Title
Computational methods to discern the genetic basis of complex disease

Author
Roytman, Megan D

Publication Date
2018

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Computational methods to discern the genetic basis of complex disease

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Bioinformatics

by

Megan Roytman

2018

ABSTRACT OF THE DISSERTATION

Computational methods to
discern the genetic basis of complex disease

by

Megan Roytman
Doctor of Philosophy in Bioinformatics
University of California, Los Angeles, 2018
Professor Bogdan Pasaniuc, Chair

Genome-wide association studies (GWAS) have identified thousands of regions in the genome containing risk variants for complex traits. Due to the correlation structure between genetic variants, there is a need for computational methods that can tease apart causal from non-causal variants in these implicated regions. This dissertation presents three statistical methods that aim to improve our detection of causal variants at risk regions and ultimately better our understanding of the genetic basis of complex disease.

The first method aims to fine-map genetic regions impacting multiple correlated traits at once, employing the Multivariate Normal (MVN) distribution to jointly model association statistics at a risk region.

The second method performs hierarchical fine-mapping on risk regions that show evidence for a SNP impacting gene expression through an epigenetic feature, such as histone modifications. It uses both the MVN as well as the Matrix-variate Normal distribution to jointly model effects from SNP to epigenetic mark to gene expression.

The third method builds on existing summary statistics imputation methods by integrating functional annotation data to improve prediction of associations at untyped SNPs.

The dissertation of Megan Roytman is approved.

Janet S Sinsheimer

Kirk Edward Lohmueller

Eleazar Eskin

Bogdan Pasaniuc, Committee Chair

University of California, Los Angeles

2018

To my parents and grandparents.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

Firstly, I would like to express my gratitude to my thesis advisor, Bogdan Pasaniuc, for his patience, encouragement and genuine interest in the success of his students. I am very lucky to have had him as my Ph.D advisor.

In addition to my advisor, I would like to thank the rest of my thesis committee: Eleazar Eskin, Kirk Lohmueller, and Janet Sinsheimer, for their insightful comments, encouragement, and willingness to help.

I would also like to thank the members of the Bogdan lab, current and past: Kathryn Burch, Malika Freund, Claudia Giambartolomei, Gleb Kichaev, Megan Major, Arunabha Majumdar, Nicholas Mancuso, Tommer Schwarz, Huwenbo Shi, Robert Smith, Valerie Arboleda, Robert Brown, Page Goddard, and Ruth Johnson for always being ready and willing to discuss problems and for the warmth and positivity they brought to lab every day.

Finally, I want to express my gratitude to my mother, Elena Roytman, and to my father, Misha Roytman, for their presence and encouragement during my time as a Ph.D. student and throughout my education. I would not be where I am today without their support and sacrifice.

VITA

2008 – 2012

B.E. (Computer Science, Bioengineering), Massachusetts Institute of Technology.

2012 – 2013

M.E. (Computer Science), Massachusetts Institute of Technology

2014 – present

Graduate Student Researcher, Bogdan Pasaniuc Lab, University of California Los Angeles.

PUBLICATIONS

Roytman M.*, Mancuso N*, Shi H., Pasaniuc B., "Leveraging functional data to improve power of GWAS summary statistic imputation." in prep

Roytman M., Kichaev G., Gusev A., Pasaniuc B.., "Methods for fine-mapping with chromatin and expression data.." PLoS Genetics 2018

Roytman M.*, Kichaev G*, Johnson R., Eskin E., Lindstrom S., Kraft P., Pasaniuc B., "Improved methods for multi-trait fine mapping of pleiotropic risk loci." Bioinformatics 2017

Kiraly O., Gong G. **Roytman M.**, Yamada Y., Samson L.D., Engelward B.P., "DNA glycosylase activity and cell proliferation are key factors in modulating homologous recombination in vivo." Carcinogenesis 2014

CHAPTER 1

Introduction

Single Nucleotide Polymorphisms, or SNPs, are the most common form of genetic variation in humans. Approximately 1 in every 300 sites in the human genome is a SNP, which constitutes about 10 million SNPs in the human genome. As SNPs underlie much of the variation in development of human disease, knowledge of their effects can be used to identify drug targets, inform personalized medicine, or for prediction and prevention of disease susceptibility. SNPs are identified through genome-wide association studies (GWAS), which scan for associations between genetic variants and a trait. GWAS have identified thousands of genetic variants that are associated with complex traits and disease. However, due of the dense correlation structure between genetic variants, it is typically difficult to discern causal from non-causal variants in a risk region. The next three chapters present methods that aim to improve our detection of causal SNPs in risk regions and better our understanding of the genetic basis of complex disease.

A common post-GWAS analysis tool is the fine-mapping study, whereby detailed genetic information is gathered at a risk region and all SNPs are associated with the trait in question. Variants are then prioritized according to probability of causality and can ultimately become candidates for functional validation studies. However, due to the costliness of fine-mapping studies, fine mapping studies are often plagued by low sample sizes. There is thus a need for methods that can improve power to detect causal variants at risk regions despite low sample size. Recent studies have found that many GWAS loci are known to be implicated in multiple traits at once. For example, breast cancer and mammographic density [52], high density lipoprotein (HDL) and low density lipoprotein (LDL)[29], or rheumatoid arthritis

and irritable bowel disease [53, 65] are all pairs of traits that share overlapping GWAS signals. In Chapter 2, we propose an integrative framework that combines association signals accross multiple correlated traits to improve prioritization of causal variants. Our approach assumes that the variants impacting both traits at a given locus are shared, but with potentially distinct effect sizes. A key advantage of our approach is that it requires only summary association data, avoiding the need to share individual-level data. In simulations we show that our method produces well-calibrated posterior probabilities for SNP causality and improves upon existing approaches by combining the strength of associations across traits and explicitly modeling LD. We validate our results by fine-mapping pleiotropic regions in a lipids GWAS. I completed this manuscript as a co-first author jointly with Gleb Kichaev. My contribution to this work was investigating the effects of incorporating multiple traits into a single fine-mapping study. This paper also presented an Important Sampling approach that Gleb Kichaev formulated and implemented as a speed-up to his previous works.

Studies have shown that the majority of non-coding GWAS hits lie in regulatory sequences in the genome [59, 60]. Given that these SNPs do not themselves lie within, or sometimes even near, a gene body, this poses the challange of identifying the target genes of these variants and the mechanisms through which they act. Studies have shown that thousands of SNPs, termed histone quantitative trait loci (hQTLs), associate with histone modifications [37, 44, 62, 22] in addition to associating with gene expression [3, 11, 30, 50]. One explanation for these observations is that regulatory variants are affecting chromatin state, which may in turn cause changes in gene expression. However, this proposed chain of causality has yet to be established, as methods to investigate the relationships between the genome, the epigenome, and expression have largely focused on just quantifying the overlap between hQTLs and eQTLs [30, 4, 85]. Since laboratory experiments are very costly, there is a need for methods that can accurately prioritize the true causal SNP and mediating mark within an implicated locus. In Chapter 3, we propose a fine-mapping framework, *pathfinder*, that models the hierarchical relationships between genome, epigenome, and gene expression to predict both the causal SNP and the causal epigenetic mark within a gene region. Our approach takes as input two sets of summary statistics (Z-scores) corresponding to SNP-

mark associations and mark-expression associations and outputs posterior probabilities for each SNP, mark, and path to be causal in the region. In simulations we demonstrate that *pathfinder* outperforms alternative approaches with respect to both fine-mapping accuracy and calibration. We validate our method using genotype, chromatin and expression data from 65 African-ancestry and 47 European-ancestry individuals, demonstrating that the top causal SNPs proposed by pathfinder tend to lie in more functional regions and disturb more regulatory motifs than expected by chance.

Imputation is another important post-GWAS tool, as GWAS only measure a limited number of markers in the genome. Genotype imputation is the process by which unmeasured genotype information is predicted using large-scale reference panels of sequenced individuals[36, 51, 7]. However, accurate genotype imputation requires significant computational resources. Recent studies have proposed methods to perform summary statistics imputation, where unmeasured GWAS summary statistics are imputed directly from measured summary statistics and LD estimated from publicly available reference panels.[49, 66]. Additionally, recent efforts to characterize functionally active regions of the genome have revealed that SNPs coinciding with certain functional features are enriched for disease heritability [20, 61, 31, 26, 27, 33, 55]. In Chapter 4 we describe a novel imputation framework to predict GWAS summary statistics by integrating functional annotation data at typed and untyped SNPs. Our approach, FIMPG, builds on the fixed-effect linear model using LD-weighted statistics [49, 66] by including prior effect-size distributions defined by functional annotations. In simulations we show that FIMPG improves on summary statistics prediction over functionally-unaware models. We also validate FIMPG using summary statistics from 27 GWASs from the UKBiobank [78, 38]. We find that, while improvements in prediction accuracy are not sustained in real data, FIMPG's predictions are less deflated than those of standard summary imputation methods, which may boost association signal at untyped SNPs.

CHAPTER 2

Improved methods for multi-trait fine mapping of pleiotropic risk loci

2.1 Introduction

Genome-wide association studies (GWAS) have identified thousands of regions in the genome containing risk variants for complex traits and diseases [29, 81, 65, 86, 54]. However, the vast majority of the GWAS reported variants are not biologically causal, but rather, correlated to the true causal variants through linkage disequilibrium (LD) [82, 34, 42]. Fine mapping studies gather detailed genetic information within the loci that have been implicated in GWAS[63, 45, 87] and statistically dissect these regions to prioritize variants according to probability of causality. The top variants resulting from this procedure may become candidates for functional validation[14, 64].

Many GWAS loci are known to be implicated in multiple related traits – a phenomenon that is observed in many phenotypic classes. For example, breast cancer and mammographic density[52], high density lipoprotein (HDL) and low density lipoprotein (LDL)[29], or rheumatoid arthritis and irritable bowel disease [53, 65] are all pairs of traits that share overlapping GWAS signals. Combining association signals at these pleiotropic regions may

This chapter is published in Kichaev, Roytman et al., Bioinformatics 2015 [41]

strengthen the signal from the causal variants that are impacting both traits. A standard approach used when combining association information across multiple studies is fixed-effects meta-analysis, which assumes that causal variants across studies share the same effect sizes. The random-effects model does allow for effect size heterogeneity, but it is poorly-suited for situations in which the variant has opposite effect sizes in the various phenotypes [74]. For this reason, multivariate analyses that jointly analyze association data from multiple phenotypes and account for effect size heterogeneity are beneficial – particularly for related traits that have opposing phenotypic consequences such as HDL and LDL[29].

Considerable effort has been put forth into characterizing the chromatin landscape across the entire spectrum of human tissues[90, 20, 46]. Most recently, the Roadmap Epigenomics consortium interrogated 111 cell types, charting histone modifications, DNA accessibility, DNA methylation, and gene expression, to produce genome-wide maps of functional elements[46]. Previous works have demonstrated that principled integration of such data can aid fine-mapping performance in the context of single and multi-population fine-mapping studies[42, 39]. Since related traits have been shown to share an underlying genetic basis[8] that localizes within similar functional classes[25], it is plausible that functional annotation data can also augment cross-trait fine-mapping.

In this work we propose a unified framework to perform integrative fine-mapping across multiple traits. We integrate the strength of association across multiple traits with functional annotation data to improve performance in the prioritization of causal variants. Our approach makes the assumption that the same variants at the risk loci impact both traits though with potentially distinct effect sizes. A key advantage of our approach is that it requires only summary association data for each trait, thus avoiding the restrictions that arise from the sharing of individual-level data. Through simulations we show that our integrative method delivers well-calibrated probabilities for SNPs to be causal and improves fine-mapping performance relative to current state-of-the-art strategies. To our knowledge, the only existing method that performs joint mapping for pleiotropy while incorporating functional annotation data is GPA[13]. We show that our approach provides superior accu-

racy to GPA, likely due to the explicit modeling of LD in our framework. We illustrate the benefit of our proposed methodologies by fine-mapping pleiotropic regions of lipid traits in a GWAS of over 180K individuals[29].

2.2 Methods

2.2.1 Overview

Here, we introduce statistical methods for fine-mapping of pleiotropic loci with functional annotation data (see Figure 2.1). We build upon previous works[42, 39, 34] that make use of a Multivariate Normal (MVN) distribution to jointly model association statistics at all SNPs at the locus. This not only allows for the possibility of multiple causal variants at any risk locus, but also avoids the need to access individual level genotype data as LD can be approximated using the appropriate population-matched reference panel[1]. We integrate relevant functional annotation data through a prior probability for SNPs to be causal. The primary output of our approach are posterior probabilities for SNPs to be casual in both traits which can subsequently be used to prioritize SNPs individually [42] or used to compute fine-mapping credible sets [57].

2.2.2 A statistical framework for fine-mapping

The standard approach to connect genotype to phenotype is through a linear model. For individual i, let y_i be the trait value and $\mathbf{g}_i$ be their vector of genotypes spanning m SNPs. The trait can be modeled as $y_i = \mathbf{g}_i^T\boldsymbol{\beta} + \epsilon_i$, where $\epsilon_i \sim N(0, \sigma_e^2)$ is random environmental noise. The vector, $\boldsymbol{\beta}$, represents the allelic effects whose entries will be non-zero only at the causal SNPs. Given N individuals with measured genotypes and trait values, the effect size at SNP j is typically estimated using standard linear regression as $\hat{\beta}^j = (\mathbf{g}_j^T\mathbf{g}_j)^{-1}\mathbf{g}_j^T\mathbf{Y}$. The

strength of association is then quantified using the Wald statistic[10]:

$$Z^j = \frac{\hat{\beta}^j}{SE(\hat{\beta}^j)} \tag{2.1}$$

which asymptotically follows a normal distribution $Z^j \sim \mathcal{N}(\lambda^j, 1)$ with mean

$$\lambda^j = \frac{\beta^j \sqrt{\mathrm{Var}(g^j)}}{\sigma_e} \sqrt{N}. \tag{2.2}$$

Here, λ^j, is referred to as the Non-Centrality Parameter (NCP) and dictates of power of finding a significant association and, by extension, the power to distinguish causal from non-causal SNPs (i.e. $\beta^j \neq 0$ vs. $\beta^j = 0$). When the j'th SNP is causal, the effect sizes are non-zero and therefore the association statistic (Z-score) corresponding to that SNP will be drawn from a non-central Normal distribution. However, LD (i.e. correlations between SNPs at each locus) will induce non-zero NCPs at non-causals variants through tagging. Therefore, neighboring non-causal SNPs will appear to be significantly associated to a trait indirectly through LD. Previous works[34, 42, 39] have shown that the NCPs at any SNP can be approximated from the NCPs at the causal SNPs:

$$\Lambda^j = \sum_c r_{j,c} \lambda^c \tag{2.3}$$

where $r_{j,c}$ denotes the Pearson correlation between SNP j and causal SNP c. If we collect all the pairwise correlations into a matrix, $\mathbf{\Sigma}$, and let $\lambda_{\mathbf{C}}$ be the vector of standardized effects sizes at the causal SNPs given by the indicator vector $\mathbf{C}$, the entire set of regional summary statistics, $\mathbf{Z}$, can be approximated by a Multivariate Normal distribution (MVN))[34, 42]:

$$\mathbf{Z} \mid \lambda_{\mathbf{C}}, \mathbf{\Sigma} \sim \mathcal{N}(\mathbf{\Sigma}\lambda_{\mathbf{C}}, \mathbf{\Sigma}) \tag{2.4}$$

However, the causal effect sizes ($\lambda_{\mathbf{C}}$) are typically unknown apriori and must be either

approximated[42, 39] or integrated out[34]. Leveraging the standard infinitesimal model[88], Hormorzdiari et al. (2014) proposed to use a normal prior on the causal NCPs which, due to conjugacy, can be conveniently integrated analytically as follows:

$$\lambda_{\mathbf{C}} \mid \mathbf{C}, \sigma^2 \sim \mathcal{N}(0, \mathbf{\Sigma_C})) \tag{2.5}$$

$$\mathbf{\Sigma_C} = \sigma^2 \, \mathrm{d}iag \, \mathbf{C} + \mathrm{d}iag \, \epsilon \tag{2.6}$$

$$\mathbf{Z} \mid \mathbf{\Sigma}, \mathbf{C} \sim \left(\int \mathcal{N}\left(\mathbf{\Sigma}\lambda_{\mathbf{C}}, \mathbf{\Sigma}\right) \mathcal{N}\left(0, \mathbf{\Sigma_C}\right) d\lambda_{\mathbf{C}} \right) P(\mathbf{C}) \tag{2.7}$$

$$= \mathcal{N}(0, \mathbf{\Sigma} + \mathbf{\Sigma}\mathbf{\Sigma_C}\mathbf{\Sigma}) P(\mathbf{C}) \tag{2.8}$$

Here the prior probability of the causal set vector ($P(\mathbf{C})$) can be set to be uniform [57], hypergeometric [[34], or can be estimated empirically using more sophisticated approaches that incorporate functional genomic data [42, 39](see Section 2.4). Chen et al (2015) made the observation that the marginal likelihood in (eq. 2.8) is approximately proportional to a Bayes Factor comparing a causal and null model which depends on the Z-scores and LD only at the causal SNPs. This effectively reduces the computational burden from cubic in the number of SNPs to cubic in the number of causal variants considered at each likelihood evaluation. This not only improves efficiency, but also improves numerical stability since a much smaller matrix is inverted thus alleviating the need for stringent regularizations. In this work, we follow the Chen et al. implementation of the likelihood computations[12, 5].

2.2.3 Fine-mapping pleiotropic loci

Next, we extend the framework to exploit pleiotropy across related traits. Given multiple phenotypic measurements across T traits, one can compute Z-scores for each trait independently. If a locus harbors a significant association for multiple traits, a reasonable assumption would be that the underlying causal variants driving this association are shared. It follows that the vectors of association statistics are conditionally independent given the causal variants ($\mathbf{C}$), thus the joint distribution for all T sets of Z-scores decomposes into product:

$$P(\mathbf{Z}_1 \dots \mathbf{Z}_T \mid \mathbf{\Sigma}, \mathbf{C}) = \prod_{t=1}^{T} P(\mathbf{Z}_t \mid \mathbf{\Sigma}_t, \mathbf{C}, \sigma_t^2) \tag{2.9}$$

To simplify notation we hereafter refer to the collection of Z-scores at a fine-mapping locus as $\mathbf{Z}_* = \{\mathbf{Z}_1 \dots \mathbf{Z}_T\}$. We assume that all trait measurements have been performed in a single population and therefore assume that $\mathbf{\Sigma}_t, = \mathbf{\Sigma}$ for all t. Importantly, we note that our formulation makes no assumptions on the coupling between effect sizes at causal SNPs across traits which allows for arbitrary levels of heterogeneity. Accommodating this effect size heterogeneity could be important for related traits that have opposing phenotypic consequences.

Under the assumption that causal variants are shared across pleiotropic loci, the marginal likelihood of the data can be written as a summation across all possible causal sets, $\mathcal{C}$:

$$L(\mathbf{Z}_* \mid \Sigma, \sigma^2) = \sum_{\mathbf{C} \in \mathcal{C}} \prod_{t=1}^{T} P(\mathbf{Z}_t \mid \mathbf{\Sigma}, \mathbf{C}, \sigma_t^2) P(\mathbf{C}) \tag{2.10}$$

We can now use this to obtain the posterior probability of any causal set with a straightforward application of Bayes' rule:

$$P(\mathbf{C} \mid \mathbf{Z}_*, \mathbf{\Sigma}) = \frac{\prod_{t=1}^{T} P(\mathbf{Z}_t \mid \mathbf{\Sigma}, \mathbf{C}, \sigma_t^2) P(\mathbf{C})}{L(\mathbf{Z}_* \mid \Sigma, \sigma^2)} \tag{2.11}$$

which can be marginalized to yield per-SNP posterior probabilities:

$$P(C^j = 1 \mid \mathbf{Z}_*, \mathbf{\Sigma}, \gamma) = \sum_{\mathbf{C}: C^j = 1} P(\mathbf{C} \mid \mathbf{Z}_*, \mathbf{\Sigma}) \tag{2.12}$$

s

2.2.4 Incorporating functional genomic data

To integrate functional annotation data within this framework, we use a logistic function to connect a SNP's functional genomic context to its causal status as follows:

$$P(C^j = 1 \mid \gamma, \mathbf{A}) = \frac{\exp(\gamma' \mathbf{A}^j)}{1 + \exp(\gamma' \mathbf{A}^j)} \tag{2.13}$$

$$P(\mathbf{C} \mid \gamma, \mathbf{A}) = \prod_{j=1}^{m} P(C^j \mid \gamma, \mathbf{A})^{C^j} \left(1 - P(C^j \mid \gamma, \mathbf{A})\right)^{1-C^j} \tag{2.14}$$

The vector $\mathbf{A}^j$ is the set of annotations corresponding to the j'th SNP and γ_k is the prior-log odds that a SNP in annotation k is causal. We note that γ can be estimated directly from the data through an Empirical Bayes approach first described in Kichaev et al. (2014). However, this restricts functional enrichment estimation to only the fine-mapping loci under investigation. Alternatively, one could exploit potentially more powerful, genome-wide approaches such as stratified LD-score regression[25] that can infer global functional genomic enrichments using only summary data. Our framework is amenable to both approaches, and we allow the user to estimate γ from all the fine-mapping loci jointly using the EM algorithm proposed in[39] or supply it from external analyses.

