STATISTICAL METHODS FOR GENETIC VARIANTS DETECTION WITH EPIGENOMIC INFORMATION

by

Constanza Rojo

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

(Statistics)

at the

UNIVERSITY OF WISCONSIN–MADISON

2019

Date of final oral examination: December 11, 2019

The dissertation is approved by the following members of the Final Oral Committee:
Sündüz Keleş, Professor, Statistics, Biostatistics and Medical Informatics
Alan D. Attie, Professor, Biochemistry
Karl Broman, Professor, Biostatistics and Medical Informatics
Michael Newton, Professor, Statistics, Biostatistics and Medical Informatics
Garvesh Raskutti, Assistant Professor, Statistics

ACKNOWLEDGMENTS

I would like to thank my advisor Dr. Sündüz Keleş. Sündüz is the kind of mentor that pushes you towards your best potential, which I greatly appreciate and admire. Her excitement about research is contagious, so much so that I even learned to love genetics, a topic that never interested me before. When other students ask me about her, I usually respond: *she's fair but tough*. It is my best attempt to summarize her in a few words but she's so much more than that. She has given me constant support, guidance, and even lifted my spirits when I was going through tough times.

I would also like to thank Dr. Qi Zhang and Dr. Pixu Shi, my main collaborators in the projects I have developed. They provided constant input and have been a vital part of my work. I am also grateful for the previous and current members of the Keleş Research Team for constant feedback. In addition, I want to express my gratitude to the members of the committee.

Lastly, I would like to thank my friends, family, and husband Jeff. They have supported me along the way through this (sometimes) not so smooth path.

This dissertation was partially supported by the Chilean National Commission for Scientific and Technological Research (CONICYT) Doctoral Fellowship program.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Genome-wide association studies (GWAS) have successfully identified thousands of genetic variants contributing to disease and other phenotypes. However, significant obstacles hamper our ability to elucidate causal variants, identify genes affected by causal variants, and characterize the mechanisms by which genotypes influence phenotypes. The increasing availability of genome-wide functional annotation data provides unique opportunities to incorporate prior information into the analysis of GWAS to better understand the impact of variants on disease etiology. Regulatory genomic information has been recognized as a potential source that can improve the detection and biological interpretation of single-nucleotide polymorphisms (SNPs) in GWAS.

Although there have been many advances in incorporating prior information into the prioritization of trait-associated variants in GWAS, functional annotation data has played a secondary role in the joint analysis of GWAS and molecular (i.e., expression) quantitative trait loci (eQTL) data in assessing evidence for association. Moreover, current methodologies that aim to integrate such annotation information focus mainly on fine-mapping and overlook the importance of its usage in earlier stages of GWAS analysis. Equally important, there is a lack of development in proper statistical frameworks that can perform selection of annotations and SNPs jointly.

To address these shortcomings, we develop two statistical models: *iFunMed* and GRAD. *iFunMed* is a novel mediation framework to integrate GWAS and eQTL data with the utilization of publicly available functional annotation data. *iFunMed* extends the scope of standard mediation analysis by incorporating information from multiple genetic variants at a time and leveraging variant-level summary statistics. GRAD integrates high-dimensional auxiliary information into high-dimensional regression. This method allows annotation information to assist the detection of important genetic variants while identifying relevant annotation simultaneously. We provide an upper bound for the estimation error of the SNP effect sizes to gain insights on what factors affect estimation accuracy.

For *iFunMed*, data-driven computational experiments convey how informative annotations improve SNP selection performance while emphasizing the robustness of the model to non-informative annotations. Applications to the Framingham Heart Study data indicates that *iFunMed* is able to boost the detection of SNPs with mediation effects that can be attributed to regulatory mechanisms.

Simulation experiments indicate that GRAD can improve the identification of genetic variants by increasing the average area under the precision-recall curve by up to 60%. Real data applications to the Framingham Heart Study show that GRAD can select relevant genetic variants while detecting several transcription factors involved in specific phenotypical changes.

1 INTRODUCTION

1.1 Genome-wide Association Studies

Genome-wide association studies (GWAS) aim to identify genetic variants associated with different phenotypical changes. They examine hundreds of thousands to millions of markers across the genome to find single nucleotide polymorphisms (SNPs) that are observed with higher (or lower) frequencies in subjects with a specific trait of interest. GWAS has successfully identified SNPs associated with complex diseases such as coronary diseases (Nikpay et al., 2015) and type 2 diabetes (Prasad and Groop, 2015), among many others. This is a vital tool for researches in the clinical field to characterize diseases and develop strategies for detection, treatment, and prevention.

A traditional GWAS analysis workflow has the following steps:

1. *Quality control and imputation:* Genotype data is first filtered based on (most commonly) three criteria. These are call rate, minor allele frequency (MAF) and violation of Hardy-Weinberg equilibrium (HWE). Call rate filters out SNPs that have an elevated proportion of missing data among patients, MAF removes rare SNPs, and deviations of HWE can potentially detect miscalled variants. This is followed by imputation of partially missing genotypes which is most successfully performed by using a reference panel to gain knowledge of the genetic structure of the variants (Marchini and Howie, 2010; Li et al., 2010). Genotypes are then recoded to an additive format (0/1/2).

2. *Association analysis:* The phenotype of interest is regressed on each SNP_i ($i = 1, \ldots, p$) individually with or without adjustment for other traits that could be relevant (e.g. age and sex). This yields p regression models that are summarized and inspected visually with Manhattan plots ($-\log_{10}$ transformed p-values for each SNP across their genomic coordinates).

3. *Selection of SNPs:* Significance at a high level is required. A threshold of 1×10^{-8} is widely used to account for the multiple testing problem. SNPs with p-values below the margin are selected and partitioned into independent regions for further inspection.

4. *Fine-mapping:* Each region with promising associations is analyzed and softwares like (Pruim et al., 2010) are used for visualization. The goal is to perfect the localization of causal variants with the use of statistical tools and/or functional methods for follow-up functional studies.

1.2 Understanding GWAS Hits

The associations that are characterized by GWAS have lead to novel discoveries (Hirschhorn et al., 2009), including pathways that were previously unsuspected (Lettre and Rioux, 2008) but many times, interpretation of SNPs that are found to be associated remains unknown. More importantly, by using strict thresholds on marginal associations we might miss candidates with moderate signals that have strong joint effects.

Current efforts by large consortia projects (Consortium, 2012; Roadmap Epigenomics Consortium, 2015) to identify functional elements are helping us to better understand the human genome. We refer to this as *functional annotation data*. The information that can be used as annotation can range from open chromatin regions, protein binding of specific proteins, or even accessibility changes among SNP alleles, and they can originate from different tissues and cell types. Findings from (Maurano et al., 2012) revealed that common disease-associated SNPs are enriched in functional DNA. This motivated scientists in the field to utilize and incorporate annotation as auxiliary information into GWAS analyses to improve SNP selection and prioritization.

To have a comprehensive understanding of the biology underlying different traits, (Nicolae et al., 2010) motivated the idea of using gene expression information starting from the basis that associated SNPs are more likely to be eQTLs. These observations suggest that GWAS SNPs modify expression levels of genes nearby which ultimately results in phenotypical changes. We can think about this biological pathway as a mediation process where gene expression acts as a mediator variable between genotype and phenotype associations.

Furthermore, nowadays, scientists are encouraged to share data that summarizes their findings. A common technique is to make results from marginal associations publicly available such as SNP effect sizes, standard errors, t-scores, or a combination of them. This is practical because of, mainly, two reasons: it doesn't yield on violations of patient privacy and summary-level data requires modest computation capacity. This data-sharing process allows us to gather information from large

cohorts in a short amount of time.

1.3 Statistical Methods for GWAS Analysis

Statistical methodologies that aim to integrate different sources of information to elucidate functionality and help to prioritize genetic variants are a focus of interest within GWAS analyses. They have the intent to answer some of the questions and challenges portrait by GWAS. We will review some of the main methods in literature and they are summarized in Figure 1.1.

Figure 1.1A displays methods that use mediation models, i.e. they provide an alternative pathway of association by assuming that phenotypical changes are observed because of changes in expression levels. The main focus of these methods in literature is to characterize gene-trait associations. TWAS (Gusev et al., 2016) provides flexibility regarding the input data, which can be raw-level or summary statistics. In both cases, it produces prediction-like gene expression that is later on tested for associations with the trait. SMR (Zhu et al., 2016) takes advantage of the Mendelian randomization framework and inputs summary data from GWAS and eQTL studies to identify associations between gene expression and traits of interest. Finally, both PrediXcan (Gamazon et al., 2015) and S-PrediXcan (Barbeira et al., 2018) compute gene-level association. PrediXcan predicts/imputes transcriptome levels from genotype and phenotype data, while S-PrediXcan does it directly from GWAS outputs.

Mediation models that have the goal of gaining information regarding direct

or mediated effects of individual SNPs are much less common. Some approaches do single-SNP mediation where they fit as many models as SNPs in the locus of interest. This technique can be useful as a form of exploratory analysis or to gain insights regarding the mediation potential of a specific gene but lack proper multiple comparison error control and do not account for joint SNP effects. A step-up from such procedures is iGWAS (Huang et al., 2015). iGWAS uses a multivariate mediation model with raw-level data and incorporates a family-based design. Yet, there are still no methodologies, to our knowledge, that incorporate epigenomic information into this models to improve SNP detection. *iFunMed* (Rojo et al., 2019) fills this hole by performing a mediation analysis with only the use of GWAS and eQTL summary statistics (t-scores) that integrates annotation information to inform non-zero status of SNP effect sizes and reports direct and mediated posterior probabilities of being non-zero. This allows the user to prioritize and select SNPs with high posterior probabilities. We accompany our analysis with an annotation selection pipeline based on enrichment values to avoid the burden of fitting hundreds to thousands of annotations that are generated nowadays, and to gain potential functionality of mediated and direct effects.

Another cluster of methods leverages annotation information into GWAS to improve SNP detection. These are illustrated in Figure 1.1B. They commonly share the way they integrate annotation into their models by using it as data-driven prior information, similarly as *iFunMed*. A portion of them assumes independence and only takes as input GWAS summaries that provides computation efficiency. One of the first ones to emerge was fgwas (Pickrell, 2014) that adopts an empirical Bayes

approach. It computes the probability of a specific block in the genome to have a non-zero SNP and within each region that harbors a non-zero SNP, it computes posterior probabilities for all of the SNPs in the locus. To do this computation, it assumes one causal variant per locus and uses a forward-backward technique to select annotations, which can be discrete (e.g. overlap vs no overlap with a TF binding region) or continuous (e.g. distance to TSS). GPA (Chung et al., 2014) takes as input marginal associations p-values and models them as null (from a uniform distribution) and non-null (from a beta distribution). It aims to discover and prioritize non-null SNPs. GPA only takes binary annotations and does not provide high-dimensional annotation selection, although it does have p-values of enrichment for one annotation at a time fits. One last method that assumes independence is RiVIERA (Li and Kellis, 2016). It provides inference of the empirical prior of a genetic variant being associated with a specific disease, which will depend on the annotation. Same as fgwas, it assumes one causal variant per-locus. It doesn't have annotation selection but they do propose a technique to recognize enrichment. It has the limitation of only using binary annotations.

A step further in these class of methods is to integrate linkage-disequilibrium (LD) into their pipelines to model the correlation structure among them. One method that has gained popularity is PAINTOR (Kichaev et al., 2014). Annotation influences the non-zero status probability through a logistic model. They relate SNPs to the observed marginal associations under a multivariate normal model. PAINTOR suggests to use one at a time fits to select top strictly binary annotations. A comparatively less popular method (by comparing the number of citations) is

CAVIARBF (Chen et al., 2016). Their model is fairly similar to *iFunMed* if we didn't have the mediation part. They start with a multivariate GWAS model and split effect sizes into zero and non-zero portions. The non-zero SNPs will have a higher prior inclusion probability that is modeled with a logistic function and a normal prior for the annotation enrichment. These classes of models follow similar ideas as earlier Bayesian variable selection models (Carbonetto and Stephens, 2012). CAVIARBF has the flexibility to adapt to multiple types of annotations (binary and continuous) and provides proper annotation selection by using penalization methods on annotation enrichment.

There are other methods that do not take advantage of meta-analyses and summary-level data. Instead, they use full raw-level data information. FM-QTL (Wen et al., 2015) proposes a multivariate model and computes posterior probabilities of SNPs being non-zero. They relate this to the Bayes Factor and use MCMC for posterior inference. They further link genomic annotations using a logistic model. FM-QTL can only handle a handful of annotations but they can be binary or continuous. DAP (Wen et al., 2016) is an extension of FM-QTL and their models are very similar. Their key difference is that DAP only uses high-priority loci and the EM algorithm instead of MCMC. Finally, bfGWAS (Yang et al., 2017) follows a similar model to (Carbonetto and Stephens, 2012) and models the joint effect of an annotation by having a prior per category. This imposes the strong assumption of non-overlapping annotations categories, which is not realistic. Moreover, it doesn't allow annotation selection.

All of these methods from Figure 1.1B calculate a form of posterior probabilities

with the use of a Bayesian hierarchical model. We propose GRAD (Rojo et al., 2020), a genome-wide regression with auxiliary data that decomposes SNP effect sizes into two components: one that is measured by the annotation information ($A\boldsymbol{\gamma}$) and an annotation-free contribution ($\boldsymbol{\eta}$). The general formulation of GRAD allows us to take advantage of regularization methods for the selection of $\boldsymbol{\gamma}$ and $\boldsymbol{\eta}$. Methods that provide annotation selection are based on forward-backward techniques (fgwas), or limited to enrichment or p-values from fittings with one annotation at a time (GPA, PAINTOR, DAP, and RiVIERA). These have the potential of wasting important information that might have joint effects that can improve SNP detection. Unlike others, except for CAVIARBF, we provide high-dimensional annotation selection. We recommend the use of GRAD for SNP selection and inspect them further with fine-mapping tools.

The rest of the document is organized as follows. In Chapter 2 I will introduce *iFunMed* to do integrative functional mediation analysis with GWAS and eQTL. In Chapter 3, GRAD is proposed to leverage annotations into GWAS analysis with a flexible model. Both chapters are accompanied by extensive simulation experiments and applications to the Framingham Heart Study (FHS) data. Lastly, I will discuss some future directions for these models in Chapter 4.

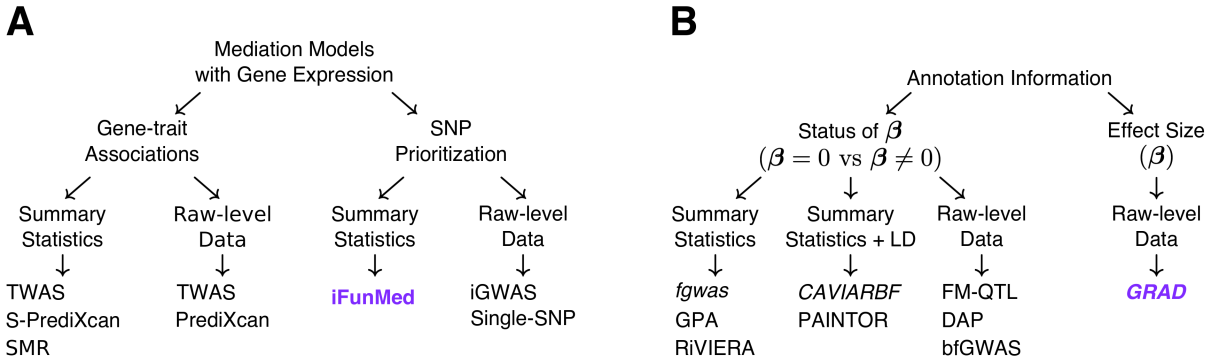

Figure 1.1: **Categorization of (main) statistical methods that perform data-integration on GWAS analysis. Methods highlighted in purple presented in this dissertation.**
(A) Methodologies based on mediation analysis can be split on what they aim to characterize: gene-trait associations or SNP prioritization. They typically take as input summary-level data or raw-level data. **(B)** Methodologies that aim to improve SNP prioritization and/or selection with the use of external annotation information. They usually use annotation to inform the status of the SNP effect $\boldsymbol{\beta}$ sizes as prior information or the effect size magnitude and take a combination of summary-level data, LD matrix, or raw-level data as input, besides the annotations. Methods in cursive indicate that the model provides tools for annotation selection.

2 IFUNMED: INTEGRATIVE FUNCTIONAL MEDIATION ANALYSIS OF GWAS AND EQTL STUDIES

2.1 Introduction

Genome-wide association studies (GWAS) and molecular quantitative trait loci (QTL) (e.g., expression QTL (eQTL), methylation QTL (meQTL)) studies are commonly used approaches in genetic research. Many studies aiming to integrate these two types of data have emerged in the study of complex diseases such as Type 2 diabetes (Zhong et al., 2010), Crohn's disease (Xiong et al., 2012), and various types of cancer (Zhang et al., 2012). These approaches capitalize on the idea that trait-associated SNPs are more likely to be functional and thus eQTLs (Nicolae et al., 2010); therefore, they aim to identify significant GWAS SNPs that are also eQTL SNPs. Although such an approach makes use of eQTLs for reranking or filtering the candidate disease SNPs, it falls short of disentangling the association of eQTL SNPs to generate mechanistic hypotheses. An emerging area to address this shortcoming is causal mediation modeling of GWAS and eQTL data to decompose the etiological mechanisms for the total genetic effect into the genetic effect on disease risk mediated through gene expression (mediation or indirect effect, Genotype $\rightarrow$ Gene Expression $\rightarrow$ Phenotype) and the genetic ef-

[1]Material in this chapter is a modified version of: Constanza Rojo, Qi Zhang & Sündüz Keleş. "iFunMed: Integrative functional mediation analysis of GWAS and eQTL studies." *Genetic Epidemiology*, 2019

fect through other biological pathways or environmental risk factors (direct effect, Genotype $\rightarrow$ Phenotype adjusted by Gene Expression). Pioneering work in this area include mediation analysis framework of (Huang et al., 2004, 2014; Chaibub-Neto et al., 2010). Although mediation analysis in genetical genomics is an active area of research with most recent methods addressing familial designs (Huang et al., 2015) or specifically aiming to identify expression-trait associations (Gamazon et al., 2015; Gusev et al., 2016; Barbeira et al., 2018), it shares the key impediment of individual GWAS and eQTL analysis: more than 90% of associated SNPs are located either intronic or intergenic regions, making their interpretation challenging.

The availability of large-scale functional annotation data through large consortia projects is enabling the annotation of non-coding SNPs that significantly associate with disease and other traits. Initial analyses from the ENCODE Consortium indicated that GWAS-identified phenotype-associated variants can be found in regulatory regions (enhancers) more often than expected by chance (Maurano et al., 2012). This and related fundamental observations led to the undertaking of phenotype-associated variant prioritization using functional annotation data. The increasing body of work in this area, such as (Kichaev et al., 2014; Wen et al., 2016; Chung et al., 2014; Pickrell, 2014; Gagliano et al., 2014; Thompson et al., 2013; Minelli et al., 2013; Chen et al., 2016), typically model univariate association statistics from GWAS as functions of annotation data. None of these methods utilize functional annotation in the joint analysis of clinical/physiological and expression phenotypes and almost all of them focus on relating univariate association statistics from GWAS to annotation data. Another major challenge in the currently

employed mediation analysis approaches, except for recently developed (Gusev et al., 2016; Zhu et al., 2016; Barbeira et al., 2018) that can only detect gene-trait associations, is that they require raw subject-level, i.e. experimental-unit level, SNP genotype, gene expression, and phenotype data. Although such level data from GWAS and eQTL studies can be available through controlled-access repositories such as dbGap (Mailman et al., 2007), information stored in these repositories are not always organized enough to enable easy interactions with the data without the involvement of data generators. Furthermore, with the ever increasing data sizes of biobanks, successful access to raw, unprocessed data requires considerable storage and computation capacity. However, GWAS and eQTL summary statistics of individual studies are often publicly available as part of the publication process.

To overcome these challenges, we introduce *iFunMed*: a mediation model that utilizes functional annotation data (**A** in Figure 2.1) as prior information and builds on summary statistics from GWAS and eQTL studies ($\mathbf{Z}_Y$, $\mathbf{Z}_\mathbf{G}$ in Figure 2.1). Specifically, *iFunMed* model leverages functional annotation information when modeling the inclusion probabilities of the SNPs, i.e., probability that a given SNP has a non-zero direct or indirect effect. As a result, it enables identification of SNPs that are associated with phenotypical changes through direct phenotype-genotype effect and/or indirect gene expression mediated phenotype-genotype effect. We further develop a functional annotation screening procedure to accompany the direct and indirect models of *iFunMed*. This is motivated by the fact that a large proportion of the annotations exhibit no to little association with the summary statistics, and annotations that associate with the GWAS summary statistics ($\mathbf{Z}_\mathbf{Y}$)

do not necessarily correlate with the eQTL summary statistics ($\mathbf{Z_G}$), and vice versa (Figure A.1). Figure 2.1 depicts an overview of the model. We derive a variational expectation-maximization (EM) algorithm that enables the fitting of the *iFunMed* model in a computationally feasible way. Our data-driven computational experiments illustrate how informative annotations improve SNP selection performance in the *iFunMed* model. These experiments also indicate that *iFunMed* is robust to non-informative annotations. Application to Framingham Heart Study (FHS) data using a large collection of publicly available annotations results in elucidation of SNPs, mediation effects of which can be attributed to regulatory mechanisms. Implementation of *iFunMed* is in `R` programming language and is freely available at `https://github.com/mcrojo/iFunMed`.

2.2 Materials and Methods

Mediation Model from Univariate Summary Statistics

Multivariate Mediation Analysis

Let $\mathbf{Y}$ be a quantitative phenotype of interest observed from n subjects, $\mathbf{G}$ the expression of a gene that is associated with phenotype $\mathbf{Y}$, $\mathbf{S}$ the $n \times p$ SNP genotype matrix, and $\mathbf{X}$ the matrix of other covariates that may be important to control for, such as age and sex. The relationship between $\mathbf{Y}$, $\mathbf{S}$, $\mathbf{G}$, and $\mathbf{X}$ can be modeled as in Figure 2.1 by the following mediation framework

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\alpha} + \mathbf{S}\boldsymbol{\beta} + \mathbf{G}\delta + \boldsymbol{\epsilon} \quad \text{and} \quad \mathbf{G} = \mathbf{X}\mathbf{a} + \mathbf{S}\mathbf{B} + \boldsymbol{\eta}, \tag{2.1}$$

where $\boldsymbol{\epsilon} \sim N(0, \sigma_\epsilon^2 \mathbf{I}_n)$ and $\boldsymbol{\eta} \sim N(0, \sigma_\eta^2 \mathbf{I}_n)$. Under model (2.1), $\boldsymbol{\beta}$ is considered as the direct effect of the genotype $\mathbf{S}$ on the phenotype $\mathbf{Y}$ adjusted by the mediator $\mathbf{G}$, $\mathbf{B}$ is the effect of the genotype on gene expression, $\mathbf{B} \cdot \delta$ is the mediated (indirect) effect of the genotype on the phenotype through gene expression $\mathbf{G}$, and $\boldsymbol{\beta} + \mathbf{B} \cdot \delta$ corresponds to the total genetic effect of the SNPs on the phenotype. In genetical genomic applications, a simplified version of this model that considers one SNP at a time is typically utilized (Yao et al., 2015). Although computationally efficient, such a model creates a prohibitive multiple testing problem and does not acknowledge the fact that, for complex diseases, each of a large number of underlying susceptibility SNPs might exhibit modest disease association, but their combined effect could contribute to a significant variation in the phenotype (Chatterjee et al., 2016).

