Pulmonary Heart Disease	Circulatory System	16848.7	4257	400046
Muscle, Ligament, & Fascia Disorders	Musculoskeletal	17726.6	4488	352949
Rheumatoid Arthritis	Musculoskeletal	17412.1	4412	325621
Cholelithiasis	Digestive	53215.7	13777	387430
Inguinal Hernia	Digestive	61024.6	15995	330268
Psoriasis	Dermatologic	8896.9	2237	389674
Breast Cancer	Neoplasms	49902.8	12898	381035
Skin Cancer	Neoplasms	53157.0	13752	394933
Asthma	Respiratory	98301.4	26332	368381
Delirium Dementia & Amnestic & Other Cognitive Disorders	Mental Disorders	7841.2	1970	397775

4.3 Results

We first describe GaMBIT (**G**WAS **a**nd **M**ulti-Omics: **B**ayesian Inference and Data **I**ntegration **T**oolkit), an open-source software implementation of the proposed methods. Next, we evaluate 1) the Type I error rates of gene-based test statistics, 2) power and specificity to identify tissue-specific enrichment, 3) performance identifying causal mechanisms underlying association, and 4) performance identifying causal genes for GaMBIT and existing methods through GWAS simulations. Finally, we discuss an application to 25 complex traits using GWAS summary statistics from the UK Biobank. We assess 1) the empirical power of GaMBIT and existing gene-based tests by comparing the numbers of independent genes identified at standard GWAS significance and FDR thresholds, 2) tissue-specific enrichment across traits, and 3) concordance with known Mendelian genes for related traits from the OMIM database using GaMBIT and existing methods.

4.3.1 Software Implementation

Figure 4.3.1: Overview of GaMBIT Method & Workflow



4.3.2 GWAS Simulations

We simulated 650 sets of whole-genome GWAS summary statistics following the procedures described in Methods 4.2.5 using LD reference data from the European subset of the 1KGP Phase 3 reference panel. To evaluate Type I error rates, we simulated 500 data sets under the null hypothesis (no genetic effects on traits). To assess the Type I error rates of eQTL enrichment statistics in the presence of association signals, we simulated 50 data sets with 5-10 coding associations and 2-5 UTR3 associations, but no causal eQTL effects. Finally, we simulated 100 traits with 1-16 eQTL

associations across 1-3 tissues, 2-10 coding associations, and 3-5 UTR associations. Each GWAS data set was simulated with GWAS sample size 100,000, and each causal locus accounted for 0.05-0.10% of trait variance.

Evaluation of Type I Error Rates using Null GWAS Simulations

We evaluated the Type I error rate of gene-and-annotation stratified gene based tests (linear combinations of z-scores for eQTLs, and the sum of squared z-scores for other annotation classes) and GaMBIT gene-based tests (the adjusted minimum p-value across stratified tests for each gene) using 500 sets of whole-genome GWAS summary statistics simulated under the null hypothesis. As expected, GaMBIT gene-based test statistics were slightly conservative due to unadjusted Bonferroni correction. All other test statistics maintained Type I error consistent with the desired alpha level (Table 4.5).

Table 4.5: Evaluation of Type I Error Rates in Simulations

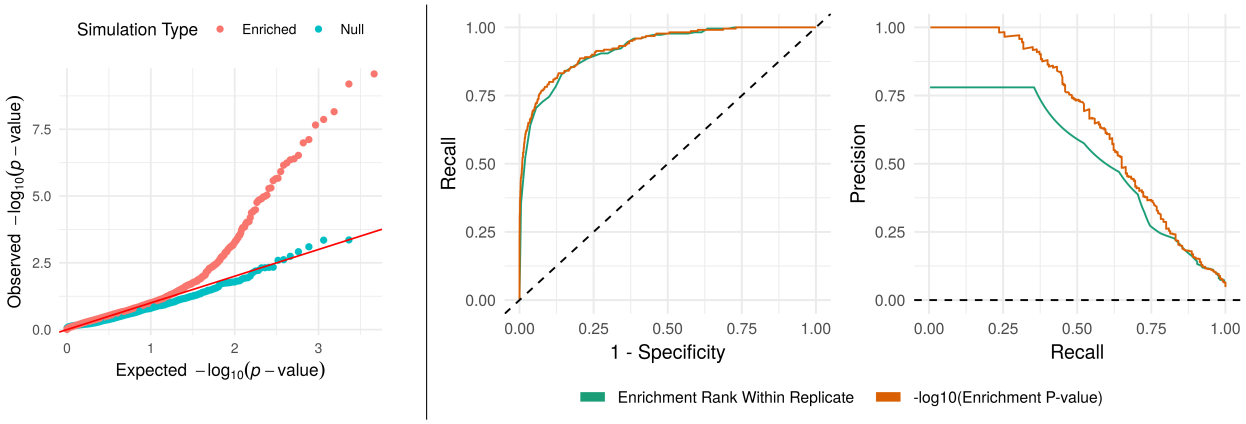
		Alpha-level Threshold							
	No. Reps	5e-08	1e-06	2.5e-06	5e-06	0.001	0.005	0.01	0.05
Coding	8,844,000	0	7.9e-07	3.4e-06	6.0e-06	0.00097	0.0048	0.0096	0.049
Enhancer	82,761,500	4.8e-08	1.3e-06	2.9e-06	5.8e-06	0.00100	0.0048	0.0096	0.048
Exon	8,349,000	0	1.8e-06	4.2e-06	7.4e-06	0.00105	0.0049	0.0096	0.047
Proximal	2,931,000	0	6.8e-07	2.0e-06	2.7e-06	0.00092	0.0048	0.0097	0.049
UTR3	6,757,000	0	1.9e-06	3.8e-06	7.0e-06	0.00099	0.0048	0.0095	0.048
UTR5	4,276,000	0	1.2e-06	2.8e-06	5.4e-06	0.00093	0.0048	0.0097	0.049
eQTL	167,598,000	6.0e-09	4.1e-07	1.5e-06	3.3e-06	0.00093	0.0048	0.0096	0.049
GaMBIT	16,679,000	0	5.4e-07	2.2e-06	4.3e-06	0.00079	0.0039	0.0076	0.037

Observed Type-I error rates for gene-based tests stratified by functional class and for GaMBIT gene-based tests (adjusted minimum p-value for each gene) across 500 GWAS traits simulated under the null hypothesis (no genetic effects on trait). Each simulated trait provides 566,350 annotation-stratified gene-based tests in total across 33,435 unique genes.

Identifying Relevant Tissues: Sensitivity & Type I Error Rate in GWAS Simulations

We assessed Type I error and sensitivity to detect relevant tissues (tissues with one or more causal eQTL gene for a given trait) in simulated GWAS data sets. Type I error was well controlled in null simulations (with causal coding or UTR3 variants, but no tissue-specific effects), and relevant tissues were detected with relatively high sensitivity in simulations with tissue-specific enrichment (Figure 4.3.2).

Figure 4.3.2: Tissue-Specific Enrichment in Simulated Data: Sensitivity & Type I Error Rate



Left: Q-Q plots of enrichment p-values for 50 null simulation replicates (no causal eQTL genes for any tissue, but causal coding and UTR3 variants) and 100 enriched simulation replicates (1-8 causal eQTL genes for each of 1-3 tissues, as well as causal coding and UTR3 variants at other loci).

Right: Performance identifying enriched tissues across 100 simulation replicates. Ranking of tissues within replicates is shown in green (the quantile of 1-sided enrichment -$\log_{10}$ p-values within each replicate; AU-ROC=0.92 and AU-PR=0.53), and aggregate performance is shown in red (raw 1-sided enrichment -$\log_{10}$ p-values across replicates; AU-ROC=0.92 and AU-PR=0.66). The overall percentage of tissues with one or more casual eQTL gene per replicate is 4.8% (range 2.2%-6.5% across replicates).