2.2.5 Simulation Setup

To mimic real genotype data, we used HAPGEN2[77] and the 1000 Genomes[1] European samples, to simulate 20,000 haplotypes for a number of randomly selected 25KB loci from chromosome 1. We filtered rare SNPs (MAP ¡ 0.01) and normalized genotypes to be mean-centered with unit variance. We overlapped our simulated regions with DNase Hypersensitivity (DHS) sites spanning 217 cell types and tissues[32]. Using these annotations, we drew causal status for each SNP according to the logistic model described previously, setting the DHS enrichment to 5.1 to reflect what was reported in[32]. Each locus harbored one causal variant in expectation, though the random assignment of causal status could yield zero or multiple casual variants for a given locus. In experiments that were done over 50 loci

simultaneously, this typically resulted in an average of 18 loci with a single causal variant and 14 loci with multiple causals. Once we established the causal SNPs, we simulated phenotypes under a linear model such that for individual i, their phenotype value Y_i was given by $Y_i = \sum_{j=1}^{N_c} \beta^j \cdot g_i^j + \epsilon_i$, where N_c is the number of causal variants, β^j is the effect size of the j'th causal SNP, and g_i^j is number of copies of the risk allele j for individual i. We drew ϵ_i for each individual from a normal distribution $\mathcal{N}(0, \sigma_e^2)$, where σ_e^2 was given by the formula $h_g^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2}$, setting σ_g^2 to the empirically observed genetic component.

We computed Z-scores for all the SNPs within causal loci by regressing the phenotype vector $\mathbf{Y}$ on each genotype vector $\mathbf{G}^j$ and then taking the Wald statistic. To simulate correlated traits, the effect sizes (β_1^c, β_2^c) at the shared causal variants were drawn from an MVN distribution:

$$\begin{bmatrix} \beta_1^c \\ \beta_2^c \end{bmatrix} \sim \mathcal{N}(0, \begin{bmatrix} h_g^2/N_c & \rho h_g^2/N_c \\ \rho h_g^2/N_c & h_g^2/N_c \end{bmatrix}) \tag{2.15}$$

where ρ represents the desired genetic correlation. We chose a ρ of 0.4, consistent with typical correlations for lipids data reported in [8].

For computational efficiency, we also performed simulations in which the vectors of association statistics where drawn directly from an MVN distribution (eq. 2.4). In this scenario the NCP ($\lambda_{\mathbf{C}}$) was set to 5 at all causal SNPs.

2.2.6 Existing methods

We compared our approach to several existing fine-mapping methods. For single-trait fine-mapping, we compared to FINEMAP and CAVIARBF[5, 12], two methods based on the CAVIAR[34] model that do not incorporate functional annotation data. We ran CAVIARBF v1.4 using the default settings, setting prior variance explained to be 0.05 and the maximum number of causal variants in the model to 3. After CAVIARBF computed Bayes factors for each SNP, we ran their model search algorithm, which outputs posterior probabilities based

on Bayes factors. In this step, we set the prior probability of each SNP being causal to $1/m$, where m is the number of variants in the locus. We ran the FINEMAP v1.1 software using default settings, allowing for 3 causal SNPs per locus with prior probabilities of (0.6, 0.3, 0.1) for 1, 2, and 3 causals respectively.

For multi-trait fine-mapping, we compared to GPA [13]. To our knowledge, GPA is the only other method that performs multi-trait fine-mapping while leveraging functional annotation data. As GPA requires p-values as input, we converted Z-scores from our simulations to p-values for each SNP. We provided GPA with the same DHS annotation data as we did for our approach. On multi-trait analyses, GPA outputs 4 posterior probabilities for each variant, indicating the probability that the SNP is causal for neither trait, Trait 1, Trait 2, or both traits. When evaluating accuracy, we considered the SNP to be deemed causal by GPA if it was implicated in both traits. In addition, we explored traditional meta-analysis techniques to combine information across traits by computing inverse variance fixed effects association statistics[24]. We then used these Z-scores in fine-mapping under the assumption of a single causal variant [57] as well as within our framework as a single trait.

2.2.7 Empirical Lipids Data

We downloaded GWAS summary data across four blood lipids phenotypes: High Density Lipoprotein, Low Density Lipoprotein, and Triglycerides [29]. For each of the traits, we used Imp-G summary[66] to impute Z-scores up to the latest version (V3) of the 1000 Genomes European reference panel[1] yielding approximately 7.6 million SNPs per trait in total. We then compiled a list of 24 pleiotropic regions which we defined as a GWAS hit that was observed in least two traits of the three traits. For each of these regions, we centered a 250KB window around the lead SNP and overlapped these regions with two functional marks derived from the Roadmap Project: Liver H3K4me1 and Liver H3K27ac[46].

2.3 Results

2.3.1 Multi-trait fine-mapping

We first sought to investigate how leveraging information across related traits as well as functional annotation data affected fine-mapping performance. We simulated two genetically correlated traits with 10K individuals where the causal variants are shared between the traits but have heterogeneous effects sizes (see Methods). To control for the effect of sample size, we also simulated a single trait with 20K individuals. We find that by borrowing information across related traits, we are able to improve fine-mapping performance with greater efficiency than just simply increasing sample size for any single trait (see Figure 2.2). In our multi-trait analysis with fastPAINTOR, we required (1.4, 12.4) SNPs per locus for follow-up in order to capture (50%, 90%) of the true causal variants, as compared with (1.9, 23.1) SNPs in a single-trait analysis. Intuitively, this is due to the fact that power to detect causal variants grows with the square root of the sample size, while growing linear with the allelic effects (see eq 2.2). Therefore leveraging multiple genetically correlated traits (i.e. traits that share casual effects) will, on average across multiple loci, be more beneficial than simply increasing the sample size for one of the traits.

We next explored principled strategies for assembling data spanning multiple traits. Our main comparator was GPA– a method specifically proposed to use pleiotropy and functional data to prioritize variants. In addition, we ran two meta-analysis approaches using fixed effects association statistics– a standard meta-analysis that assumes a single causal variant [57], as well as running fastPAINTOR using these fixed effects association statistics as a single trait, which allows for multiple causal variants. In general, our approach is more accurate and robust than previously proposed methods, requiring (1.4,12.4) SNPs per locus for follow-up in order to identify (50%, 90%) of the causal variants compared to (2.3,25.1) for fastPAINTOR with FE or (11.6,32.3) for GPA (Figure 2.2). One of the critical model assumptions of GPA is that SNPs are independent. Clearly, in the context of fine-mapping, this assumption is strongly violated which explains the sub-optimal performance. Alternatively, FE can

be viewed as simply a weighted-average of the effect sizes. In the extreme, though not implausible, scenario where causal effects are going in opposite directions, FE will provide weak evidence that a SNP is causal.

Finally, we developed our framework with the assumption that causal variants are shared across traits. This may not always hold in practice and we wanted to understand how our method responds to violations of this assumption. We performed simulations in which causal variants for the two traits were drawn independently leading to potentially distinct causal SNPs and uncorrelated effect sizes. We find that our joint fine-mapping method is robust to pleiotropic loci with differing causals, yielding relatively small mis-calibration of the credible sets on the order of 10% (see Table 2.1). We predict that, in cases where the effect sizes among distinct causal variants are correlated, the disparity between the shared causal and distinct causal cases would be even less. We can thus conclude that our proposed framework that jointly models sets of association statistics, explicitly accounts for local correlation structure, and integrates functional data prioritizes variants robustly and accurately.

2.3.2 Multi-trait fine-mapping in lipids data

In order to demonstrate that the gains in our multi-trait fine-mapping approach are realized in real data, we analyzed summary association data from a large-scale GWAS of lipids [29]. High Density Lipoprotein (HDL), Low Density Lipoprotein (LDL), and Total Triglycerides (TG) are prototypical pleiotropic traits, sharing 24 GWAS hits for at least two. To showcase our pleiotropic fine-mapping framework, we obtained GWAS data over these traits spanning 180K individuals[29] and did integrative fine-mapping across putative pleiotropic regions. Functional annotation selection was guided by the genome-wide heritability-based functional enrichments reported in Finucane et al.[25]. The authors analyzed HDL, LDL, and TG and found that the H3K4me1 mark in liver tissue had the strongest enrichment of heritability across all three traits. Their result provides strong support for the key assumption that causal variants are shared across traits in our model. In addition to liver H3K4me1, we also used the liver H3K27ac mark, which displayed strong enrichment for multiple traits. In addition

to a joint analysis, we applied our framework with and without functional data as well as on each trait independently. To quantify fine-mapping resolution we use 99% credible sets [57, 42] which are defined as the set of variants that aggregate to capture 99% of the posterior probability mass. Consistent with simulations, pleiotropic fine-mapping provided a reduction in the size of the credible set as compared with investigating individual traits alone (see Table 2.2). Additional functional data helps refine the signal, though only marginally, since exceedingly strong associations at these regions dominate the prior evidence. In conclusion, these encouraging results illustrate that carefully merging related traits can improve the resolution of statistical fine-mapping.

2.4 Discussion

In this work, we introduced a fine-mapping method that integrates several sources of genetic data to efficiently and accurately prioritize causal variants. We generalized this approach to leverage multiple traits simultaneously and demonstrated, both in simulations and real data, that this strategy can improve the ability to detect causal variants impacting both traits. As GWAS data accumulate and evidence for the abundance of pleiotropic risk loci mounts, there is a need for fine-mapping methods that can perform large-scale integrative analyses. Moreover, efforts by large consortia such as ENCODE will continue to provide genomic annotation data that will improve the accuracy of fine-mapping studies. A key advantage to our method is that it requires only summary association data, overcoming the issues that arise when sharing individual data that would otherwise limit sample sizes. In light of these developments, our proposed methodology will become increasingly applicable in the future, particularly where multiple genetically correlated traits show at least suggestive evidence of association at a locus. Furthermore, our approach could even be applied to fine-map seemingly disparate traits such as height and educational attainment, which, nonetheless, share a genetic component[8].

We conclude by highlighting some caveats and limitations of our proposed framework. The

power of our multi-trait fine-mapping framework hinges on the assumption that causal variants are shared at pleiotropic risk regions. While this notion is supported by the fact that related traits have shared functional genetic architectures[25], it is unknown whether this holds in general when doing fine-mapping. Reassuringly, we demonstrated in simulations that the coverage of the resulting credible sets is reduced by a modest 10% when this assumption is violated . Second, most large-scale GWAS have overlapping samples and the conditional independence assumption given in (eq. 2.9) may be violated. However, it is unclear whether this violation will bias the results dramatically if the underlying causal variants are shared across traits.

2.5 Tables

Method	Proportion of causals identified	SNPs selected (s.e.)
Trait 1	0.96	46.01 (0.27)
Trait 2	0.96	45.54 (0.27)
Differing causals	0.86	28.42 (0.22)
Same causals	0.97	26.00 (0.17)

Table 2.1: The performance of fastPAINTOR is largely sustained when the assumption of shared causal variants across traits is violated. As compared with fine-mapping single traits independently, the reduction in the 95% credible set size is sustained while still capturing a large proportion of the causal variants. We define an 95% confidence set as the number of SNPs we need to select in order to accumulate 95% of the total posterior probability mass per locus.

	95% Credible Set		99% Credible Set	
Annotations	-	+	-	+
HDL	4.6	4.6	4.8	5.1
LDL	5.9	5.9	14.3	11.4
TG	4.2	4.2	5.4	5.4
Multi-trait	**3.7**	**3.7**	**4.7**	**4.7**

Table 2.2: Pleiotropic fine-mapping is superior to single locus fine-mapping. Presented here are the mean number of SNPs that are in the 95 and 99% fine-mapping credible sets.

2.6 Figures

Figure 2.1: Example of input and output of fastPAINTOR at locus chr4:35Mb for LDL and TG. As input, fastPAINTOR receives an LD matrix, functional annotations, and multiple sets of Z-scores at the given locus. fastPAINTOR performs inference and outputs posterior probabilities for each SNP that quantifies the likelihood that the SNP is causal for both traits.

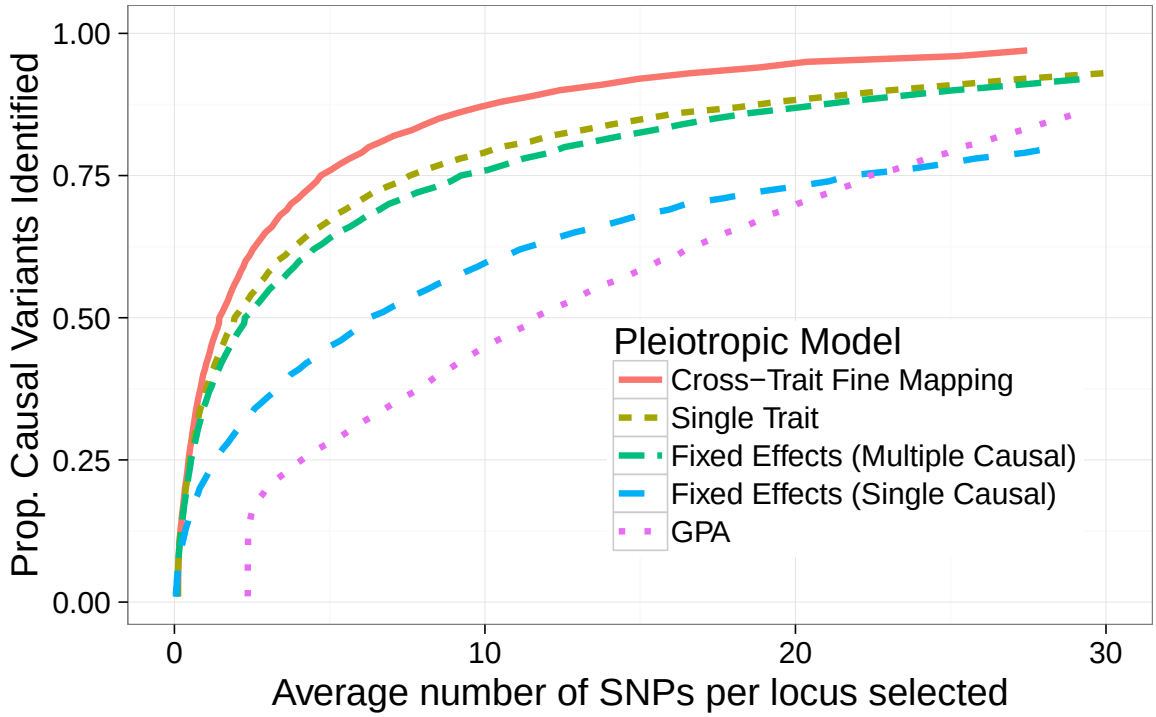


Figure 2.2: Integrative methods improve fine-mapping resolution in multiple traits. We simulated fifty 25KB loci for two traits with shared causal variants at each locus. We measure accuracy as the proportion of causal variants identified as we increase the size of our candidate SNP set.

CHAPTER 3

Methods for fine-mapping with chromatin and expression data

3.1 Introduction

Discerning the genetic and molecular basis of complex traits is a fundamental problem in biology. Genome-wide association studies have revealed that the majority of variants associated with disease lie in noncoding regulatory sequences [59, 60]. Identifying the target genes of these variants and the mechanisms through which they act remains an open problem [2]. Recent efforts to systematically characterize how genetic variation impacts more granular molecular phenotypes have yielded thousands of single nucleotide polymorphisms (SNPs) that associate with local and distal histone modifications – termed histone quantitative trait loci (hQTLs) [37, 44, 62, 22]. Furthermore, recent studies have identified many expression quantitative trait loci (eQTLs) that co-localize with hQTLs, implying there may exist a shared genetic influence on epigenetic traits and gene expression [3, 11, 30, 50]. Therefore, one proposed mechanism by which regulatory variants may affect gene expression and thereby impact traits is through changes in chromatin state [30]. However, this putative chain of causality whereby the effects of SNPs on expression are mediated by chromatin modifications has yet to be established. This is further compounded by the complex space of

This chapter is published in Roytman et al., PLOS Genetics 2018 [71]

plausible causal directions connecting transcription factor binding, DNA methylation, chromatin variation, and gene expression. Since laboratory experiments are very costly, there is a need for statistical methods that can accurately prioritize the causal SNP and chromatin mark within an implicated region under a plausible causal model. However, even if the causal direction is given, pinpointing the exact SNP and mark within a genomic region is very challenging due to the confounding effects of linkage disequilibrium (LD) among SNPs and correlations among marks [83, 42, 30, 44, 62, 85].

Methods to investigate the relationships between the genome, the epigenome, and expression have largely focused on quantifying the overlap between hQTLs and eQTLs [30, 4, 85]. Previous studies have sought to identify hQTLs by selecting the SNP with the strongest p-value for association to a local chromatin mark and to local gene expression [30, 4, 85]. Moreover, various methods exist for the fine-mapping of SNPs that may be concurrently affecting two traits, including eCAVIAR [35] and Coloc [28]. Although these methods can be applied to jointly analyze SNP, chromatin, and expression data, they do not model the causal path whereby SNPs impact expression through chromatin alteration.

Here we propose a fine-mapping framework, *pathfinder*, that explicitly models the hierarchical relationships between genome, chromatin, and gene expression to predict both the causal SNP and the causal mark within a gene region that are influencing expression of a given gene. Our framework assumes a causal model where a SNP impacts a chromatin which in turn alters gene expression. In our framework we refer to a "causal" SNP as any SNP that disrupts inter-individual variation of chromatin state either through a direct biological mechanism (e.g., chromatin accessibility) or indirectly through an unobserved biological mechanism. Similarly, we refer to a "causal" chromatin mark as either a mark that biologically alters expression or that tags an underlying epigenetic regulatory mechanism of expression. Our framework takes as input the strength of association (as quantified through the standard Z-scores) between all SNP/mark pairs and all marks to expression as measured in a given set of individuals. To explicitly account for the correlation structure among SNPs and marks, we use a Matrix-variate Normal distribution to model all Z-scores jointly. By construction,

this allows our probabilistic model to assign posterior probabilities for each SNP, mark, and path (where paths include all possible SNP-mark combinations) to be causal in the region. A key advantage of our approach is that it produces well-calibrated posterior probabilities for causality. Thus, *pathfinder* can be used to prioritize variants and marks for validation experiments.

In simulations we compare against several existing methods, demonstrating that *pathfinder* outperforms alternative approaches with respect to both accuracy and calibration. This is largely because our comparators do not take into account mark-expression associations. In some cases, these additional associations may help distinguish between two potentially causal paths that have comparable evidence for causality. For example, in cases where a SNP is associated with expression of a local gene and is also associated with two local chromatin marks, knowledge of the impact of each mark on gene expression may help distinguish between two possible paths for causality. Finally, we analyze genotype, chromatin and expression data from 65 African-ancestry and 47 European-ancestry individuals. We show that the top causal SNPs proposed by *pathfinder* tend to lie in more functional regions and disturb more regulatory motifs than expected by chance. We also present evidence that most of the top paths reported by *pathfinder* demonstrate consistency with our proposed sequential model, thus strengthening the case for our method's applicability to empirical biological data.

3.2 Results

Overview of hierarchical fine-mapping with genetic, chromatin, and gene expression data

Here we introduce a hierarchical statistical method for fine-mapping of causal SNPs and chromatin marks (e.g., histone modifications) that may be concordantly influencing gene expression within a genomic region. We build upon previous insights that a vector of Z-scores

is well-described by a Multivariate Normal (MVN) distribution parameterized by LD[42, 40, 34] to model association statistics between chromatin marks and gene expression. We analyze all chromatin peaks across four mark types (DHS, H3K4me1, H3K4me3, and H3K27ac) jointly in the same framework; we refer to a "mark" as a chromatin peak at a particular location, and "mark types" as DHS, H3K4me1, H3K4me3, and H3K27ac. To simultaneously take into account both SNP LD and the correlations between chromatin marks, we use the Matrix-variate Normal distribution to jointly model association statistics between all SNPs and marks within a region. Our method takes as input SNP-mark and mark-expression associations within a region centered around a particular gene, as well as correlations among all SNPs (LD) and correlations among all considered marks. *Pathfinder* enumerates over all possible causal paths, considering one causal SNP and one causal mark for each path, and outputs a posterior probability for each path to be causal, which can subsequently be used to prioritize SNPs and marks for validation. We compute marginal probabilities for individual SNPs (or marks) to be causal by summing the posterior probabilities over all paths that contain the SNP (or mark). For simplicity, in this work we refer to a "causal" mark as a mark that either causally drives inter-individual variation of gene expression or is correlated to an underlying causal mechanism (e.g. transcription factor binding), though it may not be biologically causal for expression.