In the following, we reformulate the individual-level data mediation model in Equation (2.1) in terms of univariate SNP-level summary statistics from GWAS and eQTL studies. Let $\mathbf{Z}_Y$ and $\mathbf{Z}_G$ denote vector of t-statistics from the univariate regression of clinical phenotype $\mathbf{Y}$ and expression phenotype $\mathbf{G}$ on the individual SNP $\mathbf{S}_j$ for $j = 1, \ldots, p$, both adjusted by covariates $\mathbf{X}$. Then, the summary statistics can be expressed as

$$\mathbf{Z}_Y = \mathbf{D}^{-1}\mathbf{S}^T(\mathbf{I} - \mathbf{H}_X)\mathbf{Y} \quad \text{and} \quad \mathbf{Z}_G = \mathbf{D}^{-1}\mathbf{S}^T(\mathbf{I} - \mathbf{H}_X)\mathbf{G}. \tag{2.2}$$

Here, $\mathbf{H}_X$ is the projection matrix onto the column space of $\mathbf{X}$, and $\mathbf{D} = \text{diag}([\mathbf{S}^T(\mathbf{I} - \mathbf{H}_X)\mathbf{S}])$. Let $\boldsymbol{\Sigma} = \mathbf{S}^T(\mathbf{I} - \mathbf{H}_X)\mathbf{S}$ denote the Pearson correlation matrix of the SNPs as a measure of the linkage disequilibrium (LD) and further define transformations $\tilde{\boldsymbol{\Sigma}} = \mathbf{D}^{-1/2}\mathbf{S}^T(\mathbf{I} - \mathbf{H}_X)\mathbf{S}\mathbf{D}^{-1/2}$, $\tilde{\mathbf{Z}}_Y = \mathbf{D}^{1/2}\mathbf{Z}_Y$, $\tilde{\mathbf{Z}}_G = \mathbf{D}^{1/2}\mathbf{Z}_G$, $\tilde{\boldsymbol{\beta}} = \mathbf{D}^{1/2}\boldsymbol{\beta}$, $\tilde{\mathbf{B}} = \mathbf{D}^{1/2}\mathbf{B}$,

$\tilde{\boldsymbol{\epsilon}} = \mathbf{D}^{-1/2}\mathbf{S}^{\mathsf{T}}(\mathbf{I}-\mathbf{H}_X)\boldsymbol{\epsilon}$, and $\tilde{\boldsymbol{\eta}} = \mathbf{D}^{-1/2}\mathbf{S}^{\mathsf{T}}(\mathbf{I}-\mathbf{H}_X)\mathbf{D}^{1/2}\boldsymbol{\eta}$. For the sake of simplifying the notation, we eliminate " ˜ " from all the notations except for $\tilde{\boldsymbol{\Sigma}}$ and obtain

$$\mathbf{Z}_Y = \tilde{\boldsymbol{\Sigma}}\boldsymbol{\beta} + \mathbf{Z}_G\delta + \boldsymbol{\epsilon} \quad \text{and} \quad \mathbf{Z}_G = \tilde{\boldsymbol{\Sigma}}\mathbf{B} + \boldsymbol{\eta}, \tag{2.3}$$

where the error covariances now depend on the SNP correlation structure as $\boldsymbol{\epsilon} \sim N(0, \tilde{\boldsymbol{\Sigma}}\sigma^2_\epsilon)$ and $\boldsymbol{\eta} \sim N(0, \tilde{\boldsymbol{\Sigma}}\sigma^2_\eta)$. Figure 2.1B is a pictorial depiction of Equation (2.3) as a mediation model, with additional components described below. Note that in Equation (2.3), $\tilde{\boldsymbol{\Sigma}}$ represents the correlation matrix between SNPs, same as the linkage disequilibrium (LD) structure, which can be approximated using large reference genome pools (e.g., 1,000 Genomes (Consortium, 2015)) or can be shared among investigators without violating the privacy of genetic data of the study participants. As a result, model (2.3) is able to recover parameter estimates from the original model in Equation (2.1) by utilizing univariate summary statistics.

Integrative Functional Mediation

The mediation formulation in Equation (2.3) is a high dimensional regression problem. We consider a Bayesian variable selection framework that can naturally administer data-driven prior information such as functional annotation. We represent such auxiliary information by a $\mathbf{A}_{p\times(K+1)}$ matrix where column j represents a length p binary vector with entries denoting whether or not an individual SNP has the j-th annotation (e.g., SNP resides within an enhancer region identified through

epigenomic studies). We define $\mathbf{s}_\beta$ and $\mathbf{s}_B$ as unobserved indicator variables as

$$s_{\beta,j} = \begin{cases} 1 & \text{if} \quad \beta_j \neq 0, \\ 0 & \text{o.w.} \end{cases}, \quad s_{B,j} = \begin{cases} 1 & \text{if} \quad B_j \neq 0, \\ 0 & \text{o.w.} \end{cases}.$$

Next, we explicitly model $\boldsymbol{\beta}$ as $\boldsymbol{\beta} = \boldsymbol{\tau}_\beta \circ \mathbf{s}_\beta$, where $\boldsymbol{\tau}_\beta \sim N(0, \nu_\beta \sigma_\epsilon^2 \mathbf{I}_p)$, $\mathbf{s}_\beta = (s_{\beta,1}, \ldots, s_{\beta,p})^T$, and $s_{\beta,j} \sim \text{Bernoulli}(\pi_{\beta,j})$. Here, $\mathbf{x} \circ \mathbf{y}$ denotes component-wise multiplication of vectors. Similarly, we model $\mathbf{B}$ as $\mathbf{B} = \boldsymbol{\tau}_B \circ \mathbf{s}_B$, where $\boldsymbol{\tau}_B \sim N(0, \nu_B \sigma_\eta^2 \mathbf{I}_p)$, $\mathbf{s}_B = (s_{B,1}, \ldots, s_{B,p})^T$, and $s_{B,j} \sim \text{Bernoulli}(\pi_{B,j})$. The key roles of $\mathbf{s}_\beta$ and $\mathbf{s}_B$ are to enable selection of SNPs with direct and gene effects, respectively. $\boldsymbol{\tau}_\beta$ and $\boldsymbol{\tau}_B$, with variances of $\nu_\beta \sigma_\epsilon^2$ and $\nu_B \sigma_\eta^2$, denote the effect sizes. $\pi_{\beta,j}$ and $\pi_{B,j}$ are SNP inclusion prior probabilities that we further link to the functional annotation information $\mathbf{A}_j$ by utilizing a logistic function

$$\pi_{\beta,j} = \frac{1}{1 + e^{-\mathbf{A}_j^T \boldsymbol{\gamma}_\beta}} \text{ and } \pi_{B,j} = \frac{1}{1 + e^{-\mathbf{A}_j^T \boldsymbol{\gamma}_B}}. \tag{2.4}$$

In Equation (2.4), $\mathbf{A}_j^T$ represents a $K + 1$ binary vector of K annotations, where $A_{j,k} = 1$ if the annotation k overlaps with SNP j, and an intercept term. Finally, we denote all the parameters of the model, including the hyperparameters, as (δ, θ) where $\theta = (\sigma_\epsilon^2, \sigma_\eta^2, \nu_\beta, \nu_B, \boldsymbol{\gamma}_\beta^T, \boldsymbol{\gamma}_B^T)$.

Likelihood of the Model

We consider the joint distribution of $(\mathbf{Z}_Y, \mathbf{Z}_G, \boldsymbol{\tau}_\beta, \boldsymbol{\tau}_B, \mathbf{s}_\beta, \mathbf{s}_B, \delta)$ as,

$$\begin{aligned} P(\mathbf{Z}_Y, \mathbf{Z}_G, \boldsymbol{\tau}_\beta, \boldsymbol{\tau}_B, \mathbf{s}_\beta, \mathbf{s}_B, \delta; \theta, \mathbf{A}, \tilde{\boldsymbol{\Sigma}}) &= P(\mathbf{Z}_Y, \boldsymbol{\tau}_\beta, \mathbf{s}_\beta, \delta | \mathbf{Z}_G; \sigma^2_\epsilon, \nu_\beta, \boldsymbol{\gamma}_\beta, \mathbf{A}, \tilde{\boldsymbol{\Sigma}}) \\ &\quad \times P(\mathbf{Z}_G, \boldsymbol{\tau}_B, \mathbf{s}_B; \sigma^2_\eta, \nu_B, \boldsymbol{\gamma}_B, \mathbf{A}, \tilde{\boldsymbol{\Sigma}}), \quad (2.5) \end{aligned}$$

and remark that right-hand side of Equation (2.5) factorizes into two parts that model the direct effect given the mediator variable ($P(\mathbf{Z}_Y, \boldsymbol{\tau}_\beta, \mathbf{s}_\beta, \delta | \mathbf{Z}_G; \sigma^2_\epsilon, \nu_\beta, \boldsymbol{\gamma}_\beta, \mathbf{A}, \tilde{\boldsymbol{\Sigma}})$), and the gene effect ($P(\mathbf{Z}_G, \boldsymbol{\tau}_B, \mathbf{s}_B; \sigma^2_\eta, \nu_B, \boldsymbol{\gamma}_B, \mathbf{A}, \tilde{\boldsymbol{\Sigma}})$). They share no hyperparameters, and can be fitted separately. We refer to these two parts as direct effect model adjusted by the mediator (DEM) and the gene effect model (GEM), respectively. Then, the log-likelihoods of GEM (L_{GEM}) and DEM (L_{DEM}) are as follows:

$$\begin{aligned} \log P(\mathbf{Z}_G, \boldsymbol{\tau}_B, \mathbf{s}_B; \sigma^2_\eta, \nu_B, \boldsymbol{\gamma}_B, \mathbf{A}, \tilde{\boldsymbol{\Sigma}}) &= \log P(\mathbf{Z}_G | \mathbf{B} \equiv \boldsymbol{\tau}_B \circ \mathbf{s}_B; \sigma^2_\eta, \tilde{\boldsymbol{\Sigma}}) + \log P(\boldsymbol{\tau}_B; \sigma^2_\eta, \nu_B) \\ &\quad + \sum_{j=1}^{p} \log P(s_{B,j}; \pi_{B,j} \equiv \text{expit}(\mathbf{A}_j^T \boldsymbol{\gamma}_B)), \quad (2.6) \end{aligned}$$

$$\begin{aligned} \log P(\mathbf{Z}_Y, \boldsymbol{\tau}_\beta, \mathbf{s}_\beta, \delta | \mathbf{Z}_G; \sigma^2_\epsilon, \nu_\beta, \boldsymbol{\gamma}_\beta, \mathbf{A}, \tilde{\boldsymbol{\Sigma}}) &= \log P(\mathbf{Z}_Y | \boldsymbol{\beta} \equiv \boldsymbol{\tau}_\beta \circ \mathbf{s}_\beta, \delta, \mathbf{Z}_G; \sigma^2_\epsilon, \tilde{\boldsymbol{\Sigma}}) + \log P(\boldsymbol{\tau}_\beta; \sigma^2_\epsilon, \nu_\beta) \\ &\quad + \sum_{j=1}^{p} \log P(s_{\beta,j}; \pi_{\beta,j} \equiv \text{expit}(\mathbf{A}_j^T \boldsymbol{\gamma}_\beta)), \quad (2.7) \end{aligned}$$

where $\text{expit}(x)$ denotes the inverse logistic link function.

We fit this model with an expectation-maximization (EM) algorithm based

on variational approximations (Ormerod and Wand, 2010; Ormerod et al., 2017) owing to the high dimensionality of the problem, i.e., the number of SNPs in a typical mediation study can fluctuate from a couple hundred to thousands. Variational methodologies have shown promising results under Bayesian approaches for large-scale genetic association studies (Carbonetto and Stephens, 2012). The variational algorithm approximates the joint posterior distribution by a product of lower dimension functions and then minimizes the Kullback-Leibler divergence between them. This approach leads to a computationally feasible algorithm while maintaining the desirable properties of the EM algorithm (Tzikas et al., 2008). The outline of the individual steps is presented in Appendix A.1.

Data-driven Simulation Experiments

To evaluate the performance of *iFunMed*, we performed a series of simulations in settings where the underlying truth is known. We used data from the Framingham Heart Study to construct the LD matrix $\tilde{\mathbf{\Sigma}}$. We utilized the 2,456 available subjects and a segment of the genome between 35,985,004 bp and 43,707,220 bp on chromosome 1, which harbored 2,000 SNPs with a wide range of LD between the loci SNPs ($r \in [-0.94, 1.00]$).

To emulate realistic SNP and annotation effect sizes, we simulated data from the hierarchical model in Equation 2.3 by leveraging estimates of δ and $\theta = (\sigma^2_\epsilon, \sigma^2_\eta, \nu_\beta, \nu_B, \gamma^T_\beta, \gamma^T_B)$ from the actual fits of the FHS data and the LD matrix $\tilde{\mathbf{\Sigma}}$. For each simulation setting, we generated the prior inclusion probabilities π_β and π_B using an annotation matrix $\mathbf{A}$ with two columns: an intercept and one

auxiliary information, which corresponded to fitting the mediation model using one annotation at a time.

We used a set of 220 annotations curated by LD Score Regression (Finucane et al., 2015) and corresponding to four histone marks (H3K4me1, H3K4me3, H3K9ac, and H3K27ac originating from the Roadmap Epigenomic Project (Roadmap Epigenomics Consortium, 2015)) across ten tissues: adrenal and pancreas, cardiovascular, central nervous system (CNS), connective and bone, gastrointestinal (GI), immune/hematopoietic, kidney, liver, skeletal muscle, and other. We further reduced this list to 209 annotations by combining replicates of each epigenetic mark from the same cell-type. To enable a wide range of simulation settings by reducing the computational time, we utilized five annotations which we selected as representatives of different proportions of SNPs with the annotation in the region, varying from 0.014 (A1: gastric) to 0.147 (A5: hippocampus middle) as depicted in Figure A.2. Individual information is detailed in Table A.1. The simulation parameters were based on the actual fits of the hierarchical model on FHS data by considering multiple loci and genes as candidate mediators. Specifically, the parameters varied as follows: $\sigma^2_\epsilon, \sigma^2_\eta \in \{0.1, 1, 5\}$, $\nu_\beta, \nu_B \in \{10, 20, 100\}$, $\delta \in \{0.05, 0.5\}$, and the parameters associated with the functional annotation effects were set as in Table 2.1 based on the FHS fits. Collectively, these combinations of parameters resulted in 54 different simulation settings for each one of the five annotations we utilized. We generated 20 simulated datasets from each scenario (54×5 settings in total) and summarized the results across these simulation replicates.

Framingham Heart Study (FHS)

We accessed FHS individual-level genotype, expression, and phenotype data of 2,456 subjects through dbGap (dbGaP: phs000007.v16.p6). Genotypes were from the SHARe sub-study that used the OMNI5M array and the whole blood RNA expression from the SABRe CVD study. A total of 1,667 patients had both expression and genotype data. We pre-processed the genotypes with PLINK v1.9 (Chang et al., 2015; Purcell et al., 2007), and imputed with IMPUTE2 v2.3.2 (Marchini and Howie, 2010; Marchini et al., 2007). This resulted in 2,244,466 SNPs (details in Appendix A.2). We utilized the 209 reduced list of LD Score Regression (Finucane et al., 2015) annotations as the functional annotations for prior construction, as in the simulations. For mediation analysis, because the expression data is from whole blood RNA, we focused on blood-related phenotypes factor VII, which is involved in the process of coagulation, red blood cell count, white blood cell count, and Von Willebrand factor, which plays a role in hemostasis. For these phenotypes, we evaluated multiple genes and selected six loci (Table 2.2) with dense and relatively strong genotype signal and evidence of potential mediation to present in detail. Further details are provided in Appendix A.3.

For each SNP in the model, we report estimates of the posterior probabilities of inclusion in the DEM and GEM fits, i.e., $P(s_{\beta,j} = 1|\mathbf{Z}_Y, \mathbf{Z}_G, \hat{\delta}; \hat{\sigma}^2_\epsilon, \hat{\nu}_\beta, \hat{\boldsymbol{\gamma}}^T_\beta, \mathbf{A}, \tilde{\boldsymbol{\Sigma}})$ and $P(s_{B,j} = 1|\mathbf{Z}_G; \hat{\sigma}^2_\eta, \hat{\nu}_B, \hat{\boldsymbol{\gamma}}^T_B, \mathbf{A}, \tilde{\boldsymbol{\Sigma}})$, with the goal of elucidating the direct and mediated SNP effects on the phenotype.

Annotation Screening

To accommodate the observation that annotations might not exhibit any association with the summary statistics and that they can be associated differently with the GWAS and eQTL statistics (Figure A.1), we carried out annotation screening. The proposed strategy starts by calculating an enrichment statistics based on the *iFunMed* fit without the use of any annotation information (every SNP has the same prior probability of inclusion). We refer to such model as the null model. In what follows, we use the notation for the gene effect model (GEM: $\mathbf{B}$) as the extension to the direct effect model (DEM: $\boldsymbol{\beta}$) follows directly. We denote the posterior probability of inclusion from the null model as $\hat{s}_{B,j} = P(s_{B,j} = 1|\mathbf{Z}_G; \hat{\sigma}^2_\eta, \hat{\nu}_B, \hat{\boldsymbol{\gamma}}_B^T, \mathbf{A}, \tilde{\boldsymbol{\Sigma}})$. We then calculate the *average* posterior probability of inclusion (e.g., evidence of the non-zero effect size) of the SNPs with the annotation k, for $k = 1, \ldots, K$, from the null model as:

$$ave(\hat{s}_{B,k}) = \frac{\sum_{j:A_{j,k}=1} \hat{s}_{B,j}}{\sum_{j=1}^{p} A_{j,k}}.$$

Next, we evaluate the significance of this enrichment statistic $ave(\hat{s}_{B,k})$ with a permutation approach. The overall procedure is summarized as follows.

1. Fit the *iFunMed* null model and compute the average posterior probability of inclusion, $ave(\hat{s}_{B,k})$, for each annotation k, for $k = 1, \ldots, K$. This is the main annotation-level summary statistic for quantifying enrichment.

2. Within each annotation k, permute the binary annotation values $\mathbf{A}_k$ and compute *permuted* $ave(\hat{s}_{B,k})$.

3. Repeat step 2 N times and compute $ave(\hat{s}^n_{B,k})$ for $n = 1, \ldots, N$ as the permuted average posterior probability of inclusion.

4. Obtain the per annotation enrichment p-value as:

$$\hat{p}_{B,k} = \frac{1}{N} \sum_{n=1}^{N} \mathbf{1}\{ave(\hat{s}^n_{B,k}) \geqslant ave(\hat{s}_{B,k})\}.$$

In both the simulations and the FHS application, we set $N = 10,000$. Evaluations of this annotation screening procedure for Type I error and power are provided in Section 3.3. In the FHS application, we utilized annotations that were marginally significant at 5% and were retained after multiple testing correction with (Benjamini and Hochberg, 1995) at false discovery rate (FDR) of 10%.

2.3 Results

Evaluation with Simulations

We first evaluated power and Type I error of the annotation screening strategy (Figure 2.2, Table A.2). Since five different annotations were simultaneously considered in these simulations, we utilized Bonferroni correction for multiplicity adjustment and observed that the family-wise error rate (FWER) is well controlled at 5% (Table A.2). Results on power are further stratified with respect to the annotation that was used to simulate data (A1 to A5, representing annotations with increasing proportions of overlapping SNPs) and annotation effect sizes (mild and strong) and are presented as the proportion of times that the screening strategy identified the

correct annotation in Figure 2.2A. As expected, median power for strong effect sizes is higher (between 0.10 and 0.75 for A1 to A5) than the mild effect sizes (median power between 0.10 and 0.70 for A1 to A5). In addition, we observe that power increases with the increasing proportion of SNPs with the annotation (A1 being the lowest and A5 the highest, Table A.2). Hence, the screening strategy may not be able to guarantee acceptable power for annotations with a low proportion of SNP overlap. To circumvent this and alleviate the multiple testing burden, we exclude individual annotations with less than 5% overlap with the SNPs from the annotation set.

Area under the receiving operating characteristic curve (AUROC) comparisons across all the simulation scenarios (Figure 2.2B) support that leveraging functional annotation data improves detection of relevant SNPs, regardless of the simulation setting and proportion of the SNPs with the annotation. While annotations with lower proportion of SNPs with the annotation (A1, A2, and A3) exhibit only marginal improvement, the ones with higher proportion (A4 and A5) show greater and more stable improvements. This is consistent with the overall observations in Figure 2.2A. Specifically, the average percentage increases in the AUROC with the use of annotation A5 are 11.6% and 16.4% for the mild and strong settings, respectively. Area under the precision-recall curves (AUPR) (Figure A.3A) exhibit similar overall patterns as the AUROC curves for both the mild and strong annotation effects.

Figures 2.2C and 2.2D contrast two simulation settings in more detail by varying parameters ν_β and ν_B. In these settings, there is a weak mediation effect ($\delta = 0.05$),

moderate error variances ($\sigma^2_\epsilon = \sigma^2_\eta = 1$), and the prior probability of SNP inclusion changes from 0.01 without annotation to 0.08 with annotation (mild effect). We observe that when the effect size variances of $\boldsymbol{\beta}$ and $\mathbf{B}$ are large ($\nu_\beta = \nu_B = 100$ in panel C), easily leading to larger effect sizes, the AUROC is above 0.8 with and without annotation, albeit using annotation improves the AUROC by 4%. In contrast, for the case with weak effect sizes (panel D), the improvement due to annotation is more pronounced at 15%. ROC curves separated for the DEM and GEM fits behave similarly and are provided in Figure A.4. In addition, the precision-recall curves for these cases are available in Figures A.3C, A.3D, and A.5 and exhibit similar improvements.

In summary, these simulation results indicate that utilizing relevant annotations in multivariate mediation analysis improves SNP selection; however, not so surprisingly, the degree of improvement relies considerably on the effect sizes $\boldsymbol{\beta}$ and $\mathbf{B}$ of the SNPs. When the effect sizes are small (Figure 2.2D), the improvement due to annotations is evident. In contrast, when effect sizes are further away from zero (Figure 2.2C), the improvement is marginal. However, we observe that even for cases where false positive annotations are selected after FWER control, i.e., the annotation strategy identifies either one or more annotations for scenarios with no annotation effect (less than 5% of the cases, Figure 2.3A) or annotations different from the specific annotations used in data generation, leveraging of the annotations does not deteriorate model performance (Figure 2.3B).

In these diverse sets of simulations, we also quantified the computational requirements of *iFunMed*. All of the simulations converged within 300 or fewer

numbers of iterations (Figure A.3B) with a median of 25 iterations. On average, 25 iterations with 2,000 SNPs runs in 26 minutes with a standard deviation of 2 minutes on a MacBook Pro with 2.7 GHz Intel Core i5 processor and 2.7 GHz Intel Core i5 memory. For a more advanced machine (e.g., 64bit with AMD Opteron 6174 processor and 24 cores), it runs in 4 minutes with a standard deviation of half a minute, indicating computational feasibility of *iFunMed*. The time difference is largely attributable to the number of cores which enable parallel computations in matrix inversion.

Application to the FHS

We next utilized FHS to explore the impact of annotations on the mediation inference. We performed annotation screening for the loci listed in Table 2.2 with a subset of the annotations that overlap with at least 5% of the loci SNPs. Two of the six loci considered (loci 3 and 4, Table A.3) resulted in enriched annotations for the GEM, using red blood cell count as phenotype. In what follows, we mainly focus on these loci. Figure 2.4A and 2.4B display p-values of the annotations for loci 3 and 4, respectively. For both loci considered, we found that tissue origins of some of the annotations are supported by the known tissue-specific activities of the mediator genes. In the case of *NINJ2*, GEM enrichment p-values indicate immune/hematopoietic as the most enriched. This is well supported by high expression of *NINJ2* in lymphatic and hematopoietic organs (Araki and Milbrandt, 2000). For *EVA1C*, both of the identified annotations originate from CNS and are supported by curated tissue-gene associations for *EVA1C* (Palasca et al., 2018).

To investigate the impact of selected annotations on SNP detection, we examined the set of SNPs identified with and without annotation by thresholding their posterior probabilities of inclusion at 0.5. Figures 2.4C and 2.4D highlight SNPs with the increased estimates of posterior probability of inclusion with the use of annotations and vice versa for the DEM and GEM fits. Specifically, for the GEM fit in loci 3 and 4 with mediator genes *NINJ2* and *EVA1C*, three SNPs, with estimated effect sizes of zero in GEM, have non-zero estimated effect sizes with the use of the screened annotations. In locus 3, *rs2245906* and *rs11063749* are detected with and without annotation in the GEM fit and (Jansen et al., 2017) identified them as *cis*-eQTLs for *NINJ2* in peripheral blood based on conditional eQTL analysis. Further investigation of the SNPs that are selected only by the use of annotation (*rs76782035* for *NINJ2* and *rs2834027* for *EVA1C*) by atSNP (Zuo et al., 2015; Shin et al., 2018), a web-based tool that provides statistical evaluation of impact of SNPs on transcription factor-DNA interactions, identified both SNPs as potentially impacting binding of transcription factor SIN3A (Figures 2.5A and 2.5B). The direction of the mediation effects and gain- or loss-of-function inference by atSNP of these SNPs are consistent with each other. Specifically, *rs76782035* exhibits negative *iFunMed* estimated effect and is leading to generation of a subsequence, i.e., binding site, that SIN3A may potentially interact with, whereas *rs2834027* has positive *iFunMed* estimated effect and seems to disrupt a potential binding site that SIN3A may interact with. Furthermore, bone-marrow-specific deletion of Sin3a in a mouse model carrying Sin3a conditional knock-out alleles causes reduction of red blood cell count (Heideman et al., 2014), supporting that these SNPs could be affecting the phenotype indirectly.

In contrast, for the SNPs that were excluded from the model fit with the use of annotation, no such apparent SNP-transcription factor relationships were revealed, highlighting the potential of the annotation framework for generating mechanistic hypotheses. Further information on these SNPs is available in Table 2.4 for loci 3 and 4.

Figures 2.5C, 2.5D, 2.5E, and 2.5F display Manhattan plots of GWAS and eQTL univariate summary statistics for the set of SNPs with large changes in their posterior probabilities of inclusion with the use of annotation. We observe that these SNPs tend to be spread around the loci as opposed to being localized on small regions that harbor a leading SNP and its proxies due to high LD. The majority of the SNPs with increased posterior probability of inclusion overlap with the selected annotations used for the *iFunMed* fitting (i.e., $A_{j,k} = 1$, Table 2.4). This indicates an increase in their prior probability of inclusion, while SNPs with decreased posterior probability of inclusion due to annotations tend to not overlap with the annotations included in the model, reducing their prior probability of inclusion.

We refer to the estimated parameters of the model ($\hat{\delta}$, $\hat{\theta}$) in Table B.1 to further elucidate the impact of the annotations. Most of the parameters were estimated similarly with or without annotation, with the exception of parameters that directly involve annotations, i.e., γ_β and γ_B, and variances, ν_β and ν_B, associated with the signal strength. Varying *iFunMed* estimates of γ_β and γ_B assign a higher prior probability of inclusion to SNPs that overlap with the annotations included in the fit. Parameters ν_β and ν_B modulate the distribution of the signal strength and impact the number of SNPs with non-zero effect sizes. These parameters play a

crucial role in the selection of the SNPs.