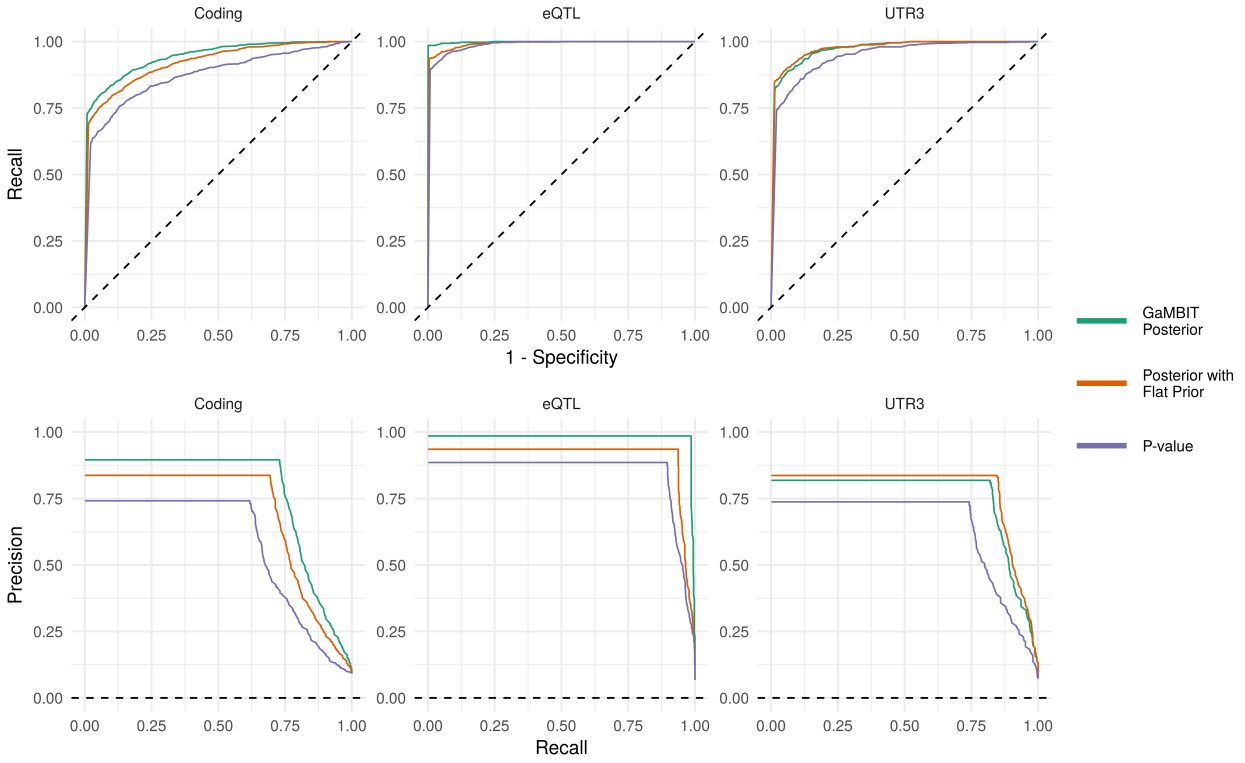
Sensitivity to Identify Causal Mechanisms Underlying Associations in GWAS Simulations

We assessed performance identifying the causal mechanism underlying association at causal genes (for example, liver-specific eVariants or protein-altering variants) in simulated data using

GaMBIT's gene-and-annotation stratified posterior probabilities (leveraging functional enrichment to re-weight annotation-stratified statistics for each gene), posterior probabilities with a flat prior (weighting all tissues and functional annotations equally), and p-values for gene-and-annotation stratified test statistics. Results include 1,064 coding-effect genes (genes for which coding variants are causal), 655 UTR3 genes (genes for which UTR3 variants are causal), and 764 eQTL genes (genes for which eVariants are causal) across 150 simulated traits.

Overall, GaMBIT posterior probabilities had the highest sensitivity to detect causal mechanisms (Figure 4.3.3). Posterior probabilities with flat prior had slightly higher sensitivity than GaMBIT posterior probabilities at UTR3 genes, reflecting the smaller numbers of UTR3 genes relative to coding-effect and eQTL genes for most simulated traits. Indeed, for simulated traits with the strongest UTR3 enrichment (43 simulated traits for which UTR3 genes make up $\geq$30% of all causal genes; a total of 185 UTR3 genes), GaMBIT outperformed flat-prior posterior probabilities at UTR3 genes (AU-ROC = 0.99, 0.98, 0.96 and AU-PR = 0.87, 0.83, 0.73 for GaMBIT, flat-prior, and p-values respectively).

Figure 4.3.3: ROC & PR Curves for Identifying Causal Mechanisms in Simulated Data



Receiver Operating Characteristic (ROC; top row) and Precision-Recall (PR; bottom row) curves for identifying causal mechanisms at 2,483 causal loci across 150 simulated GWAS traits (7-25 causal loci per trait). Performance is assessed by ranking tissue-and-annotation stratified statistics for each causal gene, where each causal gene has a single causal annotation class (and for eQTL genes, a single causal tissue). The mean number of tissue-and-annotation stratified statistics per causal gene is 13.8 (range 2-185; median = 11).

GaMBIT Posterior: Posterior probability that annotation class/tissue underlies association for a given gene, leveraging tissue-specific enrichment in empirical Bayes priors (AU-ROC=0.96 and AU-PR=0.78).

Posterior with Flat Prior: Posterior probability that annotation class/tissue underlies association for a given gene, assigning equal prior weight to each tissue and functional annotation class (AU-ROC=0.95 and AU-PR=0.72).

P-value: P-values for tissue-and-annotation stratified tests for each gene (AU-ROC=0.92 and AU-PR=0.61).

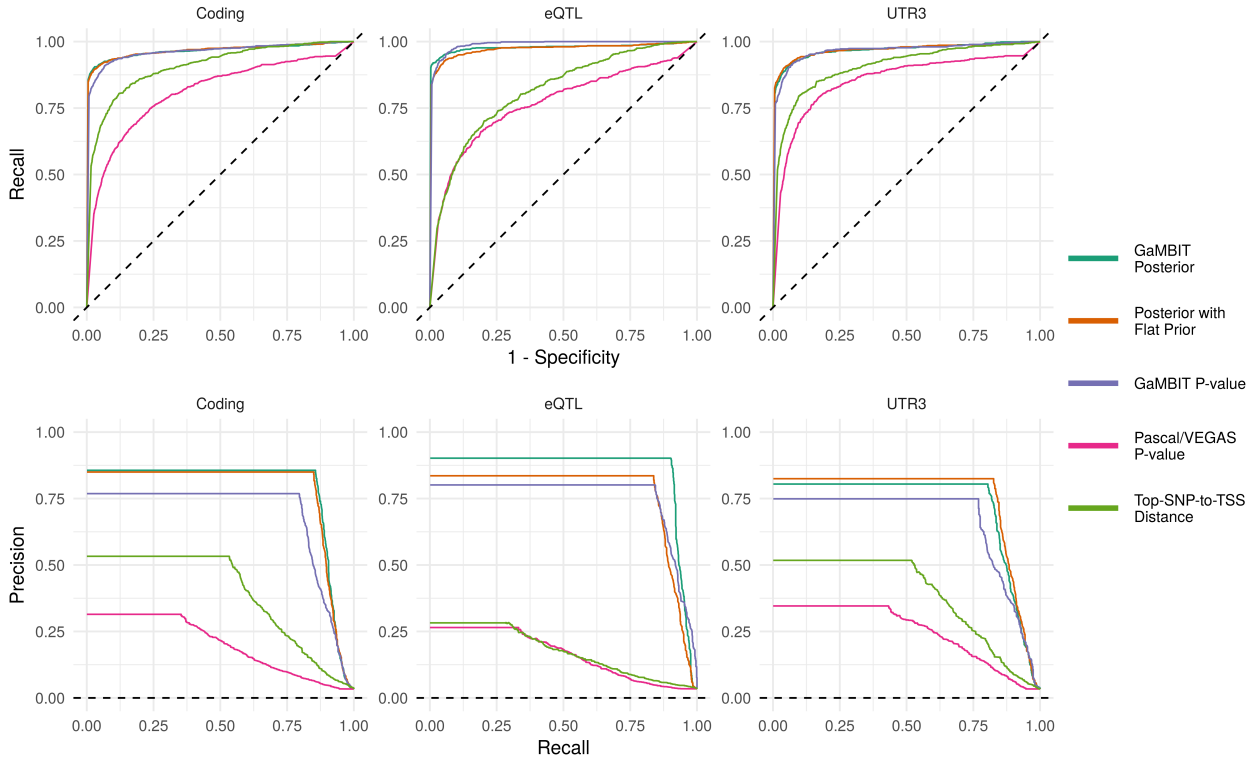
Sensitivity to Identify Causal Genes in GWAS Simulations