The advantage of our method over existing approaches is that it integrates mark-expression associations which may help distinguish between two paths with otherwise comparable evidence for causality. We illustrate a scenario in Figure 3.1. Consider a genetic region where SNP g_1 has a strong association with two local marks h_1 and h_2, as well as a significant association with gene expression. Using only SNP-mark and SNP-expression effects, we are unable to discern whether SNP g_1 influences expression through mark h_1 or h_2. However, if we consider mark-expression effects, we see that mark h_1 has a strong association with gene expression where mark h_2 does not. This additional information helps support the hypothesis that there is a causal path from SNP g_1 to mark h_1 to gene expression.

3.2.1 *Pathfinder* improves fine-mapping performance

We used simulations to compare *pathfinder*'s performance against alternative methods with respect to SNP-, mark-, and path-finding efficiency as well as the calibration of its posterior probabilities. We generated genetic, chromatin, and gene expression data for 10,000 50kb regions, each centered around a single gene, over 100 individuals, using SNP LD and mark correlations derived from 65 Yoruban (YRI) individuals (see Methods). We define a "mark" as an individual peak location for any mark type in the dataset (DHS, H3M4me1, H3K4me3, or H3K27ac). For each gene, we randomly assigned a single causal pathway from one SNP to one mark to gene expression. We then ran our methods on all regions individually and assessed their ability to correctly prioritize the true causal path in each region (Methods).

We compare against an independent fine-mapping approach (whereby we fine-map SNP-mark associations and mark-expression associations independently and take the product of the resulting probabilities to produce posterior probabilities for paths), a Bayesian network analysis[73], a naive ranking (where we rank SNP-expression and mark-expression associations to prioritize SNPs and marks within a region; for path-finding, we rank the product of these two), a formal colocalization method[28], and finally, against overlaps between eQTLs and hQTLs within a region centered around a gene of interest (see Methods). Unlike the first four approaches, the overlap methods do not produce rankings, but yield candidate sets of causal SNPs, marks, and paths. For this reason, we present these results in a separate analysis using an alternative metric for comparison.

We find that *pathfinder* has consistently better performance than the other ranking approaches with respect to all three features – SNP-, mark-, and path-mapping within a region (Figure 3.2). For example, association ranking, Coloc, Bayesian network analysis, and independent fine-mapping accumulate 55%, 62%, 47%, and 13% of the top paths on average in order to recapture 90% of the causal paths, whereas our method only requires 8% of the top paths. Note that SNP-expression association ranking is equivalent to running a basic eQTL analysis, which does not take into account chromatin data, in order to identify causal SNPs.

Next, we evaluated *pathfinder*'s performance compared against standard analyses that investigate overlaps between hQTLs and eQTLs within a genomic region. In such experiments, the variant with the strongest association to each local chromatin mark is selected, as well as the variant with the strongest association to local gene expression. In addition, marks are filtered to ensure a 10% FDR (see Methods). This produces a set of candidate marks, as well as one candidate SNP per mark, and one SNP deemed causal for gene expression in the region. Implicitly, the overlap of these variants suggests a set of candidate SNPs, marks, and paths for the region. For the same set sizes, *pathfinder* identifies 96% of the causal marks versus 74% in the standard overlap approach (Figure 3.3). SNP-finding accuracy is comparable between the two methods.

We next assessed the calibration of the posterior probabilities for causality output by *pathfinder*. Our method has slightly deflated credible sets for SNP- and path-finding, but well-calibrated credible sets for mark-finding (Figure 3.4). In contrast, the independent fine-mapping approach has consistently inflated credible sets – that is, it captures more causal paths than expected, but also has drastically larger credible set sizes. For example, when accumulating 90% of the posterior probabilities over all regions, *pathfinder* captures 88% of the true causal paths within the top 380 candidate paths, whereas independent fine-mapping captures 94% of the causal paths within the top 1026 candidate paths. Overall, *pathfinder*'s credible sets are less biased and narrower than those obtained through the independent fine-mapping approach.

Finally, we investigated the effects of simulation and method parameters on *pathfinder*'s accuracy. Firstly, we varied the causal SNP and mark effect sizes such that the variance explained of mark and gene expression ranged from 0.1 to 0.5. As anticipated, increased heritability leads to better performance (See Figure 4.3A-C). Secondly, in order to assess the impact of SNP LD and mark correlations on SNP- and mark-finding performance, we stratified our existing simulations based on the mean correlation of the causal SNP or mark to all other SNPs or marks (See Figures 4.3D-I). We grouped our simulations into three categories: low, medium, and high correlations. As anticipated, SNP-finding performance

decreases slightly as SNP LD increases. Notably, mark-finding performance is actually improved at higher SNP LD. This is due to the redundancy in information about SNP-mark associations at the causal mark when these effects are exhibited across multiple correlated SNPs. SNP- and mark-finding performance, however, do not seem to be significantly affected by mark correlations in our simulations - at least not at the level of variation exhibited in our data. Next, we evaluated the effect of the prior variance tuning parameter on fine-mapping performance (See Figure 4.3J-L). The prior variance is an estimate of the variance explained by the causal SNP and mark in the region, as we do not know a priori what the causal effect sizes are. We show that the optimal range for the prior variance parameters is between 5 and 10, in simulations with a variance explained of 0.25 on both levels. Overall, performance does not seem to change drastically in response to variations in the prior variance, even significantly outside of this optimal range.

3.2.2 Violations of the model

Our hierarchical model makes several key assumptions that may sometimes be violated in empirical data. Firstly, *pathfinder* assumes that a single causal SNP and a single causal mark are driving the associations within a region, where in reality there may exist multiple true causal SNPs or marks [42, 34]. Secondly, *pathfinder* assumes that SNP effects on gene expression are mediated by a chromatin mark, which may not be the case in real data. We therefore assessed the performance of our method when these two assumptions are violated in various ways, diagrammed in Figure 3.6.

First, we investigate violations 1-3, which include multiple causal pathways throughout the region. Path-mapping accuracy, measured by the proportion of causal paths identified, is reduced in all three scenarios (Figure 3.6). Note that the number of causals identified does not necessarily decrease, but rather the proportion, as there are more causal paths in each region. SNP- and mark-finding accuracy under these violations are also compromised, but with two notable exceptions. In the multi-causal-SNP scenario, mark-finding accuracy increased in comparison with the single-SNP simulations; for example, only 8% of marks were selected

(versus 18% in the single causal simulations) to capture 90% of the causal marks. In the multi-causal-mark scenario, SNP-finding accuracy increased. Intuitively, this is due to the redundancy in the signal that is captured by the Matrix-variate Normal distribution.

We next investigate violations 4-5, in which an additional SNP or mark influences gene expression directly. We observe in these two scenarios that performance is reduced for SNP-, mark-, and path-finding, but not drastically. For example, in order to capture 90% of the causal paths, *pathfinder* must select on average 25% and 28% of paths under violations 4 and 5, respectively (compared with 15% under standard simulations).

Finally, we discuss *pathfinder*'s performance under violations where the causal order is modified (violations 6-7). Under violation 6, where a single causal SNP affects gene expression directly, which in turn affects a single mark, *pathfinder* actually captures a higher proportion of the affected marks and overall paths. For example, in order to capture 90% of the causal paths, *pathfinder* must select on average only 3% of the top-ranked paths (compared with 15% under standard simulations). In violation 7, where the SNP has independent effects on the mark and the gene expression, we show that *pathfinder*'s accuracy in finding the causal mark and path is significantly reduced. Note that in this case, the "path" is not truly a path but a SNP/mark pair, as effects of the SNP on mark and gene expression are independent. Our power in distinguishing between these two models depends on the prior variance explained parameter. Under violation 7, the variance explained in gene expression by the causal mark is much smaller than expected, thus reducing our confidence in the true causal configuration. We conclude that under the SNP→expression→mark violation, *pathfinder* will identify causal paths very confidently even if they do not follow the assumed SNP→mark→expression model. Therefore a high posterior probability for a path may not be sufficient evidence for causality. On the other hand, when SNP effects on mark and expression are independent, *pathfinder* is less likely to produce false positives. For these reasons, we recommend a pre- or post-filtering step to retain only those regions that show some prior evidence for the SNP→mark→expression model using a conditional analysis or partial correlation approach (Methods).

3.2.3 Empirical Data Analyses

We evaluated the behavior of our hierarchical fine-mapping method when applied to empirical data. We performed these analyses on data from 65 YRI individuals whose genotypes were obtained through 1000 Genomes, and whose PEER-corrected H3K4me1, H3K4me3, H3K27ac, DHS, and RNA expression levels in lymphoblastoid cell lines (LCLs) were obtained from [30]. In each region, we analyzed all four mark types jointly (H3K4me1, H3K4me3, H3K27ac, and DHS) by including all peaks spanning the region for each mark type. Each peak of each mark type was therefore treated as a single chromatin mark. We filtered the 14,669 regions using a two-step regression analysis to yield 1,317 regions that showed evidence for the sequential model of SNPs affecting histone marks which in turn affect gene expression (see Methods).

In Table 3.1, we report the average 50%, 90%, and 99% credible set sizes produced when running *pathfinder* on real data. We compare against basic eQTL mapping, where we fine-map SNPs to gene expression ignoring chromatin data. We show that the credible set sizes are significantly narrower when running *pathfinder* with all three levels of data, consistent with our findings in simulations. For example, eQTL mapping requires an average of 45.3 SNPs in order to capture 90% of the posterior probability for SNP causality, whereas *pathfinder* only requires 28.4 SNPs. If we define a gene to be fine-mapped if 99% of the posterior probability mass for SNP causality is contained within the top 10 SNPs or fewer, then standard eQTL mapping fine-maps 46 of the genes in our data, whereas *pathfinder* fine-maps 73 of the genes. Notably, *pathfinder* also requires only 1.8 marks on average in order to capture 90% of the posterior probability for causal marks. In 82% of the regions where the top two marks capture more than 90% of the posterior probability, these two marks are two distinct peaks of the same mark type.

The mean variance explained observed in the top path chosen by *pathfinder*, across all conforming regions, were 0.38 (s.e. 0.01) for the SNP-mark effect and 0.20 (s.e. 0.01) for the mark-expression effect. These effects are reasonably consistent with the 25% variance

explained we used in simulations at each level (see Simulations). The correlation between the SNP-mark and mark-expression effect size magnitudes in the top selected paths across all regions was 0.03 (p = 0.400). That is, the strength of the SNP-mark effect did not seem to correlate with the strength of the mark-expression effect. We assessed the relative impacts of each type of histone mark by computing the proportion of probability mass assigned to each mark type in aggregate over all regions. H3K4me3 is the most informative mark type in this data, capturing 31% of the total probability mass despite being the least prevalent of all four mark types, constituting only 13% of all marks.

As our pre-filtering step was designed to preserve only regions in which SNP effects on gene expression are mediated by chromatin, we expected a large majority of the analyzed regions to show evidence for this mechanism. To confirm this, we investigated whether the top paths prioritized by our method demonstrate consistency with this causal model. We defined a set of top paths as those which were ranked first in a region and whose posterior probabilities for causality were assigned by *pathfinder* to be greater than 0.1. This resulted in 480 total top paths. Out of 480 top paths, only 12 had a significant ($p < 0.05/480$) partial correlation between SNP and gene expression after controlling for chromatin. However, 193 paths had a significant partial correlation between SNP and chromatin after controlling for gene expression. This finding suggests that the top paths are more consistent with the SNP→mark→expression model than with a SNP→expression→mark model.

Next we examined the relationship between the product of the effect sizes between SNP-mark and mark-expression against the overall SNP-expression association (Figure 3.7). We expect this relationship to be correlative; if truly mediated by the mark in question, the overall SNP-expression effect size should be proportional to the product of the two contributing effect sizes. Note that we weight our correlation by the reported posterior probability for each path, such that the paths we have more confidence in will contribute more to this metric. We find a high correlation (r = 0.91) between these effect size vectors for our top paths, as compared with a correlation of r = 0.36 when running the same analysis on random paths within each region. This result indicates that *pathfinder* is identifying many pathways

that are likely to be following its causal model.

In Table 3.2, we list the top ten paths prioritized by *pathfinder* across all real data regions. Most SNPs implicated in these paths are known to alter several regulatory motifs and often lie in an enhancer region or a promoter region of the genes whose expression they affect. 59% (s.e. 2%) of the SNPs implicated in the top paths fall into active ChromHMM states (1-7) in LCLs, including active TSS, flanking active TSS, transcription at gene 5' and 3', strong transcription, weak transcription, genic enhancers, and enhancers. Only 47% (s.e. 2%) of random paths fall into these active states ($p = 0.001834$). Moreover, on average, SNPs in the top paths disturbed 5.35 (s.e. 0.26) regulatory motifs, whereas random SNPs chosen at the same regions only disturbed 4.40 (s.e. 0.20) motifs on average ($p < 0.001$). We did not, however, observe a similar change in transcription factor binding affinity at these motifs ($\delta = 5.26$ vs $\delta = 5.27$, ($p = 0.511$)). As an example, in Figure 3.8A-D, we display the genomic context for the top region reported by *pathfinder*, including average mark signals for DHS, H3K4me1, H3K4me3, and H3K27ac, stratified by genotype, in a 4kb region centered around the TSS of the NDUFA12 gene. The implicated SNP lies within the NDUFA12 TSS. Figure 3.8E plots the gene expression signal against that of the top mark, stratified by genotype.

Next we examined the spatial relationships between the SNP, mark, and TSS implicated in the top paths reported by *pathfinder* (Figure 3.9). SNP to mark and mark to TSS distances were significantly lower in our selected paths compared with randomly chosen paths at the same regions. The average distance from SNP to mark in *pathfinder*'s top paths was approximately 11.7kb, compared to 15.3kb in randomly chosen paths ($p < 0.001$). The average distance from mark to TSS in selected paths was approximately 8.6kb, compared to 9.7kb in randomly chosen paths ($p = 0.026$). SNP to TSS distances were not significantly different in top versus random paths ($p = 0.108$), with top SNPs lying on average 11.7kb away from the TSS and random SNPs lying 12.4kb away. 5% of top SNPs lied within 2kb of the TSS while 15% lied within 2kb of the corresponding peak. 23% of peaks in the top paths lied within 2kb of the gene TSS.

To further validate the top paths chosen by *pathfinder*, we determined the extent to which

SNPs in these paths overlap with eQTLs that have been identified in LCLs using the larger scale Geuvadis data set [48]. 21% of the top paths contained SNPs that were also identified as eQTLs from the Geuvadis data set. In comparison, when randomly choosing paths at the same regions, only 11% overlapped with eQTLs ($p < 0.001$). Simply choosing the SNP with the highest association with gene expression in each region (equivalent to standard eQTL-mapping) resulted in an overlap of 24% with existing eQTLs. These results contradict the improvement in accuracy demonstrated in simulations when using *pathfinder*. We suspect this discrepancy is due either to imperfect locus ascertainment (i.e., a number of loci may include SNPs that directly affect gene expression rather than indirectly through chromatin) or the fact that the Geuvadis eQTLs were also selected using standard fine-mapping approaches and we may thus expect a stronger agreement between the two resulting eQTL sets.

We also investigated the extent to which *pathfinder*'s top SNPs overlap with eQTLs that have been experimentally validated through differential expression in an LCL dataset [80]. Here, we define the set of validated eQTLs to be those whose p-values for differential expression passed a threshold of 0.01. We find that 2.2% (or 13) of *pathfinder*'s top SNPs overlap with this validated set, where choosing the SNP with the highest association with gene expression in each region resulted in an overlap of 2.3% (also 13 SNPs).

Finally, we investigated whether any of the top paths reported by *pathfinder* could be found within GWAS hit regions for various autoimmune diseases, as our data were collected from LCLs. These autoimmune diseases included Celiac disease, Crohn's disease, PBC (Primary Biliary Cirrhosis), SLE (Systemic Lupus Erythematosus), MS (Multiple Sclerosis), RA (Rheumatoid Arthritis), IBD (Irritable Bowel Disease), and UC (Ulcerative Colitis). We restricted to GWAS hits with variants associated to the trait with $p < 5 \times 10^{-8}$. We found that 19 of our 480 top paths were contained in a GWAS-implicated region. In Table 3.3, we report the paths that localized within autoimmune GWAS regions. In order to determine whether our top paths are truly enriched in GWAS regions, we established how many of these paths appear in an equivalent number of random regions that have not been implicated by an autoimmune GWAS. We centered each random region around a SNP that

was matched for a similar MAF and LD score as the GWAS tag SNP. We ran this analysis 100 times to define a null distribution for the number of top paths found in a background region. We found that 19 out of 480 top paths was not a significant enrichment ($p = 0.44$).

3.3 Discussion

In this work we proposed a hierarchical fine-mapping framework that integrates three levels of data - genetic, chromatin, and gene expression - to pinpoint SNPs and chromatin marks that may be concordantly influencing gene expression. A key contribution of our approach is the ability to model the correlation structure in the association statistics using a Matrix-variate Normal distribution. Our approach is superior to existing methods, demonstrating the advantage of using a probabilistic approach that takes into account the full sequential model. Moreover, *pathfinder* produces well-calibrated posterior probabilities, and is thus a reliable method for the prioritization of SNPs and marks for functional validation.

We conclude by addressing some of the limitations of our method. Most notably, our method is based upon the SNP→mark→expression assumption. In many genomic regions that show simultaneous evidence for SNP to mark and SNP to gene expression effects, this model will not necessary hold true. In simulations, we show that under the SNP→expression→mark violation, *pathfinder* may identify causal paths very confidently, leading to false positives under the proposed model. When a SNP is in fact independently influencing a mark and gene expression, *pathfinder* is less likely to produce false positives. However, the risk of mis-appropriating our method in this way can be reduced by requiring genomic regions to show evidence for our causal model. We recommend a pre-filtering step before running *pathfinder* on real data that we outline in Methods. In our empirical data analyses, we demonstrate that this two-step regression robustly filters out non-conforming regions. We also acknowledge that, though there are multiple lines of evidence for SNPs influencing expression through local hQTLs, recent works have also emphasized the importance of interactions with distal hQTLs. Thus, developing a systematic way to incorporate data in distal regions with evidence for

interactions with a local eQTL would be a fruitful direction. Moreover, *pathfinder* assumes that the true causal SNP and mark within a region are present in the data, which may not always be the case. In this scenario, *pathfinder* will instead place its confidence in the SNP or mark that best correlates with the missing causal SNP or mark in question. Similarly, many epigenetic marks are not themselves causal for gene expression, but are simply correlated to a causal event (e.g., transcription factor binding). It is also often the case that multiple marks at promoter and enhancer regions are concordantly acting to impact gene expression. In these cases, individual marks are not necessarily causal in themselves, but may be viewed as a cause for inter-individual variation or simply correlated to a causal factor. In this light, *pathfinder* aims to identify the epigenetically modifying region so that it can be tested experimentally and/or characterized functionally (for example, to identify the effector transcription factor). We also note that *pathfinder* currently uses an approximation whereby the observed Z-score at the causal SNP is used to estimate the true NCP at the causal SNP (Methods). We leave this to be addressed in future work; this correction will likely further improve the calibration of our method's credible sets. We note that *pathfinder* only uses individuals for which we simultaneously have genetic, chromatin, and gene expression measurements, thus ignoring eQTL data that has been measured in larger sample sizes. However, eQTL data from larger samples could potentially be used as a prior for expectation of SNP causality or perhaps for validation after running *pathfinder* on real data. Finally, although our analyses showed that H3K4me3 marks are the most informative for fine-mapping, small data set sizes analyzed in this work prohibit us in making definitive conclusions on which mark is most useful leaving such avenues for future work.

3.4 Materials and Methods

3.4.1 Model and Likelihood

For each individual, let h be the signal value for the causal histone mark and G be their vector of genotypes at a region containing s SNPs. Let E be the individual's mRNA expression

level for the gene at this region and H be a vector representing all t marks at the region, which contains h. Here we analyze all individual peak locations across all available mark types in a joint framework. As such, each of t individual marks represents one peak location for a particular mark type. Our causal framework can be modeled as:

$$h = \mathbf{G}\beta_{\mathbf{g}} + \epsilon_g \tag{3.1}$$

$$E = \mathbf{H}\beta_{\mathbf{h}} + \epsilon_h \tag{3.2}$$

where $\epsilon_g \sim \mathcal{N}(0, 1-\sigma_g^2)$ and $\epsilon_h \sim \mathcal{N}(0, 1-\sigma_h^2)$. The vector β_g represents the allelic effects on the causal histone mark whose entries will be non-zero only at the causal SNP. The vector β_h represents the histone mark effects on expression levels whose entries will be non-zero only at the causal histone mark. σ_g^2 and σ_h^2 represent the variance explained at the SNP-mark and mark-expression levels.