Finally, for the loci where the screening did not identified any enriched annotations, fits of the null model are summarized in Figure A.6 and Table A.4. For loci 1 and 6, four of the SNPs with non-zero DEM estimated effect sizes have reported associations with Factor VII and Von Willebrand Factor, respectively, based on independent studies. In locus 1, *rs2181540* (Williams et al., 2013) and *rs488703* (Smith et al., 2010) are selected in the DEM while, for locus 6, *rs8176704* (Desch et al., 2013) and *rs505922* (Williams et al., 2017) are identified. In locus 2, *rs330787* is detected in the GEM null fit and its association with the mediator gene *MSH6* is further supported by (Jansen et al., 2017).

Mediation Analysis for Other FHS Phenotypes

To expand our analysis to other phenotypes where gene expression from whole blood, where the bulk of the RNA comes largely from peripheral blood mononuclear cells, may not be a relevant mediator, we considered identifying potential mediators as tissue-specific genes from tissues that may be directly related to the phenotypic variation using reference expression datasets of the GTEx Project (Carithers et al., 2015). Specifically, we considered two phenotypes: fasting glucose and HDL, and utilized GTEx pancreas and liver datasets, respectively. For fasting glucose, out of all the pancreatic eQTL genes, only *IL32* and *P2RX1* had median pancreatic and whole blood gene expression at least two-fold larger than average median expression across all GTEx tissues, suggesting their specificity for pancreas and whole blood. Same procedure was carried out for HDL using liver eQTL genes and

identified six genes (*CDA*, *PSD4*, *IL1RAP*, *ASGR2*, *IGFLR1*, and *APMAP*) as liver and whole blood specific. Information on these loci are provided in Table A.5 and results of the annotation screening are available in Tables A.6 and A.7.

Results for *P2RX1* are summarized in Figure A.7. Figure A.7A highlights annotations from cardiovascular and GI tissues as the most enriched. Data integration from many different technologies and sources has found cardiovascular tissue to be associated with *P2RX1* (Palasca et al., 2018) and, furthermore, *P2RX1* is also highly expressed in midgut (GI) and its associated cells (Edgar et al., 2013). SNPs identified by thresholding posterior probabilities of inclusion at 0.5 are highlighted in their respective Manhattan plots (Figures A.7B and A.7C). *rs8076916* is selected by models both with and without annotation and is also a *cis*-eQTL for *P2RX1* (Jansen et al., 2017). Further investigation of the SNPs that are included only by the use of annotation (*rs76395158*, *rs117071988*, and *rs1050997*) by atSNP identified *rs76395158* and *rs117071988* as potentially impacting binding of transcription factors SRF and NR5A2, respectively. SRF is known to be linked to insulin resistance (Jin et al., 2011) and NR5A2 is associated with increase in glucose levels (Bolado-Carrancio et al., 2014) (Figures A.8A and A.8B).

For the rest of the loci, we found that when *IL32* is a candidate mediator for fasting glucose, different types of T cells underlie the most enriched annotations for the DEM and established links exist between glucose and T cells (MacIver et al., 2008). For HDL with *IL1RAP* as mediator, the most enriched annotation for the DEM is CD19, which acts as a biomarker for B lymphocytes that have been associated with HDL (Kaji, 2013). For *IL32*, *rs1075581* is the only SNP selected with the

use of annotation. atSNP analysis indicates this SNP as a loss-of-function candidate for transcription factor NFE2L1 (Figure A.8C). *iFunMed* estimates a positive effect size for *rs1075581*, leading to higher values of fasting glucose with the SNP allele. Interestingly, NFE2L1 has been linked with glucose levels since its deficiency disrupts glucose metabolism and impairs insulin secretion (Zheng et al., 2015). Further details for the SNPs selected with each candidate mediator are presented in Table A.8.

2.4 Conclusion

Mediation analysis is often used to identify and account for potential mechanisms that underlie an observed association between genetic variants and a phenotype through a mediator variable, e.g., eQTL gene. *iFunMed* extends the existing mediation methods originating from the framework of (Baron and Kenny, 1986) by considering effects of multiple genetic variants on the trait mediated by a single mediator and integrating informative epigenome and regulation-based large scale functional annotation into mediation analysis. This framework complements other areas of analysis of genome-wide association studies that utilize auxiliary annotation information (Kichaev et al., 2014; Wen et al., 2016; Chung et al., 2014; Pickrell, 2014; Gagliano et al., 2014; Thompson et al., 2013; Minelli et al., 2013; Chen et al., 2016; Finucane et al., 2015) and goes one step further from current mediation-based techniques (Gamazon et al., 2015; Gusev et al., 2016; Barbeira et al., 2018) by allowing variant-level identification. *iFunMed* model is fit in a computationally feasible

way by taking advantage of variational methodologies and can operate even when only GWAS and eQTL summary statistics are available. The key output of *iFunMed* includes posterior probabilities of inclusion for each SNP for both the direct and the mediation model, and effect size estimates. While our current application focused on gene expression as a mediator, *iFunMed* can conceptually accommodate other types of mediators.

Evaluation of *iFunMed* with data-driven simulations indicate that relevant annotation information improves SNP detection for both the direct and indirect effects in the mediation analysis and highlights the robustness of *iFunMed* to the use of irrelevant annotations. Our analysis with the FHS data focused on blood-related phenotypes and provided comparisons of *iFunMed* fits that integrates regulatory information and with those that do not. Use of annotation information identified a number of additional SNPs that are missed in the mediation analysis without annotation but well-supported by independent studies. Furthermore, several of them are potentially impacting transcription factors binding. Follow-up investigation of these transcription factors (e.g., SIN3A for red blood cell count and SRF and NR5A2 for fasting glucose) could reveal new potential regulatory roles and diagnostic biomarkers for diseases associated with high/low levels of red blood cell count or fasting glucose.

The choice of an informative prior is an integral part of the *iFunMed* framework. Besides the potential to boost SNP signals with its multivariate model as we have shown in our simulation experiments, it can also facilitate hypotheses generating for the underlying mechanisms of association. Our current applications focused

on binary epigenomic annotations; however, other types of annotations such as impact of SNPs on transcription factor or RNA binding protein-DNA interactions as measured by allele-specific analysis of ChIP-seq or eCLIP-seq experiments (Zhang and Keleş, 2017) can be easily accommodated either as continuous or categorical annotations without further computational cost. In addition, although we presented a well-calibrated and powered annotation selection framework for *iFunMed*, an interesting extension includes adaptively selecting informative priors from a set of noisy prior information by imposing a variable selection framework on the prior model.

iFunMed focuses on scenarios where GWAS and eQTL summary statistics are available from the same set of study subjects and treats multiple genetic variants as instrumental variables, akin to practice in Mendelian randomization (Davey Smith and Ebrahim, 2003). Generalizations of instrumental variable analysis that combine instrumental measurements, exposure and outcome (i.e., phenotype) effects of which are measured on different study populations, have been recently addressed by (Zhao et al., 2019). Although Mendelian randomization techniques employ the strong assumption that all the genetic effect on the phenotype is being mediated by the exposure variable - an assumption that can certainly be violated in our framework when other cellular/genomic events beyond gene expression is considered, it is still worth noting the important discussion of (Zhao et al., 2019) with regard to the use of heterogeneous samples: they can lead to biased estimators and are less robust to model misspecifications. Since, in practice, *iFunMed* can combine summary statistics from meta-analysis studies and LD structure from a reference

panel (e.g., 1,000 Genomes (Consortium, 2015)), it is imperative that the latter is as close as possible to the study population of the summary statistics. If the reference panel does not approximate the study population well, LD structure estimated from this reference might have profound effects on SNP prioritization and mediation effect quantification. Previous work in this area has proposed to use a shrinkage estimator on the reference panel LD matrix (Zhu and Stephens, 2017) and showed that this can improve inference. Further investigation of these approximations and their impact on *iFunMed* are part of our current work.

Web Resources

iFunMed `https://github.com/mcrojo/iFunMed`;

GTEx Portal (release v.7) `http://www.gtexportal.org`;

LD Score Annotations `https://data.broadinstitute.org/alkesgroup/LDSCORE/`;

atSNP `http://atsnp.biostat.wisc.edu`;

PLINK v1.9 `https://www.cog-genomics.org/plink2`;

IMPUTE2 v2.3.2 `https://mathgen.stats.ox.ac.uk/impute/impute_v2.html#download`.

Data Availability

The data that support the findings of this study (Framingham Heart Study (Kannel et al., 1979)) are available through dbGap (phs000007.v16.p6). Ge-

netic data is under study number phs000342.v14.p10 (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000342.v14.p10) and expression data under study number phs000363.v13.p10 (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000363.v13.p10).

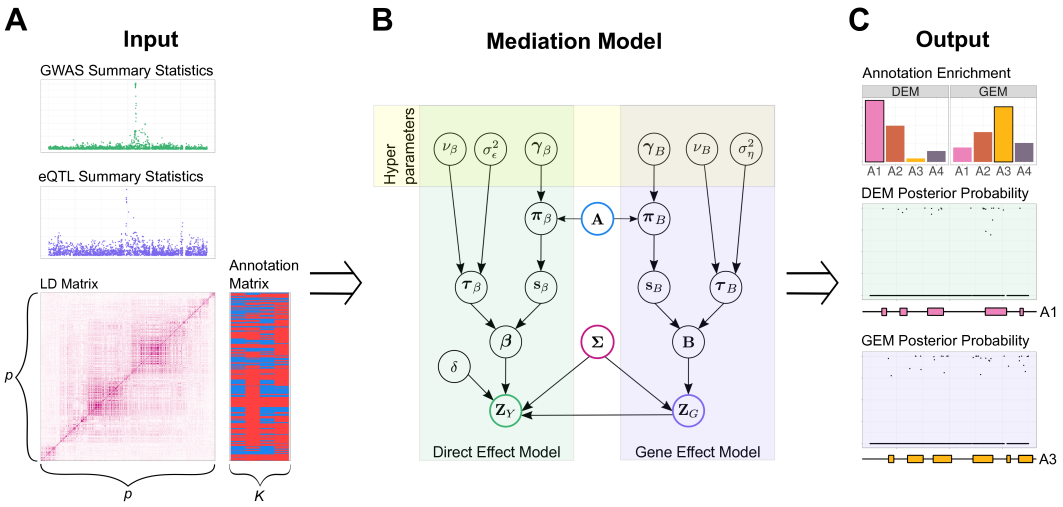

Figure 2.1: **Overview of *iFunMed* modeling framework.**
(A) *iFunMed* input consists of four different types of summary data: GWAS ($\mathbf{Z_Y}$) and eQTL ($\mathbf{Z_G}$) summary statistics, LD matrix ($\mathbf{\Sigma}$), and annotation matrix ($\mathbf{A}$). **(B)** A graphical representation of the proposed hierarchical mediation model where annotation information is integrated through priors for the model parameters $\boldsymbol{\beta}$ and $\mathbf{B}$ with $\mathbf{Z_G}$ as the mediator variable, $\mathbf{\Sigma}$ as the set of independent variables, and $\mathbf{Z_Y}$ a dependent variable. **(C)** *iFunMed* output provides results of the annotation screening for the direct and gene effect models and posterior probabilities of inclusion for each SNP, in addition to other estimated parameters of the model.

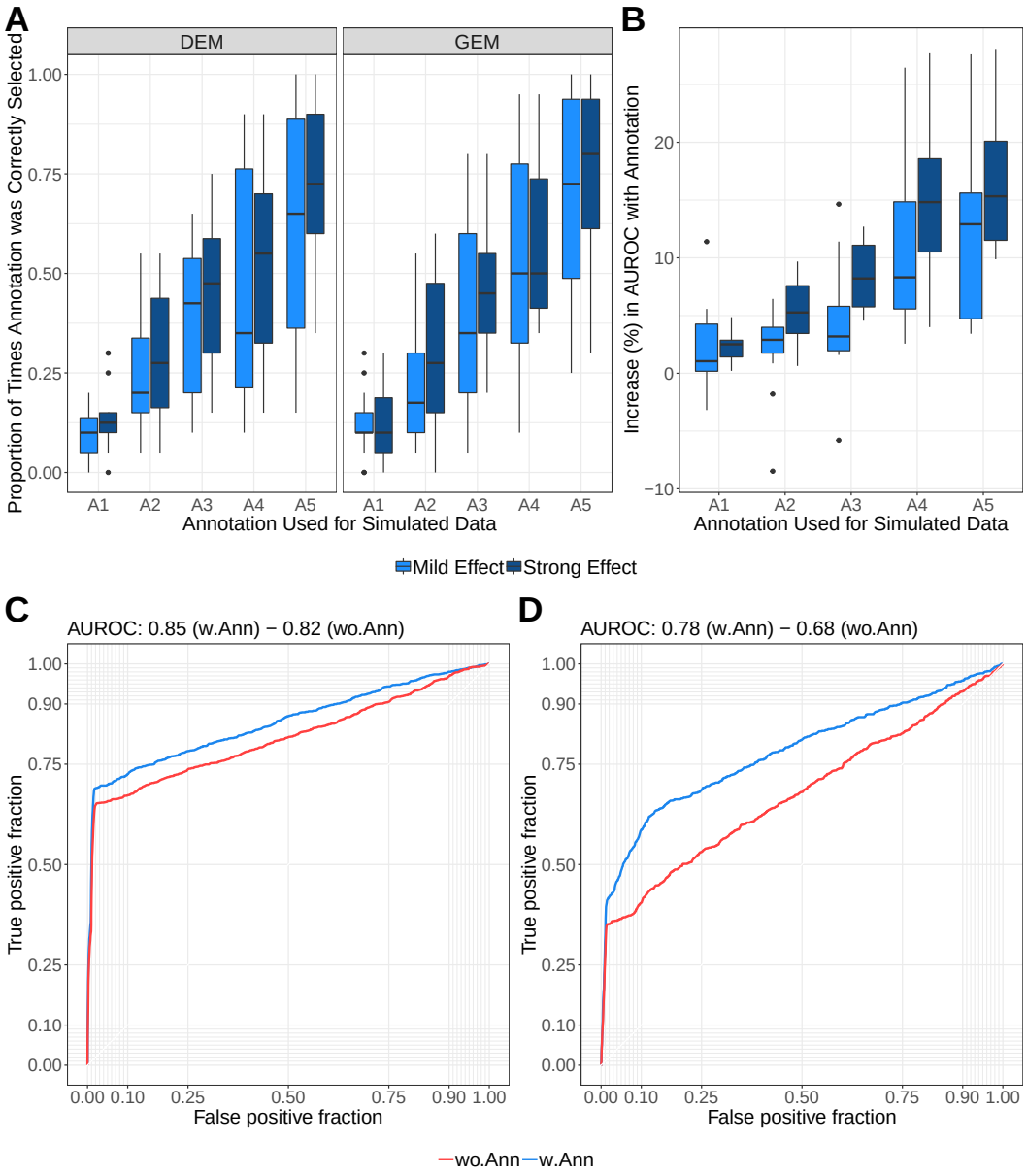

Figure 2.2: **Simulations for comparing *iFunMed* fits with (w.Anno) and without annotation data (wo.Anno).**
(A) Power for annotation screening at family-wise error rate of 0.05. Simulations are stratified with respect to the utilized annotation (A1 to A5 with proportion of SNPs with the annotation increasing from 1.4% to 14.7%) and the annotation effect sizes (mild or strong). For each simulated dataset, all five annotations were evaluated with the Bonferroni adjustment at 5% level. **(B)** Percentage change in the area under the ROC curves across fits for all the 54 simulation settings with the use of annotation. The total set of annotations (54 $\times$ 5 settings) are stratified by the annotation effect sizes $\boldsymbol{\gamma_\beta}$ and $\boldsymbol{\gamma_B}$. ROC curves are obtained by thresholding the total effect estimates. **(C, D)** ROC curves for simulation scenarios with a mild annotation effect ($\boldsymbol{\gamma_\beta}, \boldsymbol{\gamma_B} = (-4.5, 2)$), $\sigma^2_\epsilon = \sigma^2_\eta = 1$ and $\delta = 0.05$, using annotation A5, and varying effect size variances. **(C)**: $\nu_\beta = \nu_B = 100$ for strong and **(D)**: $\nu_\beta = \nu_B = 20$ for weak effect sizes of the SNPs.

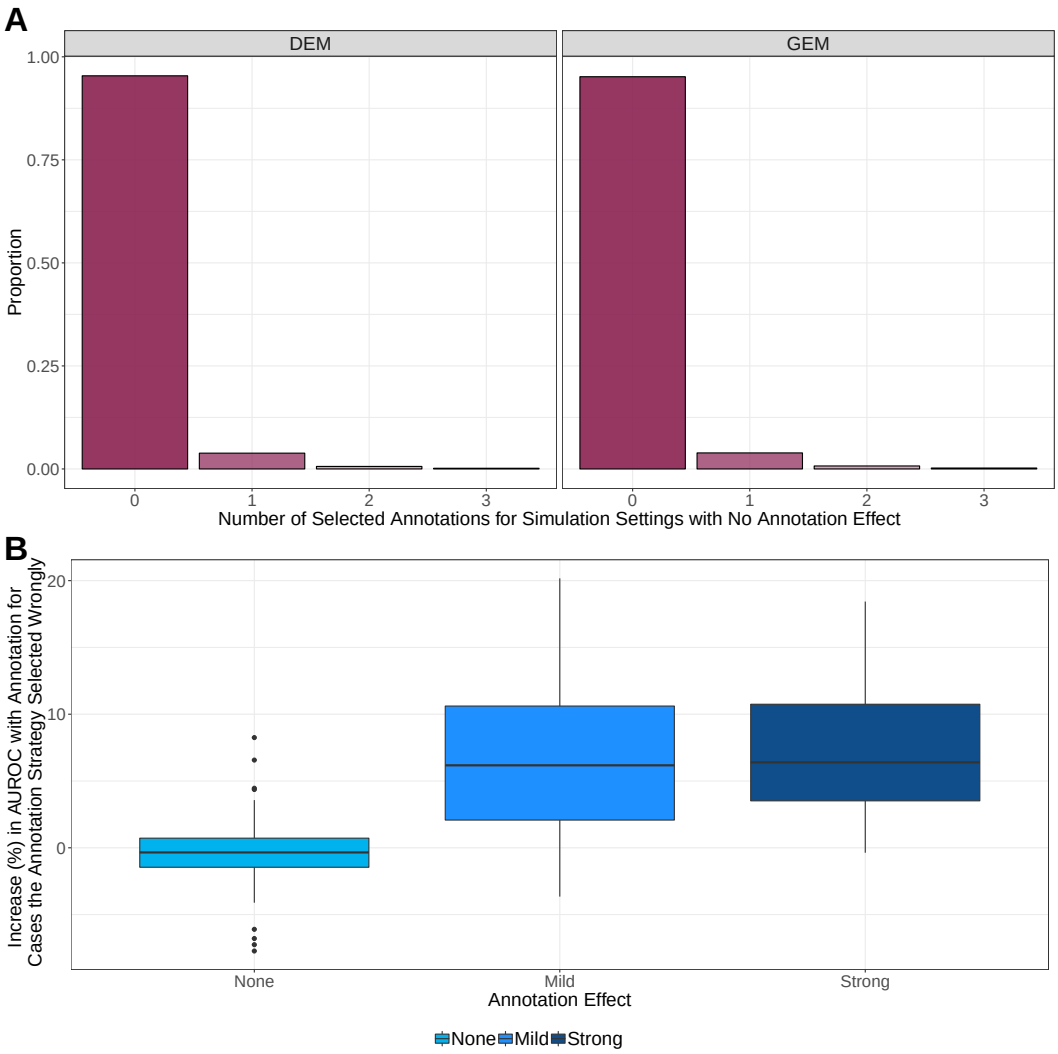

Figure 2.3: **Evaluations of the effect of false positive annotations resulting from annotation screening.**
(A) Proportion of times that the annotation screening strategy identified incorrect numbers of annotations with Bonferroni adjustment at significance level of 5% across simulation scenarios with zero annotation effect sizes (18×5 settings in total). **(B)** Percentage change in the area under the ROC curves across fits where the annotation screening strategy selected one or more incorrect annotation. Incorrect identification when the annotation effect size is zero ("None" category), considers cases where there was at least one selected annotation, whereas "Mild" and "Strong" settings include cases where only false positive annotations were selected. ROC curves are obtained by thresholding the total effect estimates.

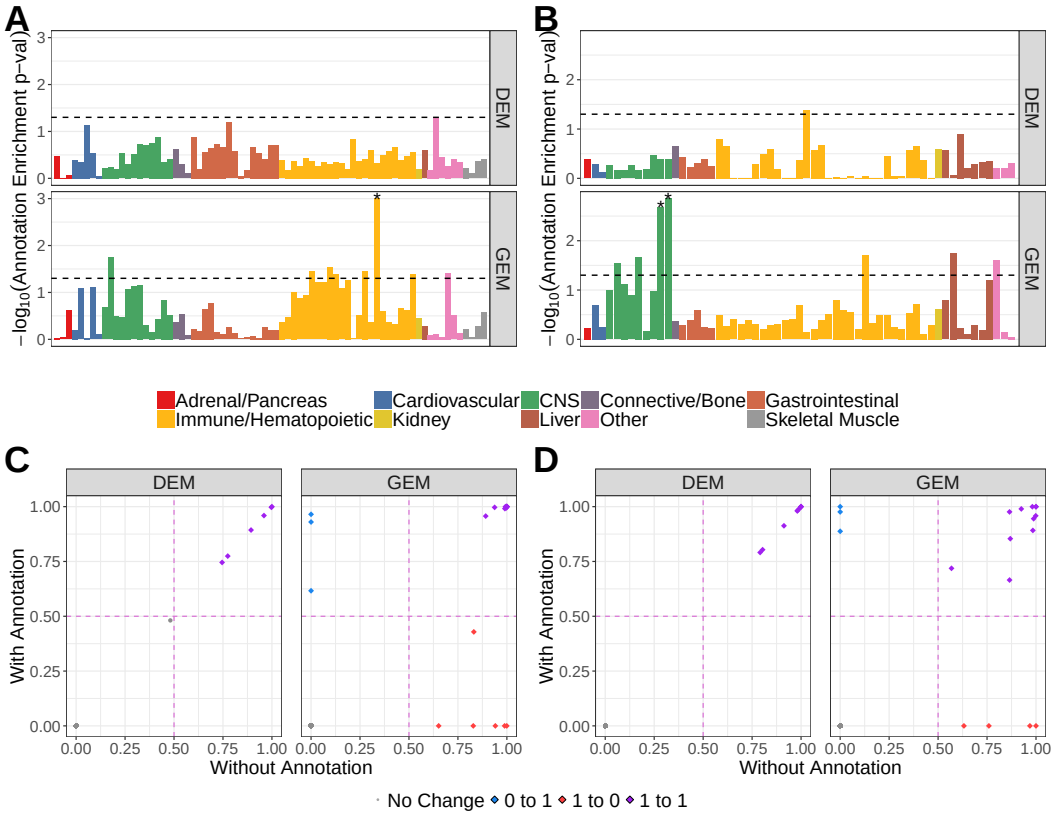

Figure 2.4: **Red Blood Cell Count with *NINJ2* and *EVA1C* as mediators: enriched annotations and identified SNPs.**
(A, B) Enrichment p-values of the annotations (after $-\log_{10}$ transformation) with at least 5% overlap with the loci SNPs. Dashed line represents marginal significance level of 5%. Annotations used for the fits are significant at FDR of 10% and are marked with asterisks. **(A)** *NINJ2* as mediator and **(B)** *EVA1C* as mediator. **(C, D)** Estimated posterior probabilities of inclusion from *iFunMed* for DEM ($P(s_{\beta,j} = 1|\mathbf{Z_Y}, \mathbf{Z_G}, \hat{\delta}; \hat{\sigma}^2_\epsilon, \hat{\nu}_\beta, \hat{\gamma}^T_\beta, \mathbf{A}, \tilde{\mathbf{\Sigma}}) = 1,\ j = 1, \ldots, p$) and GEM ($P(s_{B,j} = 1|\mathbf{Z_G}; \hat{\sigma}^2_\eta, \hat{\nu}_B, \hat{\gamma}^T_B, \mathbf{A}, \tilde{\mathbf{\Sigma}}) = 1,\ j = 1, \ldots, p$) across the two fits (with and without annotation). Dashed line represents the posterior probability cut-off at 0.5. Majority of the SNPs are clustered around values of 0 in the plot. **(C)** *NINJ2* as mediator and **(D)** *EVA1C* as mediator.