We assessed performance identifying causal genes at associated loci using GaMBIT gene-based test p-values, GaMBIT gene-based posterior probabilities (leveraging functional enrichment to re-weight

annotation-stratified statistics), gene-based posterior probabilities with a flat prior (weighting each tissue and functional annotation class equally, but otherwise equivalent to GaMBIT posterior probabilities), Pascal/VEGAS gene-based p-values (the sum of squared z-scores for all variants within 50 Kbp of each gene), and top-SNP-to-TSS distance (ranking each gene at a locus by the genomic distance between its TSS and the most significant independent GWAS variant). Here, posterior probabilities are calculated by summing across gene-and-annotation stratified posterior probabilities for each gene. We evaluate performance for each method by ranking genes within each causal locus (defined as the region +/- 1Mbp of a causal gene), and aggregating across all causal loci and simulated traits.

GaMBIT posterior probabilities generally provided the highest sensitivity to identify causal genes, although flat-prior posterior probabilities had slightly higher sensitivity at UTR3 loci. Similar to the previous section, this trend reflects UTR3 being the least-enriched annotation class for 75/150 simulated traits. For simulated traits with the strongest UTR3 enrichment ($\geq$30% of all causal genes have causal UTR3 variants), GaMBIT posterior had the highest performance identifying causal genes at UTR3 loci (AU-ROC = 0.97, 0.97, 0.97, 0.81, 0.88 and AU-PR = 0.78, 0.76, 0.69, 0.20, 0.28 for GaMBIT posterior, flat-prior posterior, GaMBIT p-value, Pascal/VEGAS p-value, and top-SNP-to-TSS distance respectively).

Figure 4.3.4: ROC & PR Curves for Identifying Causal Genes in Simulated Data



Receiver Operating Characteristic (ROC; top row) and Precision-Recall (PR; bottom row) curves for identifying causal genes at across 150 simulated GWAS traits with 7-25 causal loci per trait. Performance is assessed by ranking genes within each causal locus for each trait, where a causal locus is defined as the set of genes +/- 1Mbp from a causal gene. Further details are provided in Table 4.6.

GaMBIT Posterior: Posterior probability that a gene is causal, leveraging tissue-specific enrichment in empirical Bayes priors.

Posterior with Flat Prior: Posterior probability that a gene is causal, assigning equal prior weight to each tissue and functional annotation class.

GaMBIT P-value: Adjusted minimum p-value across gene-based tests stratified by tissue and functional annotation class for each gene.

Pascal/VEGAS P-value: Annotation-agnostic SKAT p-value calculated by aggregating all variants +/- 50 Kbp from each gene.

Top-SNP-to-TSS Distance: Distance from TSS to nearest top independent GWAS variant (most significant p-value in sliding 2 Mbp window).

Table 4.6: Performance Identifying Causal Genes in Simulated Data

Predictor	AU-PRC	AU-ROC	Top-Ranked
Causal Class: *cis*-eVariants (765 Loci)			
GaMBIT P-value	0.75	**0.99**	0.84
GaMBIT Posterior	**0.85**	0.98	**0.90**
Posterior with Flat Prior	0.76	0.97	0.84
Pascal/VEGAS P-value	0.17	0.77	0.35
Top-SNP-to-TSS Distance	0.18	0.81	0.28
Causal Class: Coding Variants (1,065 Loci)			
GaMBIT P-value	0.68	0.96	0.80
GaMBIT Posterior	**0.78**	0.96	**0.86**
Posterior with Flat Prior	0.77	**0.97**	0.85
Pascal/VEGAS P-value	0.20	0.81	0.37
Top-SNP-to-TSS Distance	0.39	0.91	0.53
Causal Class: UTR3 Variants (655 Loci)			
GaMBIT P-value	0.66	0.96	0.77
GaMBIT Posterior	0.72	**0.97**	0.80
Posterior with Flat Prior	**0.75**	**0.97**	**0.83**
Pascal/VEGAS P-value	0.25	0.85	0.46
Top-SNP-to-TSS Distance	0.39	0.91	0.52

Proportion of causal genes that are top-ranked within locus (defined by +/- 1Mbp of causal gene) according to each method across 2,725 causal loci from 150 simulated traits. Leveraging tissue-specific enrichment to re-weight associations (GaMBIT Posterior) improves performance identifying causal genes at eQTL loci by 7.6% relative to weighting each tissue and functional class equally (Posterior with Flat Prior).

AU-ROC: Area Under Receiver Operating Characteristic (ROC) curve.

AU-PR: Area Under Precision-Recall (PR) curve.

Top-Ranked: Proportion of loci at which top-ranked gene according to method is causal.

4.3.3 Application to the UK Biobank: Analysis of 25 Complex Traits

We selected 25 traits from the UK Biobank for primary analysis with effective sample size of at least 7,500, one or more significant GWAS association, and one or more relevant Mendelian gene (Table 4.4).

Empirical Power: Numbers of Loci Discovered Across UK Biobank Traits

We compared the numbers of significant independent loci detected for UK Biobank traits using GaMBIT gene-based tests, Pascal/VEGAS gene-based tests, and GWAS single-variant analysis. For GaMBIT and Pascal/VEGAS, we assessed the numbers of CCDS protein-coding genes reaching gene-based test significance (p-value < 2.5e-6) and discovered at 5% FDR threshold, counting only the most significant gene within a sliding 2 Mbp window. For single-variant analysis, we similarly counted the numbers of genome-wide significant independent variants (p-value < 5e-8), including only the most significant variant within a sliding 2 Mbp window.

Single-variant analysis identified more significant loci than either gene-based approach overall and almost uniformly across traits. This is unsurprising, as many single-variant associations are intergenic and not in close proximity to any protein-coding gene. Despite using fewer variants and a higher burden of multiple-testing (applied at the gene level), GaMBIT detected more significant genes than Pascal/VEGAS in total across traits and for each trait individually, and discovered more genes at 5% FDR than Pascal/VEGAS for 72% of all traits. This suggests that GaMBIT's functional annotation strategy effectively reduces noise and increases power overall, despite sacrificing a large number of variants.

We also assessed the numbers of genes detected by using gene-and-annotation stratified test statistics individually. GaMBIT gene-based tests (adjusted minimum p-value across annotation-stratified statistics for each gene) consistently detected more significant independent genes than any annotation class used individually across traits (85.6%, 17.1%, 83.3%, and 31.9% more than gene-based tests from coding variants, enhancers, UTR variants, and eQTLs respectively). This suggests that GaMBIT's stratify-and-combine approach increases power overall, despite having a higher burden of multiple testing. However, substantially more loci were discovered at 5% FDR by applying an overall FDR adjustment across all gene-and-annotation stratified tests together rather

than applying FDR adjustment to GaMBIT's aggregated gene-based tests directly (Table 4.7).