Modeling mark to expression associations

We estimate mark to expression effects with linear regression to quantify the strength of association of the kth mark through the Wald statistic:

$$Z_h^k = \frac{\hat{\beta}_h^k}{SE(\hat{\beta}_h^k)} \tag{3.3}$$

$$Z_h^k \sim \mathcal{N}(\lambda_h^k, 1) \tag{3.4}$$

$$\lambda_h^k = \frac{\beta_h^k \sqrt{Var(h^k)}}{\sigma_h}\sqrt{N} \tag{3.5}$$

Here, $\hat{\beta}_h^k$ is the estimated effect size of the causal peak on expression. λ_h^k represents the strength of our signal for causal marks [34]. However, correlations between histone marks will induce a non-zero non-centrality parameters (NCPs) at non-causal histone marks. If we collect all pairwise mark correlations into Σ_h, and let $\Lambda_{h,d}$ be the vector of NCPs for all

histone marks on expression given causal mark d, all summary statistics can be approximated by an MVN.

$$\mathbf{Z_h}|\mathbf{C_h} \sim \mathcal{N}(\mathbf{\Sigma_h \Lambda_{h,d}}, \mathbf{\Sigma_h}) \tag{3.6}$$

where $\mathbf{C_h}$ is an indicator vector containing zeros at all non-causal marks and 1 at the causal mark d, and $\mathbf{\Sigma_h \Lambda_{h,d}}$ represents the vector of induced effect sizes at non-causal marks due to inter-mark correlations.

As we do not know the causal effect size $\mathbf{\Lambda_{h,d}}$, we use a normal prior on the causal mark NCPs which can be integrated out as follows:

$$\mathbf{\Lambda_{h,d}} \mid \mathbf{C_h}, \sigma_h^2 \sim \mathcal{N}(0, \mathbf{\Sigma_{C,h}})) \tag{3.7}$$

$$\mathbf{\Sigma_{C,h}} = \sigma_h^2 \, \mathrm{d}iag \, \mathbf{C_h} + \mathrm{d}iag \, \epsilon \tag{3.8}$$

$$\mathbf{Z_h} \mid \mathbf{\Sigma_h}, \mathbf{C_h} \sim \left(\int \mathcal{N}\left(\mathbf{\Sigma_h \Lambda_{h,d}}, \mathbf{\Sigma_h}\right) \mathcal{N}\left(0, \mathbf{\Sigma_{C,h}}\right) d\mathbf{\Lambda_{h,d}} \right) P(\mathbf{C_h}) \tag{3.9}$$

$$= \mathcal{N}(0, \mathbf{\Sigma_h} + \mathbf{\Sigma_h \Sigma_{C,h} \Sigma_h}) P(\mathbf{C_h}) \tag{3.10}$$

Here the prior probabilities of the causal set vector $P(\mathbf{C_h}))$ is set to be uniform. As a parameter of the model, we set a prior variance explained σ_h^2 for the mark effects. We found the method to be fairly robust to variations in this parameter (Figure 4.3J-L), and chose a prior variance of 5 for our analyses. In practice, we add an ϵ of 0.0001 along the diagonal of $\Sigma_{C,h}$ to ensure positive semidefiniteness. Thus, the mark-expression association statistics can be expressed as:

$$\mathbf{Z_h}|\mathbf{C_h} \sim \mathcal{N}(0, \mathbf{\Sigma_h} + \mathbf{\Sigma_h \Sigma_{C_h} \Sigma_h}) \tag{3.11}$$

Modeling SNP to mark associations

As before, we estimate SNP to mark effects with linear regression to quantify the strength of association of the jth SNP on the kth mark through the Wald statistic:

$$Z_g^{j,k} = \frac{\hat{\beta_g^{j,k}}}{SE(\hat{\beta_g^{j,k}})} \tag{3.12}$$

$$Z_g^{j,k} \sim \mathcal{N}(\lambda_g^{j,k}, 1) \tag{3.13}$$

$$\lambda_g^{j,k} = \frac{\beta_g^{j,k}\sqrt{Var(g^j)}}{\sigma_g}\sqrt{N} \tag{3.14}$$

Here, $\hat{\beta_g^{j,k}}$ is the estimated effect size of the causal SNP on the causal peak. $\lambda_g^{j,k}$, the NCP, represents the strength of our signal for causal SNP-mark effects. However, LD between SNPs and correlations between marks will induce non-zero NCPs at non-causal SNP-mark pairs. We collect all pairwise SNP correlations into $\mathbf{\Sigma}_g$ and all pairwise mark correlations into $\mathbf{\Sigma}_h$, and use the Matrix-variate Normal distribution to jointly approximate the association statistics for all SNPs on all marks as:

$$\mathbf{Z_g}|\mathbf{C_g}, \mathbf{C_h} \sim \mathcal{MN}(\mathbf{M}, \mathbf{\Sigma_g}, \mathbf{\Sigma_h}) \tag{3.15}$$

Here, $\mathbf{M}$ is an $s \times t$ matrix representing association means between all s SNPs and all t marks, where each entry $\mathbf{M}_{j,k} = \mathbf{\Sigma}_g^{j,c}\mathbf{\Sigma}_h^{k,d}\lambda_{c,d}$, such that the induced NCP for SNP j on mark k is just the NCP for causal SNP c on causal mark d, attenuated by the correlation between SNPs j and c, as well as the correlation between marks k and d. Here, rather than integrating out the causal NCPs as we did with the mark-expression associations, we use the observed Z-score for the causal SNP-mark pair to approximate the $\lambda_{j,k}$ terms, as the integration is not straightforward in the matrix-variate setting.

3.4.1.1 Computing posterior probabilities for causality

The posterior probability for causality for a given path can be expressed as

$$P(\mathbf{C_h}, \mathbf{C_g}|\mathbf{Z_g}, \mathbf{Z_h}) = \frac{P(\mathbf{Z_g}, \mathbf{Z_h}|\mathbf{C_h}, \mathbf{C_g})P(\mathbf{C_h}, \mathbf{C_g})}{P(\mathbf{Z_g}, \mathbf{Z_h})} \tag{3.16}$$

A prior can be specified on the probability that a SNP or mark within a fine-mapping region is causal, informed by features like distance to TSS, which is known to correlate with causality [18, 30], or functional annotations. Here we assign this prior to be uniform:

$$P(\mathbf{C_h}, \mathbf{C_g}|\mathbf{Z_g}, \mathbf{Z_h}) = \frac{P(\mathbf{Z_g}, \mathbf{Z_h}|\mathbf{C_h}, \mathbf{C_g})}{P(\mathbf{Z_g}, \mathbf{Z_h})} \tag{3.17}$$

$$= \frac{P(\mathbf{Z_h}|\mathbf{C_h})(P(\mathbf{Z_g}|\mathbf{C_h}, \mathbf{C_g})}{P(\mathbf{Z_g}, \mathbf{Z_h})} \tag{3.18}$$

We obtain $P(\mathbf{Z_h}|\mathbf{C_h})$ from Equation 11 and $P(\mathbf{Z_g}|\mathbf{C_h}, \mathbf{C_g})$ from Equation 15. We then compute $P(\mathbf{Z_g}, \mathbf{Z_h})$ by summing over the individual likelihoods for all possible causal paths. Here our method assumes a single causal SNP and mark per region, as we restrict our enumeration to only pairwise causal SNP-mark combinations.

Simulation Framework

We simulated data for 100 individuals over 10,000 50KB regions, using genotypes and LD from 65 YRI individuals obtained through 1000 Genomes [1]. SNP and mark correlations in our simulations were taken from the true correlations exhibited in these regions derived from these individuals. To determine causal status, we randomly chose one SNP and one mark to be causal in each region, thus defining a causal path through the data. Subsequently, we standardized genotypes and simulated values for chromatin marks and gene expression over all 100 individuals.

In order to simulate correlations between histone marks as observed in our empirical data, we drew mark values from an MVN as $\mathcal{N}(H_{ind}, \epsilon_g \Sigma_h)$, where the means, $H_{ind} = H_c \Sigma_{h,c}$, represent the induced values on non-causal marks due to correlations with the causal mark. The mean mark values for the causal mark were generated for each of the 100 individuals as $H_c = \beta_g G_c$, where G_c is the genotype of the individual at the causal SNP, the effect size β_g was drawn from a normal distribution, $\mathcal{N}(0, \sigma_g^2)$, with variance set to the desired variance explained by SNPs on marks $\sigma_g^2 = 0.25$, with the error term ϵ_g set to $1 - \sigma_g^2$. Finally, the individuals' values for gene expression are computed as $E = \beta_h H_c + \epsilon_h$, where H_c is the causal mark value as computed from the MVN, the effect size β_h was set to the desired variance explained from mark to expression $\sigma_g^2 = 0.25$, with the remaining error term given by $\mathcal{N}(0, 1 - \sigma_g^2)$.

For simulations in which there were multiple causal SNPs or marks, we randomly drew m or p, the number of causal SNPs or marks, from a binomial distribution where the expected number of causals per region was set to 1. However, we only included simulations with two or more causals. For multi-causal-SNP simulations, we then randomly selected m causal SNPs in the region and simulated chromatin marks and gene expression as described previously, but drew the effect sizes of each SNP as $\mathcal{N}(0, \sigma_g^2/m)$, such that the total expected variance explained remained at 0.25. For multi-causal-mark simulations, we randomly selected p causal marks in the region and simulated chromatin marks by defining the means, H_c, of each causal mark independently as described for the single-causal simulations. We then computed gene expression by drawing the effect size, β_h, of each causal mark from $\mathcal{N}(0, \sigma_g^2/p)$ such that the total expected variance explained remained at 0.25.

3.4.2 Existing approaches

We benchmark our method against five alternative approaches. Firstly, we compare against the standard overlap analysis whereby hQTLs and eQTLs are independently identified within a region centered around a gene. We follow the protocol outlined in [85]. In this experiment, we computed the best SNP association in each region with every mark measured in the

region as well as with the gene expression value for that region. We determined adjusted p-values for each top association by performing permutation tests. We then accounted for multiple testing at the mark level by determining the minimum FDR at which each adjusted p-value would be considered significant. This was estimated via the qvalue package [76]. This procedure resulted in a set of significant SNP-mark associations, as well as one SNP-expression association within the region, as only the top SNP association is retained for each biological phenotype. We then evaluated the number of causal SNPs, marks, and paths that were ultimately included in these candidate sets.

Secondly, we compared against the approach of independently fine-mapping the two levels of data (SNP-mark and mark-expression), and multiplying together pairs of posterior probabilities to produce probabilities of causality for paths. For these independent fine-mapping experiments, we used a simple approach that assumes a single causal variant, approximating posterior probabilities for causality directly from Z-scores [58].

In addition, we compared against a basic ranking approach, where we independently computed SNP-mark, mark-expression, and SNP-expression associations for every SNP and mark within a region. For SNP and mark prioritization, we simply produced a ranking of the SNP-expression and mark-expression posterior probabilities for causality, respectively. For path prioritization, we produced a ranking of the product of SNP-mark and SNP-expression posterior probabilities.

We next compared against a bayesian network model which computes directed association strengths between all possible pairs of nodes in a given network [73]. The method takes as input raw genotype and phenotype values. As nodes, we included all SNPs and marks, as well as the gene expression value, within a region. We allowed only for node pairings directed from SNP to mark or from mark to gene expression. For SNP and mark prioritization, we ranked association strengths over all directed SNP-expression edges and mark-expression edges, respectively. For path prioritization, we produced a ranking of the product of SNP-mark and mark-expression strengths.

Finally, we compared against Coloc, which is designed to identify SNPs that are likely to be causal for multiple traits at once. Specifically, Coloc outputs a posterior probability that a SNP is causal for two arbitrary traits simultaneously. We adapted Coloc for our purposes by running the method on all SNPs independently. For each SNP, the two given traits were (1) gene expression, and (2) a mark value. Thus, we ran Coloc independently for all SNP-mark combinations. This produced a set of posterior probabilities indicating, for each SNP-mark combination, the likelihood that the SNP is causal for both the mark value and gene expression simultaneously. For path prioritization, we ranked these probabilities over all SNP and mark combinations. For SNP and mark prioritization, we marginalized over all marks and SNPs, respectively, producing posterior probabilities for each SNP and mark to be causal independently.

3.4.3 Real data

The real data analyses were done on 65 YRI individuals whose genotypes were obtained through 1000 Genomes and standardized. PEER-normalized [75] H3K4me1, H3K4me3, H3K27ac, DHS, and RNA expression marks in lymphoblastoid cell lines (LCLs) for these individuals were obtained from [30]. For each gene in the dataset, we computed associations for every SNP-mark, SNP-gene, and mark-gene pair within a 50kb window centered around the gene TSS. On average, each region contained 160 SNPs and 25 marks (across the four mark types – H3K4me1, H3K4me3, H3K27ac, and DHS – whose peak values we analyzed together in each region). Overall, from 14,669 50kb regions, we filtered for regions that exhibited evidence for our sequential model where SNPs affect chromatin marks, which in turn affect gene expression. Specifically, for each region we performed a two-stage regression where we first regressed gene expression on all chromatin marks, and (2) regressed the proportion of expression explained by the chromatin marks on each SNP. If at least one SNP had a low p-value for association ($p < 0.05/n.snps$) to the proportion of gene expression explained by chromatin data, we kept this region for our real data analysis. After this filtering procedure, we retained 1,317 regions.

We obtained motif annotations from HaploReg [84] and ChromHMM annotations from the NIH Roadmap Epigenomics Consortium [46]. When comparing annotations of top prioritized paths with those of random paths, we established corresponding background paths by choosing a random SNP/mark combination at every region where a top path was reported.

For GWAS analyses, we explored regions whose tag SNP was associated to an autoimmune trait with $p < 5 \times 10^{-8}$. Associations were obtained from recent literature for eight autoimmune phenotypes [53, 6, 19, 16, 15, 65]. For each of *pathfinder*'s top reported paths, we determined whether the corresponding SNP was contained within any of the GWAS regions in our dataset. In order establish a null distribution for this statistic, we ran the same analysis for random regions in the genome not overlapping with the GWAS regions in our dataset. Specifically, for each GWAS region, we randomly selected a SNP in the same chromosome matched for MAF ($\epsilon = 0.01$) and LD score ($\epsilon = 0.001$) with the GWAS tag SNP. We established a window around this matched SNP corresponding to the window size of the GWAS region. Finally, we determined the number of top paths that fell within these random regions. We repeated this experiment 100 times to establish the null distribution of this measurement and calculated a p-value using a Z-test.

3.5 Tables

Table 3.1: **50%, 90%, and 99% credible sets for SNP-, mark-, and path-mapping for real data analysis.** We compare *pathfinder* to basic eQTL mapping, with respect to the size of their credible sets, averaged across all regions. Standard errors are included next to each measurement.

	50% credible set			90% credible set			99% credible set		
method	SNPs	Marks	Paths	SNPs	Marks	Paths	SNPs	Marks	Paths
pathfinder	4.9 (0.2)	1.0 (0.0)	7.4 (0.3)	28.4 (1.1)	1.8 (0.1)	158.4 (6.0)	64.2 (2.4)	6.3 (0.2)	765.5 (29.0)
eQTL mapping	8.1 (0.3)	-	-	45.3 (1.7)	-	-	92.9 (3.6)	-	-

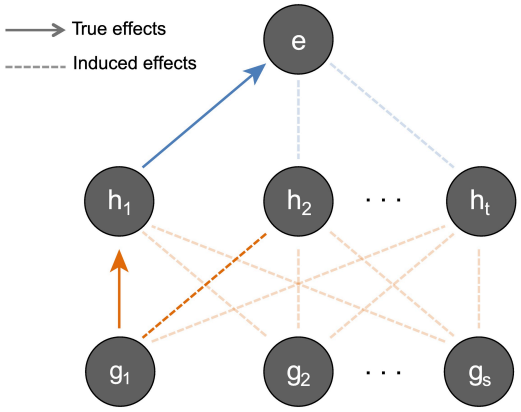
Table 3.2: **Top causal paths produced by real data analysis.** For each path, we report the chromosome, the RSID of the implicated SNP, the implicated mark type, the posterior probability we assigned to this path, three Z-scores (SNP to mark association, mark to expression association, SNP to expression association), the GENCODE gene around which this region was centered, the ChromImpute [21] annotation for the SNP, and the number of regulatory motifs altered by the SNP, as designated by HaploReg [84].

chr	rsid	mark type	posterior	SNP-mark Z	mark-exp Z	SNP-exp Z	gene	chromatin state	motifs altered
12	rs835044	H3K27ac	> 0.99	-13.05	4.97	-4.65	NDUFA12	1TssA	5
1	esv3587154	H3K4me1	> 0.99	-18.13	17.40	-14.97	GSTM1	15Quies	-
19	rs385895	H3K4me1	> 0.99	12.60	2.41	1.50	CLC	7Enh	3
15	rs8025332	H3K4me1	> 0.99	-12.07	2.11	-2.35	CELF6	15Quies	1
5	rs1217817	H3K4me1	> 0.99	-14.59	5.58	-4.52	MAP1B	7Enh	4
1	rs7417106	DHS	> 0.99	-8.62	-0.16	-0.54	C1orf170	4Tx	22
1	rs111900551	H3K4me3	> 0.99	-8.82	2.26	-2.95	CLCNKA	15Quies	18
3	rs57339700	H3K4me1	> 0.99	-9.66	2.37	-2.29	CAND2	14ReprPCWk	5
6	rs9349050	H3K4me3	> 0.99	-12.47	10.80	-8.19	MDGA1	11BivFlnk	2
3	rs6763025	H3K4me1	> 0.99	10.59	-2.21	-2.18	PRSS50	7Enh	4

Table 3.3: **Top causal paths reported in real data analysis that localized within GWAS regions for 8 autoimmune diseases.** For each path, we report the chromosome, the RSID of the implicated SNP, the implicated mark type, the posterior probability we assigned to this path, three Z-scores (SNP to mark association, mark to expression association, SNP to expression association), the GENCODE gene around which this region was centered, the ChromHMM [21] annotation for the SNP, and the number of regulatory motifs altered by the SNP, as designated by HaploReg [84].

chr	rsid	GWAS	mark type	posterior	SNP-mark	mark-exp	SNP-exp	gene	chrom state	motifs altered
2	rs2975781	UC, IBD	H3K27ac	1.00	-9.00	5.33	-4.96	GPR35	7Enh	9
8	rs2618481	SLE	H3K27ac	0.94	-6.04	6.59	-3.99	BLK	2TssAFlnk	0
16	rs9927129	Crohn's, IBD	H3K4me1	0.66	-7.82	-0.79	1.59	RP11-1348G14.2	15Quies	1
6	rs2071889	UC, SLE, MS, RA, IBD	DHS	0.61	6.51	-3.23	-1.78	TAPBP	4Tx	2
16	rs394502	Crohn's, IBD	H3K4me1	0.44	9.96	-1.59	-2.62	EIF3CL	15Quies	4
1	rs57126490	UC, MS, RA, IBD	DHS	0.43	4.65	-0.14	0.04	PANK4	5TxWk	0
6	rs915654	UC, SLE, Crohn's, PBC, MS, RA, IBD	H3K4me3	0.42	3.48	5.98	3.51	LTA	7Enh	5
1	rs114312440	Crohn's	H3K4me3	0.41	-4.54	3.44	-2.79	MTX1	5TxWk	2
3	rs71155551	SLE	H3K27ac	0.39	4.73	3.20	1.27	COPG1	5TxWk	2
1	rs34769708	Crohn's	H3K4me3	0.39	-4.86	2.13	-2.71	ASH1L	7Enh	3
6	rs13197384	MS	H3K4me3	0.35	6.68	4.44	3.84	AHI1	1TssA	16
6	rs147085011	UC, SLE, PBC, MS, RA, IBD	H3K4me3	0.32	5.11	-0.27	-0.42	RPP21	1TssA	16
16	rs243332	PBC, MS	DHS	0.28	4.45	2.26	0.74	SOCS1	1TssA	9
6	rs575034	RA	H3K4me1	0.23	3.73	3.51	0.85	SLC35B2	1TssA	1
2	rs737231	Crohn's, Celiac	H3K4me1	0.22	3.59	3.13	2.10	SLC9A4	15Quies	6
5	rs17097187	MS	H3K4me3	0.22	-2.94	6.27	-4.93	PCDHGA1	9Het	4
2	rs737231	IBD	H3K4me1	0.22	3.59	3.13	2.10	SLC9A4	15Quies	6
1	rs2641116	UC, IBD	H3K4me3	0.20	4.57	0.59	1.08	PARK7	4Tx	1
20	rs6115319	MS	H3K27ac	0.11	-5.58	6.39	-4.13	FAM182B	15Quies	0

3.6 Figures



$$\mathbf{Z_h}|\mathbf{C_h} \sim \mathcal{N}(\mathbf{\Sigma_h \Lambda_{h,d}}, \mathbf{\Sigma_h})$$

$$\mathbf{Z_g}|\mathbf{C_g}, \mathbf{C_h} \sim \mathcal{MN}(\mathbf{M}, \mathbf{\Sigma_g}, \mathbf{\Sigma_h})$$

Figure 3.1: **Schematic of hierarchical model whereby SNPs affect histone marks, which in turn affect gene expression.** We illustrate a scenario where SNP g_1 and mark h_1 are causal. All other induced correlations, such as the effect of g_1 on h_2, are an effect of LD and/or correlations among marks. To the right we show our mathematical model for this hierarchical framework. On the top level, we model mark-expression associations with a Multivariate Normal (MVN) distribution. On the bottom, we jointly model all associations between all SNPs and marks with a Matrix Variate Normal distribution (see Methods).