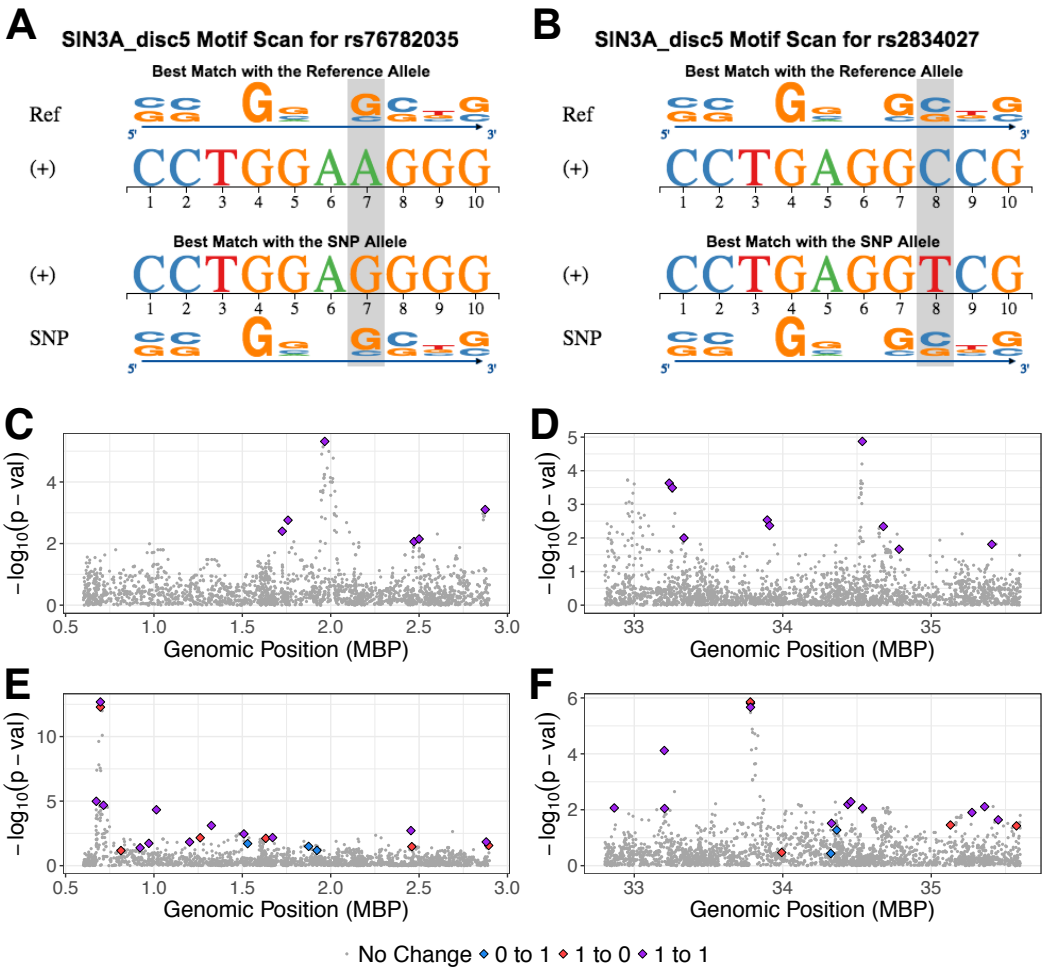

Figure 2.5: **Red Blood Cell Count with *NINJ2* and *EVA1C* as mediators: atSNP results and Manhattan plots.**
(A, B) atSNP composite logo plots for SNPs that are identified only by the use of annotation. The composite logo plots compare the best matches of SIN3A motif to the DNA sequences overlapping the SNP positions with the reference and SNP alleles to suggest potential gain- or loss-of-function with atSNP p-value cutoff of $\leqslant 1e^{-7}$. **(A)** *rs76782035*-SIN3A pair from the model using *NINJ2* as mediator; **(B)** *rs2834027*-SIN3A pair from the model using *EVA1C* as mediator. **(C-F)** Manhattan plots for the GWAS **(C, D)** and eQTL **(E, F)** input summary statistics. SNPs highlighted in blue/red represent SNPs with large changes in their posterior probabilities of inclusion across the two *iFunMed* fits (with and without annotation). Blue SNPs are selected with the use of annotation whereas red SNPs are excluded, and the status of the purple (selected) and gray SNPs (not selected) do not vary between the two fits at the posterior probability of inclusion threshold of 0.5. **(C, E)** *NINJ2* as mediator and **(D, F)** *EVA1C* as mediator.

Table 2.1: Prior probabilities of inclusion as defined in Equation (4) with the annotation effects (γ_β and γ_B) considered in simulations. The prior inclusion probability without annotation is computed with $\mathbf{A}_j^T = (1, 0)$ and the prior inclusion probability with annotation is with $\mathbf{A}_j^T = (1, 1)$.

		Prior Inclusion Probability	
	$\gamma_\beta = \gamma_B$	wo. Annotation $\mathbf{A}_j^T = (1, 0)$	w. Annotation $\mathbf{A}_j^T = (1, 1)$
No Effect	(-4, 0)	0.018	0.018
Mild Effect	(-4.5, 2)	0.011	0.076
Strong Effect	(-3, 2)	0.047	0.269

Table 2.2: Details of loci considered for the mediation analysis.

	Phenotype	Mediator Gene	Chrom	Start	End	# of SNPs	# of Subjects
Locus 1	Factor VII	*TMCO3*	chr13	112,505,203	114,498,328	1,894	1,500
Locus 2	White Blood Cell Count (log)	*MSH6*	chr2	47,001,834	49,698,778	2,745	1,258
Locus 3	Red Blood Cell Count	*NINJ2*	chr12	601,584	2,897,864	2,174	1,255
Locus 4	Red Blood Cell Count	*EVA1C*	chr21	32,802,778	35,599,366	2,593	1,255
Locus 5	Red Blood Cell Count	*ITSN1*	chr21	32,802,778	35,599,366	2,593	1,255
Locus 6	Von Willebrand Factor	*RALGDS*	chr9	134,500,059	137,499,448	2,869	1,500

Table 2.3: Information on the annotations that were identified by the screening strategy.

Mediator Gene	Model	Tissue	Mark	Cell-type	Enrichment p-value
NINJ2	GEM	Immune/Hematopoietic	H3K27ac	CD3 primary	0.001
EVA1C	GEM	CNS	H3K27ac	Cingulate gyrus	0.002
EVA1C	GEM	CNS	H3K27ac	Substantia nigra	0.001

Table 2.4: List of SNPs selected in the analysis of Red Blood Cell Count with *NINJ2* and *EVA1C* as mediators. SNPs are labeled as 0 to 1 ((↑) direction) if they are selected only with the use of annotation and as 1 to 0 ((↓) direction) if they are excluded from the *iFunMed* fit with the use of annotation. SNPs selected with and without annotation are labeled as 1 to 1 (−). Details of the annotations included for both models (DEM and GEM) at the individual SNP level are also displayed.

NINJ2	Location	Model	Direction
rs3825386	1,725,753	DEM	1 to 1 (−)
kgp9894929	1,758,398	DEM	1 to 1 (−)
rs758162	1,966,289	DEM	1 to 1 (−)
kgp1280039	2,471,266	DEM	1 to 1 (−)
rs7968680	2,500,203	DEM	1 to 1 (−)
kgp9131341	2,873,827	DEM	1 to 1 (−)

				Annotation
NINJ2	Location	Model	Direction	CD3 Primary
rs2245906	673,788	GEM	1 to 1 (−)	1
rs11063749	697,095	GEM	1 to 1 (−)	0
rs11503082	697,158	GEM	1 to 0 (↓)	0
kgp9542890	714,576	GEM	1 to 1 (−)	0
kgp2779595	813,385	GEM	1 to 0 (↓)	0
kgp4731528	921,616	GEM	1 to 1 (−)	1
rs2286007	971,291	GEM	1 to 1 (−)	1
kgp4334187	1,013,954	GEM	1 to 1 (−)	1
kgp2348645	1,201,772	GEM	1 to 1 (−)	0
rs117759283	1,261,573	GEM	1 to 0 (↓)	0
kgp27840668	1,324,952	GEM	1 to 1 (−)	0
kgp3224402	1,508,979	GEM	1 to 1 (−)	0
kgp27666261	1,529,921	GEM	0 to 1 (↑)	0
rs1859389	1,632,668	GEM	1 to 0 (↓)	0
rs11061851	1,671,180	GEM	1 to 1 (−)	0
kgp2822459	1,875,163	GEM	0 to 1 (↑)	1
rs76782035	1,922,305	GEM	0 to 1 (↑)	1
kgp6212365	2,453,853	GEM	1 to 1 (−)	0
rs73035417	2,458,950	GEM	1 to 0 (↓)	0
kgp6671610	2,879,808	GEM	1 to 1 (−)	1
rs16929977	2,893,650	GEM	1 to 0 (↓)	0

EVA1C	Location	Model	Direction
kgp8102103	33,235,336	DEM	1 to 1 (−)
kgp3044871	33,256,005	DEM	1 to 1 (−)
kgp5757773	33,334,632	DEM	1 to 1 (−)
kgp1163247	33,895,682	DEM	1 to 1 (−)
kgp4934738	33,910,920	DEM	1 to 1 (−)
kgp349380	34,535,884	DEM	1 to 1 (−)
rs2834178	34,677,391	DEM	1 to 1 (−)
kgp2131229	34,783,522	DEM	1 to 1 (−)
kgp6697616	35,407,829	DEM	1 to 1 (−)

				Annotation	
EVA1C	Location	Model	Direction	Cingulate gyrus	Substantia nigra
kgp5202566	32,865,315	GEM	1 to 1 (−)	0	0
kgp4337540	33,202,254	GEM	1 to 1 (−)	0	0
kgp5881618	33,204,096	GEM	1 to 1 (−)	0	0
rs4817488	33,781,596	GEM	1 to 0 (↓)	0	1
kgp7246838	33,782,785	GEM	1 to 1 (−)	1	1
rs2211789	33,782,887	GEM	0 to 1 (↑)	1	1
rs113131388	33,992,844	GEM	1 to 0 (↓)	0	0
rs2834027	34,323,524	GEM	0 to 1 (↑)	1	1
kgp8275553	34,328,258	GEM	1 to 1 (−)	1	1
rs73200447	34,363,524	GEM	0 to 1 (↑)	1	1
kgp6028639	34,438,734	GEM	1 to 1 (−)	1	1
kgp244201	34,458,657	GEM	1 to 1 (−)	0	0
kgp3444218	34,537,929	GEM	1 to 1 (−)	0	0
kgp23287607	35,129,049	GEM	1 to 0 (↓)	0	0
kgp917601	35,274,135	GEM	1 to 1 (−)	1	0
kgp23258579	35,359,515	GEM	1 to 1 (−)	0	0
kgp5648029	35,450,938	GEM	1 to 0 (↓)	0	0
rs141547866	35,573,813	GEM	1 to 0 (↓)	0	0

Table 2.5: Estimated *iFunMed* parameters for the Red Blood Cell Count phenotype.

	Locus 3/*NINJ2*		Locus 4/*EVA1C*	
	wo. Annotation	w. Annotation	wo. Annotation	w. Annotation
$\hat{\gamma}_\beta$	-5.915	-5.915	-5.722	-5.722
$\hat{\gamma}_B$	-4.840	(-5.496, 2.518)	-5.260	(-6.011, 2.698, 0.404)
$\hat{\delta}$	-0.044	-0.044	0.011	0.011
$\hat{\sigma}^2_\epsilon$	0.529	0.529	0.440	0.440
$\hat{\sigma}^2_\eta$	0.473	0.476	0.417	0.419
$\hat{\nu}_\beta$	21.367	21.367	22.677	22.677
$\hat{\nu}_B$	24.138	28.173	19.982	18.735

3 HIGH DIMENSIONAL SPARSE REGRESSION WITH AUXILIARY DATA ON THE FEATURES

3.1 Introduction

Genome-wide association studies (GWAS) have become the standard tool in the scientific community to identify evidence of association between genetic variants and traits of interest. These studies commonly perform univariate analyses of single-nucleotide polymorphisms (SNPs) to identify loci of association and draw conclusions. Although widely-adopted, this approach completely overlooks SNPs with smaller associations (e.g., subthreshold SNPs) and their potential joint contribution to variations in the phenotype. Loci identified can typically explain only a small fraction of the variance in complex traits and do not directly provide suggestions for functional mechanisms of association (Boyle et al., 2017).

Current efforts from large consortia (Encyclopedia of DNA Elements (ENCODE) (Consortium, 2012) and Roadmap Epigenomic Project (Roadmap Epigenomics Consortium, 2015), among others) to collect epigenomic information from a wide array of tissues and cell types are facilitating the interpretation of noncoding associated genetic variants. The integration of epigenomic information into GWAS pipelines has the potential of improving SNP detection and unravel regulatory mechanisms.

Material in this chapter is a modified version of: Constanza Rojo, Pixu Shi, Ming Yuan & Sündüz Keleş. "High dimensional sparse regression with auxiliary data on the features." *(Under Preparation)*

The body of work in this area typically assumes that annotation informs a latent variable that represents the non-zero status of SNP effects and adopts statistical frameworks for variable selection (e.g., Bayesian variable selection regression (Guan et al., 2011)) that compute posterior inclusion probabilities (Carbonetto and Stephens, 2012) and leverage annotation information as a data-driven prior to determining underlying causal variants (Chen et al., 2016; Chung et al., 2014; Kichaev et al., 2014; Li and Kellis, 2016; Rojo et al., 2019; Pickrell, 2014; Wen et al., 2015, 2016; Yang et al., 2017). Many of these models underutilize the data by considering only binary annotations, e.g., wether or not SNP resides in a region with a particular histone modification or transcription factor (TF) binding, and assume that all the variants with the same annotation share the same prior probability of having a non-zero effect (Chung et al., 2014; Kichaev et al., 2014; Li and Kellis, 2016; Rojo et al., 2019; Yang et al., 2017). This treatment of annotations overlooks the sequence dependency of TF binding (Slattery et al., 2014) and the fact that noncoding genetic variants that reside within the same TF can associate differently with a trait. By utilizing annotations that predict accessibility changes among SNP alleles (Alipanahi et al., 2015; Kelley et al., 2016; Shin et al., 2018; Zhou and Troyanskaya, 2015), we gain nucleotide-level refinement to measure the potential impact of annotation on SNP effect sizes. Moreover, annotation data is usually high-dimensional and potentially noisy. Proper selection, except for (Chen et al., 2016) that takes advantage of regularization methodologies on the annotation, is usually nonexistent (Li and Kellis, 2016; Wen et al., 2015, 2016; Yang et al., 2017) or limited to one at a time fitting that misses the joint annotation effect (Chung et al., 2014; Kichaev et al., 2014)

and forward-backward selection (Pickrell, 2014). Methods that allow annotations to inform the magnitude of SNP effect sizes are scarce and typically do not focus on SNP selection or prioritization (Finucane et al., 2015; Reshef et al., 2018).

In this paper, we introduce GRAD, **G**enome-wide **R**egression with **A**uxiliary **D**ata, a statistical framework that leverages external functional annotations by modeling their specific impact on SNP effect sizes. GRAD enables high-dimensional simultaneous SNP and annotation selection by integrating genotype and annotation information in the model using the Lasso (Tibshirani, 1996) and control the per-comparison error rate (PCER) of feature selection through stability selection (Meinshausen and Bühlmann, 2010). Figure 3.1 depicts an overview of the model. A theoretical analysis of GRAD model yields upper bound for the SNP effect estimation errors and provide the requirements concerning genotype and annotation data for optimal estimation. Unlike other methods that focus on whole-genome fine-mapping, GRAD carries out an annotation-informed multivariate SNP selection to determine loci of interest that can be followed up by other fine-mapping techniques to determine causal status. With a wide range of data generation schemes in our real data-driven simulation experiments, we demonstrate that the use of informative annotation improves SNP selection and shows how GRAD outperforms other competing methodologies, such as GPA (Chung et al., 2014). Applications of GRAD to the Framingham Heart Study (FHS) with annotations that reflect the signed effect of SNPs on transcription factor binding leads to the discovery of SNPs that could potentially affect specific phenotypes by disrupting a number of TFs.

3.2 Materials and Methods

GRAD Model

Statistical Model

We denote $Y_{n\times 1}$ as the phenotype vector for n subjects, $G_{n\times p}$ as the genotype matrix with columns corresponding to SNPs in the study, $X_{n\times k}$ as the matrix of explanatory variables (e.g., age, sex) that may be important to adjust for, and $A_{p\times m}$ as the annotation matrix with columns representing annotation information for p SNPs (Figure 3.1A). We assume a linear model:

$$Y = X\boldsymbol{\alpha} + G\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{3.1}$$

where $\boldsymbol{\epsilon}_{n\times 1}$ is a vector of independent random errors with variance σ^2, $\boldsymbol{\beta}_{p\times 1}$ is the effect of genotype on the phenotype, and $\boldsymbol{\alpha}_{k\times 1}$ is the effect of other features. The genotype effect $\boldsymbol{\beta}_{p\times 1}$ is further partitioned based on annotation:

$$\boldsymbol{\beta} = A\boldsymbol{\gamma} + \boldsymbol{\eta}, \tag{3.2}$$

where the effect size $\boldsymbol{\beta}$ is decomposed into a linear combination of the annotation effects $\boldsymbol{\gamma}$ and annotation-free component $\boldsymbol{\eta}$. The effect of SNP i on the phenotype through the influence of annotation j is quantified by $A_{ij}\gamma_j$, where the sign of A_{ij} determines whether the effect of annotation j is positively or negatively reflected in SNP i, and the magnitude of A_{ij} determines the strength of the annotation j for

SNP i. The annotation-free effect of SNP i is captured by η_i. Combining (3.1) and (3.2) as

$$Y = X\boldsymbol{\alpha} + (GA)\boldsymbol{\gamma} + G\boldsymbol{\eta} + \boldsymbol{\epsilon}, \tag{3.3}$$

provides a more intuitive interpretation of this model and justification of the partition in (3.2). Here, each column of GA can be considered as an aggregated SNP with the corresponding annotation as weights. As a result, SNPs with the same (or different) signs in annotation j are accumulated (or subtracted) to create the jth aggregated SNP. γ_j quantifies the effect explained by the jth aggregated SNP, and $\boldsymbol{\eta}$ quantifies the annotation-free effect of the SNPs. If an annotation does not provide any information on how the genotype associates with phenotype, then the corresponding aggregated SNP will have no effect and $\gamma_j = 0$. If none of the annotations has any influence, then $\boldsymbol{\gamma} = 0$ and the model becomes standard polygenic model $Y = X\boldsymbol{\alpha} + G\boldsymbol{\beta} + \boldsymbol{\epsilon}$. The overall genotype effect $\boldsymbol{\beta}$, annotation effect $\boldsymbol{\gamma}$, and the annotation-free genotype effect $\boldsymbol{\eta}$ are all fixed effects and will be estimated in our model.

Model Estimation

Our goal is to enable the simultaneous selection of relevant genomic variants that affect the phenotype of interest and their relevant annotations. Because $p \gg n$ in most GWAS studies, ordinary least square cannot provide a unique solution for the estimation of the proposed model. For such situations, it is common to use regularization regression methods such as ridge regression (Hoerl and Kennard, 1970), Lasso (Tibshirani, 1996), SCAD (Fan and Li, 2001), MCP (Zhang et al., 2010),

among others. We adopt the Lasso estimator due to its ability to select variables, its easy implementation, and widespread popularity. Specifically, we obtain the estimated parameters with the following convex optimization:

$$\begin{aligned}(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\eta}}, \widehat{\boldsymbol{\beta}}) &= \underset{\substack{\boldsymbol{\alpha},\boldsymbol{\gamma},\boldsymbol{\eta},\\ \boldsymbol{\beta}=A\boldsymbol{\gamma}+\boldsymbol{\eta}}}{\operatorname{argmin}} \left\{ \frac{1}{2n}\|Y - X\boldsymbol{\alpha} - G\boldsymbol{\beta}\|_2^2 + \lambda\left(\|\boldsymbol{\gamma}\|_1 + \|\boldsymbol{\eta}\|_1\right)\right\} \\ &= \underset{\substack{\boldsymbol{\alpha},\boldsymbol{\gamma},\boldsymbol{\eta},\\ \boldsymbol{\beta}=A\boldsymbol{\gamma}+\boldsymbol{\eta}}}{\operatorname{argmin}} \left\{ \frac{1}{2n}\|Y - X\boldsymbol{\alpha} - (GA)\boldsymbol{\gamma} - G\boldsymbol{\eta}\|_2^2 + \lambda\left(\|\boldsymbol{\gamma}\|_1 + \|\boldsymbol{\eta}\|_1\right)\right\}, \quad (3.4)\end{aligned}$$

where $\|\cdot\|_2$ and $\|\cdot\|_1$ are the ℓ_2 and ℓ_1 norms of vectors respectively. The ℓ_1 penalty on $\boldsymbol{\gamma}$ and $\boldsymbol{\eta}$ induces variable selection on both annotation and the remaining annotation-free genotype effects. Without loss of generality, we assume that the columns of X, G, and GA are standardized to have mean zero and variance one so that the ℓ_1 penalty on variables are at comparable levels and an intercept for the regression can be fitted separately. The tuning parameter λ controls the number of variables selected. To avoid the burden of choosing a single optimal value of λ using cross-validation or scaled lasso (Sun and Zhang, 2012), we use stability selection (Meinshausen and Bühlmann, 2010). Stability selection adopts a subsampling aggregation approach that is virtually insensitive to the choice of λ. Combined with high dimensional selection algorithms, it yields on a bound for the expected number of false selections, hence providing per-comparison error rate (PCER) control. Specifically, we estimate (3.4) with a range of values of λ in one hundred bootstrapped samples with halves of the observations and record the frequency of each variable being selected among the one hundred runs (Figure 3.1B). Variables with selection frequency exceeding a certain cutoff are kept in the final model (Figure 3.1C).

Theoretical Analysis of the GRAD Model

We provide an upper bound for the estimation error of $\boldsymbol{\beta} = A\boldsymbol{\gamma} + \boldsymbol{\eta}$ in Theorem 1 and Theorem 2 in Appendix B.1 to gain insights on what factors affect the estimation accuracy. These theorems reveal the necessary conditions to be within the derived bound.

Based on these results, leveraging annotations yields improved estimation accuracy of $\boldsymbol{\beta}$ when its non-zero components are mostly accounted for $A\boldsymbol{\gamma}$ instead of $\boldsymbol{\eta}$, i.e., when the effect of genotype on phenotype is largely through annotations. In addition, Theorem 1 also provides a guide on how to filter the genotype and annotation information to improve the model estimation. Specifically, the condition in Theorem 1 requires: (i) weak LD structure in genotype matrix G, (ii) a well conditioned annotation matrix A, (iii) annotations with enough non-zero entries, and (iv) non-degenerate annotations where values for individual SNPs are not too similar.

Framingham Heart Study Data and Annotation Data

We used data from the Framingham Heart Study with individual-level genotypes (SHARe substudy) and phenotypes of 2,456 subjects (dbGaP: phs000007.v16.p6). We used *iFunMed* (Rojo et al., 2019) preprocessed genotypes. We considered 382 signed and continuous annotations curated by signed LD profile (SLDP) (Reshef et al., 2018). These annotations were derived from ChIP-seq experiments (75 transcription factors and 84 distinct cell lines) from ENCODE (Consortium, 2012) using the Basset software (Kelley et al., 2016). The resulting nucleotide-level annotation

matrix reflects the signed effect of SNPs on transcription factor binding. A total number of 847,491 FHS SNPs were present in the SLDP annotations and had the same minor allele. We proceeded with these for our analyses.

We focused our analysis on four phenotypes that were selected based on their genotype signal strength: factor VII, von Willebrand factor, fasting glucose, and height. For each phenotype, we further screened genotype and annotation data to meet the requirements of Theorem 1 and Theorem 2. We reduced the number of candidate SNPs to P by keeping SNPs with minor allele frequency (MAF) $\geqslant$5% and the lowest p-values from its univariate associations (adjusted by age and sex). Based on exploratory analyses, we use $P = 1,000$ along with the results. This subset of SNPs allows us to have a wide range of signals while removing noisy (low signal) SNPs from the model fitting. Requirement (i) of Theorem 1 is met because the LD structure of SNPs is greatly weakened. We later removed annotations that were highly correlated (pairwise Pearson's correlation magnitude $\geqslant 0.95$ and kept one or the other) and that overlap with less than 3 out of the 1,000 SNPs for requirements (ii) and (iii) of Theorem 1. Requirement (iv) is not an issue with our annotations since they are measured at the nucleotide level and are highly sparse. Information regarding the data considered for each phenotype is listed in Table 3.1.

Simulation Experiments

We design simulation studies to measure the performance of the GRAD model under diverse scenarios. In all of the simulations, we utilized data from the Framingham Heart Study to construct the genotype matrix G and generated 100 datasets

per scenario, i.e., simulation replicates. We followed the same procedure as in the data applications (Section 3.2) to obtain G using von Willebrand factor as the phenotype. This resulted in 2,163 subjects and 1,000 SNPs.

Our simulations can be divided into two categories. One is purely designed to evaluate the performance of GRAD for fits with and without annotation in SNP and annotation selection. The other category compares the selection of SNPs between GRAD and GPA (Chung et al., 2014), a model that prioritizes GWAS results (univariate p-values) by using annotation information. Since GPA model can only use binary annotations, instead of utilizing the SLDP annotations, we used the annotations curated by LD Score Regression (Finucane et al., 2015).

Evaluation of GRAD

We leveraged two data generation schemes to measure the impact of annotation information on SNP selection. In the first scheme (linear partition), we simulate data directly from the model setting of GRAD as in equation (3.3), making it the most favorable scenario. In the second scheme (model misspecification), we simulate from a misspecified model, where the parameter $\boldsymbol{\beta}$ is drawn from a Laplace distribution centered in zero and scale depending on $A\boldsymbol{\gamma} + \boldsymbol{\eta}$, i.e. $\boldsymbol{\beta} \sim \mathtt{Laplace}(0, |A\boldsymbol{\gamma} + \boldsymbol{\eta}|)$, and the response phenotype generated as in equation (3.1). This resembles the treatment of annotations of (Finucane et al., 2015) where SNP effect sizes have mean zero and the variance depends on functional categories.