Table 4.7: Empirical Power: Number of Independent Loci Discovered for UK Biobank Traits

Trait	Single-Variant	Pascal/VEGAS	GaMBIT	Stratified
Cancer				
Skin cancer	35	15 (60)	25 (66)	134
Breast cancer	24	11 (34)	17 (40)	72
Endocrine/Metabolic				
Hypothyroidism NOS	51	28 (159)	45 (178)	328
Type 2 diabetes	46	24 (189)	39 (185)	306
Lipoid Metabolism Disorders	37	27 (134)	32 (127)	255
Hyperlipidemia	38	27 (137)	33 (133)	254
Hypercholesterolemia	34	24 (119)	30 (124)	249
Gout	7	4 (11)	6 (13)	23
Disorders of mineral metabolism	4	2 (5)	3 (12)	38
Overweight, obesity and other hyperalimentation	9	5 (28)	7 (31)	71
Circulatory System				
Ischemic Heart Disease	40	21 (133)	33 (142)	256
Myocardial infarction	21	13 (49)	18 (47)	100
Angina pectoris	20	10 (65)	19 (55)	111
Pulmonary heart disease	9	4 (5)	8 (10)	17
Atrial fibrillation and flutter	36	27 (73)	30 (82)	149
Circulatory disease NEC	6	4 (9)	6 (10)	20
Digestive				
Inguinal hernia	22	13 (52)	21 (67)	128
Cholelithiasis	17	15 (48)	19 (41)	92
Musculoskeletal				
Rheumatoid arthritis	6	1 (13)	5 (18)	63
Muscle, ligament, & fascia disorders	23	13 (33)	21 (34)	67
Other				
Asthma	31	22 (126)	28 (122)	258
Psoriasis	7	4 (12)	6 (22)	91

Single-Variant: Number of independent genome-wide significant (p-value $\leq 5 \times 10^{-8}$) variants.

Pascal/VEGAS: Number of significant independent genes (gene-based p-value $\leq 2.5 \times 10^{-6}$) and number of independent genes discovered by Pascal/VEGAS gene-based tests at 5% FDR (in parenthesis).

GaMBIT: Number of significant independent genes (gene-based p-value $\leq 2.5 \times 10^{-6}$) and number of independent genes discovered by GaMBIT gene-based tests at 5% FDR (in parenthesis).

Stratified: Number of independent genes discovered at 5% FDR using stratified gene-and-annotation GaMBIT test statistics.

Tissue-Specific Enrichment

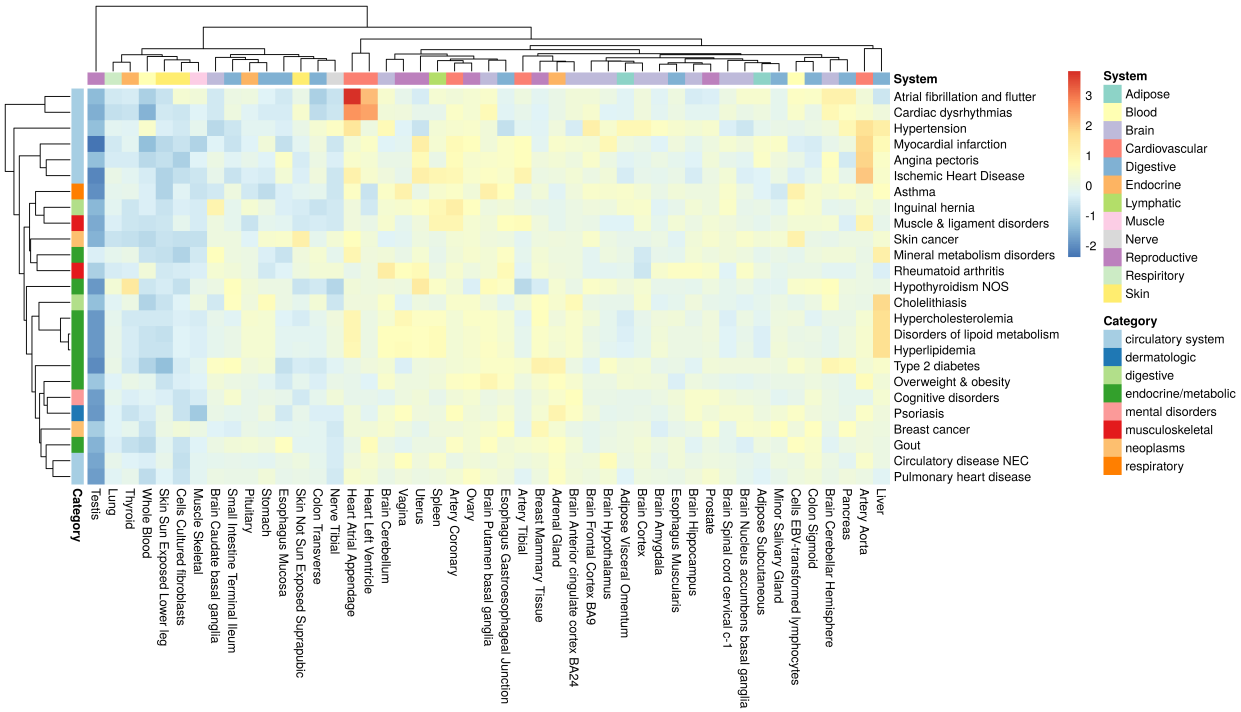
We assessed tissue-specific eQTL enrichment across the 25 selected complex traits from the UK Biobank. We detected strong enrichment in heart tissues for atrial fibrillation and other cardiovascular traits, and enrichment in liver for multiple endocrine and metabolic traits (Table 4.8). However, relatively few traits exhibited significant tissue-specific enrichment overall. This may reflect the imprecision of eQTL weights (due to the limited sample size of GTEx and imbalance across tissues), limited statistical power (due to the limited effective sample size for many traits), polygenic effects of small magnitude dispersed across many genes (which are less likely to be detected by GaMBIT), or perhaps genuine genetic etiology. Here, it is also important to note that our simulations assumed eQTL effects are measured without error, and therefore likely overestimate power to detect enrichment in real data.

Table 4.8: Tissue-Specific eQTL Enrichment across UK Biobank Traits

Trait	Term	Coef. (SE)	Pval
Circulatory			
Angina Pectoris	Artery Aorta	1.24 (0.35)	2e-04
	Liver	1.45 (0.44)	5e-04
Atrial Fib. & Flutter	Heart Atrial Appendage	1.52 (0.24)	9e-11
	Heart Left Ventricle	1.09 (0.28)	5e-05
	Brain Cerebellar Hemisphere	0.97 (0.36)	3e-03
Cardiac Dysrhythmias	Heart Atrial Appendage	1.28 (0.35)	1e-04
	Heart Left Ventricle	1.09 (0.38)	2e-03
Hypertension	Artery Aorta	1.30 (0.52)	6e-03
Ischemic Heart Disease	Artery Aorta	0.93 (0.26)	2e-04
	Liver	1.07 (0.35)	1e-03
Myocardial Infarction	Artery Aorta	0.92 (0.33)	3e-03
	Liver	1.15 (0.43)	4e-03
Pulmonary Heart Disease	Vagina	1.83 (0.60)	1e-03
Endocrine/Metabolic			
Hypercholesterolemia	Liver	0.85 (0.34)	6e-03
Overweight, Obesity, & Other Hyperalimentation	Liver	1.20 (0.51)	9e-03
Type 2 Diabetes	Liver	0.74 (0.31)	9e-03
Mental Disorders			
Delirium Dementia,	Adrenal Gland	5.64 (1.60)	2e-04
Amnestic, & Other	Esophagus Mucosa	5.00 (1.67)	1e-03
Cognitive Disorders	Nerve Tibial	4.85 (1.67)	2e-03
	Brain Anterior Cingulate Cortex BA24	5.48 (2.32)	9e-03
Musculoskeletal			
Muscle, Ligament, & Fascia Disorders	Artery Aorta	1.01 (0.38)	4e-03
Neoplasms			
Breast Cancer	Testis	1.38 (0.56)	7e-03
Respiratory			
Asthma	Prostate	0.58 (0.23)	6e-03
	Liver	0.59 (0.24)	7e-03

eQTL enrichment estimates and p-values for tissues with marginally significant p-values ($p < 0.01$) across traits.