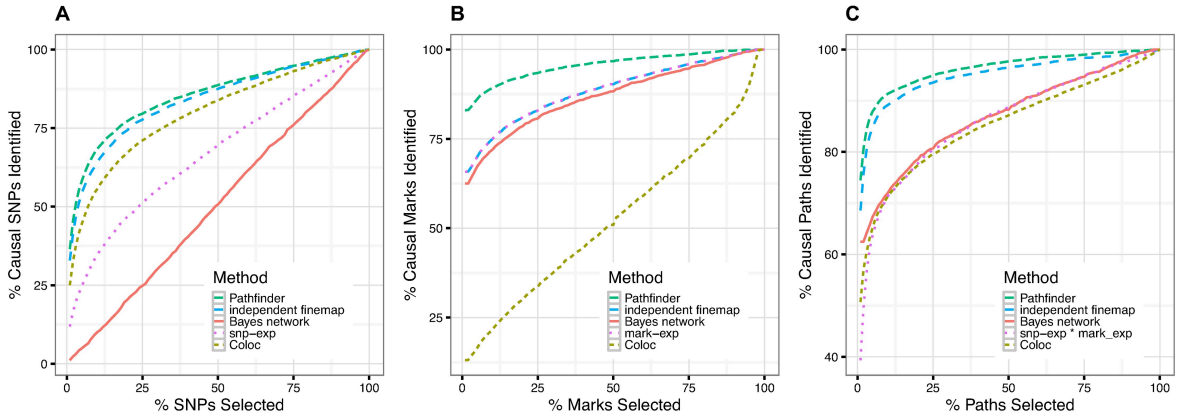


Figure 3.2: **Comparison of our method against four potential competitors - independent fine-mapping, a simple ranking of associations, Coloc, and Bayesian network analysis.** We measure performance as the number of simulated causal SNPs, marks, and paths that each method is able to recapture, while varying the number of SNPs, marks, or paths considered.

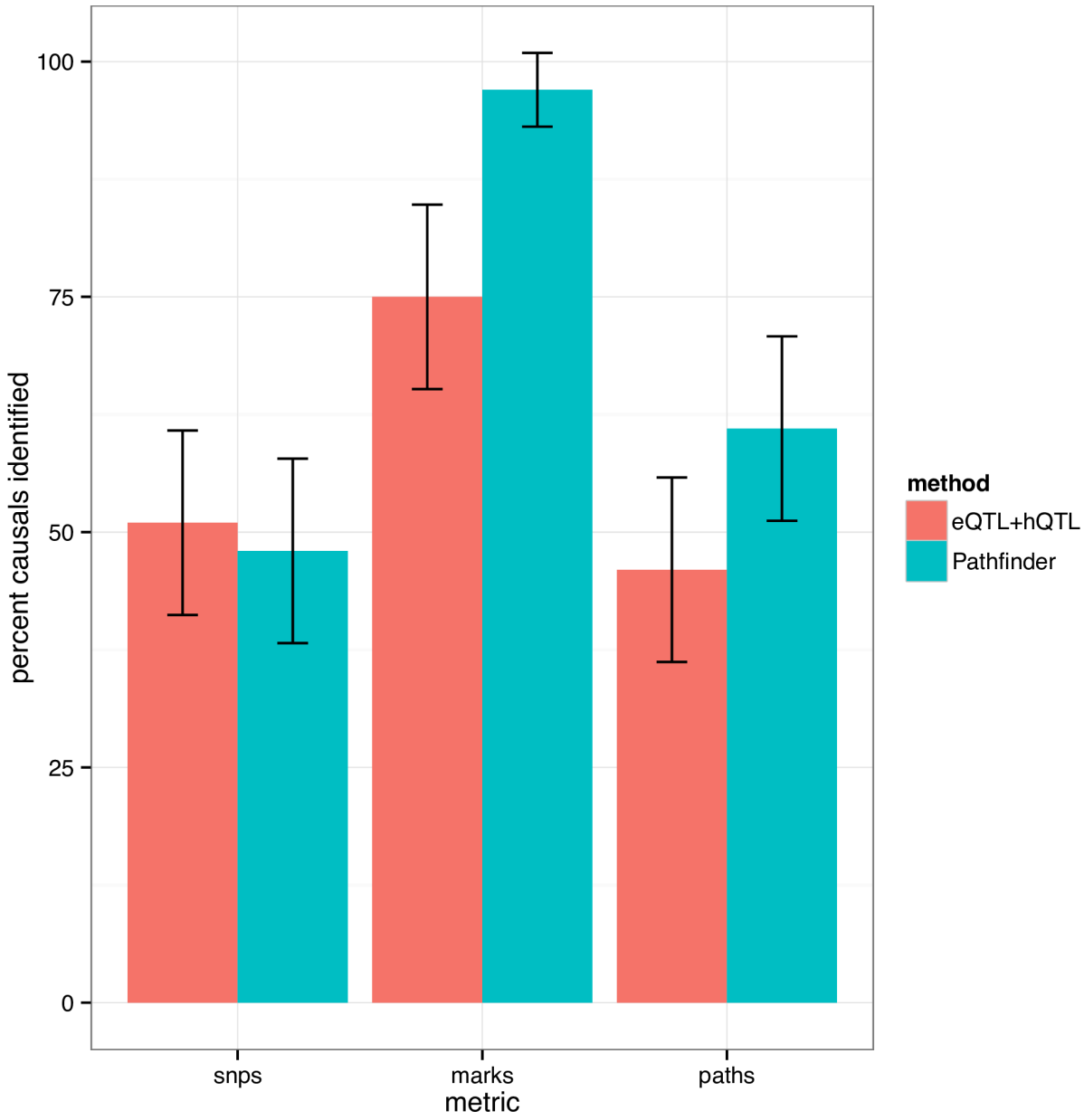


Figure 3.3: **Comparison of our method to standard eQTL + hQTL overlap analyses** In overlap analyses, only the top SNP for association to each histone mark and gene expression is considered. We demonstrate significant gains in our method with respect to mark-finding accuracy, where SNP-mapping performance is comparable between the two methods.

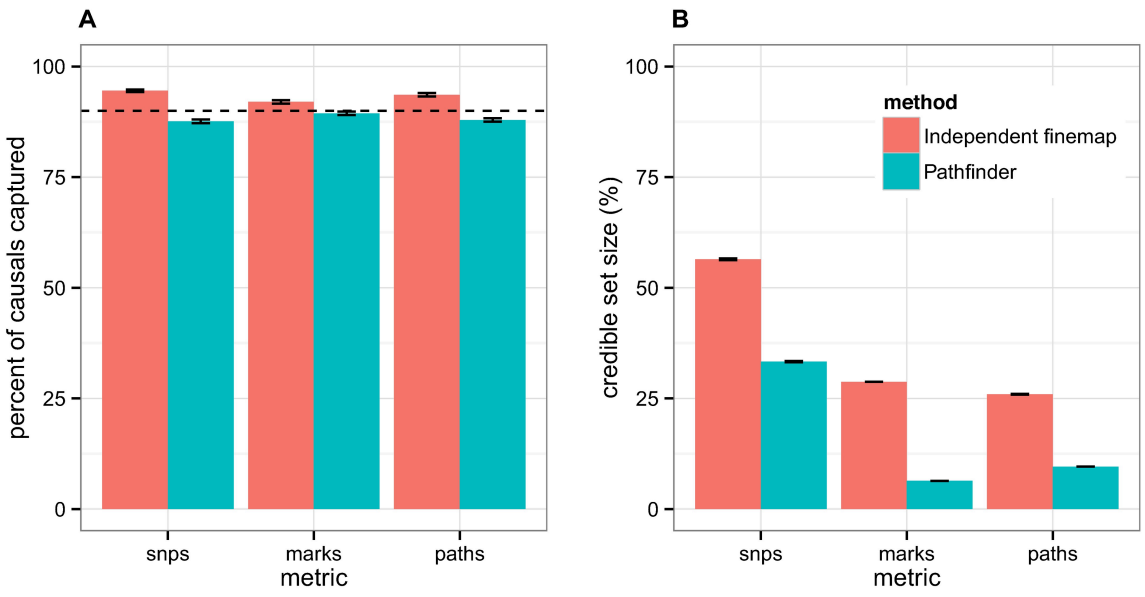


Figure 3.4: **90% credible sets for SNP-, mark-, and path-mapping.** We compare *pathfinder* to the technique of independently fine-mapping the two levels of data, with respect to (A) the calibration of their credible sets and (B) the size of their credible sets. In (A), we compare the proportion of causal variants that were captured in the 90% credible sets using *pathfinder* vs. independent fine-mapping against the expected proportion (represented by the dotted line). In (B), we display the corresponding sizes of these credible sets.

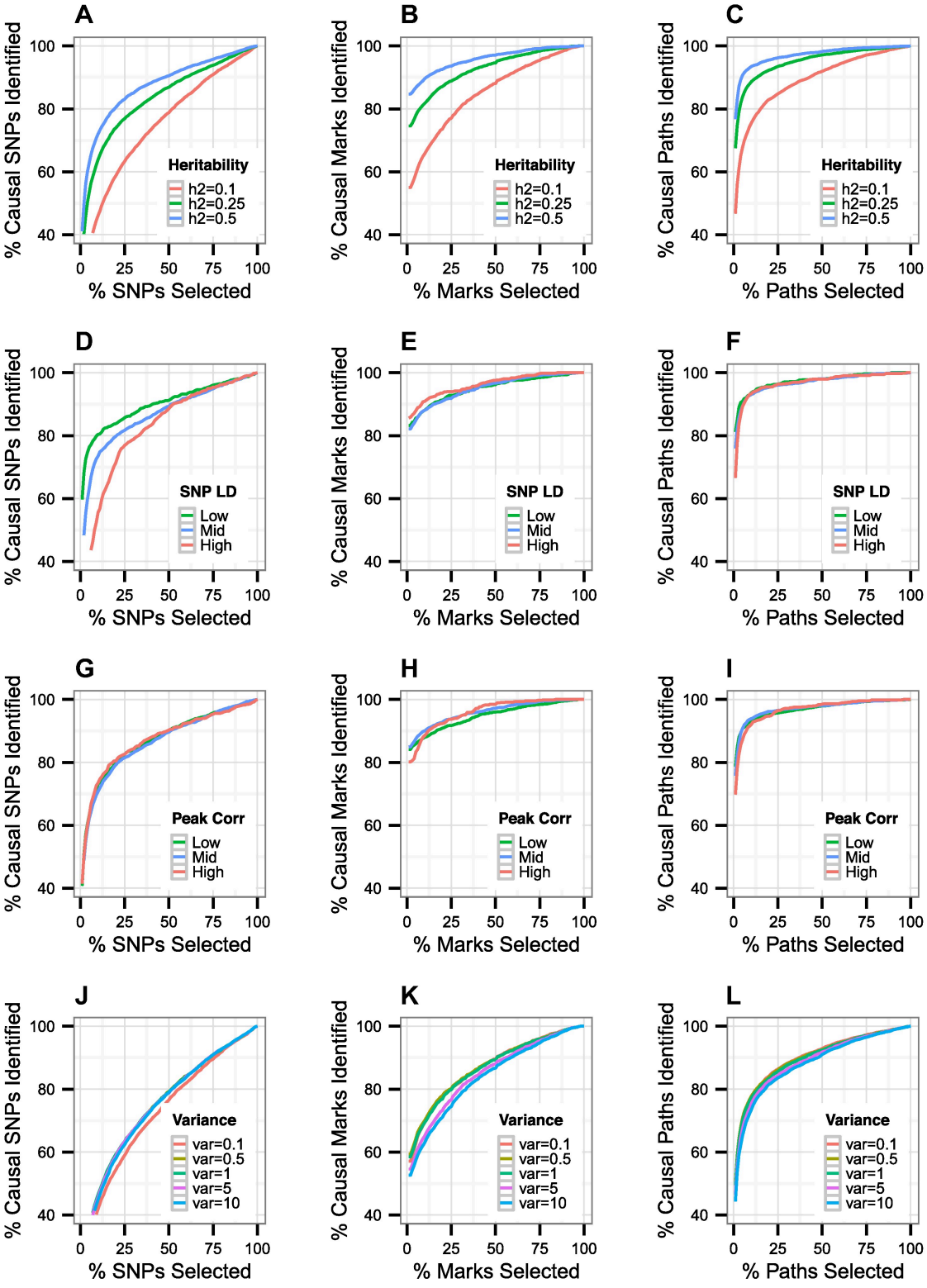


Figure 3.5: **Performance of our method as we vary levels of variance explained, SNP LD, mark correlations, and the prior variance parameter.** (A-C) We simultaneously vary the variance explained by SNP and mark from 0.1 to 0.5 per region. (D-I) We stratified based on mean SNP/mark correlations at the causal SNP/mark. (J-L) We show that *pathfinder* is not sensitive to variations in our prior variance parameter.

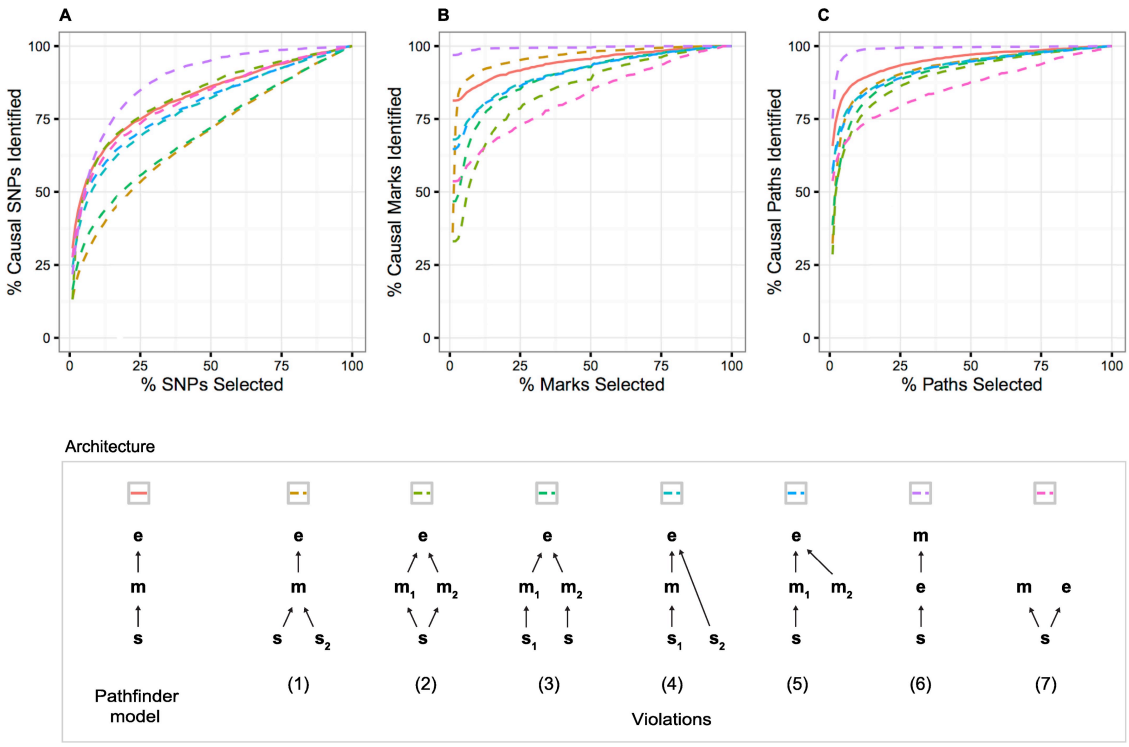


Figure 3.6: **Performance of our method under violations of the causal model.** (A-C) *pathfinder*'s SNP-, mark-, and path-mapping accuracy for standard simulations compared with seven model violations. (D) The model violations include the following scenarios: (1) multiple causal SNPs impact a single causal mark, which affects gene expression, (2) a single SNP impacts multiple causal marks, which both affect gene expression, (3) two SNPs affect two marks (respectively), which both impact gene expression, (4) a single causal SNP impacts a single causal mark that affects gene expression, with an additional SNP also impacting gene expression directly, (5) a single causal SNP impacts a single causal mark that affects gene expression, with an additional mark also impacting gene expression, (6) a single causal SNP affects gene expression directly, which in turn affects a single mark, and (7) a single causal SNP has independent effects on a single mark and gene expression

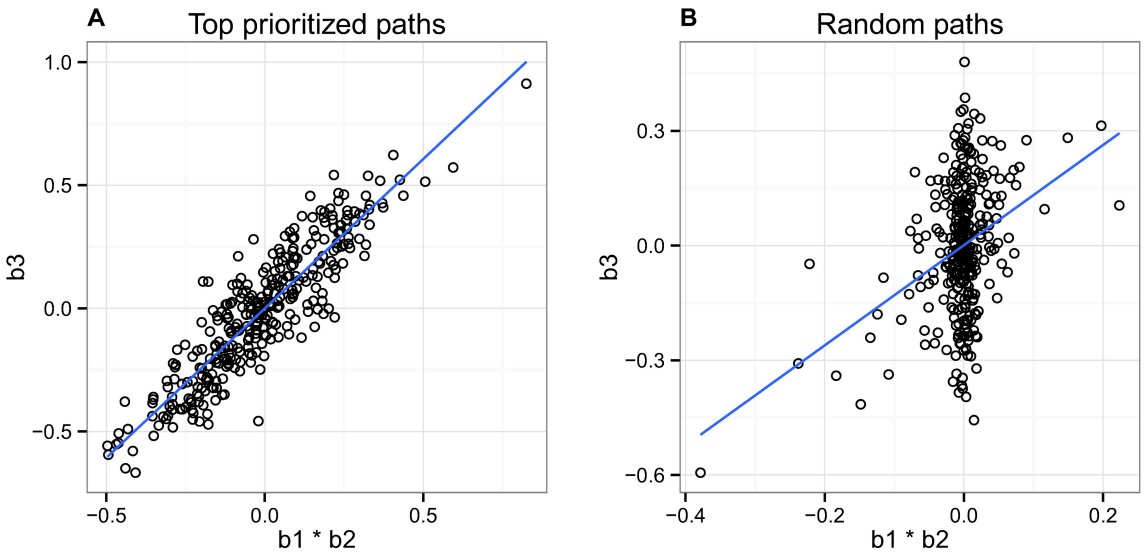


Figure 3.7: **Relationship between the product of the SNP-mark and mark-expression effect sizes against the overall SNP-expression effect size.** (A) We observe a high correlation (r = 0.91) between these effect size vectors, indicating that our method is identifying many pathways that are likely to be following our causal model. Here we included only the top paths whose posterior probabilities for causality were assigned to be greater than 0.1. (B) We show that a significant correlation does not exist for randomly chosen paths.