In both schemes, we simulate data by leveraging model parameters $\boldsymbol{\gamma}$, $\boldsymbol{\eta}$, and

σ^2. We varied error variance $\sigma^2 \in \{100, 150, 200, 250, 300\}$. We simulated $\boldsymbol{\eta}$ using

$$\eta_i = \begin{cases} N(5,1) & \text{w. p. } p_{\eta\neq 0}/2, \\ -N(5,1) & \text{w. p. } p_{\eta\neq 0}/2, \\ 0 & \text{w. p. } 1 - p_{\eta\neq 0}. \end{cases} , i = 1, \ldots, p,$$

where the proportion of non-zero values $p_{\eta\neq 0} \in \{0.01, 0.02, 0.04, 0.08, 0.1, 0.15, 0.2\}$. $\boldsymbol{\gamma}$ is generated using

$$\gamma_j = \begin{cases} N(\mu_\gamma, 100^2) & \text{w. p. } p_{\gamma\neq 0}/2, \\ -N(\mu_\gamma, 100^2) & \text{w. p. } p_{\gamma\neq 0}/2, \\ 0 & \text{w. p. } 1 - p_{\gamma\neq 0}. \end{cases} , j = 1, \ldots, m,$$

where the informativeness of annotation is controlled in two different ways: the proportions of informative annotations $p_{\gamma\neq 0} \in \{0.02, 0.05, 0.08, 0.1, 0.15, 0.2\}$, and the magnitude of the mean of individual annotation effects sizes $\mu_\gamma \in \{200, 500, 1000\}$ (low, mild, and strong). Both $p_{\eta\neq 0}$ and $p_{\gamma\neq 0}$ control the sparsity of parameters $\boldsymbol{\eta}$ and $\boldsymbol{\gamma}$, respectively. These values are set based on actual fits of the model with the FHS data.

Comparison Between GRAD and GPA

To elicit the best performance of GPA, we followed the data generation procedure of GPA for simulations using a linear mixed model and the liability threshold model, which we will refer to as GPA-LTM. We used the same parameters as in the GPA experiments and only introduced a few differences: we used the genotype

matrix from the FHS data, kept a continuous response variable, and simulated fifty annotations instead of one, among which $S_\gamma \in \{2, 5, 8, 10\}$ are set to be informative. We varied the proportion of risk SNPs as $p_{\beta \neq 0} \in \{0.01, 0.02, 0.04, 0.08, 0.1, 0.15, 0.2\}$.

We added a second simulation experiment because GPA-LTM simulates binary annotations after obtaining the effect sizes and they have more non-zero entries when there are more risk SNPs, which means that the sparsity of the annotations depends on the sparsity of the risk SNPs. The second simulation experiment aims to explore a scenario where annotation informs the non-zero status of the SNP, a common design to account for annotation. We used the *iFunMed* hierarchical model (Rojo et al., 2019) and created 27 simulation settings (details in Table B.1) using the same combination of parameters as in the *iFunMed* experiments that vary variances for the overall SNP effect and error, and individual annotation impact. We generated data with one informative binary annotation at a time (randomly selected out of fifty binary LD Score Regression annotations) with different prior inclusion probabilities (none, mild, and strong) of the SNPs based on the annotation information.

Since GPA does not provide a direct pipeline for annotation selection, we run it with one annotation at a time, tested individual annotation enrichment (with the `aTest` method) in each run, and retained the significant annotations using multiple testing with FDR controlled at 10% (Benjamini and Hochberg, 1995). We re-fitted GPA with the selected annotations to obtain the final set of selected SNPs.

3.3 Results

Evaluation of GRAD with Simulations

Figure 3.2A compares the area under the precision-recall curves (AUPR) and evaluates how annotation information impacts SNP selection. We observe that the use of annotation always improves the detection of relevant SNPs. Such improvement is present regardless of the generative model, although is weaker for cases with model misspecification. For example, the average increase in the AUPR is up to 58.7% for the linear partition model and 37.8% for model misspecification. Both maximums occur for cases when the proportion of non-zero $\boldsymbol{\gamma}$ and $\boldsymbol{\eta}$ are fixed at 0.2 and 0.01, respectively. These values bring up two important patterns. For a fixed value of $p_{\eta \neq 0}$, the improvement on SNP selection due to the annotation increases with $p_{\gamma \neq 0}$. This tendency is not surprising since we expect to have a better SNP selection when there are more informative annotations. On the other hand, for fixed values of $p_{\gamma \neq 0}$, when the proportion of risk SNPs $p_{\eta \neq 0}$ is small we observe greater improvement compared to cases with larger $p_{\eta \neq 0}$. This can be attributed, potentially, to the error bound in Theorem 1 increasing for denser $\boldsymbol{\eta}$ (s_η), i.e., the accuracy of our estimation is better for a sparse $\boldsymbol{\eta}$. A similar pattern arises when looking the area under the receiver operating characteristic (AUROC) curve in Figure B.1B.

Comparisons of fits with and without annotation in terms of partial area under the receiver operating characteristic curve (pAUC) for assessing SNP selection performance are shown in Figure 3.2B. When the false positive rate (FPR) is below 0.1, the pAUC reaches values of up to 0.09 and displays higher values for the fits with

annotation and stronger annotation effect magnitudes, for both simulation schemes. In addition, the fit with annotation showed an increased average sensitivity across simulations at 90% specificity, i.e. FPR 0.1, reaching its maximum sensitivity of 0.92 for the linear partition simulation scheme with strong annotation effect sizes (Figure B.6A).

Individual precision-recall curves for SNP selection with a specific parameter combination ($p_{\eta \neq 0} = 0.01$, $p_{\gamma \neq 0} = 0.05$, and $\sigma^2 = 100$) are displayed in Figure 3.2C. When comparing the three annotation effect sizes μ_γ, we observe better AUPR for stronger magnitudes. Since annotation has an impact on the SNP effect size, either by being a linear combination of it (linear partition) or by influencing its variance (model misspecification), both simulation schemes will have larger effect sizes when the annotation effect magnitude becomes greater, leading to better SNP selection performance and a more moderate improvement with the use of annotation, as shown in Table 3.2. Moreover, the curve from the fit with annotation achieves a power of 73% compared to 16% for the fit without annotation when controlling FDR at 10% for these parameters (Table B.2). Overall power at FDR 10% are presented in Figure B.6B. Other simulation settings exhibit comparable patterns for AUPR (Figure B.3) and AUROC (Figure B.4).

Evaluations of annotation selection in Figure 3.2D reveal multiple contributors to its performance, which we summarize into two sources: how much of the SNP effect signal $\boldsymbol{\beta}$ is explained by the annotation $A\boldsymbol{\gamma}$ in contrast to the annotation free parameter $\boldsymbol{\eta}$, and how large the correlation among annotations is. First, for fixed values of $p_{\gamma \neq 0}$, AUPR increases with μ_γ because Lasso has a better selection of

relevant annotations when the signal of annotation effects $\boldsymbol{\gamma}$ becomes stronger. Second, for a fixed value of μ_γ, the AUPR changes with $p_{\gamma\neq 0}$ in a convex fashion due to the joint effect of signal to noise ratio and correlation among annotation. On one hand, a larger $p_{\gamma\neq 0}$ brings in denser signals from the annotations that improve the AUPR. On the other hand, a larger $p_{\gamma\neq 0}$ makes it more possible to have a strong correlation between informative and non-informative annotations, in which case Lasso will pick one randomly, worsening the AUPR. For small values of μ_γ, $\boldsymbol{\beta}$ is dominated by $\boldsymbol{\eta}$ instead of $A\boldsymbol{\gamma}$ (Figure B.8), so annotation selection benefits more from denser signals than being hurt by more annotation correlation, making the convex trend of AUPR increasing. For large values of μ_γ, the effect of more annotation correlation outstrips the effect of denser signals, so the convex trend of AUPR becomes to behave more decreasing. In addition, Figure B.5 shows a decreasing trend in AUPR of annotation selection when $p_{\eta\neq 0}$ increases, because larger $p_{\eta\neq 0}$ makes annotation effect $A\boldsymbol{\gamma}$ less dominant in $\boldsymbol{\beta}$. A larger $p_{\eta\neq 0}$ also makes denser signals of $\boldsymbol{\gamma}$ to be more prominent than the annotation correlation when $p_{\gamma\neq 0}$ increases, so the convex trend of AUPR becomes more increasing. For the cases with model misspecification, we never observe decreasing tendencies and there is virtually no difference in annotation selection across different annotation effect magnitudes. When annotations impact the variance of the effect sizes, its selection suffers and the only improvement we observe is attributed to a denser annotation parameter $\boldsymbol{\gamma}$ that allows for more informative annotations.

GPA Comparisons with Simulations

When we evaluated SNP selection in simulations that aim to compare GRAD with GPA in Figure 3.2E and 3.2F, we observe that GRAD outperforms GPA. For cases when the GPA-LTM is used to simulate the data (Figure 3.2E), GPA tends to improve its performance as the proportion of risk SNPs increases while the opposite pattern occurs with GRAD. This is due to GPA's over selection of SNPs. For smaller proportions of risk SNPs, GPA displays an elevated number of false positives. When the proportion increases, the false positives decreases in favor of true positives. On the other hand, since GRAD's estimation accuracy increases with the sparsity of $\boldsymbol{\beta}$ (s_β in Theorem 2), selects fewer false positives for smaller proportions of risk SNPs. When contrasting their AUROC in Figure B.1C, performance of both methods decreases with $p_{\eta \neq 0}$. This can be attributed to the comparable number of correctly identified SNPs (true positives) for both methods that increase with the proportion of risk SNPs. When comparing the two methods in terms of annotation selection (Figure B.9A and B.9B), GPA does better under the GPA-LTM especially when $p_{\eta \neq 0}$ is large. Under the GPA-LTM, the binary annotations are simulated after generating the SNP effect sizes and they have more overlap, i.e. non-zero entries, for larger values of $p_{\eta \neq 0}$. This leads us to believe that GPA selects annotations with higher overlap, regardless of how informative they truly are. When *iFunMed* is used to simulate data in Figure 3.2F, GPA is unable to select SNPs properly while GRAD reaches average AUPR values above 0.9. This is due to the small percent of risk SNPs ($< 6\%$) under this simulation scheme. Moreover, GRAD lacks the ability to select the correct annotation out of fifty under this scenario (Figure B.9C and

B.9D) while GPA is slightly better.

Real Data Results

We run GRAD for the phenotypes specified in Table 3.1. All of the cases considered resulted in selected annotations. For simplicity, we will mainly focus on von Willebrand factor and detailed results from other phenotypes are available on the Supplementary Material. Figures 3.3A and 3.3B display stability selection frequency paths with cutoff selection probability of 0.75 for the parameters considered on fits without and with annotation, respectively. The fit without annotation identified one non-zero SNP whereas the fit with annotation identified two annotation free SNP ($\hat{\boldsymbol{\eta}} \neq 0$) and 20 additional SNPs ($A\hat{\boldsymbol{\gamma}} \neq 0$) that are attributed to four non-zero annotations ($\hat{\boldsymbol{\gamma}} \neq 0$). Comparisons of estimated SNP effect sizes for models with and without annotation in Figures 3.3C-E provide information about the source and strength of associations. We observe that SNPs that have the strongest univariate associations (Figure 3.3C) tend to be selected by both the fit without annotation and the annotation free SNP effect $\boldsymbol{\eta}$ in the fit with annotation. In contrast, non-zero SNPs with lower univariate associations can be captured by the annotation contribution $A\boldsymbol{\gamma}$ but are missed by the fit without annotation. Manhattan plots in Figure B.13 display similar observations and provide information about the genomic location of these SNPs, from which the annotation free non-zero SNPs are localized in small regions and the ones boosted by the use of annotation are more spread around the genome. These patterns are similar when other phenotypes are considered (Figures B.10-B.12; B.14-B.16).

Further details of the detected SNPs and annotations are provided in Table 3.3. Five SNPs in chromosome 9 and 12 are identified as GWAS SNPs for von Willebrand factor or blood protein levels by the GWAS Catalog (Buniello et al., 2018) in studies with populations of majority European descent. From these SNPs, three out of five are selected because of the annotation contribution parameter and display overlap with the selected annotations. Three of these SNPs (*rs8176749*, *rs505922*, and *rs579459*) are located in chromosome 9 and are considered to be part of the *ABO* blood group locus, which has a relationship with hemostasis that influences von Willebrand factor (Peyvandi et al., 2011; Franchini et al., 2007, 2014). *rs8176749* and *rs505922* do not overlap with any of the non-zero annotations. *rs8176749* is detected by the model with annotation and is associated with von Willebrand factor antigen levels and highly determines the ABO blood group (Desch et al., 2013) while *rs505922* is detected with and without annotation and it has been found to be strongly associated with von Willebrand factor (Williams et al., 2013, 2017). According to our fitting, *rs579459* has a positive effect on Pol2 TF binding and (Emilsson et al., 2018) identified it as a protein single-nucleotide polymorphism (pSNP) for the von Willebrand factor protein in *trans*. On chromosome 12, both *rs1063857* and *rs1063856* have an effect on c-Myc TF binding and reside within the von Willebrand factor gene. *rs1063857* is coding synonymous and *rs1063856* coding nonsynonymous SNP and they are highly associated with von Willebrand factor levels (Smith et al., 2010; Desch et al., 2013). This specific locus displays a low GWAS signal in the FHS data ($-\log_{10}(p) < 8$ in Figure B.13) but it has clear importance on von Willebrand factor. This highlights the importance of the use of

annotations that can boost the signal of relevant SNPs that would be overlooked otherwise.

The rest of the selected SNPs are all captured by the fit with annotation and have an effect through at least one of the selected TF. We found well supported TF-phenotype associations in literature for two of the selected annotations. Specifically, Stat3 modulates hemostasis signals (Zhou et al., 2013; Aleva et al., 2018) and upregulation of c-Myc increases expression of von Willebrand factor (Xiang and Hwa, 2016). For the other phenotypes considered, we also found evidence of potential TF-phenotype associations. One of the TF selected in factor VII, Mxi1, is induced by hypoxia, as well as factor VII (Corn et al., 2005; Koizume and Miyagi, 2015). For fasting glucose (log), all of the TF (Elf-1, MafK, Ccn-T2, and C/EBPβ) showed relationship with either diabetes related traits or adipose tissue: Elf-1 is related to NKT cells in mice that plays a role in diabetes (Choi et al., 2011) and related to O-GlcNAc that perturbs insulin levels (Lim and Chang, 2010); *MAFK* is a potential target gene for impaired fasting glucose (Cui et al., 2016) and MafK negatively regulates β−cell function in mice (Nomoto et al., 2015); *CCNT2* gene may play a role in development of adipose tissue (Broholm et al., 2016); C/EBPβ promotes adipose tissue inflammation and insulin resistance (Rahman et al., 2012). Moreover, the selected annotations are not necessarily the ones with the highest overlap with the SNPs but they do tend to be the ones that overlap with SNPs that have the strongest univariate associations (Figure B.17 and B.18).

We further examined how the selected SNPs correlate with each other to elucidate how LD impacts mechanisms of association (Figure B.19). Most of them

do not display high LD except for 2 pairs in chromosome 12 and 15 that were less than 200 base pairs away from each other. In particular, *rs1063857* and *rs1063856* have an LD of 0.99 and impact c-Myc TF, as discussed previously. By incorporating nucleotide-level annotation that informs in both magnitude and direction, we are able to separate the SNP effects for high LD pairs. Having nucleotide-precision measurements is especially beneficial for our model to differentiate high LD SNPs by assigning different effect size estimates.

We also provide results using 5-fold cross-validation (CV) for model estimation instead of stability selection (Figure B.20-B.23). Because CV chooses the tuning parameter λ that minimizes prediction error, it emphasizes more on prediction accuracy instead of offering family-wise error control like stability selection. As a result, the model chosen by CV is generally much denser, with hundreds of selected SNPs that are mostly captured by the annotation-free effect $\boldsymbol{\eta}$.

3.4 Conclusion

The integration of external auxiliary data, e.g. epigenomic information, into GWAS analyses is an important step to further understand underlying mechanisms of association and to better prioritize SNPs loci. Recently emerging methodologies usually model annotation as a data-driven prior that informs the non-zero status of SNP effects and are limited to binary annotations in most cases. In this study, we develop GRAD, a flexible statistical method that models the impact of annotation information on SNP effect $\boldsymbol{\beta}$ by assuming $\boldsymbol{\beta} = A\boldsymbol{\gamma} + \boldsymbol{\eta}$. This model allows the

utilization of continuous-valued annotation information at the nucleotide-level to inform the magnitude and direction of TF accessibility changes.

By adopting Lasso (Tibshirani, 1996) as the estimation method of our model, we are able to perform variable selection on both annotation and genotype information. We provide two techniques to deal with the tuning parameter in Lasso: stability selection and cross-validation. We mainly focus on the results yielded by stability selection because of its error control (PCER). In our simulation experiments, we found cross-validation to have a tendency to over-select SNPs and an inflated number of false positives. Even though stability selection has been deemed as overly conservative in the context of GWAS when no external information is used (Alexander and Lange, 2011), we found in our extensive simulation experiments that by leveraging annotation information into our analyses, we increase the number of relevant selected SNPs and reduce the number of false positives, especially when the true proportion of risk SNPs is small ($< 6\%$). Ultimately, it should be up to the users which technique to use and our code is flexible enough to provide both options.

In this work, we assume one common tuning parameter λ for both $\boldsymbol{\eta}$ and $\boldsymbol{\gamma}$ in the convex optimization problem in equation (3.4). To make the penalties comparable on $\boldsymbol{\eta}$ and $\boldsymbol{\gamma}$, we standardized columns of $\{GA, G\}$ together instead of standardizing G before its multiplication by A. This standardization procedure is interpretable because GA can be seen as "meta SNPs" with annotation as weights, and the "meta SNPs" GA are standardized to have the same scale as the original SNPs G. Our initial explorations considered different tuning parameters, i.e. λ_1 for $\boldsymbol{\gamma}$ and λ_2 for

$\boldsymbol{\eta}$, and we used cross-validation to choose them. No considerable improvement was seen compared to the current approach and computing cost was significantly increased with two tuning parameters. It is also feasible to extend stability selection into the setting of two tuning parameters, although this has not been seen in the current literature, to our knowledge. This could be considered in the future if substantial improvement is detected compared to one common tuning parameter.

The variable selection performance of Lasso can be greatly undermined by high collinearity among features, which is typically observed in GWAS because of linkage disequilibrium patterns among SNPs in close genomic proximity. When high LD is present, Lasso has the tendency of picking one single variant instead of the entire LD block. For such cases, elastic net (Zou and Hastie, 2005) can be an attractive alternative but initial inspections within our simulations failed to uphold its use compared to Lasso. To overcome potential drawbacks that might arise from highly correlated SNPs and/or annotations, we provide a guideline on how to filter genotypes and annotations for optimal estimations by calculating the upper bound of the SNP effect estimation errors. It is part of our current work to inspect how much correlation our method can handle without compromising the quality of the selected SNPs, and whether we can expand our analysis to smaller regions with stronger LD structures. It could also be beneficial to inspect other sequence-based algorithms (Alipanahi et al., 2015; Zhou and Troyanskaya, 2015) to construct signed annotations besides (Kelley et al., 2016) and check their correlation structure.

One important strength of our method is its ability to pick up SNPs with traditionally low signal from univariate associations ($-\log_{10}(p) < 8$) mainly because

of the following reasons: (1) annotation information puts more emphasis on the SNPs with changes in TF accessibility, regardless of the univariate signal strength of the SNP; (2) the multivariate modeling of our method inspect the joint effects of SNPs instead of marginal effect. For example, in our analysis with the FHS data, in addition to recognizing GWAS SNPs that are validated by independent studies and TF-phenotype associations well supported in literature, we are also able to identify a high LD pair in chromosome 15 for the von Willebrand factor (*rs1063857* and *rs1063856*) that impacts c-Myc with different magnitude and opposite direction, and have weak univariate associations ($-\log_{10}(p) \sim 3.6$).

Web Resources

Signed LD Profile (SLDP) Annotations `https://data.broadinstitute.org/alkesgroup/SLDP/annots/`;

LD Score Annotations `https://data.broadinstitute.org/alkesgroup/LDSCORE/`;

GPA v1.1 `https://github.com/dongjunchung/GPA`.

Data Availability

The data that support the findings of this study (Framingham Heart Study (Kannel et al., 1979)) are available through dbGap (phs000007.v16.p6). Genetic data is under study number phs000342.v14.p10 (`https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000342.v14.p10`) and

expression data under study number phs000363.v13.p10 (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000363.v13.p10).

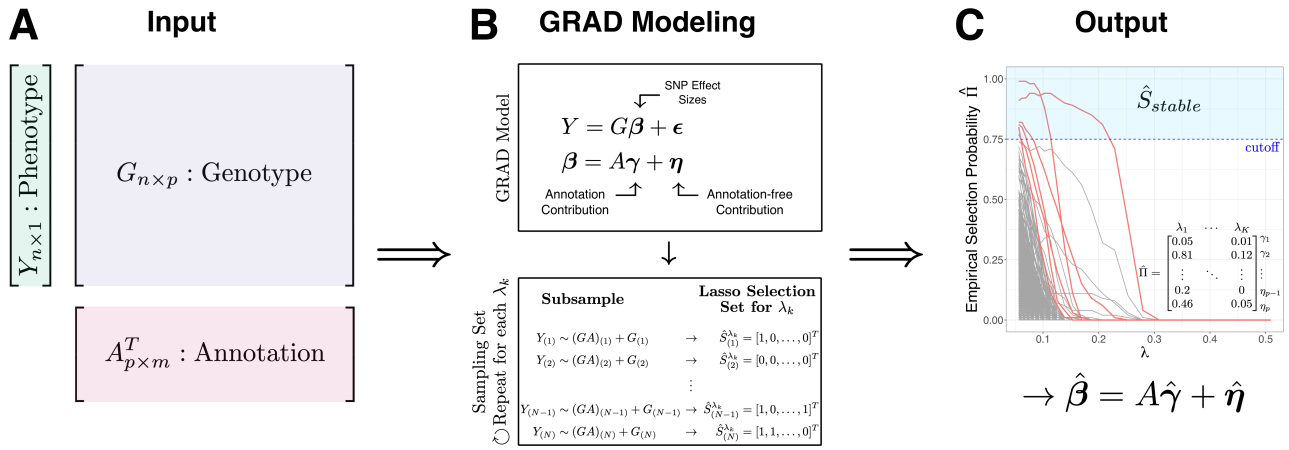

Figure 3.1: **Overview of GRAD modeling framework.**
(A) GRAD input consists of three different types of data: phenotype ($Y_{n\times 1}$), genotype ($G_{n\times p}$), and annotation matrix ($A_{m\times p}$). **(B)** The proposed model partitions SNP effects sizes $\boldsymbol{\beta}$ into an annotation contribution ($A\boldsymbol{\gamma}$) and an annotation-free contribution ($\boldsymbol{\eta}$). Selection of the features ($\boldsymbol{\eta}$ and $\boldsymbol{\gamma}$) is performed with stability selection. For each value of λ_k ($k = 1,\ldots,K$), N subsamples with halves of the observations are followed by lasso to obtain a selection set $\hat{S}^{\lambda_k}_{(N)}$ to be later on aggregated into empirical selection probabilities. **(C)** GRAD output provides results of empirical selection probability for $\boldsymbol{\eta}$ and $\boldsymbol{\gamma}$ from stability selection for each λ_k value. Model parameters with selection probability above a certain cutoff for at least one λ_k are selected ($\hat{S}_{stable}$) and highlighted in red. Estimates for $\hat{\boldsymbol{\eta}}$ and $\hat{\boldsymbol{\gamma}}$ result on selected SNPs $\hat{\boldsymbol{\beta}} = A\hat{\boldsymbol{\gamma}} + \hat{\boldsymbol{\eta}}$.