Figure 4.3.5: Heatmap of Tissue-Specific Enrichment

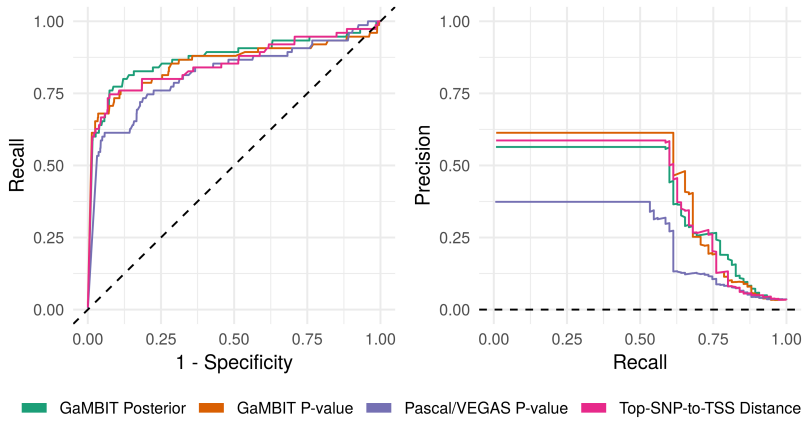


Clustering of tissues and traits from tissue-specific enrichment estimates.

Concordance with OMIM Genes

To assess GaMBIT's performance identifying causal genes in real data, we first extracted lists of known Mendelian genes for related traits the OMIM database for each of the 25 UK Biobank traits. We next filtered the lists of OMIM genes, retaining only genes within 1 Mbp of a significant GWAS association (single-variant p-value < 5e-8) or a significant gene-based test (Pascal/VEGAS or GaMBIT p-value < 2.5e-6) for the corresponding UK Biobank trait. After filtering, we retained 55 unique OMIM genes across traits from the original list of 353 unique genes.

Figure 4.3.6: ROC and PR Curves for OMIM Genes in the UK Biobank



Receiver Operating Characteristic (ROC; left) and Precision-Recall (PR; right) curves for 55 OMIM genes in total across 25 traits in the UK Biobank. Only OMIM genes within +/- 1 Mbp of a genome-wide significant association (p-value $<$ 5e-8) or significant Pascal/VEGAS or GaMBIT gene-based test (p-value $<$ 2.5e-6) are included (55 OMIM genes from a total of 404). Performance is assessed by ranking genes within each OMIM locus for each trait, where an OMIM locus is defined as the set of genes +/- 1Mbp from an OMIM gene.

GaMBIT Posterior: Posterior probability that gene causal (AU-ROC=0.87, AU-PR=0.41).

GaMBIT P-value: Adjusted minimum p-value across gene-based tests stratified by tissue and functional annotation class for each gene (AU-ROC=0.86, AU-PR=0.44).

Pascal/VEGAS P-value: Annotation-agnostic SKAT p-value calculated by aggregating all variants +/- 50 Kbp from each gene (AU-ROC=0.85, AU-PR=0.26).

Top-SNP-to-TSS Distance: Distance from TSS to nearest top independent GWAS variant (most significant p-value in sliding 2 Mbp window) (AU-ROC=0.85, AU-PR=0.42).

Table 4.9: Concordance at OMIM Loci in UK Biobank Data: Quantile Rank of OMIM Genes

	Mean Locus Quantile Rank of OMIM Genes for:				
	GaMBIT		Top SNP to	Pascal/VEGAS	No.
Trait Category	Post. Pr.	P-value	TSS Dist.	P-value	Genes
Neoplasms	0.83	0.83	0.73	0.74	8
Endocrine/metabolic	0.90	0.87	0.90	0.88	26
Circulatory System	0.91	0.88	0.96	0.82	7
Respiratory	0.88	0.67	0.77	0.86	3
Digestive	0.97	1.00	1.00	0.70	3
Dermatologic	0.75	0.76	0.72	0.71	3
Musculoskeletal	0.64	0.75	0.42	0.73	5
Overall	0.86	0.84	0.83	0.82	55

Average quantile of OMIM genes within each locus (+/- 1Mbp of each OMIM gene) across traits in each category for each method. Quantiles of OMIM genes that appear for multiple traits in a given category are first averaged across traits so that each OMIM gene contributes equally to the total. The average quantile of OMIM genes is approximately equal to AU-ROC.

Table 4.10: Concordance at OMIM Genes in UK Biobank Data: Proportion of Top-Ranked OMIM Genes

	Proportion of Top-Ranked OMIM Genes for:				
	GaMBIT		Top SNP to	Pascal/VEGAS	No.
Trait Category	Post. Pr.	P-value	TSS Dist.	P-value	Genes
Neoplasms	0.50	0.62	0.25	0.25	8
Endocrine/metabolic	0.64	0.65	0.62	0.62	26
Circulatory System	0.57	0.43	0.71	0.29	7
Respiratory	0.33	0.33	0.33	0.33	3
Digestive	0.67	1.00	1.00	0.67	3
Dermatologic	0.33	0.00	0.33	0.33	3
Musculoskeletal	0.20	0.40	0.20	0.00	5
Overall	0.54	0.56	0.53	0.44	55

Proportion of OMIM genes that are top-ranked within each locus (+/- 1Mbp of each OMIM gene) across traits in each category for each method. For OMIM genes that correspond with multiple traits, proportions are averaged across traits within each category.

4.4 Discussion

Here, we described novel Bayesian methods to identify likely causal genes, pathways, and mechanisms underlying GWAS associations. We also outlined our efforts to assemble a comprehensive compendium of gene-centric regulatory and functional annotations by aggregating databases derived from Roadmap/ENCODE (Cao et al. 2017; Bernstein et al. 2010; ENCODE Project Consortium 2012), FANTOM5 (Marbach et al. 2016; Cao et al. 2017; Lizio et al. 2015), and GTEx (GTEx Consortium 2015; Gamazon et al. 2015; Barbeira et al. 2016). To apply our approach with GWAS summary statistics, we used our approach described Chapter 2 to estimate LD from reference panels on-the-fly for gene-based association statistics. Finally, we presented a novel, open-source computational toolkit, GaMBIT, that allows researchers to simultaneously examine a range of potential regulatory and functional perterbations underlying GWAS association signals, and permits both Frequentist and Bayesian inference.

4.5 Acknowledgments

I thank William Wen for many useful discussions and suggestions regarding Bayesian inference, algorithms, finemapping, and GTEx data. I thank Sayantan Das and Alan Kwong for their help estimating predictive weights using GTEx. I thank Gonçalo Abecasis, Michael Boehnke, and Hyun Min Kang for feedback and suggestions regarding algorithms and evaluating causal genes. I thank Sarah Gagliano, Jonas Nielsen, and Cristen Willer for assistance with GWAS summary statistics and the UK Biobank resource. This research has been conducted using the UK Biobank Resource under Application Number 24460.