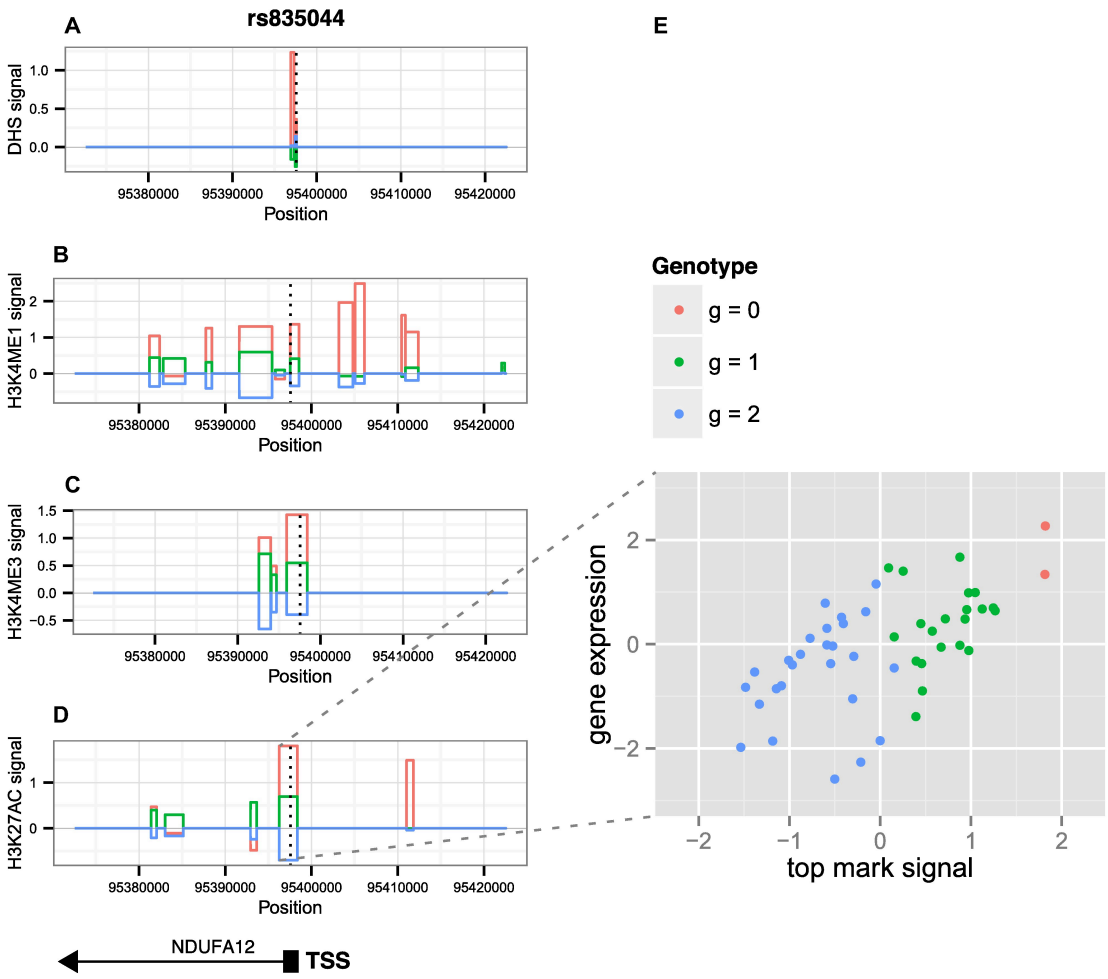


Figure 3.8: **Genomic context of top path reported by *pathfinder* in real data.** (A-D) Mark signals for DHS, H3K4me1, H3K4me3, H3K27ac in a 4kb region centered around the NDUFA12 TSS, stratified by genotype. The implicated SNP, signified by the vertical dotted line, lies 6bp downstream of the gene TSS, and falls within an H3K27ac peak, which is also the top mark reported by *pathfinder*. The posterior probability for causality for this peak was greater than 0.999. (E) Relationship between the H3K27ac peak signal and gene expression, stratified by genotype.

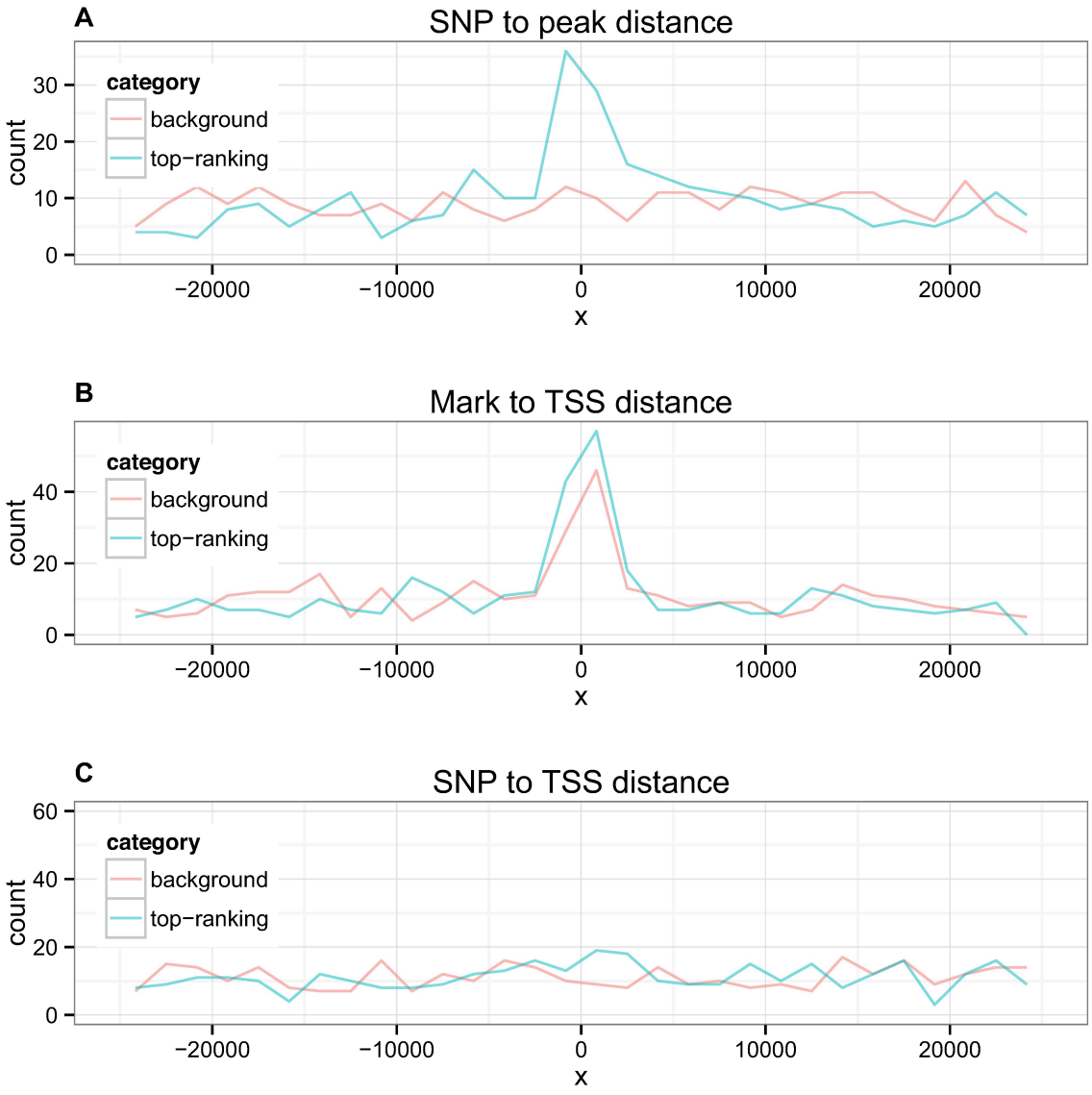


Figure 3.9: **Spatial relationships between SNP, mark, and TSS in top paths reported by *pathfinder* vs random paths.** (A) Distances from SNP to mark (B) Distances from mark to TSS (C) Distances from SNP to TSS.

CHAPTER 4

Leveraging functional data to improve power of GWAS summary statistic imputation

4.1 Introduction

Genome-wide association studies (GWASs) have identified thousands of genetic variants robustly associated with complex traits and disease. With few exceptions, GWASs typically measure individual genotypes using affordable array-based technologies that capture a limited number of markers. To increase statistical power, GWASs have relied on genotype imputation, where unmeasured genotype information is predicted using large-scale reference panels of sequenced individuals[36, 51, 7]. While genotype imputation using individual-level data results in highly accurate genotypes, it requires significant computational resources [7]. Recent studies have proposed methods to impute directly unmeasured GWAS summary statistics[49, 66]. The primary source of information enabling summary-based imputation is linkage-disequilibrium (LD) estimated from publicly available reference genotype panels. Indeed, methods typically model statistics at missing markers as a weighted combination of the measured statistics, where weights are determined by regional LD. Summary-based imputation has been shown to be computationally scalable, accurate, and has the added benefit of not requiring individual-level genotyping data[66].

This chapter is being prepared for submission.

Collaborative efforts to identify functionally active regions in the genome have resulted in a rich categorization of putative activity for non-coding genetic variation[20, 47]. A large body of work integrating functional genomics with GWAS has revealed that single-nucleotide polymorphisms (SNPs) coinciding with certain functional features are enriched for disease heritability[20, 61, 31, 26, 27, 33, 55]. This insight has inspired computational methods to incorporate functional information together with GWAS to increase statistical power for association testing, boost performance for statistical fine-mapping, and dissect SNP-heritability [70, 23, 17, 72, 67, 43, 56, 79, 89, 38]. Thus, a natural extension of these findings would be to incorporate functional information into summary-based imputation methods.

In this work we describe a novel computational framework to impute GWAS summary statistics by leveraging functional annotation data at typed and untyped SNPs. Our approach, FIMPG, extends the fixed-effect linear model based on LD-weighted statistics[49, 66] by including prior effect-size distributions defined by functional annotations. We performed exhaustive simulations using real genotype data and various trait architectures and find that FIMPG improves summary statistics prediction at higher rates of SNP missingness for a single-causal model, and across a wide range of SNP missingness under the infinitesimal model. Lastly, we validate FIMPG using publicly available summary data from 27 GWASs performed using the UKBiobank[78, 38]. Overall, we find that while improvements in prediction accuracy are not sustained in real data, FIMPG's predicted statistics are consistently less deflated than those of functionally-unaware methods and may boost signal at missing statistics.

4.2 Results

4.2.1 Overview of methods

We propose FIMPG, a summary statistics imputation method that integrates functional annotation data to improve prediction of associations at untyped SNPs. Our approach, like

earlier works[49, 66], models summary statistics under a linear model; however, we make use of a random-effects model where SNP effect sizes are drawn from a normal distribution with variance defined by functional categories[43, 26]. Specifically, given GWAS summary statistics $\mathbf{z}_o$, linkage disequilibrium $\mathbf{\Sigma}$, and variance estimates for functional categories $\mathbf{D}$, we model unobserved summary data $\mathbf{z}_u$ under a conditional normal as,

$$\mathbf{z}_u|\mathbf{z}_o \sim \mathcal{N}(\mathbf{V}_{u,o}\mathbf{V}_{o,o}^{-1}\mathbf{z}_o, \mathbf{V}_{u,u} - \mathbf{V}_{u,o}\mathbf{V}_{o,o}^{-1}\mathbf{V}_{o,u}),$$

where $\mathbf{V}_{u,o} = \mathbf{\Sigma}_{u,o}+\mathbf{\Sigma}_{u,u}\mathbf{D}_{u,u}\mathbf{\Sigma}_{u,o}+\mathbf{\Sigma}_{u,o}\mathbf{D}_{o,o}\mathbf{\Sigma}_{o,o}$ and $\mathbf{V}_{o,o} = \mathbf{\Sigma}_{o,o}+\mathbf{\Sigma}_{o,o}\mathbf{D}_{o,o}\mathbf{\Sigma}_{o,o}+\mathbf{\Sigma}_{o,u}\mathbf{D}_{u,u}\mathbf{\Sigma}_{u,o}$ capture uncertainty due to finite-sample size ($\mathbf{\Sigma}$), and tagged effect-size uncertainty explained by functional annotations ($\mathbf{\Sigma D \Sigma}$). We note that the FIMPG model recovers the IMPG model as a degenerate case when the prior variance parameters are zero. To impute summary data $\mathbf{z}_u$ under our model, we require the relevant functional categories and their corresponding variance parameters $\mathbf{D}$. We infer variance parameters for 53 baseline functional annotations using stratified LD-Score regression and prune non-significant results [9, 26]. To avoid overfitting we use a leave-one-chromosome-out approach to fitting LD-Score regression models, therefore ensuring that observed summary statistics at a given locus are not used twice for inference (see Methods).

4.2.2 FIMPG accurately imputes GWAS summary statistics

We first sought to assess the performance of FIMPG using simulated GWAS summary statistics. Briefly, we simulated GWAS summary data by sampling Z-scores directly under multiple genetic architectures at each region, while varying the proportion of missing SNPs (see Methods). For completeness, we compare our approach with the functionally unaware method, IMPG[66]. We find that FIMPG outperforms IMPG under the infinitesimal simulation setup across all proportions of SNP retention (Figure 4.1a). For example, under the infinitesimal model and averaged across all rates of SNP retention, the mean R^2 between the true and predicted Z-scores is 0.97 for FIMPG and 0.96 for IMPG ($p < 0.001$, Wilcoxon test). For

single causal simulations, FIMPG appears to outperform IMPG only at lower rates of SNP retention (Figure 4.1b).

In order to assess FIMPG's performance at null loci, we simulated summary statistics from regions where no SNPs were causal. Under the single-causal model ,FIMPG appears to be slightly inflated, as the mean λ_{gc} reported across all proportions of retained SNPs is significantly different from 1 (mean = 1.19, p¡0.001) (Figure 4.2). A similar trend was observed in infinitesimal simulations.

Next, we investigated the effects of simulation parameters on FIMPG's accuracy under the single-causal model, fixed at 50% SNP retention. Firstly, we varied the number of annotations used in simulations, randomly sampling 1-3 of the three available annotations at each locus. Secondly, we varied the enrichment level for all annotations. To simulate varying enrichments, we multiplied the original enrichments of each annotation by various scalar values. Under both scenarios, FIMPG's improvement over IMPG does not appear to significantly change (Figure 4.3A-B). Thirdly, we predicted that FIMPG's accuracy would increase with GWAS sample size. We show that this trend holds across varying sample sizes between 25k and 200k (Figure 4.3C). In addition, we explored the effect of σ^2 on performance, anticipating that higher σ^2 values will lead to better accuracy. To simulate this effect, we varied σ_0^2, which represents the baseline variance contributed by any causal SNP. As the variance contributed by each additional annotation (σ^2) is determined by multiplying σ_0^2 by the annotation's enrichment, each change in σ_0^2 also affected σ^2 values for all annotations. The results confirmed that FIMPG's accuracy improves as σ^2 increases (Figure 4.3D). Finally, we vary the number of causal variants drawn at each locus. As expected, FIMPG's performance improves with the number of causal variants, as higher numbers of causal variants will increasingly resemble the infinitesimal genetic architecture (Figure 4.3E).

4.2.3 FIMPG performance is stable under model mis-specifications

In order to model mis-specifications with respect to the annotations and their enrichments, we performed two analyses. As a weak violation, we inferred enrichments that varied to a degree from simulated enrichments (drawn from the annotation-specific distribution of LD-score estimates reported in [26]) (Figure 4.4A). As a strong violation, we randomly omitted one of three simulated annotations from the inference step (Figure 4.4B). Under both violations, FIMPG's performance does not fall below that of IMPG. For example, under the weak violation, the average squared correlations of FIMPG and IMPG are 0.937 and 0.936 (p=0.49, Wilcoxon test). Under the strong violation, the average squared correlations of FIMPG and IMPG are 0.936 and 0.935 (p=0.37, Wilcoxon test).

4.2.4 Application to real data

We applied FIMPG to 27 UKBiobank traits, each including approximately 337K European-ancestry individuals. For each trait, we ran LDScore regression on 53 annotations in the baselineLD model [27] and removed annotations with low enrichment significance ($|Z| < 1.96$) independently for each trait, leaving on average 7 annotations per trait. At each locus, we masked a random 90% of SNPs and compared FIMPG's and IMPG's predictions against the masked statistics.

We first investigated the difference in prediction performance between FIMPG and IMPG. We show that, at 10% SNP retention, FIMPG's and IMPG's performance with respect to squared correlation are comparable (FIMPG: $r^2 = 0.687 \pm 0.014$, IMPG: $r^2 = 0.686 \pm 0.014$, 95% CI).

We also investigated whether incorporating functional information as in FIMPG boosts association signal. For every locus, we computed the ratio of the most significant true association to FIMG's and IMPG's corresponding predictions, respectively. Averaged over all loci at all traits, this ratio for FIMPG (21.2) was significantly lower than for IMPG (21.9),

(p ¡ 0.001, paired Wilcoxon test). This finding suggests that FIMPG's predictions are less deflated than IMPG's and incorporating functional information may boost signal at missing statistics.

4.3 Discussion

We have introduced a summary statistics imputation method that leverages functional annotation data to improve imputation accuracy. We demonstrated in simulations that our method consistently improves imputation accuracy under the infinitesimal model, and at higher rates of SNP missingness under the single causal model. However, this finding was not sustained in real data. Rather, the benefit of integrating functional information as in FIMPG comes from a boost in signal at missing statistics. We showed in real data that FIMPG's predictions are consistently less deflated than those of traditional summary imputation methods, which could lead to the detection of significant associations not identified using traditional methods.

Moreover, FIMPG's advantage over functionally-unaware methods is bounded by the limitations in heritability partitioning by functional category. As many annotations were filtered out prior to imputation due to high standard errors, FIMPG could not leverage the full functional model, which hindered prediction accuracy. We anticipate that advancements in the accuracy of functional heritability partitioning may further strengthen FIMPG's improvements with respect to both imputation accuracy and power.

4.4 Methods

4.4.1 Model for a polygenic trait

We define a quantitative polygenic trait for n individuals $\mathbf{y}$ as a linear function of p centered and standardized genotype values, given by $n \times p$ matrix $\mathbf{X}$, their respective effects $\boldsymbol{m}\beta$,

and environmental noise $\boldsymbol{m}\epsilon$. We assume that genotypic effects $\boldsymbol{m}\beta$ are random with their variance determined by the k functional categories a given SNP falls in. We model which functional categories SNP i falls in by a 0-1 $p \times k$ matrix $\mathbf{A}$. We define this formally as

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

$$\boldsymbol{m}\beta_i \mid \sigma^2(\mathbf{A}_i) \sim \mathcal{N}(0, \sigma^2(\mathbf{A}_i)) \tag{4.2}$$

$$\boldsymbol{m}\epsilon \mid \sigma_e^2 \sim \mathcal{N}(0, \mathbf{I}_n\sigma_e^2), \tag{4.3}$$

where $\sigma^2(\mathbf{A}_i) = \sum_k \sigma_k^2 \mathbf{A}_{i,k}$. Therefore, the sampling distribution for $\mathbf{y}$ is given by

$$\mathbf{y} \mid \mathbf{X}, \mathbf{D}, \sigma_e^2 \sim \mathcal{N}(\mathbf{0}, \mathbf{X}\mathbf{D}\mathbf{X}^\intercal + \mathbf{I}_n\sigma_e^2), \tag{4.4}$$

where $\mathbf{D}_{i,i} = \sigma^2(\mathbf{A}_i)$ and 0 elsewhere.

4.4.2 Imputation of summary statistics using reference LD

The association strength of the ith SNP is defined as,

$$z_i = \frac{1}{\sqrt{n}\sigma_e}\mathbf{X}_i^\intercal y \tag{4.5}$$

$$= \frac{1}{\sqrt{n}\sigma_e}\mathbf{X}_i^\intercal\mathbf{X}\boldsymbol{m}\beta + \frac{1}{\sqrt{n}\sigma_e}\mathbf{X}_i^\intercal\boldsymbol{m}\epsilon, \tag{4.6}$$

which can be extended to p SNPs by

$$\mathbf{z} = \frac{\sqrt{n}}{\sigma_e}\boldsymbol{\Sigma}\boldsymbol{m}\beta + \frac{1}{\sqrt{n}\sigma_e}\mathbf{X}^\intercal\boldsymbol{m}\epsilon, \tag{4.7}$$

where $\boldsymbol{\Sigma} = n^{-1}\mathbf{X}^\intercal\mathbf{X}$ is the SNP correlation matrix (i.e. linkage disequilibrium). Thus, the sampling distribution of $\mathbf{z}$ is characterized by

$$\mathbf{z} \mid \boldsymbol{\Sigma}, \mathbf{D}, \sigma_e^2, n \sim \mathcal{N}(\mathbf{0}, \frac{n}{\sigma_e^2}\boldsymbol{\Sigma}\mathbf{D}\boldsymbol{\Sigma} + \boldsymbol{\Sigma}). \tag{4.8}$$

To model missing associations statistics, we partition $\mathbf{z}$ into observed $\mathbf{z}_o$ and unobserved $\mathbf{z}_u$. Our goal is to predict or impute the unobserved summary statistics $\mathbf{z}_u$ given $\mathbf{z}_o$. A natural choice is the expectation of the conditional distribution, which has the added benefit of being the "best linear unbiased predictor" (i.e. BLUP[69]) for our missing data. Without loss of generality, we partition the LD ($\boldsymbol{\Sigma}$) and prior effect variance matrices ($\mathbf{D}$) into observed and unobserved blocks as

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{u,u} & \boldsymbol{\Sigma}_{u,o} \\ \boldsymbol{\Sigma}_{o,u} & \boldsymbol{\Sigma}_{o,o} \end{bmatrix} \text{ and } \mathbf{D} = \begin{bmatrix} \mathbf{D}_{u,u} & \mathbf{0} \\ \mathbf{0} & \mathbf{D}_{o,o} \end{bmatrix}. \tag{4.9}$$

The conditional distribution of $\mathbf{z}_u \mid \mathbf{z}_o$ is defined as

$$\mathbf{z}_u|\mathbf{z}_o, \boldsymbol{\Sigma}, \mathbf{D}, \sigma_e^2, n \sim \mathcal{N}(\mathbf{V}_{u,o}\mathbf{V}_{o,o}^{-1}\mathbf{z}_o, \mathbf{V}_{u,u} - \mathbf{V}_{u,o}\mathbf{V}_{o,o}^{-1}\mathbf{V}_{o,u}), \tag{4.10}$$

where $\mathbf{V}_{u,o} = \boldsymbol{\Sigma}_{u,o}+\boldsymbol{\Sigma}_{u,u}\mathbf{D}_{u,u}\boldsymbol{\Sigma}_{u,o}+\boldsymbol{\Sigma}_{u,o}\mathbf{D}_{o,o}\boldsymbol{\Sigma}_{o,o}$ and $\mathbf{V}_{o,o} = \boldsymbol{\Sigma}_{o,o}+\boldsymbol{\Sigma}_{o,o}\mathbf{D}_{o,o}\boldsymbol{\Sigma}_{o,o}+\boldsymbol{\Sigma}_{o,u}\mathbf{D}_{u,u}\boldsymbol{\Sigma}_{u,o}$. Therefore, our functional-BLUP for untyped summary statistics is $\mathbf{z}_u^* = \mathbf{E}[\mathbf{z}_u|\mathbf{z}_o, \boldsymbol{\Sigma}, \mathbf{D}, \sigma_e^2, n]$.