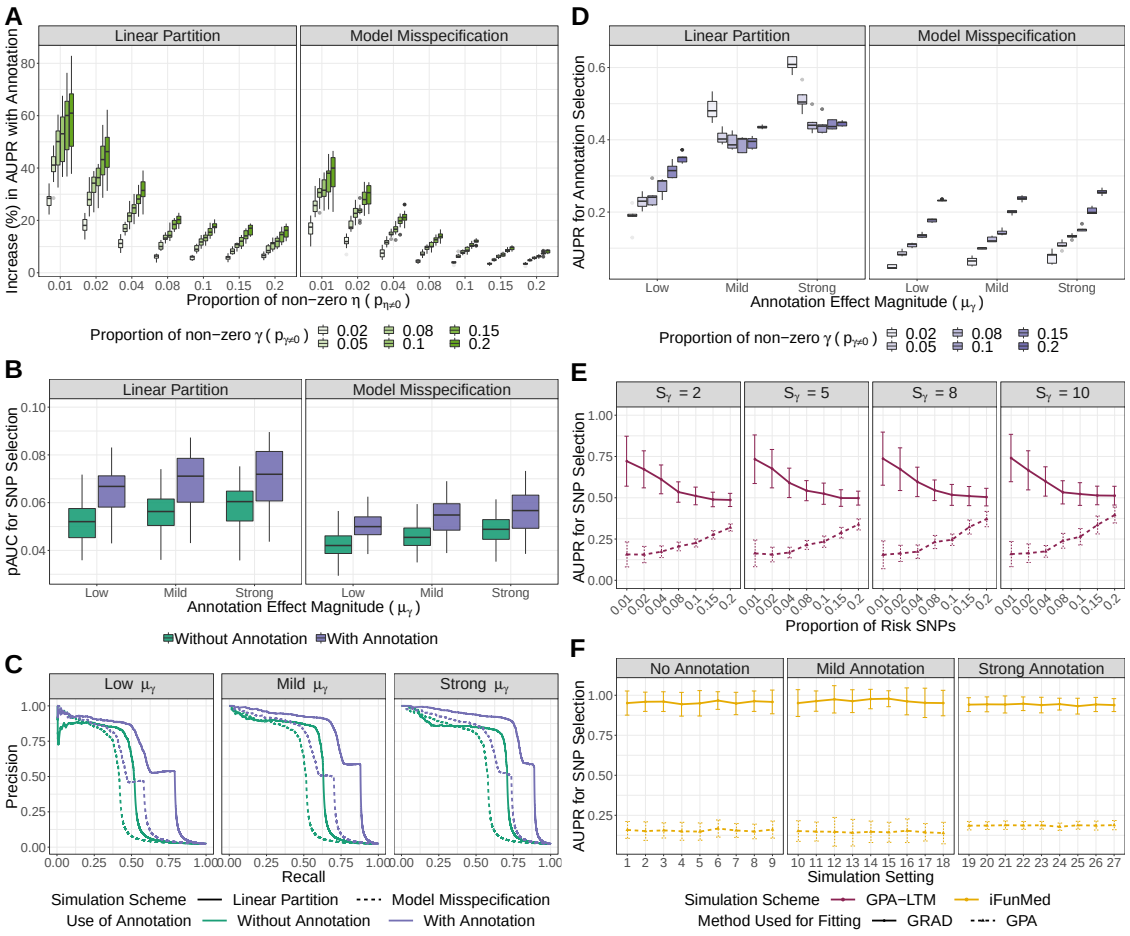

Figure 3.2: **Simulation results comparing fits with and without annotation.** **(A)** Percentage change in the area under the precision-recall curves (AUPR) for SNP selection across fits with the use of annotation comparing simulations generated by linear partition and model misspecification for different proportions of non-zero $\boldsymbol{\eta}$ and $\boldsymbol{\gamma}$ ($p_{\eta\neq 0}$ and $p_{\gamma\neq 0}$, respectively). **(B)** Partial area under the ROC curve (pAUC) for SNP selection for false positive rate below 0.1 for fits with and without annotation comparing simulations generated by linear partition and model misspecification schemes for simulation scenarios with low, mild, and strong annotation effect magnitude μ_γ. **(C)** SNP selection precision-recall curves for fits with and without annotation comparing simulations generated by linear partition and model misspecification schemes for simulation scenarios with $p_{\eta\neq 0} = 0.01$, $p_{\gamma\neq 0} = 0.05$, and $\sigma^2 = 100$ for low, mild, and strong annotation effect magnitude μ_γ. **(D)** Area under the precision-recall curve (AUPR) for annotation selection ($\boldsymbol{\gamma}$) when the proportion of risk SNPs is 0.04 ($p_{\eta\neq 0} = 0.04$) across fits comparing simulations generated by linear partition and model misspecification for different proportions of non-zero $\boldsymbol{\gamma}$ ($p_{\gamma\neq 0}$). **(E, F)** GPA comparisons: average area under the precision-recall curves (AUPR) for SNP selection across 100 simulation replicates and their corresponding error bars (mean $\pm$ standard deviation) for GRAD and GPA. **(E)** Data generated using the GPA liability threshold model (GPA-LTM) . Results are divided by the number of risk annotations $S_\gamma \in \{2, 5, 8, 10\}$. **(F)** Data generated using the *iFunMed* model. Results are divided by their prior inclusion probabilities with the use of annotation (no annotation effect, mild annotation effect, and strong annotation effect).

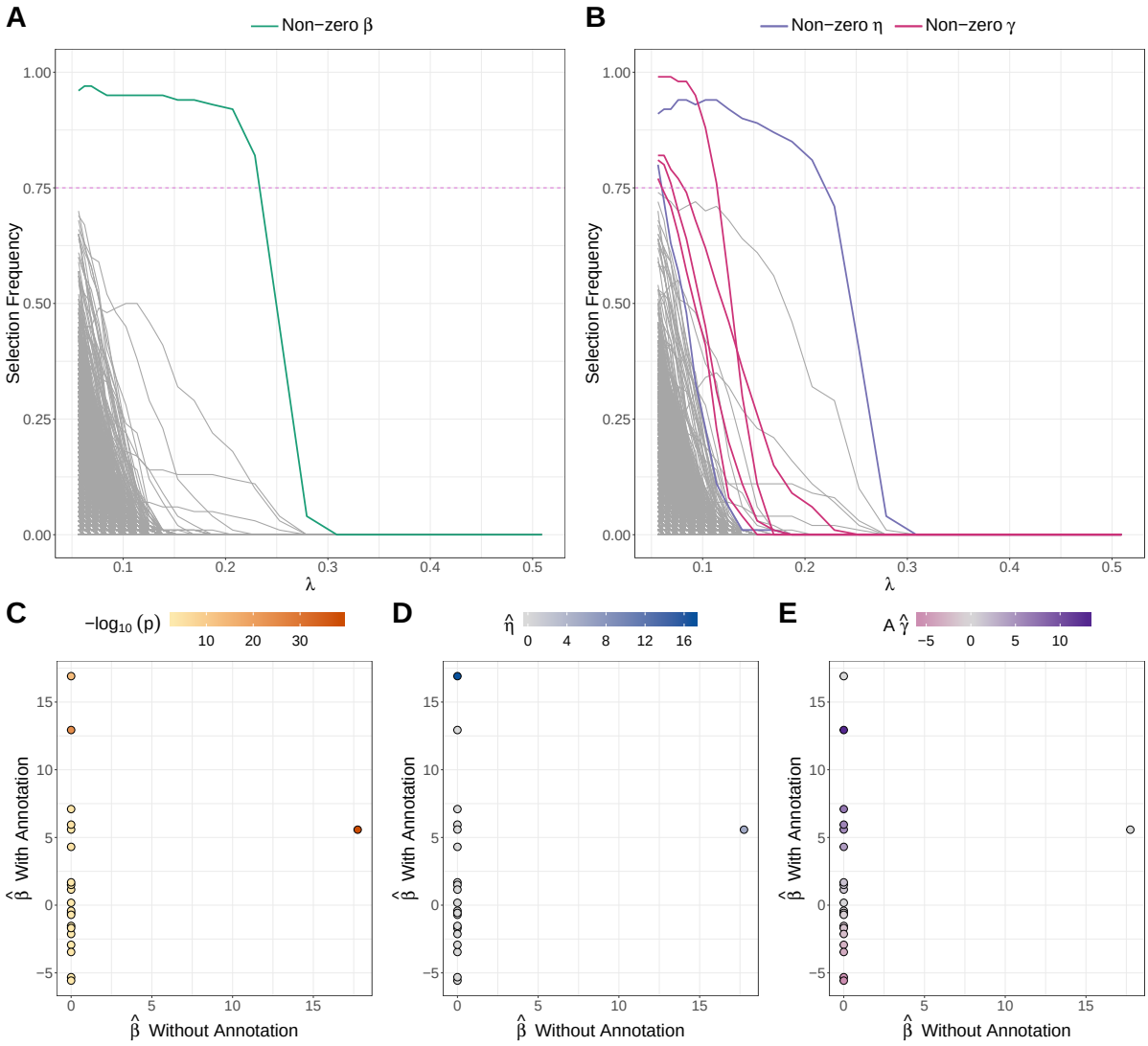

Figure 3.3: **Stability selection results for von Willebrand factor.**
(A, B) Stability paths for each parameter included in the model. Colored paths indicate non-zero estimated parameters. Dashed line represents selection frequency cutoff of 0.75. **(A)** Without annotation and **(B)** with annotation. **(C, E)** Estimated SNP effect sizes across fits with and without annotation. SNPs with effect sizes exactly equal to zero with and without annotation are omitted. SNPs are colored by **(C)** $-\log_{10}$ transformed p-values from univariate GWAS associations, **(D)** strength of the annotation free contribution $\hat{\boldsymbol{\eta}}$ from the model with annotation, and **(E)** strength of the annotation contribution $A\hat{\boldsymbol{\gamma}}$ from the model with annotation.

Table 3.1: Details of cases considered for the analysis

Phenotype	# of Annotations	# of Subjects	Smallest $-\log_{10}(p)$ (chromosome)	Largest $-\log_{10}(p)$ (chromosome)
Factor VII	64	2,162	3.158 (chr12)	39.782 (chr13)
von Willebrand factor	70	2,163	3.164 (chr20)	38.468 (chr9)
Fasting glucose (log)	69	2,070	2.964 (chr2)	7.458 (chr1)
Height	78	2,268	4.150 (chr1)	9.140 (chr16)

Table 3.2: Area under the precision-recall curves (AUPR) stratified by annotation effect magnitude μ_γ (low, mild, and strong) for simulation scenarios displayed in Figure 3.2C ($p_{\eta\neq 0} = 0.01$, $p_{\gamma\neq 0} = 0.05$, and $\sigma^2 = 100$) with and without annotation, and their respective improvements due to annotation.

	Low μ_γ			Mild μ_γ			Strong μ_γ		
	Without Annotation	With Annotation	Improvement (%)	Without Annotation	With Annotation	Improvement (%)	Without Annotation	With Annotation	Improvement (%)
Linear Partition	0.446	0.659	47.824	0.596	0.838	40.641	0.647	0.843	30.318
Model Misspecification	0.440	0.544	23.638	0.463	0.580	25.411	0.541	0.655	21.042

Table 3.3: List of SNPs selected in GRAD for von Willebrand factor. SNP signals refers to the direction of the estimated SNP effect sizes for fits without and with annotation. Details of the annotations included at the individual SNP level are also displayed. Bold SNPs have evidence of association in the GWAS Catalog.

		SNP Signal		Annotation			
SNP ID	Location	wo. Ann	w. Ann	JunD	Pol2	STAT3	c-Myc
rs12565220	chr1:61522880	0	−	0	−	0	−
rs2144555	chr1:101760889	0	+	0	0	+	0
rs12041138	chr1:190149198	0	−	0	0	−	0
rs1257019	chr2:97617140	0	−	0	0	0	−
rs1730122	chr2:97630540	0	−	0	0	0	−
rs4142942	chr3:4841657	0	−	0	0	−	0
rs2280630	chr3:39195964	0	−	0	−	0	0
rs3732610	chr3:124691470	0	−	0	0	0	−
rs261126	chr5:4375160	0	+	0	0	+	0
rs29775	chr5:172483023	0	−	0	0	0	−
rs10984077	chr9:121284915	0	+	0	0	+	0
rs8176749	**chr9:136131188**	**0**	+	**0**	**0**	**0**	**0**
rs505922	**chr9:136149229**	+	+	**0**	**0**	**0**	**0**
rs579459	**chr9:136154168**	**0**	+	**0**	+	**0**	**0**
rs1063857	**chr12:6153514**	**0**	−	**0**	**0**	**0**	−
rs1063856	**chr12:6153534**	**0**	+	**0**	**0**	**0**	+
rs3912393	chr12:94594035	0	−	−	0	0	0
rs8027767	chr15:99297503	0	+	0	0	+	0
rs8041224	chr15:99297665	0	+	0	0	+	0
rs2663849	chr18:55464523	0	+	+	0	0	0
rs1206808	chr20:45688440	0	−	0	0	−	0
rs2154592	chr22:23947352	0	+	+	0	0	0

4 DISCUSSION

4.1 Limitations from Using Summary-level Data

The number of methodologies that use summary-level data has increased in the last couple of years. These tools provide a convenient and practical way to analyze thousands of genetic variants without the hassle of applications to controlled-access repositories, extensive data cleaning and processing, and potential problems from limited computing resources.

Ready to use data generally consist of results from univariate associations where where patient privacy is not violated. The majority of statistical models developed to use summary-level data typically require a combination of p-values, estimated effect sizes, standard deviations, and linkage-disequilibrium (LD). A large proportion of the data applications for these methods use information coming from meta-analyses, which is a common practice to increase sample size where summarized information from different studies is combined. One of the biggest meta-analysis explored data on blood lipids-related phenotypes (HDL, LDL, and triglycerides) data (Teslovich et al., 2010). They combined summary statistics from over 20 different studies and cohorts with individuals of majority European ancestry to reach sample sizes close to 100,000. For treatment of the LD matrix, it is usually approximated using a reference panel, e.g. 1,000 Genomes (Consortium, 2015), with matching ancestry. Both approximations induce potential error from using multiple reference population results that are generated by different scientists with diverse treatment of the data and mismatched LD from the study and the reference. Moreover, al-

though meta-analyses increase sample size, there are still only 2,504 subjects in the data available from 1,000 Genomes. Four hundred sixty five samples are of European ancestry which is used in the majority of larger GWAS studies.

When using the lipids data (Teslovich et al., 2010) and the LD matrix from 465 subjects with European ancestry from 1,000 Genomes, *iFunMed* showed convergence problems, even after strict matching to the reference alleles. The only solution we found was to severely filter by minor-allele frequency which considerably reduced the number of SNPs and resulted in a very small number of discoveries.

To have a better understanding of *iFunMed* operative mechanisms, we inspected the SNPs with the highest univariate associations within each loci for HDL. Our premise was simple: if two SNPs are highly correlated (close to perfect LD) then their summary statistics should be relatively close in magnitude.

Figure 4.1 displays information of two SNPs: *rs12678919* and *rs9600212* (4.1A and 4.1B, respectively). They are located in different regions of the genome and both display high marginal association with HDL. We compare the LD of the specific SNP with the rest of the ones in the locus and their corresponding summary statistics. *rs12678919* (Figure 4.1A) has a t-score of 22.24 and as expected, the overall tendency is increasing and linear. SNPs with elevated summary statistics have stronger correlations with *rs12678919* and weakly correlated SNPs are concentrated in a cloud around zero. There are two SNPs highlighted in the figure that break the linear pattern. Both of them have almost perfect LD with *rs12678919* and yet, their summary statistics magnitude is below 5. More importantly, one of them has opposite direction (red in Figure 4.1A). In Figure 4.1B, *rs9600212*'s t-score

is 15.69 and the tendency is linear but vertical around zero (GWAS summary statistics around zero). The three highlighted SNPs display even weaker marginal associations (close to zero in one case) for high LD with *rs9600212*. These are clear examples of mismatched LD between the study and the reference which will most likely have an impact on different models that use this data. This was not an isolated event, it was observed in multiple loci with varying patterns of off-diagonal SNPs.

As a next step, we intended to reduce the sources of potential error by using summary statistics from only one study (FHS) and approximate the LD matrix. Many SNP pairs have small LD values (close to zero) but opposite magnitudes when comparing FHS with the reference panel. Our results with this input data were much more stable but there were still very few discoveries that fluctuated between zero to two for regions with 1,000 to 2,000 SNPs and high univariate summary statistics.

The success of these models that use summary-level data is undeniable. Strong examples are LD Score Regression (Finucane et al., 2015) used to partition heritability and TWAS (Gusev et al., 2016) used to find gene-trait associations. That being said, there are still few studies that measure the impact of different sources of approximation have on model estimates and results. The influence and the effect on the results will vary across methodologies and could potentially lead to elevated numbers of false discoveries that are not accounted for. One alternative proposed by (Zhu and Stephens, 2017) is to use a shrinkage estimator on the reference panel LD matrix. As scientists, we should promote such inspections when approximations are being used. More importantly, there is an increasing need to not only share

summary-level data but to also share LD structures.

4.2 Error Control

Traditional approaches to control for the multiple comparisons problem that arises from univariate associations threshold p-values by 1×10^{-8}. This is based on a conservative Bonferroni correction for one million *independent* variants that controls familywise error rate (FWER) at $\leqslant \alpha = 0.01$. Due to high correlations along the human genome, the key independence assumption is violated. Regions with high and specific patterns of linkage-disequilibrium, i.e. LD blocks, might generate an elevated, and not accounted for, number of false negatives which can decrease power.

Majority of methods that leverage functional annotation data report as a final output some sort of posterior inclusion probability, i.e. the posterior probability for each SNP of being non-zero. With a lack of p-values, posterior probabilities are usually thresholded based on their distributions, calibrations of the values by using independent datasets (Pickrell, 2014), or utility functions (Kichaev et al., 2014). For *iFunMed*, we used a 0.5 threshold. This value was based on our observations of simulation results and real data applications. Our posterior probabilities behaved as bimodal around zero and one with few to none values in the middle. We recommend to proceed with care with SNPs around 0.5 as they could easily lead to false positives or false negatives.

If we expect a reasonable proportion of rejections, i.e., SNPs with an association

to the trait, it might be more appropriate to control the false discovery rate (FDR). Methodologies that assume independence among SNPs (Chung et al., 2014), can use a *direct posterior probability approach* (Newton et al., 2004) to control global false discovery rates for Bayesian hierarchical models. There is another class of methods within this realm that partition the genome into blocks that are assumed to be independent between each other but not within. They control FDR at the locus-level to identify loci of interest by calculating a common posterior probability within each locus (Wen, 2016; Wen et al., 2016) to be followed up by the *direct posterior probability approach* (Newton et al., 2004). This technique seems appropriate, compared to others, because it takes information that we know about the human genome and GWAS studies and treats each locus as the unit in the analysis, but does not guarantee FDR control within each locus.

Unlike others, GRAD assumes that the SNP effect sizes can be linearly partitioned into an annotation and annotation-free contribution. By adopting the Lasso within stability selection for feature selection, we provide per-comparison error rate (PCER) control that showed promising results within our analyses. A technique that aims for FDR control within the Lasso called SLOPE (Bogdan et al., 2015) has a strict orthogonal restriction for the design matrix. When used in GWAS data (Brzyski et al., 2017), the genotype matrix follows strict procedures to pre-select SNPs. Further inspections of this technique are part of our current plan.

4.3 GWAS Advances and Future Directions

After more than a decade of GWAS and thousands of associated SNPs, our understanding of complex diseases keeps improving, but there is still significant progress to be made. Early-stage analyses only focus on univariate associations with obsolete thresholds to account for multiple testing. Such an approach overlooks the potential gain from incorporating biological knowledge and the joint contribution of SNPs on a trait. Many statistical advances are aiming to incorporate auxiliary data and biological processes to better characterize genetic variants. These are extremely relevant because not only they refine and improve detection but they also take into account biological mechanisms of association by, for example, adopting multivariate mediation models like *iFunMed*. Methodologies that use summarizy-level data usually complement it with LD structures and annotation information (Chen et al., 2016; Kichaev et al., 2014; Rojo et al., 2019). This is important to characterize joint effects and boost signals of SNPs in regulatory regions.

Methods that impose strict assumptions on the data such as independence (Chung et al., 2014), one causal variant per locus (Li and Kellis, 2016; Pickrell, 2014), or non-overlapping annotations (Yang et al., 2017) are typically computationally-efficient and they could provide a good first exploration of the data but their unrealistic assumptions might hinder true discoveries. With increasing advances in GWAS data, the need for flexible models will become greater. The majority of existing studies mainly focus on European populations and have modest sample sizes. For example, after filtering, we only have a little over 2,000 individuals in the FHS that varies for different phenotypes after removing missing data. We believe

that with larger sample sizes, both the proposed models will benefit. *iFunMed*'s input summary statistics will be more stable and less dependent on specific subjects with rare variants, especially if the LD is approximated with a reference panel, without sacrificing computational time once summary statistics are calculated. GRAD's computational cost will be greater compared to *iFunMed* but it will only depend on the computation of X^TX and X^TY within each round of the Lasso.

Also, it is important to point out that we are limited to the epigenomic information that is available to us and our discoveries might be related to other features within the human genome. With the development of software that predict signed effects of a SNP on transcription factor binding (Alipanahi et al., 2015; Kelley et al., 2016; Zhou and Troyanskaya, 2015) and higher quality annotations, the demand for models that can integrate continuous auxiliary data in an efficient way (Chen et al., 2016; Rojo et al., 2020) will increase. Annotation sources can also be extended to single-cell data that hasn't been explored for this context. For example, we could construct annotations from ATAC-seq data and summarize weather or not SNPs overlap with a peak for different cell-types or even accessibility quantifications of the SNP region within a cell-type.

When more ethnic groups are collected, there will be a need to adapt current methods or create new ones that account for population structure while utilizing annotation information. With the flexibility and increasing advancements on the Lasso, an extension of GRAD to a mixed-effect model to correct for the grouping structure could be possible (Schelldorfer et al., 2011).

As for now, GWAS seems like only the beginning and the tip of the iceberg. Once

there are more advances in technology, study design, sample sizes, and epigenomic information the potential discoveries and possibilities could be endless.

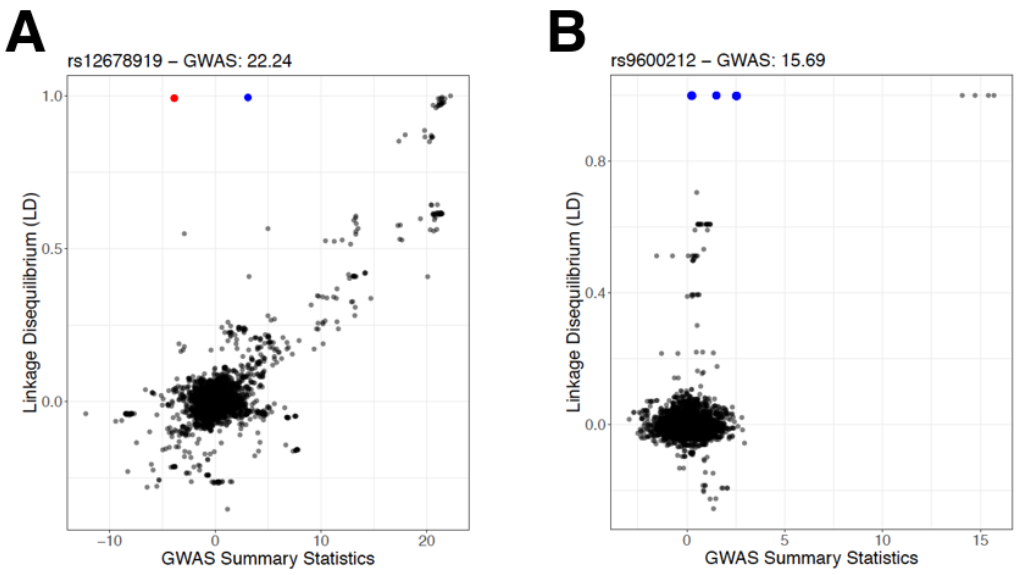

Figure 4.1: **Pairwise LD versus their corresponding univariate GWAS summary statistics for HDL (Teslovich et al., 2010) for two SNPs with high marginal associations using a European ancestry reference panel for LD computations.** **(A)** *rs12678919* with a GWAS summary statistics of 22.24 and **(B)** *rs9600212* with a GWAS summary statistics of 15.69

A APPENDIX A

A.1 Fitting *iFunMed* with Variational EM

In this appendix, we provide details on the derivation of the variational EM algorithm for the *iFunMed* model.