4.6 Appendix: Supplementary Tables & Figures

Supplementary Table 4.1: Ranking Genes at OMIM Loci for UK Biobank Traits

		GaMBIT		Top SNP to	Pascal/VEGAS	No.
	Gene	Post. Pr.	P-value	TSS Dist.	P-value	Traits
Neoplasms	CDKN2A	0.86	0.90	0.95	0.95	1
	CDSN	0.77	0.64	0.13	0.79	1
	IRF4	1.00	1.00	1.00	1.00	1
	MC1R	1.00	1.00	0.96	0.25	1
	SLC45A2	1.00	1.00	1.00	0.82	1
	TUBB	0.09	0.12	0.49	0.25	1
	TYR	1.00	1.00	0.89	1.00	1
	CHEK2	0.95	1.00	0.40	0.85	1
Endocrine/metabolic	ABCC8	0.91	0.91	0.95	0.86	1
	GCKR	1.00	0.89	1.00	0.84	1
	HNF1A	1.00	1.00	0.97	1.00	1
	HNF1B	0.96	1.00	1.00	0.83	1
	IGF2BP2	1.00	1.00	0.94	1.00	1
	INS	0.84	0.23	0.55	0.65	1
	INSR	0.94	0.74	0.49	0.83	1
	IRS1	1.00	1.00	1.00	0.44	1
	KCNJ11	1.00	1.00	1.00	1.00	1
	MTNR1B	1.00	1.00	1.00	1.00	1
	RETN	0.13	0.29	0.67	0.58	1
	SLC30A8	1.00	1.00	1.00	1.00	1
	TCF7L2	1.00	1.00	1.00	1.00	1
	WFS1	1.00	1.00	1.00	1.00	1
	APOB	1.00	1.00	1.00	1.00	3
	HMGCR	0.92	1.00	1.00	1.00	3
	LDLR	1.00	1.00	1.00	1.00	3
	LIPC	1.00	1.00	1.00	1.00	3
	LPL	1.00	1.00	1.00	1.00	3
	PCSK9	1.00	1.00	1.00	1.00	3
	SCARB1	1.00	1.00	1.00	1.00	3
	SORT1	0.99	0.97	0.93	1.00	3
	LDLRAP1	0.03	0.06	0.30	0.08	2
	DYRK1B	0.66	0.50	0.68	0.70	1
	SLC17A3	0.94	0.96	0.98	0.98	1
	MC4R	1.00	1.00	1.00	1.00	1
Circulatory System	APOE	1.00	1.00	1.00	1.00	1
	MMP3	0.83	0.77	0.81	0.19	3
	NOS3	0.60	0.71	1.00	0.97	1
	MYH6	1.00	1.00	1.00	0.97	2
	NKX2-5	1.00	1.00	1.00	1.00	1
	SCN5A	0.92	0.88	0.92	0.96	1
	TAB2	1.00	0.78	1.00	0.67	1
Respiratory	CHI3L1	0.93	0.96	0.93	0.86	1
	IL13	1.00	1.00	1.00	1.00	1
	TNF	0.71	0.03	0.39	0.73	1
Digestive	ABCB4	0.92	1.00	1.00	1.00	1
	ABCG8	1.00	1.00	1.00	1.00	1
	SLC10A2	1.00	1.00	1.00	0.11	1
Dermatologic	HLA-C	0.99	0.92	0.99	0.33	1
	PSMB8	0.25	0.45	0.16	0.81	1
	TRAF3IP2	1.00	0.92	1.00	1.00	1
Musculoskeletal	LTA	0.38	0.72	0.02	0.72	1
	NFKBIL1	0.38	0.72	0.03	0.71	1
	PTPN22	1.00	1.00	1.00	0.84	1
	IL10	0.96	1.00	0.64	0.84	1
	PLEC	0.49	0.32	0.42	0.53	1

4.7 References

Barbeira, A et al. (2016). "MetaXcan: Summary statistics based gene-level association method infers accurate PrediXcan results. bioRxiv: 045260". In:

Benner, Christian et al. (2016). "FINEMAP: efficient variable selection using summary data from genome-wide association studies". In: *Bioinformatics* 32.10, pp. 1493–1501.

Bernstein, Bradley E et al. (2010). "The NIH roadmap epigenomics mapping consortium". In: *Nature biotechnology* 28.10, pp. 1045–1048.

Cao, Qin et al. (2017). "Reconstruction of enhancer-target networks in 935 samples of human primary cells, tissues and cell lines". In: *Nature genetics* 49, p. 1428.

Conneely, Karen N and Michael Boehnke (2007). "So many correlated tests, so little time! Rapid adjustment of P values for multiple correlated tests". In: *The American Journal of Human Genetics* 81.6, pp. 1158–1168.

Davies, Robert B (1980). "Algorithm AS 155: The distribution of a linear combination of χ 2 random variables". In: *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 29.3, pp. 323–333.

ENCODE Project Consortium (2012). "An integrated encyclopedia of DNA elements in the human genome". In: *Nature* 489.7414, pp. 57–74.

Farh, Kyle Kai-How et al. (2015). "Genetic and epigenetic fine mapping of causal autoimmune disease variants". In: *Nature* 518.7539, pp. 337–343.

Gamazon, Eric R et al. (2015). "PrediXcan: Trait Mapping Using Human Transcriptome Regulation". In: *bioRxiv*, p. 020164.

GTEx Consortium (2015). "The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans". In: *Science* 348.6235, pp. 648–660.

Gusev, Alexander et al. (2016). "Integrative approaches for large-scale transcriptome-wide association studies". In: *Nature genetics* 48.3, pp. 245–252.

Hauberg, Mads Engel et al. (2017). "Large-Scale Identification of Common Trait and Disease Variants Affecting Gene Expression". In: *The American Journal of Human Genetics*.

Hoeting, Jennifer A et al. (1999). "Bayesian model averaging: a tutorial". In: *Statistical science*, pp. 382–401.

Huang, Hailiang et al. (2017). "Fine-mapping inflammatory bowel disease loci to single-variant resolution". In: *Nature* 547.7662, pp. 173–178.

Kang, HM (2014). "Efficient and parallelizable association container toolbox (EPACTS)". In: *University of Michigan Center for Statistical Genetics. Accessed* 6, p. 16.

Kwak, Il-Youp and Wei Pan (2015). "Adaptive gene-and pathway-trait association testing with GWAS summary statistics". In: *Bioinformatics* 32.8, pp. 1178–1184.

Lee, Seunggeun, Michael C Wu, and Xihong Lin (2012). "Optimal tests for rare variant effects in sequencing association studies". In: *Biostatistics* 13.4, pp. 762–775.

Lee, Yeji et al. (2018). "Bayesian Multi-SNP Genetic Association Analysis: Control of FDR and Use of Summary Statistics". In: *bioRxiv*, p. 316471.

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

Liu, Dajiang J et al. (2014). "Meta-analysis of gene-level tests for rare variant association". In: *Nature genetics* 46.2, pp. 200–204.

Liu, Huan, Yongqiang Tang, and Hao Helen Zhang (2009). "A new chi-square approximation to the distribution of non-negative definite quadratic forms in non-central normal variables". In: *Computational Statistics & Data Analysis* 53.4, pp. 853–856.

Liu, Jimmy Z et al. (2012). "Dense fine-mapping study identifies new susceptibility loci for primary biliary cirrhosis". In: *Nature genetics* 44.10, pp. 1137–1141.

Lizio, Marina et al. (2015). "Gateways to the FANTOM5 promoter level mammalian expression atlas". In: *Genome biology* 16.1, p. 22.

Marbach, Daniel et al. (2016). "Tissue-specific regulatory circuits reveal variable modular perturbations across complex diseases". In: *Nature methods* 13.4, pp. 366–370.

Morrison, Alanna C et al. (2013). "Whole-genome sequence-based analysis of high-density lipoprotein cholesterol". In: *Nature genetics* 45.8, p. 899.

Sham, Pak C and Shaun M Purcell (2014). "Statistical power and significance testing in large-scale genetic studies". In: *Nature Reviews Genetics* 15.5, pp. 335–346.

Varin, Cristiano, Nancy Reid, and David Firth (2011). "An overview of composite likelihood methods". In: *Statistica Sinica*, pp. 5–42.

Welter, Danielle et al. (2013). "The NHGRI GWAS Catalog, a curated resource of SNP-trait associations". In: *Nucleic acids research* 42.D1, pp. D1001–D1006.

Wen, Xiaoquan et al. (2016). "Efficient integrative multi-SNP association analysis via deterministic approximation of posteriors". In: *The American Journal of Human Genetics* 98.6, pp. 1114–1129.