Given our functionally-aware predictive model for summary statistics, we compute prediction accuracy using a generalized measure of R^2_{pred}. For reference, prediction accuracy for a model of fixed effects at the ith untyped marker is computed by $R^2_{\text{pred}}(i) \triangleq (\boldsymbol{\Sigma}_{u,o}\boldsymbol{\Sigma}_{o,o}^{-1}\boldsymbol{\Sigma}_{o,u})_{i,i}$ which will always be bounded between 0 and 1 when $\boldsymbol{\Sigma}$ is a full-rank correlation matrix[66]. If we naively replace $\boldsymbol{\Sigma}$ partitions with $\mathbf{V}$ partitions, we cannot guarantee estimates bounded between 0 and 1, due to tagged prior variance terms. To compute bounded prediction accuracy while accounting for variance due to random effects we propose

$$\begin{aligned} R^2_{\text{blup}}(i) &\triangleq 1 - \frac{\mathbf{V}[\mathbf{z}_u - \mathbf{z}_u^*]_{i,i}}{\mathbf{V}[\mathbf{z}_u]_{i,i}} && (4.11)\\ &= 1 - \frac{\mathbf{E}[\mathbf{V}[\mathbf{z}_u \mid \mathbf{z}_o]]_{i,i}}{\mathbf{V}[\mathbf{z}_u]_{i,i}} && (4.12)\\ &= 1 - \frac{(\mathbf{V}_{u,u} - \mathbf{V}_{u,o}\mathbf{V}_{o,o}^{-1}\mathbf{V}_{o,u})_{i,i}}{(\mathbf{V}_{u,u})_{i,i}} && (4.13)\\ &= \frac{(\mathbf{V}_{u,o}\mathbf{V}_{o,o}^{-1}\mathbf{V}_{o,u})_{i,i}}{(\mathbf{V}_{u,u})_{i,i}}, && (4.14) \end{aligned}$$

where we drop the conditioned parameters $\boldsymbol{\Sigma}, \mathbf{D}, \sigma_e^2, n$ to simplify notation. We see this definition also recovers R^2_{pred} in the limit of $\mathbf{D} \to \text{d}iag(\mathbf{0})$.

4.4.3 Fitting functional variance terms

Algorithm to fit model:

1. Fit variance terms using leave-one-chromosome-out functional LDSC regression (i.e. $\boldsymbol{m}\tau$).

2. Predict unobserved z-scores $\mathbf{z}_u$ at independent LD blocks using $\mathbf{z}_u^* = \mathbf{E}[\mathbf{z}_u|\mathbf{z}_o, \boldsymbol{\Sigma}, \mathbf{D}, \sigma_e^2, n]$ where $\mathbf{D}_{i,i} = n\boldsymbol{m}\tau_i$ obtained from step 1

4.4.4 Simulation Pipeline

We simulated summary statistics for SNPs under both an infinitesimal model and a non-infinitesimal model across 100 independent loci on chromosome 1. Z-scores were drawn according to the sampling distribution described in Methods (Equation 8), assuming a GWAS sample size of 50,000. Under the non-infinitesimal model, a single causal was drawn at each locus, informed by the available annotations, according to the logistic function described in [43]. The default baseline σ^2 was 0.0001 for infinitesimal simulations, and 0.1 for single-causal simulations. We used three annotations (promoter, DHS, intron), with enrichments of 2.81, 1.70, and 1.19 respectively, in accordance with [26]. Annotation-specific σ^2 values were obtained by multiplying the baseline σ^2 by the annotation's enrichment. In order to simulate missing statistics, we partition the generated Z-scores into observed and un-observed blocks by randomly sampling SNPs according to the desired proportion of SNPs kept. LD and the annotation matrix D are partitioned accordingly (Equation 9). Finally, Z-scores are inferred according to Equation 10. To assess accuracy across each condition, we measure the squared correlations between our predictions and the true statistics for missing SNPs, averaged across all 100 loci. We run FIMPG and IMPG on each locus with various proportions of retained

SNPs and 10 trials at each proportion, where a new set of retained SNPs is sampled at each trial.

For null simulations, we perform the above procedure where the covariance of the Z-scores is simulated directly from LD rather than from the matrix D as in Equation 8. We assess performance by averaging the resulting λ_{gc} over all trials.

For all simulations in which model parameters were varied, we fixed the proportion of retained SNPs to 50%. For simulations in which we vary the number of simulated annotations, we randomly select 1-3 of 3 annotations for each locus. When varying enrichment levels themselves, we kept the baseline variance fixed and scaled the original annotation-specific enrichments by 5 different multipliers (1.0, 3.0, 5.0, 7.0, 9.0). For experiments in which σ^2 was varied, we vary the baseline σ^2, which in turn affects σ^2 for each annotation according to its enrichment.

For simulations in which one annotation was omitted from the inference, 1 of 3 annotations was randomly omitted at each trial. When enrichment levels were misspecified, we simulate with the enrichments reported in [26], but draw enrichments during inference using the annotation-specific mean and standard deviation reported in [26].

4.4.5 Real Data

For real data analysis, we applied FIMPG and IMPG to the 27 UKBiobank traits, which include 337K European-ancestry individuals. We ran LD-score on 53 annotations in the baselineLD model [27], using 1000 Genomes Phase 3 SNPs. We filter out annotations with low enrichment significance ($|Z| < 1.96$), leaving on average 7 annotations per trait. Summary statistics were obtained for 133 independent regions on chromosome 1 [68], computed using BOLT-LMM [38]. At each locus, we randomly masked a 90% of SNPs and ran both FIMPG and IMPG on the region. Predictions were then compared with the true statistics.

4.5 Figures

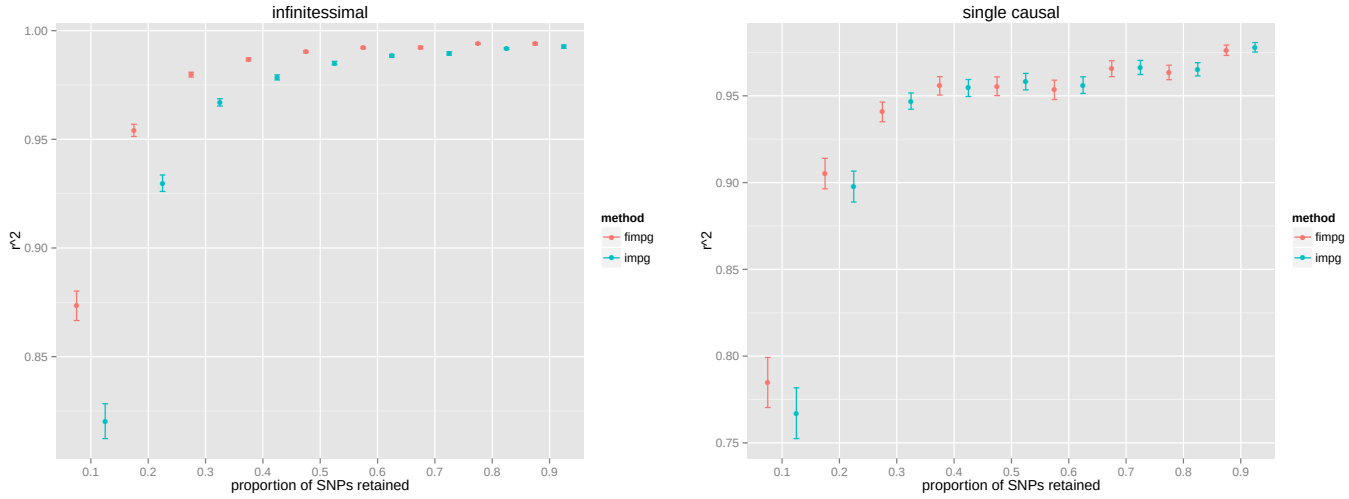


Figure 4.1: **Performance of FIMPG in simulations.** We included three annotations and ran FIMPG and IMPG across various proportions of retained SNPs and under both the (A) infinitesimal model and (B) single-causal model. The squared correlation between simulated and predicted Z-scores are averaged across 100 independent loci on chromosome 1, with 10 trials at each locus, where a different set of SNPs is retained at each trial.

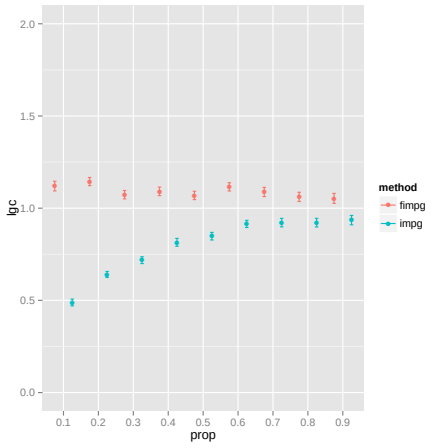


Figure 4.2: **FIMPG is slightly biased under the null.** Average λ_{gc} under null simulations where no SNPs are causal for FIMPG and IMPG, varied across the proportion of retained SNPs. FIMPG is slightly inflated under the null model, with a mean λ_{gc} of 1.19.

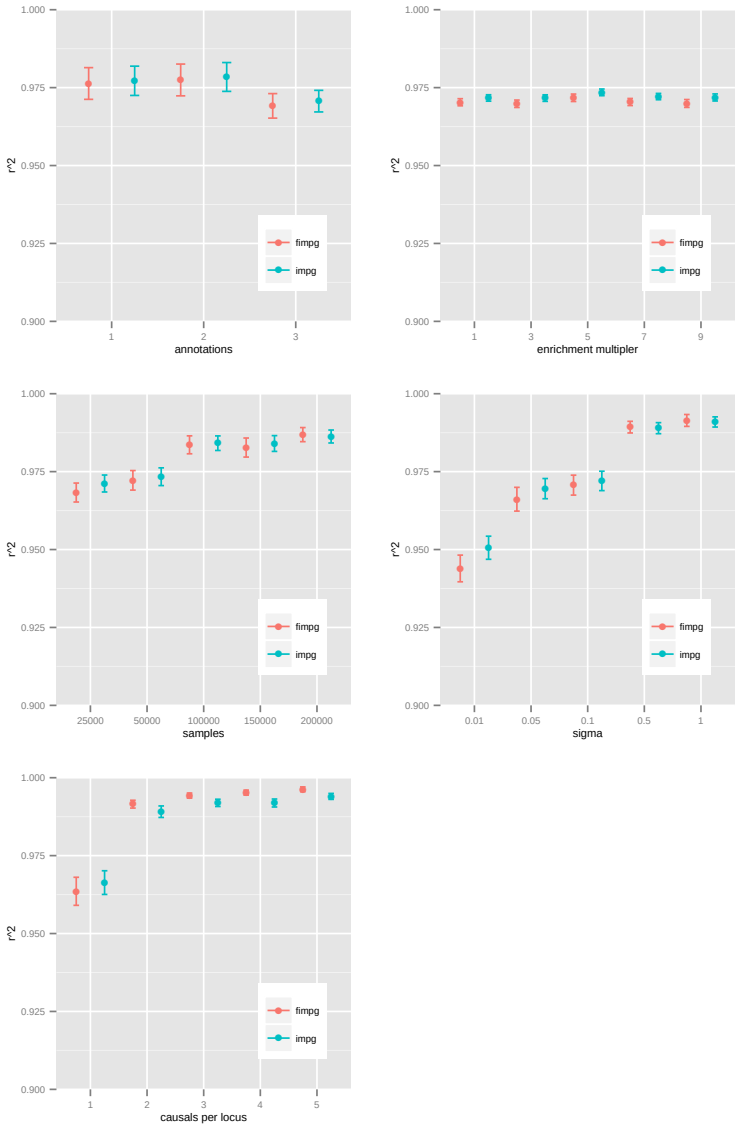


Figure 4.3: **Behavior of FIMPG as we vary simulation parameters.** Squared correlation for FIMPG vs IMPG, across a number of conditions, including (A) the number of annotations, (B) the enrichment multiplier, (C) sample size, (D) the simulated σ^2, (E) the number of causal variants per locus.

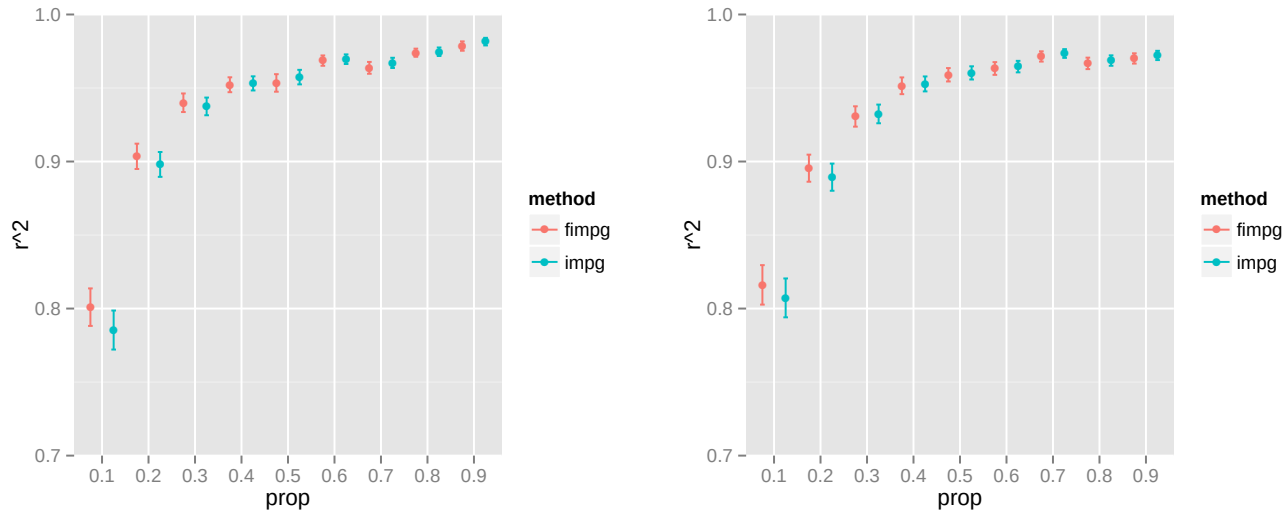


Figure 4.4: **Performance under model violations.** Squared correlation for FIMPG vs IMPG under (A) weak violation, where annotation enrichments are misspecified, and (B) strong violation, where one of the simulated annotations is randomly omitted from the inference step. Under both violations, averaged across all proportions of retained SNPs, FIMPG's performance does not fall below that of IMPG.

REFERENCES

[1] ' 1000 Genomes Project Consortium et al. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65, 2012.

[2] Frank W Albert and Leonid Kruglyak. The role of regulatory variation in complex traits and disease. *Nature Reviews Genetics*, 16(4):197–212, 2015.

[3] William J Astle, Heather Elding, Tao Jiang, Dave Allen, Dace Ruklisa, Alice L Mann, Daniel Mead, Heleen Bouman, Fernando Riveros-Mckay, Myrto A Kostadima, et al. The allelic landscape of human blood cell trait variation and links to common complex disease. *Cell*, 167(5):1415–1429, 2016.

[4] Nicholas E Banovich, Xun Lan, Graham McVicker, Bryce Van de Geijn, Jacob F Degner, John D Blischak, Julien Roux, Jonathan K Pritchard, and Yoav Gilad. Methylation qtls are associated with coordinated changes in transcription factor binding, histone modifications, and gene expression levels. *PLoS Genet*, 10(9):e1004663, 2014.

[5] Christian Benner, Chris CA Spencer, Samuli Ripatti, and Matti Pirinen. Finemap: Efficient variable selection using summary data from genome-wide association studies. *bioRxiv*, page 027342, 2015.

[6] James Bentham, David L Morris, Deborah S Cunninghame Graham, Christopher L Pinder, Philip Tombleson, Timothy W Behrens, Javier Martín, Benjamin P Fairfax, Julian C Knight, Lingyan Chen, et al. Genetic association analyses implicate aberrant regulation of innate and adaptive immunity genes in the pathogenesis of systemic lupus erythematosus. *Nature genetics*, 2015.

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

[8] Brendan Bulik-Sullivan, Hilary K Finucane, Verneri Anttila, Alexander Gusev, Felix R Day, Po-Ru Loh, Laramie Duncan, John RB Perry, Nick Patterson, Elise B Robinson, et al. An atlas of genetic correlations across human diseases and traits. *Nature genetics*, 47(11):1236, 2015.

[9] Brendan Bulik-Sullivan, Hilary K. Finucane, Verneri Anttila, Alexander Gusev, Felix R. Day, Po-Ru Loh, Consortium ReproGen, Consortium Psychiatric Genomics, Consortium Genetic Consortium for Anorexia Nervosa of the Wellcome Trust Case Control, Laramie Duncan, John R. B. Perry, Nick Patterson, Elise B. Robinson, Mark J. Daly, Alkes L. Price, and Benjamin M. Neale. An atlas of genetic correlations across human diseases and traits. *Nat Genet*, 47(11):1236–1241, 2015.

[10] Adolf Buse. The likelihood ratio, wald, and lagrange multiplier tests: An expository note. *The American Statistician*, 36(3a):153–157, 1982.

[11] Lu Chen, Bing Ge, Francesco Paolo Casale, Louella Vasquez, Tony Kwan, Diego Garrido-Martín, Stephen Watt, Ying Yan, Kousik Kundu, Simone Ecker, et al. Genetic drivers of epigenetic and transcriptional variation in human immune cells. *Cell*, 167(5):1398–1414, 2016.

[12] Wenan Chen, Beth R Larrabee, Inna G Ovsyannikova, Richard B Kennedy, Iana H Haralambieva, Gregory A Poland, and Daniel J Schaid. Fine mapping causal variants with an approximate bayesian method using marginal test statistics. *Genetics*, pages genetics–115, 2015.

[13] Dongjun Chung, Can Yang, Cong Li, Joel Gelernter, and Hongyu Zhao. Gpa: a statistical approach to prioritizing gwas results by integrating pleiotropy and annotation. *PLoS genetics*, 2014.

[14] Melina Claussnitzer, Simon N Dankel, Kyoung-Han Kim, Gerald Quon, Wouter Meuleman, Christine Haugen, Viktoria Glunk, Isabel S Sousa, Jacqueline L Beaudry, Vijitha Puviindran, et al. Fto obesity variant circuitry and adipocyte browning in humans. *New England Journal of Medicine*, 373(10):895–907, 2015.

[15] International Multiple Sclerosis Genetics Consortium, Wellcome Trust Case Control Consortium 2, et al. Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. *Nature*, 476(7359):214–219, 2011.

[16] Heather J Cordell, Younghun Han, George F Mells, Yafang Li, Gideon M Hirschfield, Casey S Greene, Gang Xie, Brian D Juran, Dakai Zhu, David C Qian, et al. International genome-wide meta-analysis identifies new primary biliary cirrhosis risk loci and targetable pathogenic pathways. *Nature communications*, 6, 2015.