Variational E-Step

The variational algorithm approximates the joint posterior distribution by a product of lower dimensional functions using factorized distributions as

$$q(\boldsymbol{\tau}_\beta, \mathbf{s}_\beta, \delta) = q(\delta)q(\boldsymbol{\tau}_\beta)\prod_{j=1}^{p} q(s_{\beta,j}) \quad \text{and} \quad q(\boldsymbol{\tau}_B, \mathbf{s}_B) = q(\boldsymbol{\tau}_B)\prod_{j=1}^{p} q(s_{B,j}),$$

where $q(\cdot)$ is an arbitrary density function that generates a $q-$dependent lower bound on the marginal likelihood by minimizing the Kullback-Leibler divergence between the posterior density and q. It considers the expectation with respect to the parameters in their corresponding factorized form of the full log posterior distribution for minimizing. In what follows, $E_{a,-b}(L)$ represents the expectation of L over distribution a, excluding the distribution of variable b, and $q^{(t)}(\cdot)$ is the current variational estimate of the posterior distributions of each component. Moreover, the parameters $(\sigma_\epsilon^2, \sigma_\eta^2, \nu_\beta, \nu_B, \boldsymbol{\gamma}_\beta^T, \boldsymbol{\gamma}_B^T)$ are in fact estimates from the previous iteration, but we will drop the superscript (t) for the ease of notation. Next, we layout the updating steps of each component, which will iterate until convergence. For the DEM, we update $\boldsymbol{\tau}_\beta$, δ, and $\mathbf{s}_\beta$. Let $\mathbf{w}_\beta^{(t)} = \left(w_{\beta,1}^{(t)}, \ldots, w_{\beta,p}^{(t)}\right)^T$

for $j = 1, \ldots, p$ denote the posterior inclusion probability of the elements of $\boldsymbol{\beta}$ from the last iteration, and $\mathbf{W}_\beta^{(t)} = \text{diag}\left(\mathbf{w}_\beta^{(t)}\right)$. The resulting posterior distribution of $\boldsymbol{\tau}_\beta$ from $E_{q^{(t)},-\tau_\beta}(L_{DEM})$ follows a normal distribution

$$\boldsymbol{\tau}_\beta | q^{(t)} \sim N(\boldsymbol{\mu}_{\tau_\beta}, \mathbf{V}_{\tau_\beta}),$$

where $\boldsymbol{\mu}_{\tau_\beta} = \mathbf{K}_\beta^{-1}\mathbf{W}_\beta^{(t)}(\mathbf{Z}_Y - \mu_\delta \mathbf{Z}_G)$, $\mathbf{V}_{\tau_\beta} = \sigma_\epsilon^2 \mathbf{K}_\beta^{-1}$, and $\mathbf{K}_\beta = \mathbf{W}_\beta^{(t)}\tilde{\boldsymbol{\Sigma}}\mathbf{W}_\beta^{(t)} + \text{diag}(\tilde{\boldsymbol{\Sigma}})\left[\mathbf{W}_B^{(t)} - \left(\mathbf{W}_B^{(t)}\right)^2\right] + \nu_\beta^{-1}\mathbf{I}_p$. Here, μ_δ is the posterior mean of δ from the last iteration and $\mathbf{K}_\beta$ is a $p \times p$ matrix. Next, from the computation of $E_{q^{(t)},-\delta}(L_{DEM})$, we update the variational posterior distribution of δ as

$$\delta \sim N(\mu_\delta, \sigma_\delta^2) \text{ where } \sigma_\delta^2 = \frac{\sigma_\epsilon^2}{\mathbf{Z}_G^T \tilde{\boldsymbol{\Sigma}}^{-1} \mathbf{Z}_G} \text{ and } \mu_\delta = \frac{\mathbf{Z}_G^T \tilde{\boldsymbol{\Sigma}}^{-1} \mathbf{Z}_Y - \mathbf{Z}_G^T \mathbf{W}_\beta^{(t)} \boldsymbol{\mu}_{\tau_\beta}}{\mathbf{Z}_G^T \tilde{\boldsymbol{\Sigma}}^{-1} \mathbf{Z}_G}.$$

Finally, the variational posterior distribution of the variable $s_{\beta,j}$ derived from $E_{q^{(t)},-s_{\beta,j}}(L_{DEM})$ is a Bernoulli distribution with $P(s_{\beta,j} = 1) \equiv w_{\beta,j}^{(t+1)}$, where

$$\begin{aligned} \text{logit}\left(w_{\beta,j}^{(t+1)}\right) &= \text{logit}(\pi_{\beta,j}) - \frac{1}{2\sigma_\epsilon^2}\left\{\tilde{\Sigma}_{j,j}\left(\mu_{\tau_\beta,j}^2 + V_{\tau_\beta,j,j}\right) - 2\mu_{\tau_\beta,j}\left[Z_{Y,j} - \mu_\delta Z_{G,j}\right.\right. \\ &\quad \left.\left. - \tilde{\boldsymbol{\Sigma}}_{j,-j}\left(\boldsymbol{\mu}_{\tau_\beta,-j} \circ \mathbf{w}_{\beta,-j}^{(t)}\right)\right] + 2\left(\tilde{\boldsymbol{\Sigma}}_{j,-j} \circ \mathbf{V}_{\tau_\beta,j,-j}\right)\mathbf{w}_{\beta,-j}^{(t)}\right\}. \end{aligned}$$

For the GEM, we update $\boldsymbol{\tau}_B$ and $\mathbf{s}_B$. Similar to the quantities in the DEM, let $\mathbf{w}_B^{(t)} = \left(w_{B,1}^{(t)}, \ldots, w_{B,p}^{(t)}\right)^T$ for $j = 1, \ldots, p$ denote the posterior inclusion probability of the elements of $\mathbf{B}$ from the last iteration, and $\mathbf{W}_B^{(t)} = \text{diag}\left(\mathbf{w}_B^{(t)}\right)$. We first update the posterior distribution of $\boldsymbol{\tau}_B$ by computing $E_{q^{(t)},-\tau_B}(L_{GEM})$, which follows a

normal variational posterior

$$\boldsymbol{\tau}_B|q^{(t)} \sim N(\boldsymbol{\mu}_{\tau_B}, \mathbf{V}_{\tau_B}),$$

where $\boldsymbol{\mu}_{\tau_B} = \mathbf{K}_B^{-1}\mathbf{W}_B^{(t)}\mathbf{Z}_G$, $\mathbf{V}_{\tau_B} = \sigma_\eta^2\mathbf{K}_B^{-1}$, and $\mathbf{K}_B = \mathbf{W}_B^{(t)}\tilde{\boldsymbol{\Sigma}}\mathbf{W}_B^{(t)} + \text{diag}(\tilde{\boldsymbol{\Sigma}})\left[\mathbf{W}_B^{(t)} - \left(\mathbf{W}_B^{(t)}\right)^2\right] + \nu_B^{-1}\mathbf{I}_p$ is a $p \times p$ matrix. From $E_{q^{(t)},-s_{B,j}}(L_{GEM})$, we update the variational posterior distribution of the signal $s_{B,j}$ which is a Bernoulli distribution with $P(s_{B,j} = 1) \equiv w_{B,j}^{(t+1)}$, where

$$\begin{aligned}\text{logit}\left(w_{B,j}^{(t+1)}\right) &= \text{logit}(\pi_{B,j}) - \frac{1}{2\sigma_\eta^2}\left\{\tilde{\Sigma}_{j,j}\left(\mu_{\tau_B,j}^2 + V_{\tau_B,j,j}\right) - 2\mu_{\tau_B,j}\left[Z_{G,j}\right.\right. \\ &\quad \left.\left. -\tilde{\boldsymbol{\Sigma}}_{j,-j}\left(\boldsymbol{\mu}_{\tau_B,-j} \circ \mathbf{w}_{B,-j}^{(t)}\right)\right] + 2(\tilde{\boldsymbol{\Sigma}}_{j,-j} \circ \mathbf{V}_{\tau_B,j,-j})\mathbf{w}_{B,-j}^{(t)}\right\}.\end{aligned}$$

Variational M-Step

Following the variational E-step, we obtain point estimates of the hyperparameters in the variational M-step as:

$$\begin{aligned}\sigma_\eta^2 &= \frac{1}{p - |\mathbf{w}_B|_1 - 1}\left\{\left(\boldsymbol{\mu}_{\tau_B} \circ \mathbf{w}_B\right)^T\tilde{\boldsymbol{\Sigma}}\left(\boldsymbol{\mu}_{\tau_B} \circ \mathbf{w}_B\right) - 2\mathbf{Z}_G^T\left(\boldsymbol{\mu}_{\tau_B} \circ \mathbf{w}_B\right) + \mathbf{Z}_G^T\tilde{\boldsymbol{\Sigma}}^{-1}\mathbf{Z}_G\right. \\ &\quad \left. + \mathbf{w}_B^T\left(\tilde{\boldsymbol{\Sigma}} \circ \mathbf{V}_{\tau_B}\right)\mathbf{w}_B + \text{diag}(\tilde{\boldsymbol{\Sigma}})^T\left(\mathbf{W}_B - \mathbf{W}_B^2\right)\left[\boldsymbol{\mu}_{\tau_B}^2 + \text{diag}\left(\mathbf{V}_{\tau_B}\right)\right]\right\}, \\ \nu_B &= \frac{\sum_{j=1}^p w_{B,j}\left(\mu_{\tau_{B,j}}^2 + V_{\tau_B,j,j}\right)}{\sigma_\eta^2|\mathbf{w}_B|_1},\end{aligned}$$

where $|\mathbf{x}|_1$ denotes number of non-zero elements of $\mathbf{x}$. We update $\boldsymbol{\gamma}_B$ by maximizing

$$Q\left(\boldsymbol{\gamma}_B\,|q^{(t+1)}\right) = \sum_{j=1}^p\left[w_{B,j}\mathbf{A}_j^T\boldsymbol{\gamma}_B - \log\left(1 + \exp\left(\mathbf{A}_j^T\boldsymbol{\gamma}_B\right)\right)\right].$$

Similarly, we get point estimates of σ_ϵ^2, ν_β and $\boldsymbol{\gamma}_\beta$ as follows:

$$\begin{aligned}
\sigma_\epsilon^2 &= \frac{1}{p-|\mathbf{w}_\beta|_1}\left\{\left(\boldsymbol{\mu}_{\tau_\beta}\circ\mathbf{w}_\beta\right)^T\tilde{\boldsymbol{\Sigma}}\left(\boldsymbol{\mu}_{\tau_\beta}\circ\mathbf{w}_\beta\right)-2\mathbf{Z}_Y^T\left(\boldsymbol{\mu}_{\tau_\beta}\circ\mathbf{w}_\beta\right)+\mathbf{Z}_Y^T\tilde{\boldsymbol{\Sigma}}^{-1}\mathbf{Z}_Y\right.\\
&\quad+\mathbf{w}_\beta^T\left(\tilde{\boldsymbol{\Sigma}}\circ\mathbf{V}_{\tau_\beta}\right)\mathbf{w}_\beta+\operatorname{diag}\left(\tilde{\boldsymbol{\Sigma}}\right)^T\left(\mathbf{W}_\beta-\mathbf{W}_\beta^2\right)\left[\boldsymbol{\mu}_{\tau_\beta}^2+\operatorname{diag}\left(\mathbf{V}_{\tau_\beta}\right)\right]\\
&\quad\left.+\left(\mu_\delta^2+\sigma_\delta^2\right)\mathbf{Z}_G^T\tilde{\boldsymbol{\Sigma}}^{-1}\mathbf{Z}_G+2\mu_\delta\mathbf{Z}_G^T\mathbf{W}_\beta\boldsymbol{\mu}_{\tau_\beta}-2\mu_\delta\mathbf{Z}_G^T\tilde{\boldsymbol{\Sigma}}^{-1}\mathbf{Z}_Y\right\},\\
\nu_\beta &= \frac{\sum_{j=1}^p w_{\beta,j}^{(t+1)}\left(\mu_{\tau_{\beta,j}}^2+V_{\tau_\beta,j,j}\right)}{\sigma_\epsilon^2|\mathbf{w}_\beta|_1}.
\end{aligned}$$

We update $\boldsymbol{\gamma}_B$ by maximizing

$$Q\left(\boldsymbol{\gamma}_\beta\,|q^{(t+1)}\right)=\sum_{j=1}^p\left[w_{\beta,j}\mathbf{A}_j^T\boldsymbol{\gamma}_\beta-\log\left(1+\exp\left(\mathbf{A}_j^T\boldsymbol{\gamma}_\beta\right)\right)\right].$$

A.2 Pre-processing of Framingham Heart Study Data

Genotypes, expression levels, and phenotypes were acquired from the Framingham Heart Study (FHS) using project number 8158 and dbGaP study accession phs000007. Genotypes were obtained from the SHARe substudy (phs000342) that used the Illumina HumanOmni5M-4v1 array for genome-wide genotyping array of 4,271,233 SNPs from the human genome version GRCh37 (hg19). FHS included expression data from whole blood RNA for different cohorts from the Systems Approach to Biomarker Research in Cardiovascular Disease (SABRe CVD) study (phs000363) that utilized Affymetrix Human Exon 1.0 ST Array. 284,558 core probe sets (exons) were annotated using the Affymetrix annotation file resulting in 17,873 hg18 transcripts, 15,004 of which were successfully mapped to hg19.

A total of 1,667 subjects had both expression and genotype data. All of these subjects are from the offspring cohort and their phenotype information was acquired from clinical exam 5 in the case of factor VII and von Willebrand factor, exam 7 for fasting glucose, exam 8 for HDL, and exam 9 for red and white blood cell count. Based on initial exploratory analysis, we log transformed fasting glucose, HDL, and white blood cell count measurements.

Preprocessing of the genotypes was performed with PLINK v1.9 (Chang et al., 2015; Purcell et al., 2007) and the SNPs were filtered following the guidelines in (Roshyara et al., 2014). SNPs with call rates $\leqslant$ 95%, that were discordant with Hardy-Weinberg equilibrium (HWE p-value $\leqslant 10^{-6}$), and with minor allele frequency (MAF) small than 1% were filtered. After removing non-autosomal chromosomes, indel and repeated SNPs, a total of 2,478,340 SNPs remained for imputation. IMPUTE2 v2.3.2 (Marchini and Howie, 2010; Marchini et al., 2007) with one phased reference panel from 1,000 Genomes (phase 3) and a probability of 0.9 as threshold for calling genotypes resulted in 2,244,466 SNPs. Genotypes were recoded to an additive format (0/1/2) using the `--recode A` option from PLINK.

A.3 Procedure for Identifying Candidate Mediators

In order to define potential mediator genes, we followed the guidelines provided in (Baron and Kenny, 1986) with some modifications to adapt to the fact that we consider multiple SNPs at the same time. Specifically, we applied the following three steps:

1. *Identify SNPs significantly associated with the phenotype.* We assessed univariate association between the phenotype of interest and the genotype as $\mathbf{Y} \sim \mathrm{SNP}_j, \;\; j = 1, \ldots p$. This translated to considering regions of the genome that showed moderate to high GWAS signal (at least one SNP with $-\log_{10}(\text{p-value}) \leqslant 4$). We formed windows around such SNPs to consider regions of size 2 Mb approximately.

2. *Identify potential mediator variables significantly associated with the phenotype.* For each gene within regions from step 1, we calculated univariate association between the expression of the gene and the SNPs in that region as $\mathbf{G} \sim \mathrm{SNP}_j, \;\; j = 1, \ldots p$. We reduced the set of candidate genes by considering only those that showed high and dense signal (at least a couple of SNPs with $-\log_{10}(\text{p-value}) \leqslant 8$).

3. *Identify potential mediator variables that significantly associate with the phenotype adjusted for the genotype effect*. For the genes with eQTL signal from step 2, we fitted $\mathbf{Y} \sim \mathrm{SNP}_j + \mathbf{G}, \;\; j = 1, \ldots p$ and required that $\mathbf{G}$ remained significant, for at least one SNP within the region.

Finally, we remark that the final selection of mediator genes was further subjected to visual inspection by paying attention to cases that were at the boundary but didn't pass the threshold from step 2 and setting a more liberal significance level in step 3, if necessary. We further note that majority of the candidate mediator genes arising from this procedure (Table 2.2) also exhibited gene-trait associations based on TWAS results of (Gusev et al., 2016) available at `http://twas-hub.org/`.

A.4 Supplementary Figures for "iFunMed: Integrative Functional Mediation Analysis of GWAS and eQTL Studies"

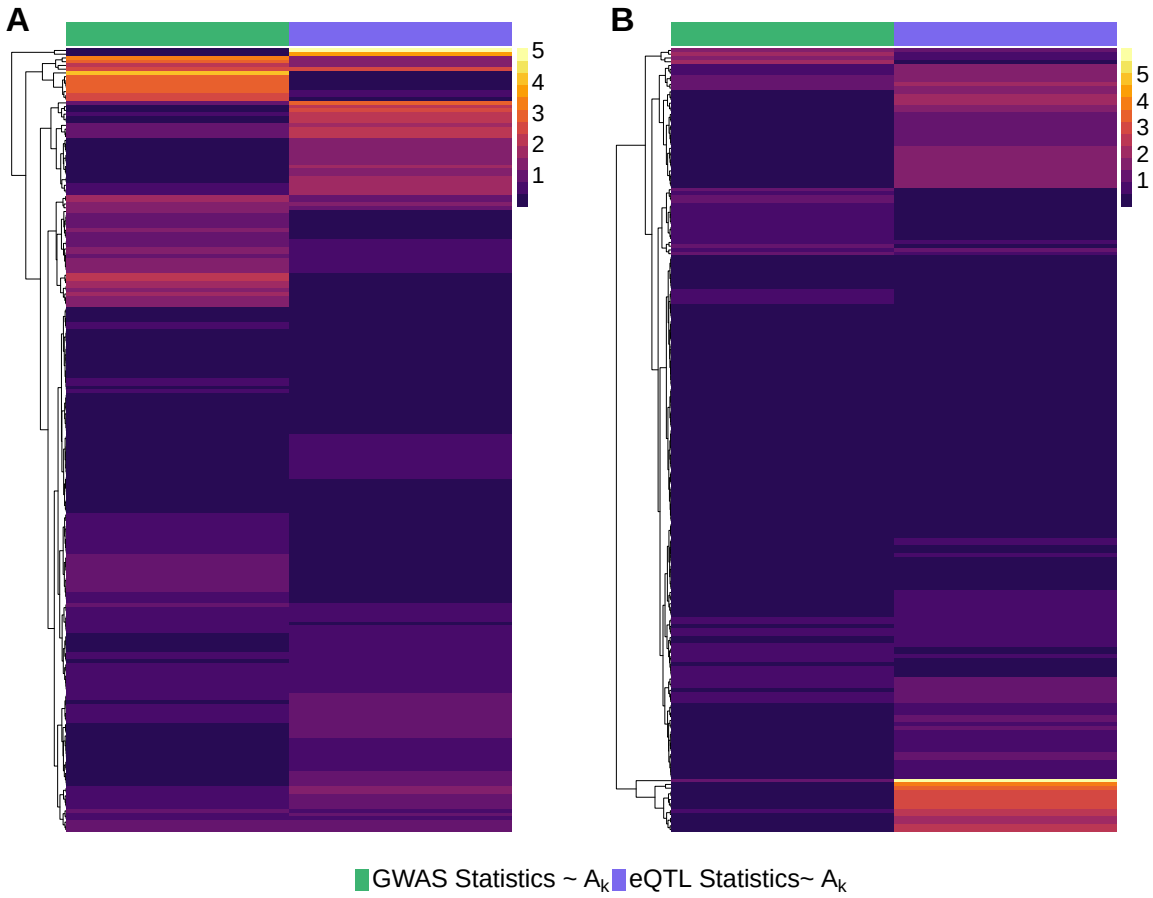

Figure A.1: **Heatmaps for $-\log_{10}$ transformed p-values from the univariate association analysis of GWAS and eQTL summary statistics with individual annotations.**
Rows depict a list of 209 epigenomic annotations from 4 activation histone marks from the Roadmap Epigenomic Project (Roadmap Epigenomics Consortium, 2015). Left column for each panel corresponds to p-values ($-\log_{10}$ transformed) from univariate association analysis of GWAS summary statistics and individual annotations, i.e., $\mathbf{Z}_Y \sim \mathbf{A}_k$, and right column to univariate association analysis of eQTL summary statistics and individual annotations, i.e., $\mathbf{Z}_G \sim \mathbf{A}_k$ ($k = 1, \ldots, 209$). Results depicted are for Red Blood Cell Count as phenotype. **(A)** *NINJ1* as mediator and **(B)** *EVA1C* as mediator.

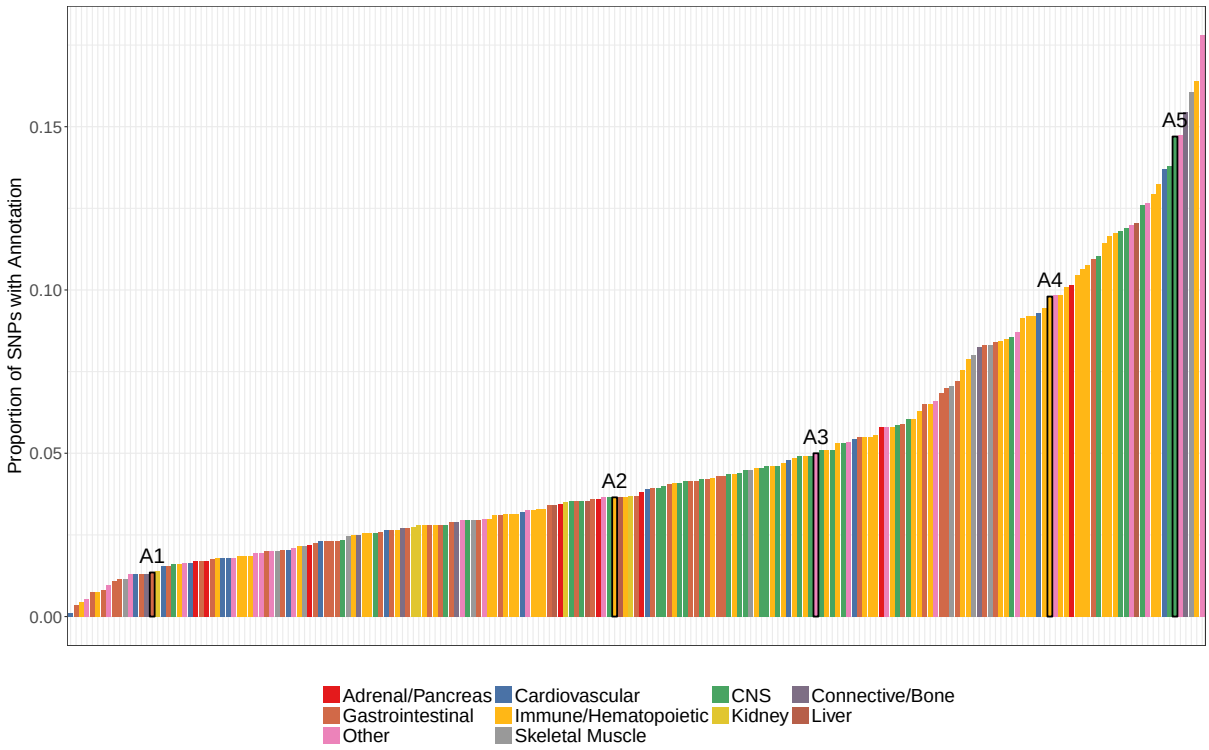

Figure A.2: **Proportion of SNPs with annotations.**
Proportion of SNPs with annotations, i.e., with corresponding entry of the **A** matrix equal to 1, across the 209 annotations considered. Annotations used in the simulations are boxed in black with their corresponding labels.

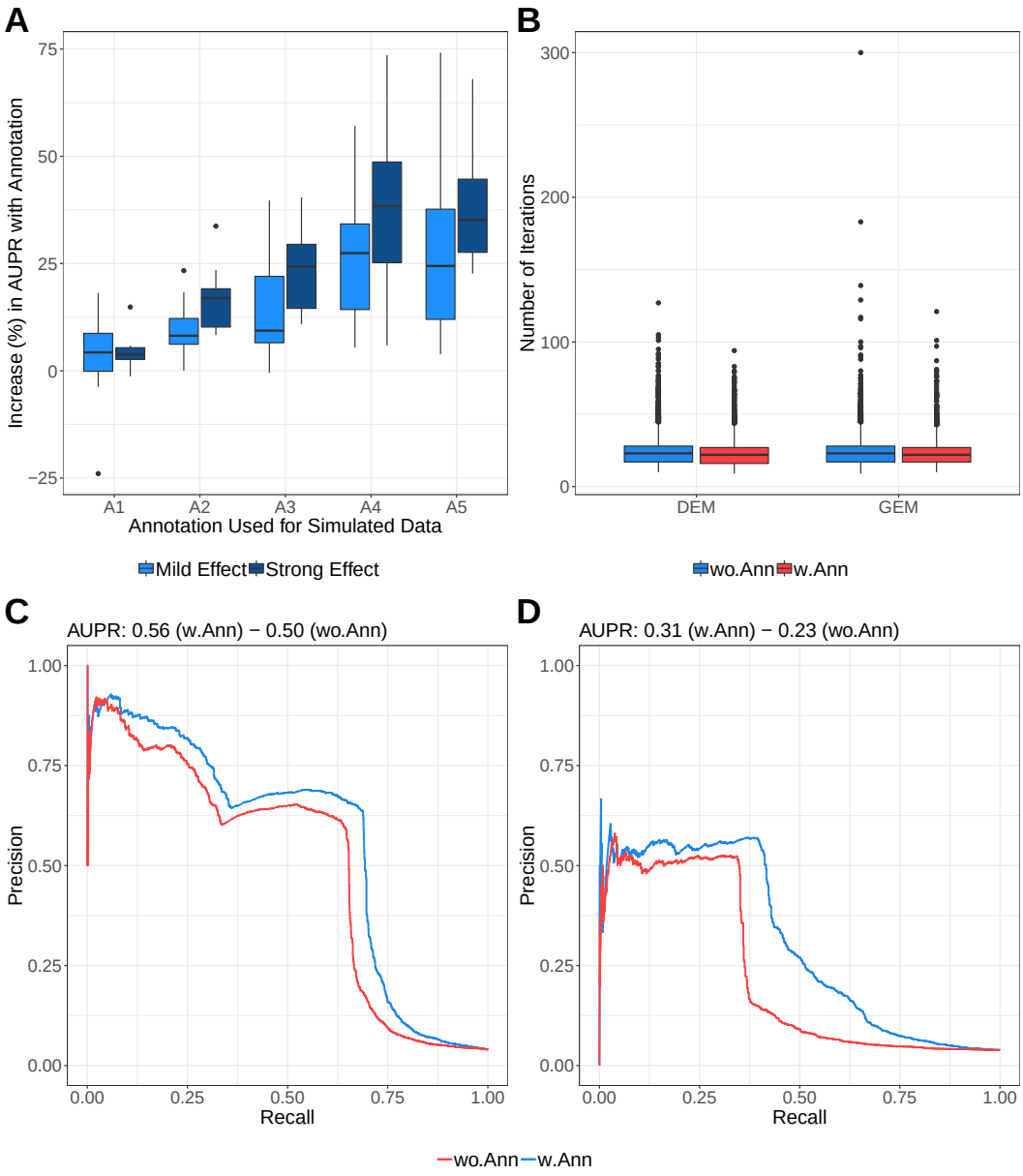

Figure A.3: **Simulations for comparing *iFunMed* fits with (w.Anno) and without annotation data (wo.Anno).**
(A) Percentage change in the area under the precision-recall (AUPR) curves with the use of annotation across fits for all the 54 simulation settings. The total set of annotations (54×5 settings) are stratified by the annotation effect sizes γ_β and γ_B. PR curves are obtained by thresholding the total effect estimates. **(B)** Boxplots of numbers of iterations until convergence across simulation replicates. **(C, D)** PR curves for simulation scenarios with a mild annotation effect ($\gamma_\beta, \gamma_B = (-4.5, 2)$), $\sigma^2_\epsilon = \sigma^2_\eta = 1$ and $\delta = 0.05$, using annotation A5, and varying effect size variances. **(C)** $\nu_\beta = \nu_B = 100$ for strong and **(D)** $\nu_\beta = \nu_B = 20$ for weak effect sizes of the SNPs.