Supplementary Table 4.2: OMIM Genes & Traits Used for Analysis of UK Biobank Traits

Gene	OMIM Traits	UK Biobank Traits	Detected
MMP3	Coronary Heart Disease	Angina Pectoris; Cardiovascular Traits	GWAS
CHI3L1	Asthma-related Traits	Asthma	GWAS
IL13	Asthma	Asthma	GWAS,Pascal,GaMBIT
TNF	Asthma; Dementia	Asthma; Delirium Dementia	Pascal
MYH6	Atrial Septal Defect	Atrial Fibrillation; Cardiac Dysrhythmias	Pascal,GaMBIT
NKX2-5	Atrial Septal Defect; Heart Malformations; Hypothyroidism	Cardiovascular Traits	Pascal,GaMBIT
CHEK2	Breast And Colorectal Cancer; Breast Cancer	Breast Cancer; Skin Cancer	GaMBIT
ABCB4	Gallbladder Disease	Cholelithiasis	GWAS,Pascal,GaMBIT
ABCG8	Gallbladder Disease	Cholelithiasis	GWAS,Pascal,GaMBIT
SLC10A2	Bile Acid Malabsorption	Cholelithiasis	GaMBIT
APOB	Hypercholesterolemia	Lipid Traits	GWAS,Pascal,GaMBIT
HMGCR	LDL Level Qtl 3; Statins	Lipid Traits	GWAS,Pascal,GaMBIT
LDLR	Hypercholesterolemia; Ldl Cholesterol Level Qtl2	Lipid Traits	GWAS,Pascal,GaMBIT
PCSK9	Hypercholesterolemia; LDL Level Qtl	Lipid Traits	GWAS,Pascal,GaMBIT
SCARB1	HDL Level Qtl6	Lipid Traits	GWAS,GaMBIT
SORT1	LDL Level Qtl6	Lipid Traits	GWAS,Pascal,GaMBIT
LIPC	T2D; HDL Level Qtl 12	Lipid Traits; Type 2 Diabetes	GaMBIT
LPL	Combined Hyperlipidemia; HDL Level Qtl 11	Lipid Traits; Type 2 Diabetes	Pascal,GaMBIT
PLEC	Muscular Dystrophy	Muscle/Ligament Disorders	GWAS
NFKBIL1	Rheumatoid Arthritis	Muscle/Ligament Disorders; Gout; Rheumatoid Arthritis	Pascal,GaMBIT
PTPN22	Diabetes; Rheumatoid Arthritis	Musculoskeletal Traits; Type 2 Diabetes	GaMBIT
SLC17A3	Gout Susceptibility	Gout	GWAS,GaMBIT
LTA	Myocardial Infarction; Psoriatic Arthritis	Gout; Myocardial Infarction; Rheumatoid Arthritis	Pascal,GaMBIT
APOE	Coronary Artery Disease	Ischemic Heart Disease	GWAS,Pascal,GaMBIT
NOS3	Coronary Artery Spasm 1; Hypertension; Ischemic Stroke	Ischemic Heart Disease	GWAS
IRS1	Coronary Artery Disease; T2D	Ischemic Heart Disease; Type 2 Diabetes	GWAS,GaMBIT
IRS1	Coronary Artery Disease; T2D	Ischemic Heart Disease; Type 2 Diabetes	GWAS
MC4R	Obesity	Overweight, Obesity	GWAS,GaMBIT
HLA-C	Psoriasis Susceptibility	Psoriasis	GWAS,Pascal,GaMBIT
PSMB8	Autoinflammation	Psoriasis	GWAS,Pascal
TRAF3IP2	Psoriasis Susceptibility	Psoriasis	GWAS,Pascal,GaMBIT
MC1R	Uv-induced Skin Damage	Psoriasis; Skin Cancer	GWAS,Pascal,GaMBIT
MC1R	Uv-induced Skin Damage	Psoriasis; Skin Cancer	GWAS
CDKN2A	Pancreatic Cancer/melanoma Syndrome	Skin Cancer	GWAS,Pascal
CDSN	Peeling Skin Syndrome	Skin Cancer	GWAS
IRF4	Skin/hair/eye Pigmentation	Skin Cancer	GWAS,Pascal,GaMBIT
SLC45A2	Skin/hair/eye Pigmentation 5	Skin Cancer	GWAS,GaMBIT
TUBB	Symmetric Circumferential Skin Creases	Skin Cancer	GWAS
TYR	Skin/hair/eye Pigmentation 3	Skin Cancer	GWAS,Pascal,GaMBIT
ABCC8	T2D; Hyperinsulinemic Hypoglycemia	Type 2 Diabetes	GWAS,GaMBIT
GCKR	Fasting Plasma Glucose Level Qtl 5	Type 2 Diabetes	GWAS,Pascal
HNF1A	T2D	Type 2 Diabetes	GaMBIT
HNF1B	T2D; Renal Cysts And Diabetes Syndrome	Type 2 Diabetes	GWAS
IGF2BP2	T2D	Type 2 Diabetes	GWAS,Pascal,GaMBIT
INS	T2D; Hyperproinsulinemia; MODY	Type 2 Diabetes	GWAS
INSR	T2D; Hyperinsulinemic Hypoglycemia	Type 2 Diabetes	GWAS
KCNJ11	Diabetes; T2D; Hyperinsulinemic Hypoglycemia; MODY	Type 2 Diabetes	GWAS,GaMBIT
MTNR1B	T2D	Type 2 Diabetes	GWAS,Pascal
RETN	T2D; Hypertension	Type 2 Diabetes	GWAS
SLC30A8	T2D	Type 2 Diabetes	GWAS,GaMBIT
TCF7L2	T2D	Type 2 Diabetes	GWAS,Pascal,GaMBIT
WFS1	T2D	Type 2 Diabetes	GWAS,GaMBIT

Chapter 5

Discussion

5.1 Summary

In this dissertation, we assessed strategies to improve imputation and statistical power for GWAS of populations that are underrepresented in current imputation reference panels, presented methods to expedite LD computations with increasingly large sample sizes, and developed a statistical framework to identify likely causal genes and mechanisms in post-GWAS analysis leveraging expanding regulatory genomic annotation data. Here, we review these works, discuss their limitations and prospects in the ever-evolving landscape of human genomics, and suggest possible directions for future research.

5.2 Sequencing & Imputation in the Age of Massive Reference Panels

In Chapter 2, we assessed sequencing-and-imputation as a strategy to improve genotype imputation and increase power in GWAS populations that are underrepresented in current imputation reference

panels. While imputation reference panels have increased in size since the release of the 1KGP, they have not always increased in diversity. The largest current imputation reference panels, e.g. HRC and UK10K (UK10K Consortium 2015; McCarthy et al. 2016), are predominantly European, and provide limited imputation quality in non-European, admixed, and isolate populations (Deelen et al. 2014; Lencz et al. 2017). For these populations, we found that sequencing a subset of participants can substantially increase genomic coverage and power to detect association.

Sequencing to construct augmented or population-specific reference panels is expected to be most impactful for populations that are least represented in current imputation reference panels (Van Leeuwen et al. 2015; Roshyara and Scholz 2015). By contrast, our analysis in Chapter 2 was limited to four populations in which relatively large sequencing studies have already been conducted (African Americans, Latino Americans, Sardinians, and Finns). However, we are hopeful that our results can serve as a guidepost for the design and planning of GWAS in other populations with comparable demographic histories but more limited representation in current imputation reference panels.

With larger and more representative imputation reference panels, array-and-imputation based genotyping provides a closer approximation to genotyping by deep sequencing. Forthcoming reference panels, e.g. from the TOPMed WGS Program (NHLBI TOPMed WGS 2018), are substantially larger and more diverse than the HRC. These resources will allow far more accurate and comprehensive imputation for a wider range of populations, so that imputation coverage and accuracy more closely approximate WGS, thereby lessening the utility of study-specific sequencing for many populations and traits. However, the utility of sequencing and imputation for extreme genetic architectures is less clear; e.g., *de novo* mutations cannot be imputed regardless of reference panel size. Thus, the efficacy of population-based sequencing and imputation studies, family-based studies, and other strategies under extreme genetic architectures is an area for further research.