[17] Gregory Darnell, Dat Duong, Buhm Han, and Eleazar Eskin. Incorporating prior information into association studies. *Bioinformatics*, 28(12):i147–i153, 2012.

[18] Jacob F Degner, Athma A Pai, Roger Pique-Regi, Jean-Baptiste Veyrieras, Daniel J Gaffney, Joseph K Pickrell, Sherryl De Leon, Katelyn Michelini, Noah Lewellen, Gregory E Crawford, et al. Dnase [thinsp] i sensitivity qtls are a major determinant of human expression variation. *Nature*, 482(7385):390–394, 2012.

[19] Patrick CA Dubois, Gosia Trynka, Lude Franke, Karen A Hunt, Jihane Romanos, Alessandra Curtotti, Alexandra Zhernakova, Graham AR Heap, Róza Ádány, Arpo Aromaa, et al. Multiple common variants for celiac disease influencing immune gene expression. *Nature genetics*, 42(4):295–302, 2010.

[20] ' ENCODE Project Consortium et al. An integrated encyclopedia of dna elements in the human genome. *Nature*, 489(7414):57–74, 2012.

[21] Jason Ernst and Manolis Kellis. Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. *Nature biotechnology*, 33(4):364–376, 2015.

[22] Jason Ernst, Pouya Kheradpour, Tarjei S Mikkelsen, Noam Shoresh, Lucas D Ward, Charles B Epstein, Xiaolan Zhang, Li Wang, Robbyn Issner, Michael Coyne, et al. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, 473(7345):43–49, 2011.

[23] Eleazar Eskin. Increasing power in association studies by using linkage disequilibrium structure and molecular function as prior information. *Genome research*, 18(4):653–660, 2008.

[24] Evangelos Evangelou and John PA Ioannidis. Meta-analysis methods for genome-wide association studies and beyond. *Nature Reviews Genetics*, 14(6):379–389, 2013.

[25] Hilary K Finucane, Brendan Bulik-Sullivan, Alexander Gusev, Gosia Trynka, Yakir Reshef, Po-Ru Loh, Verneri Anttila, Han Xu, Chongzhi Zang, Kyle Farh, et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nature genetics*, 47(11):1228–1235, 2015.

[26] Hilary K. Finucane, Brendan Bulik-Sullivan, Alexander Gusev, Gosia Trynka, Yakir Reshef, Po-Ru Loh, Verneri Anttila, Han Xu, Chongzhi Zang, Kyle Farh, Stephan Ripke, Felix R. Day, Consortium ReproGen, Consortium Schizophrenia Working Group of the Psychiatric Genomics, Raci Consortium The, Shaun Purcell, Eli Stahl, Sara Lindstrom, John R. B. Perry, Yukinori Okada, Soumya Raychaudhuri, Mark J. Daly, Nick Patterson, Benjamin M. Neale, and Alkes L. Price. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat Genet*, 47(11):1228–1235, 2015.

[27] Steven Gazal, Hilary K Finucane, Nicholas A Furlotte, Po-Ru Loh, Pier Francesco Palamara, Xuanyao Liu, Armin Schoech, Brendan Bulik-Sullivan, Benjamin M Neale, Alexander Gusev, et al. Linkage disequilibrium–dependent architecture of human complex traits shows action of negative selection. *Nature genetics*, 49(10):1421, 2017.

[28] Claudia Giambartolomei, Damjan Vukcevic, Eric E Schadt, Lude Franke, Aroon D Hingorani, Chris Wallace, and Vincent Plagnol. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet*, 10(5):e1004383, 2014.

[29] ' Global Lipids Genetics Consortium et al. Discovery and refinement of loci associated with lipid levels. *Nature genetics*, 45(11):1274–1283, 2013.

[30] Fabian Grubert, Judith B Zaugg, Maya Kasowski, Oana Ursu, Damek V Spacek, Alicia R Martin, Peyton Greenside, Rohith Srivas, Doug H Phanstiel, Aleksandra Pekowska, et al. Genetic control of chromatin states in humans involves local and distal chromosomal interactions. *Cell*, 162(5):1051–1065, 2015.

[31] Alexander Gusev, S Hong Lee, Gosia Trynka, Hilary Finucane, Bjarni J Vilhjálmsson, Han Xu, Chongzhi Zang, Stephan Ripke, Brendan Bulik-Sullivan, Eli Stahl, et al. Par-

titioning heritability of regulatory and cell-type-specific variants across 11 common diseases. *The American Journal of Human Genetics*, 95(5):535–552, 2014.

[32] Alexander Gusev, S Hong Lee, Gosia Trynka, Hilary Finucane, Bjarni J Vilhjálmsson, Han Xu, Chongzhi Zang, Stephan Ripke, Brendan Bulik-Sullivan, Eli Stahl, et al. Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. *The American Journal of Human Genetics*, 95(5):535–552, 2014.

[33] Alexander Gusev, Huwenbo Shi, Gleb Kichaev, Mark Pomerantz, Fugen Li, Henry W Long, Sue A Ingles, Rick A Kittles, Sara S Strom, Benjamin A Rybicki, et al. Atlas of prostate cancer heritability in european and african-american men pinpoints tissue-specific regulation. *Nature Communications*, 7, 2016.

[34] Farhad Hormozdiari, Emrah Kostem, Eun Yong Kang, Bogdan Pasaniuc, and Eleazar Eskin. Identifying causal variants at loci with multiple signals of association. *Genetics*, 198(2):497–508, 2014.

[35] Farhad Hormozdiari, Martijn van de Bunt, Ayellet V Segre, Xiao Li, Jong Wha J Joo, Michael Bilow, Jae Hoon Sul, Sriram Sankararaman, Bogdan Pasaniuc, and Eleazar Eskin. Colocalization of gwas and eqtl signals detects target genes. *The American Journal of Human Genetics*, 99(6):1245–1260, 2016.

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

[37] Maya Kasowski, Sofia Kyriazopoulou-Panagiotopoulou, Fabian Grubert, Judith B Zaugg, Anshul Kundaje, Yuling Liu, Alan P Boyle, Qiangfeng Cliff Zhang, Fouad Zakharia, Damek V Spacek, et al. Extensive variation in chromatin states across humans. *Science*, 342(6159):750–752, 2013.

[38] Gleb Kichaev, Gaurav Bhatia, Po-Ru Loh, Steven Gazal, Kathryn Burch, Malika Freund, Armin Scoech, Bogdan Pasaniuc, and Alkes Price. Leveraging polygenic functional enrichment to improve gwas power. *bioRxiv*, page 222265, 2017.

[39] Gleb Kichaev and Bogdan Pasaniuc. Leveraging functional-annotation data in trans-ethnic fine-mapping studies. *The American Journal of Human Genetics*, 97(2):260–271, 2015.

[40] Gleb Kichaev, Megan Roytman, Ruth Johnson, Eleazar Eskin, Sara Lindstroem, Peter Kraft, and Bogdan Pasaniuc. Improved methods for multi-trait fine mapping of pleiotropic risk loci. *Bioinformatics*, page btw615, 2016.

[41] Gleb Kichaev, Megan Roytman, Ruth Johnson, Eleazar Eskin, Sara Lindstrm, Peter Kraft, and Bogdan Pasaniuc. Improved methods for multi-trait fine mapping of pleiotropic risk loci. *Bioinformatics*, 33(2):248–255, 2017.

[42] Gleb Kichaev, Wen-Yun Yang, Sara Lindstrom, Farhad Hormozdiari, Eleazar Eskin, Alkes L Price, Peter Kraft, and Bogdan Pasaniuc. Integrating functional data to prioritize causal variants in statistical fine-mapping studies. *PLoS genetics*, 10(10):e1004722, 2014.

[43] Gleb Kichaev, Wen-Yun Yang, Sara Lindstrom, Farhad Hormozdiari, Eleazar Eskin, Alkes L Price, Peter Kraft, and Bogdan Pasaniuc. Integrating functional data to prioritize causal variants in statistical fine-mapping studies. *PLoS genetics*, 10(10):e1004722, 2014.

[44] Helena Kilpinen, Sebastian M Waszak, Andreas R Gschwind, Sunil K Raghav, Robert M Witwicki, Andrea Orioli, Eugenia Migliavacca, Michaël Wiederkehr, Maria Gutierrez-Arcelus, Nikolaos I Panousis, et al. Coordinated effects of sequence variation on dna binding, chromatin structure, and transcription. *Science*, 342(6159):744–747, 2013.

[45] Zsofia Kote-Jarai, Edward J Saunders, Daniel A Leongamornlert, Malgorzata Tymrakiewicz, Tokhir Dadaev, Sarah Jugurnauth-Little, Helen Ross-Adams, Ali Amin Al Olama, Sara Benlloch, Silvia Halim, et al. Fine-mapping identifies multiple prostate cancer risk loci at 5p15, one of which associates with tert expression. *Human molecular genetics*, 22(12):2520–2528, 2013.

[46] Anshul Kundaje, Wouter Meuleman, Jason Ernst, Misha Bilenky, Angela Yen, Alireza Heravi-Moussavi, Pouya Kheradpour, Zhizhuo Zhang, Jianrong Wang, Michael J Ziller, et al. Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317–330, 2015.

[47] Anshul Kundaje, Wouter Meuleman, Jason Ernst, Misha Bilenky, Angela Yen, Alireza Heravi-Moussavi, Pouya Kheradpour, Zhizhuo Zhang, Jianrong Wang, Michael J Ziller, et al. Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317, 2015.

[48] Tuuli Lappalainen, Michael Sammeth, Marc R Friedländer, Peter AC?t Hoen, Jean Monlong, Manuel A Rivas, Mar Gonzalez-Porta, Natalja Kurbatova, Thasso Griebel, Pedro G Ferreira, et al. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, 501(7468):506, 2013.

[49] Donghyung Lee, T Bernard Bigdeli, Brien P Riley, Ayman H Fanous, and Silviu-Alin Bacanu. Dist: direct imputation of summary statistics for unmeasured snps. *Bioinformatics*, 29(22):2925–2927, 2013.

[50] Yang I Li, Bryce van de Geijn, Anil Raj, David A Knowles, Allegra A Petti, David Golan, Yoav Gilad, and Jonathan K Pritchard. Rna splicing is a primary link between genetic variation and disease. *Science*, 352(6285):600–604, 2016.

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

[52] Sara Lindström, Deborah J Thompson, Andrew D Paterson, Jingmei Li, Gretchen L Gierach, Christopher Scott, Jennifer Stone, Julie A Douglas, Isabel dos Santos-Silva, Pablo Fernandez-Navarro, et al. Genome-wide association study identifies multiple loci associated with both mammographic density and breast cancer risk. *Nature communications*, 5, 2014.

[53] Jimmy Z Liu, Suzanne van Sommeren, Hailiang Huang, Siew C Ng, Rudi Alberts, Atsushi Takahashi, Stephan Ripke, James C Lee, Luke Jostins, Tejas Shah, et al. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nature genetics*, 47(9):979–986, 2015.

[54] Adam E Locke, Bratati Kahali, Sonja I Berndt, Anne E Justice, Tune H Pers, Felix R Day, Corey Powell, Sailaja Vedantam, Martin L Buchkovich, Jian Yang, et al. Genetic studies of body mass index yield new insights for obesity biology. *Nature*, 518(7538):197–206, 2015.

[55] Qiongshi Lu, Ryan Lee Powles, Qian Wang, Beixin Julie He, and Hongyu Zhao. Integrative tissue-specific functional annotations in the human genome provide novel insights on many complex traits and improve signal prioritization in genome wide association studies. *PLoS genetics*, 12(4):e1005947, 2016.

[56] Qiongshi Lu, Xinwei Yao, Yiming Hu, and Hongyu Zhao. Genowap: Gwas signal prioritization through integrated analysis of genomic functional annotation. *Bioinformatics*, 32(4):542–548, 2015.

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

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

[59] TA Manolio, FS Collins, NJ Cox, and DB Goldstein. Finding the missing heritability of complex diseases. *Nature*, 2009.

[60] Matthew T Maurano, Richard Humbert, Eric Rynes, Robert E Thurman, Eric Haugen, Hao Wang, Alex P Reynolds, Richard Sandstrom, Hongzhu Qu, Jennifer Brody, et al. Systematic localization of common disease-associated variation in regulatory dna. *Science*, 337(6099):1190–1195, 2012.

[61] Matthew T. Maurano, Richard Humbert, Eric Rynes, Robert E. Thurman, Eric Haugen, Hao Wang, Alex P. Reynolds, Richard Sandstrom, Hongzhu Qu, Jennifer Brody, Anthony Shafer, Fidencio Neri, Kristen Lee, Tanya Kutyavin, Sandra Stehling-Sun, Audra K. Johnson, Theresa K. Canfield, Erika Giste, Morgan Diegel, Daniel Bates,

R. Scott Hansen, Shane Neph, Peter J. Sabo, Shelly Heimfeld, Antony Raubitschek, Steven Ziegler, Chris Cotsapas, Nona Sotoodehnia, Ian Glass, Shamil R. Sunyaev, Rajinder Kaul, and John A. Stamatoyannopoulos. Systematic localization of common disease-associated variation in regulatory dna. *Science*, 337(6099):1190–1195, 2012.

[62] Graham McVicker, Bryce van de Geijn, Jacob F Degner, Carolyn E Cain, Nicholas E Banovich, Anil Raj, Noah Lewellen, Marsha Myrthil, Yoav Gilad, and Jonathan K Pritchard. Identification of genetic variants that affect histone modifications in human cells. *Science*, 342(6159):747–749, 2013.

[63] Kerstin B Meyer, Martin OReilly, Kyriaki Michailidou, Saskia Carlebur, Stacey L Edwards, Juliet D French, Radhika Prathalingham, Joe Dennis, Manjeet K Bolla, Qin Wang, et al. Fine-scale mapping of the fgfr2 breast cancer risk locus: putative functional variants differentially bind foxa1 and e2f1. *The American Journal of Human Genetics*, 93(6):1046–1060, 2013.

[64] Kiran Musunuru, Alanna Strong, Maria Frank-Kamenetsky, Noemi E Lee, Tim Ahfeldt, Katherine V Sachs, Xiaoyu Li, Hui Li, Nicolas Kuperwasser, Vera M Ruda, et al. From noncoding variant to phenotype via sort1 at the 1p13 cholesterol locus. *Nature*, 466(7307):714–719, 2010.

[65] Yukinori Okada, Di Wu, Gosia Trynka, Towfique Raj, Chikashi Terao, Katsunori Ikari, Yuta Kochi, Koichiro Ohmura, Akari Suzuki, Shinji Yoshida, et al. Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature*, 506(7488):376–381, 2014.

[66] Bogdan Pasaniuc, Noah Zaitlen, Huwenbo Shi, Gaurav Bhatia, Alexander Gusev, Joseph Pickrell, Joel Hirschhorn, David P Strachan, Nick Patterson, and Alkes L Price. Fast and accurate imputation of summary statistics enhances evidence of functional enrichment. *Bioinformatics*, page btu416, 2014.

[67] Joseph K Pickrell. Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *The American Journal of Human Genetics*, 94(4):559–573, 2014.

[68] Joseph K. Pickrell, Tomaz Berisa, Jimmy Z. Liu, Laure Segurel, Joyce Y. Tung, and David A. Hinds. Detection and interpretation of shared genetic influences on 42 human traits. *Nat Genet*, advance online publication, 2016.

[69] George K Robinson. That blup is a good thing: the estimation of random effects. *Statistical science*, pages 15–32, 1991.

[70] Kathryn Roeder, Bernie Devlin, and Larry Wasserman. Improving power in genome-wide association studies: weights tip the scale. *Genetic epidemiology*, 31(7):741–747, 2007.

[71] Megan Roytman, Gleb Kichaev, Alexander Gusev, and Bogdan Pasaniuc. Methods for

fine-mapping with chromatin and expression data. *PLoS genetics*, 14(2):e1007240, 2018.

[72] Andrew J Schork, Wesley K Thompson, Phillip Pham, Ali Torkamani, J Cooper Roddey, Patrick F Sullivan, John R Kelsoe, Michael C O'Donovan, Helena Furberg, Nicholas J Schork, et al. All snps are not created equal: genome-wide association studies reveal a consistent pattern of enrichment among functionally annotated snps. *PLoS genetics*, 9(4):e1003449, 2013.

[73] Marco Scutari. Learning bayesian networks with the bnlearn r package. *arXiv preprint arXiv:0908.3817*, 2009.

[74] Nadia Solovieff, Chris Cotsapas, Phil H Lee, Shaun M Purcell, and Jordan W Smoller. Pleiotropy in complex traits: challenges and strategies. *Nature Reviews Genetics*, 14(7):483–495, 2013.

[75] Oliver Stegle, Leopold Parts, Matias Piipari, John Winn, and Richard Durbin. Using probabilistic estimation of expression residuals (peer) to obtain increased power and interpretability of gene expression analyses. *Nature protocols*, 7(3):500–507, 2012.

[76] John D Storey and Robert Tibshirani. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, 100(16):9440–9445, 2003.

[77] Zhan Su, Jonathan Marchini, and Peter Donnelly. Hapgen2: simulation of multiple disease snps. *Bioinformatics*, 2011.

[78] Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, et al. Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine*, 12(3):e1001779, 2015.

[79] Gardar Sveinbjornsson, Anders Albrechtsen, Florian Zink, Sigurjón A Gudjonsson, Asmundur Oddson, Gísli Másson, Hilma Holm, Augustine Kong, Unnur Thorsteinsdottir, Patrick Sulem, et al. Weighting sequence variants based on their annotation increases power of whole-genome association studies. *Nature genetics*, 48(3):314, 2016.

[80] Ryan Tewhey, Dylan Kotliar, Daniel S Park, Brandon Liu, Sarah Winnicki, Steven K Reilly, Kristian G Andersen, Tarjei S Mikkelsen, Eric S Lander, Stephen F Schaffner, et al. Direct identification of hundreds of expression-modulating variants using a multiplexed reporter assay. *Cell*, 165(6):1519–1529, 2016.

[81] Asian Genetic Epidemiology Network Type, South Asian Type, Diabetes SAT2D Consortium, Mexican American Type, Diabetes MAT2D Consortium, Anubha Mahajan, Min Jin Go, Weihua Zhang, Jennifer E Below, Kyle J Gaulton, et al. Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. *Nature genetics*, 46(3):234–244, 2014.

[82] Peter M Visscher, Matthew A Brown, Mark I McCarthy, and Jian Yang. Five years of

gwas discovery. *The American Journal of Human Genetics*, 90(1):7–24, 2012.

[83] Peter M Visscher, Matthew A Brown, Mark I McCarthy, and Jian Yang. Five years of gwas discovery. *The American Journal of Human Genetics*, 90(1):7–24, 2012.

[84] Lucas D Ward and Manolis Kellis. Haploreg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic acids research*, 40(D1):D930–D934, 2012.

[85] Sebastian M Waszak, Olivier Delaneau, Andreas R Gschwind, Helena Kilpinen, Sunil K Raghav, Robert M Witwicki, Andrea Orioli, Michael Wiederkehr, Nikolaos I Panousis, Alisa Yurovsky, et al. Population variation and genetic control of modular chromatin architecture in humans. *Cell*, 162(5):1039–1050, 2015.

[86] Andrew R Wood, Tonu Esko, Jian Yang, Sailaja Vedantam, Tune H Pers, Stefan Gustafsson, Audrey Y Chu, Karol Estrada, Jian'an Luan, Zoltán Kutalik, et al. Defining the role of common variation in the genomic and biological architecture of adult human height. *Nature genetics*, 46(11):1173–1186, 2014.

[87] Ying Wu, Lindsay L Waite, Anne U Jackson, Wayne HH Sheu, Steven Buyske, Devin Absher, Donna K Arnett, Eric Boerwinkle, Lori L Bonnycastle, Cara L Carty, et al. Trans-ethnic fine-mapping of lipid loci identifies population-specific signals and allelic heterogeneity that increases the trait variance explained. *PLoS genetics*, 9(3):e1003379, 2013.

[88] Jian Yang, Teri A Manolio, Louis R Pasquale, Eric Boerwinkle, Neil Caporaso, Julie M Cunningham, Mariza de Andrade, Bjarke Feenstra, Eleanor Feingold, M Geoffrey Hayes, et al. Genome partitioning of genetic variation for complex traits using common snps. *Nature genetics*, 43(6):519–525, 2011.

[89] Jingjing Yang, Lars G Fritsche, Xiang Zhou, Goncalo Abecasis, International Age-Related Macular Degeneration Genomics Consortium, et al. A scalable bayesian method for integrating functional information in genome-wide association studies. *The American Journal of Human Genetics*, 101(3):404–416, 2017.

[90] Vicky W Zhou, Alon Goren, and Bradley E Bernstein. Charting histone modifications and the functional organization of mammalian genomes. *Nature Reviews Genetics*, 12(1):7–18, 2011.