Finally, individual-level genotype data are not publicly available for many of the largest imputation reference panels, e.g. the HRC and UK10K, which must be accessed indirectly through imputation servers (UK10K Consortium 2015; McCarthy et al. 2016). Thus, direct augmentation with other reference panels, e.g. study-specific reference data, is often difficult or impossible. Other strategies to combine results from multiple reference panels include meta-imputation, which strategically combines results from multiple reference panels (Sayantan Das 2017), and a distributed reference panel approach, in which sets of imputed dosages from multiple reference panels are combined for association analysis (Zhou et al. 2017). A more comprehensive evaluation of these and other strategies to indirectly combine multiple reference panels, as well as tools to facilitate analysis with multiple reference panels (e.g., on imputation servers), will be important for future sequencing-and-imputation studies.

5.3 Efficient Computation with Human Genetic Data

In Chapter 3, we developed efficient methods to estimate linkage disequilibrium (LD) with large sample sizes, exploiting the natural sparsity and high redundancy of genetic data to increase computational efficiency. Efficient methods to calculate LD will be critical for the analysis of GWAS summary statistics with large sample sizes, which will require correspondingly precise and comprehensive LD estimates, including for rare and low-frequency variants (Benner, Havulinna, et al. 2017). In addition, because individual-level genotype data are not publicly available for many of the largest WGS datasets (e.g., McCarthy et al. 2016; NHLBI TOPMed WGS 2018), the development of web-based utilities for querying LD without compromising genetic privacy will be important to allow researchers to utilize these resources for analyses involving LD (Quick et al. 2018).

One area for further research in this domain is the development of compressed data formats

that allow efficient LD querying and estimation while avoiding quadratic storage costs, and do not compromise genetic privacy. In addition, sparsity and haplotype structure can be used to improve efficiency in a variety of other contexts; e.g., compact genotype data formats (e.g., M3VCF format; Das and Abecasis 2015; Sayantan Das et al. 2016), and efficient computation for routine tasks and analyses (e.g., sparse representations have been used to efficiently calculate single-variant association tests; Dey et al. 2017).

5.4 Post-GWAS Methods for the Omics Age

In Chapter 4, we developed an integrative Bayesian model and software toolkit, GaMBIT, to identify causitive genes, pathways, and biological mechanisms underlying GWAS associations by leveraging regulatory and functional genomics databases. We demonstrated through simulations that GaMBIT has high precision to identify causal genes and mechanisms given accurate and comprehensive functional genomic annotations. However, despite substantial progress in regulatory genomics, current functional genomic annotations remain incomplete and often imprecise (Bodea et al. 2018; e.g., GTEx sample sizes are limited and imbalanced across tissues [GTEx Consortium 2015; Hao et al. 2018]). Thus, we expect the utility of the proposed methods to increase as functional genomic annotations improve through new technologies and larger studies. In particular, improved methods to infer relationships between regulatory elements and target genes will be important to facilitate the interpretation of non-coding associations (Marbach et al. 2016).

Despite the imprecision of current annotation data, our analysis of 25 traits using GWAS summary association statistics from the UK Biobank identified biologically relevant tissues and showed high concordance with known Mendelian genes. We also demonstrated that LD and pleiotropy can confound gene-based tests, and showed that our framework can provide more reliable

and interpretable gene-based analysis by accounting for multiple possible mechanisms underlying association. Finally, we found that our aggregated annotation-stratified gene-based testing approach has substantially higher power than annotation-agnostic gene-based tests and any individual class of annotation-stratified gene-based test (e.g., TWAS/PrediXcan tests across one or more tissues, Gusev et al. 2016; Gamazon et al. 2015).

GaMBIT's core methods involve a number of simplifying assumptions, which could be relaxed in a variety of possible extensions and generalizations. For example, GaMBIT's Bayesian model assumes that each GWAS association region harbors at most one causal gene and annotation class. This assumption primarily serves to simplify computation, as the number of possible causal configurations is exponential in the number of genes and annotation classes. However, a variety of approaches have been used to effectively reduce the search space over causal configurations in single-variant fine-mapping (e.g., Wen et al. 2016; Benner, Spencer, et al. 2016), which could also be adapted in the proposed framework.

Another natural extension of the GaMBIT framework is to incorporate gene-based annotations, e.g. biological pathways and molecular functions, to improve gene prioritization and detect gene-set enrichment. Gene-based annotations can be incorporated by introducing an additional logistic prior at the gene level, similar in form to the priors used for annotation-stratified gene-based associations. This extension will be explored in a forthcoming work.

5.5 References

Benner, Christian, Aki S Havulinna, et al. (2017). "Prospects of Fine-Mapping Trait-Associated Genomic Regions by Using Summary Statistics from Genome-wide Association Studies". In: *The American Journal of Human Genetics* 101.4, pp. 539–551.

Benner, Christian, Chris CA Spencer, et al. (2016). "FINEMAP: efficient variable selection using summary data from genome-wide association studies". In: *Bioinformatics* 32.10, pp. 1493–1501.

Bodea, Corneliu A et al. (2018). "Phenotype-specific information improves prediction of functional impact for noncoding variants". In: *bioRxiv*, p. 083642.

Das, S and G Abecasis (2015). *M3vcftools 1.0.1*.

Das, Sayantan (2017). "Next Generation of Genotype Imputation Methods". In:

Das, Sayantan et al. (2016). "Next-generation genotype imputation service and methods". In: *Nature genetics* 48.10, pp. 1284–1287.

Deelen, Patrick et al. (2014). "Improved imputation quality of low-frequency and rare variants in European samples using the 'Genome of The Netherlands'". In: *European Journal of Human Genetics* 22.11, pp. 1321–1326.

Dey, Rounak et al. (2017). "A fast and accurate algorithm to test for binary phenotypes and its application to PheWAS". In: *The American Journal of Human Genetics* 101.1, pp. 37–49.

Gamazon, Eric R et al. (2015). "PrediXcan: Trait Mapping Using Human Transcriptome Regulation". In: *bioRxiv*, p. 020164.

GTEx Consortium (2015). "The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans". In: *Science* 348.6235, pp. 648–660.

Gusev, Alexander et al. (2016). "Integrative approaches for large-scale transcriptome-wide association studies". In: *Nature genetics* 48.3, pp. 245–252.

Hao, Xingjie et al. (2018). "Identifying and exploiting trait-relevant tissues with multiple functional annotations in genome-wide association studies". In: *PLoS genetics* 14.1, e1007186.

Lencz, Todd et al. (2017). "High-depth whole genome sequencing of a large population-specific reference panel: Enhancing sensitivity, accuracy, and imputation". In: *bioRxiv*, p. 167924.

Marbach, Daniel et al. (2016). "Tissue-specific regulatory circuits reveal variable modular perturbations across complex diseases". In: *Nature methods* 13.4, pp. 366–370.

McCarthy, Shane et al. (2016). "A reference panel of 64,976 haplotypes for genotype imputation". In: *Nature genetics* 48.10, p. 1279.

NHLBI TOPMed WGS (Jan. 2018). *Whole Genome Sequencing in the NHLBI Trans-Omics for Precision Medicine*. URL: http://www.nhlbiwgs.org/.

Quick, Corbin et al. (2018). "emeraLD: Rapid Linkage Disequilibrium Estimation with Massive Data Sets". In: *Bioinformatics*.

Roshyara, Nab Raj and Markus Scholz (2015). "Impact of genetic similarity on imputation accuracy". In: *BMC genetics* 16.1, p. 90.

UK10K Consortium (2015). "The UK10K project identifies rare variants in health and disease". In: *Nature* 526.7571, pp. 82–90.

Van Leeuwen, Elisabeth M et al. (2015). "Population-specific genotype imputations using minimac or IMPUTE2". In: *Nature protocols* 10.9, p. 1285.

Wen, Xiaoquan et al. (2016). "Efficient integrative multi-SNP association analysis via deterministic approximation of posteriors". In: *The American Journal of Human Genetics* 98.6, pp. 1114–1129.

Zhou, Wei et al. (2017). "Improving power of association tests using multiple sets of imputed genotypes from distributed reference panels". In: *Genetic epidemiology* 41.8, pp. 744–755.