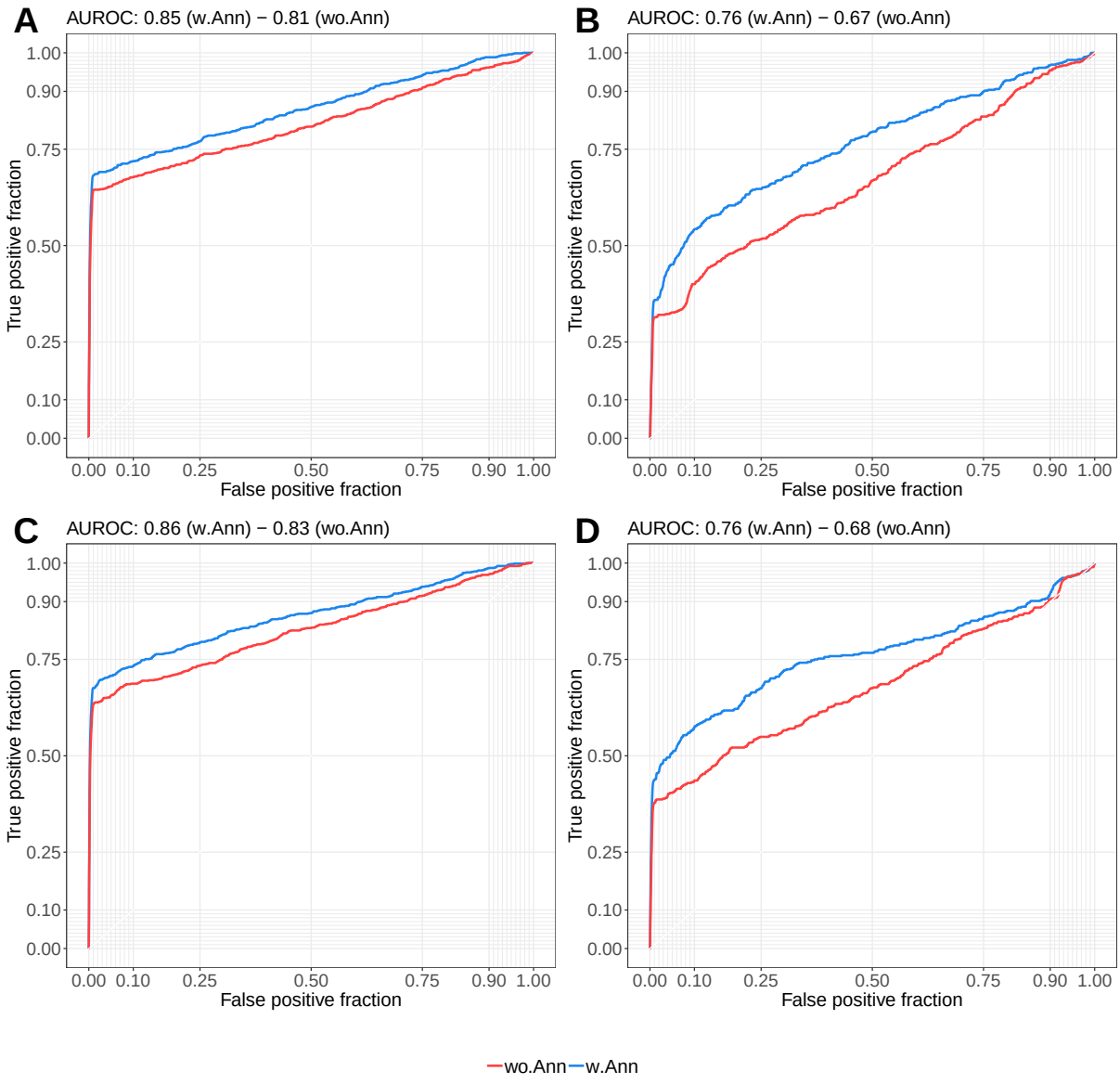

Figure A.4: **Area under the DEM and GEM ROC curves from simulation settings with a mild annotation effect ($\gamma_\beta, \gamma_B = (-4.5, 2)$), $\sigma^2_\epsilon = \sigma^2_\eta = 1$ and $\delta = 0.05$, using annotation A5, and varying effect size variances.**
(A, B) ROC curves for the direct effect model (DEM). **(C, D)** ROC curves for the gene effect model (GEM). **(A, C)** $\nu_\beta = \nu_B = 100$. **(B, D)** $\nu_\beta = \nu_B = 20$.

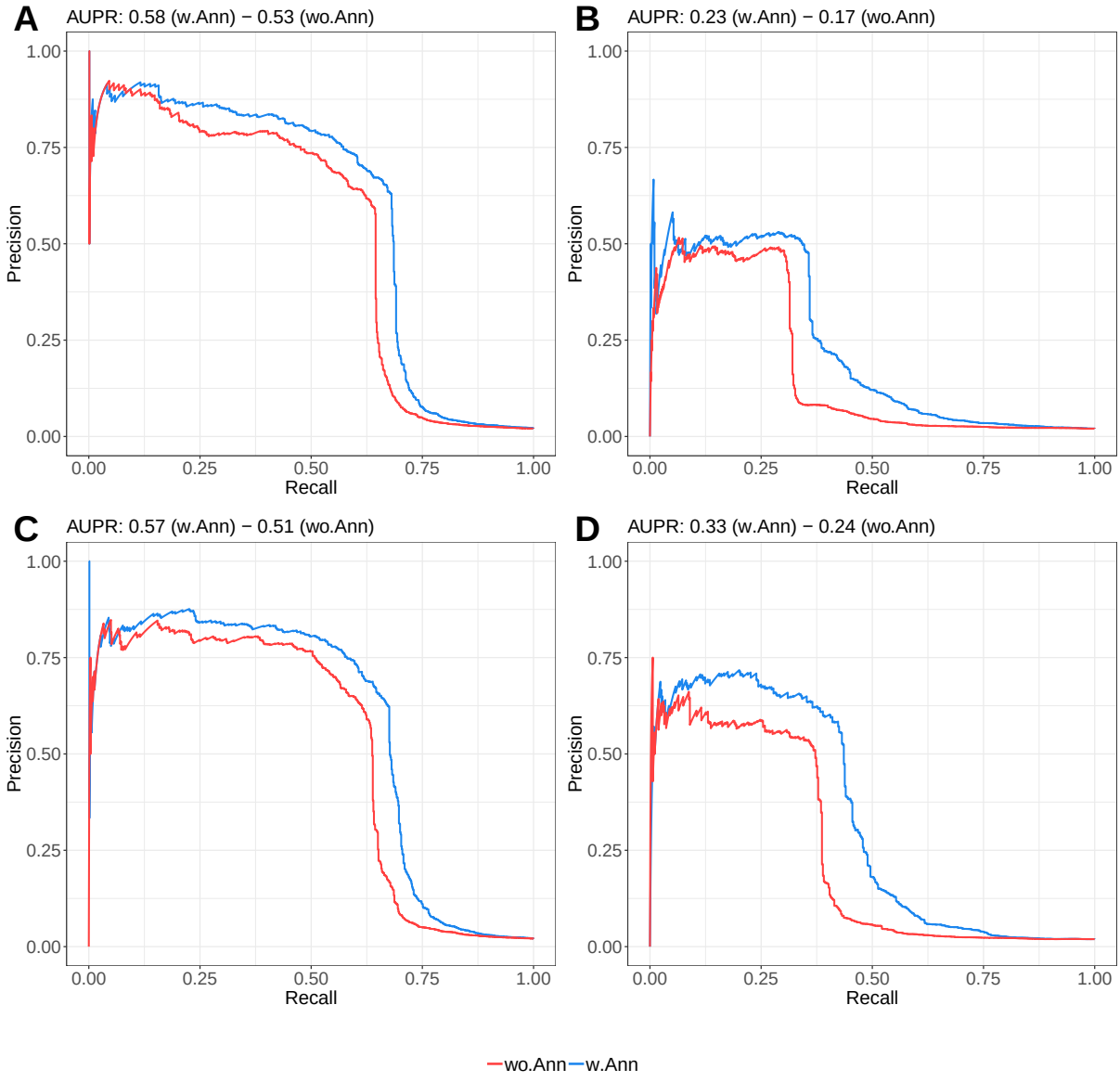

Figure A.5: **Area under the DEM and GEM PR curves from simulation settings with a mild annotation effect ($\gamma_\beta, \gamma_B = (-4.5, 2)$), $\sigma^2_\epsilon = \sigma^2_\eta = 1$ and $\delta = 0.05$, using annotation A5, and varying effect size variances.**
(A, B) PR curves for the direct effect model (DEM). **(C, D)** PR curves for the gene effect model (GEM). **(A, C)** $\nu_\beta = \nu_B = 100$. **(B, D)** $\nu_\beta = \nu_B = 20$.

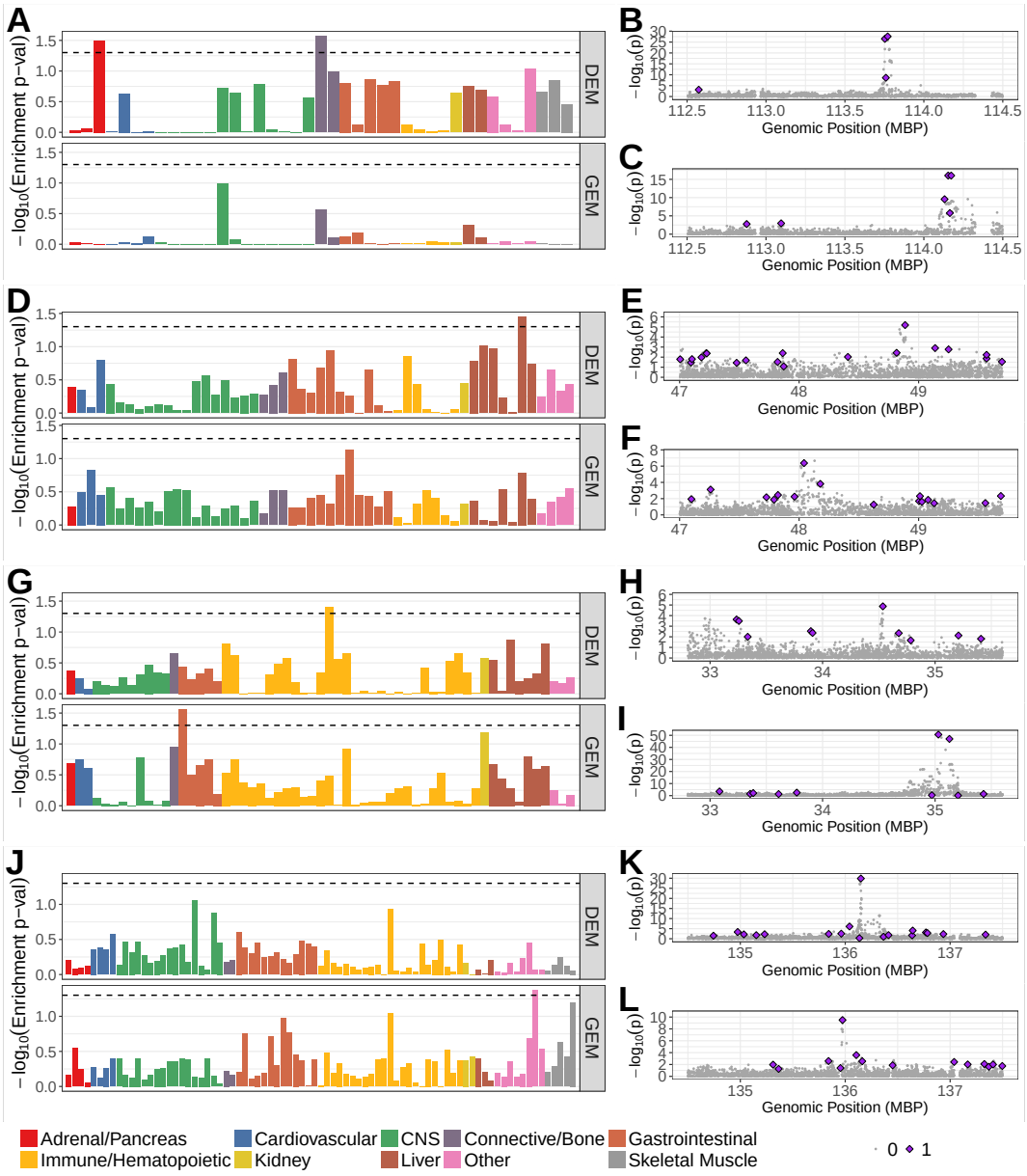

Figure A.6: ***iFunMed* results for fits that the annotation screening did not identify any enriched annotations.**
(A, D, G, J) $-\log_{10}$ transformed enrichment p-values for annotations with more than 5% of loci SNPs with the annotation. Dashed line represents marginal significance level of 5%. **(A)** *TMCO3* as mediator, **(D)** *MSH6* as mediator, **(G)** *ITSN1* as mediator, and **(J)** *RALGDS* as mediator. **(B, C, E, F, H, I, K, L)** Manhattan plots for the GWAS **(B, E, H, K)** and eQTL **(C, F, I, L)** input summary statistics. SNPs highlighted in purple are selected by the null model whereas gray SNPs are not selected using posterior probability of inclusion cut-off at 0.5. **(B, C)** *TMCO3* as mediator, **(E, F)** *MSH6* as mediator, **(H, I)** *ITSN1* as mediator, and **(K, L)** *RALGDS* as mediator.

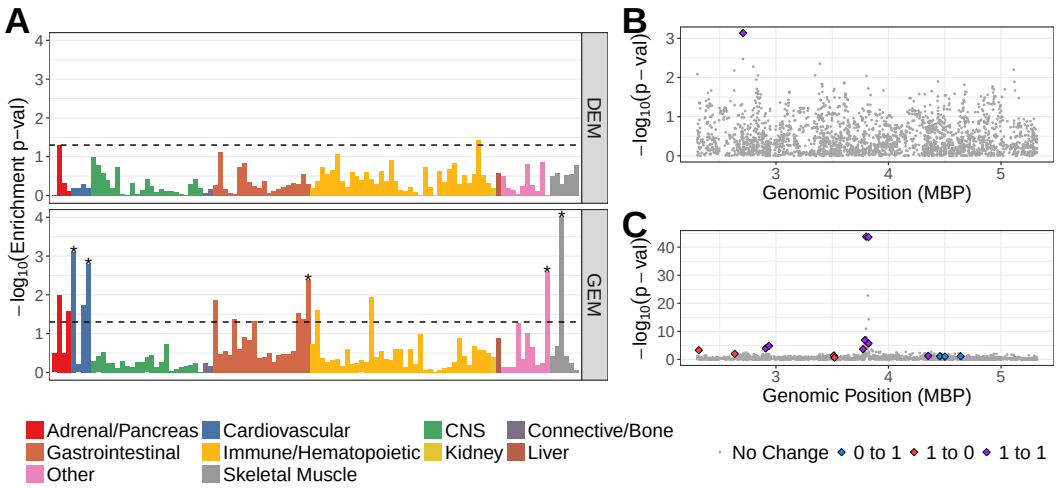

Figure A.7: ***iFunMed* results for log transformed Fasting Glucose with *P2RX1* as mediator.**
(A) $-\log_{10}$ transformed enrichment p-values for annotations with more than 5% of loci SNPs with the annotation. Dashed line represents marginal significance level of 5%. Annotations used for the fits are significant at FDR of 10% and are marked with asterisks. **(B, C)** Manhattan plots for the GWAS and eQTL input summary statistics, respectively. SNPs highlighted in blue/red represent SNPs with large changes in their posterior probabilities of inclusion across the two *iFunMed* fits (with and without annotation). Blue SNPs are selected with the use of annotation whereas red SNPs are excluded, and the status of the purple (selected) and gray SNPs (not selected) do not vary between the two fits.

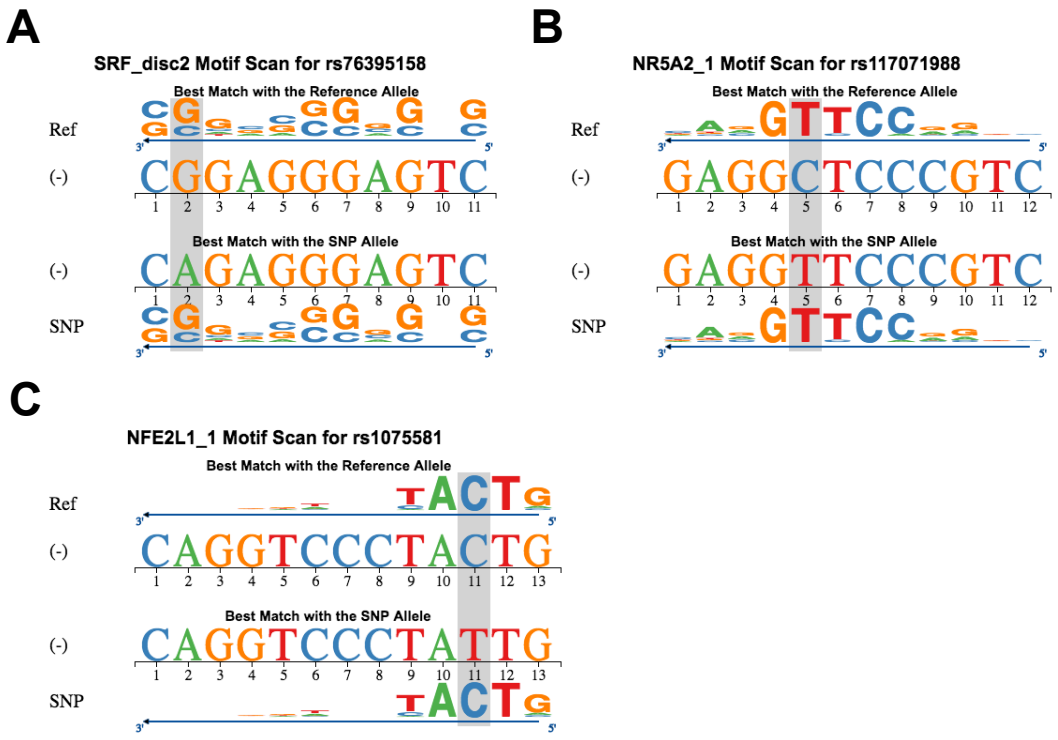

Figure A.8: **atSNP (Shin et al., 2018) composite logo plots for SNPs that are identified only by the use of annotation.**
The composite logo plots compare the best matches of TF motifs to the DNA sequences overlapping the SNP positions with the reference and SNP alleles to hypothesize potential gain- or loss-of-function with atSNP p-value cutoff of $\leqslant 1e^{-7}$. **(A)** *rs76395158*-SRF pair from the model using *P2RX1* as mediator, suggesting potential loss-of-function. **(B)** *rs117071988*-NR5A2 pair from the model using *P2RX1* as mediator, suggesting potential gain-of-function. **(C)** *rs1075581*-NFE2L1 pair from the model using *IL32* as mediator, suggesting potential loss-of-function.

A.5 Supplementary Tables for "iFunMed: Integrative Functional Mediation Analysis of GWAS and eQTL Studies"

Label	Cell-type	Histone Mark	Tissue Group	Proportion
A1	Gastric	H3K4me3	GI	0.014
A2	Peripheralblood Mononuclear Primary	H3K9ac	Immune/Hematopoietic	0.036
A3	Adipose Nuclei	H3K4me3	Other	0.050
A4	CD8 Memory Primary	H3K4me1	Immune/Hematopoietic	0.098
A5	Hippocampus Middle	H3K4me1	CNS	0.147

Table A.1: Details of the annotations used in the simulations. "Proportion" refers to proportion of SNPs residing in the peak regions, i.e., candidate regulatory regions, of the underlying histone mark.

	Type I Error	Power									
		A1		A2		A3		A4		A5	
		Mild	*Strong*	*Mild*	*Strong*	*Mild*	*Strong*	*Mild*	*Strong*	*Mild*	*Strong*
DEM	0.047	0.089	0.136	0.256	0.286	0.383	0.450	0.447	0.539	0.617	0.722
GEM	0.049	0.125	0.108	0.206	0.300	0.378	0.475	0.522	0.578	0.672	0.736

Table A.2: Evaluation of annotation screening for Type I error control and power with simulations for direct and gene effect models. The null hypothesis (H_0 : No annotation effect) considered 18 simulation settings where annotation effect sizes were set to 0 ($\boldsymbol{\gamma}_\beta, \boldsymbol{\gamma}_B = (-4, 0)$). For the remaining 36 simulation settings, the alternative hypothesis was true and included scenarios with a non-zero annotation effect (mild or strong). Within each simulation setting, we used five different annotations and generated 20 datasets. For each dataset, we calculated enrichment p-values for all annotations used for the simulation for direct ($\hat{\mathbf{p}}_\beta$) and gene ($\hat{\mathbf{p}}_B$) effect models and thresholded the Bonferroni corrected p-values at 5%.

	Phenotype	Mediator Gene	# Annotations after Filtering	# of Selected Annotations DEM	GEM
Locus 1	Factor VII	*TMCO3*	41	0	0
Locus 2	White Blood Cell Count (log)	*MSH6*	53	0	0
Locus 3	Red Blood Cell Count	*NINJ2*	73	0	1
Locus 4	Red Blood Cell Count	*EVA1C*	59	0	2
Locus 5	Red Blood Cell Count	*ITSN1*	59	0	0
Locus 6	Von Willebrand Factor	*RALGDS*	81	0	0

Table A.3: Results of the annotation screening, including the total numbers of candidate annotations for each locus after filtering out annotations with less than 5% of overlap with the locus SNPs.

TMCO3	Location	Model
kgp10717896	112,573,998	DEM
rs2181540	113,753,164	DEM
kgp9453357	113,759,526	DEM
rs488703	113,770,876	DEM
kgp1387232	112,877,768	GEM
rs1536678	113,095,947	GEM
kgp12297185	114,131,289	GEM
kgp2036972	114,154,230	GEM
kgp4645137	114,164,811	GEM
kgp525215	114,173,204	GEM
MSH6	Location	Model
kgp3063860	47,004,949	DEM
kgp2682188	47,095,636	DEM
kgp2427793	47,103,998	DEM
kgp11431633	47,182,740	DEM
rs17481182	47,223,522	DEM
kgp11283959	47,226,640	DEM
kgp2688517	47,477,578	DEM
kgp4611764	47,555,232	DEM
kgp12183625	47,819,972	DEM
rs6716984	47,863,075	DEM
rs17504691	47,871,470	DEM
rs2348719	48,408,487	DEM
kgp6493984	48,815,977	DEM
kgp10514201	48,887,051	DEM
kgp9736999	49,138,080	DEM
kgp7392282	49,250,623	DEM
kgp154300	49,568,675	DEM
kgp10248994	49,568,756	DEM
rs10865241	49,697,619	DEM
rs7586009	47,098,968	GEM
kgp10038202	47,259,497	GEM
kgp738041	47,726,765	GEM
rs1863334	47,790,611	GEM
kgp10844215	47,823,379	GEM
kgp255139	47,961,712	GEM
rs330787	48,041,377	GEM
rs4583515	48,177,487	GEM
kgp8128499	48,624,433	GEM
kgp9184078	49,003,681	GEM
kgp9692670	49,011,274	GEM
kgp4392460	49,027,304	GEM
kgp1901800	49,079,247	GEM
rs7563889	49,129,890	GEM
kgp10252206	49,558,562	GEM
kgp4555932	49,688,295	GEM
ITSN1	Location	Model
kgp8102103	33,235,336	DEM
kgp3044871	33,256,005	DEM
kgp5757773	33,334,632	DEM
kgp1163247	33,895,682	DEM
kgp4934738	33,910,920	DEM
kgp349380	34,535,884	DEM
rs2834178	34,677,391	DEM
kgp2131229	34,783,522	DEM
kgp12140722	35,207,719	DEM
kgp6697616	35,407,829	DEM
kgp9934392	33,083,774	GEM
kgp564488	33,356,500	GEM
rs8134098	33,383,368	GEM
kgp4450304	33,609,279	GEM
kgp1811605	33,770,584	GEM
kgp9149520	34,971,300	GEM
kgp2881149	35,029,572	GEM
kgp11178917	35,127,360	GEM
kgp7266158	35,204,420	GEM
kgp8478798	35,430,916	GEM
RALGDS	Location	Model
kgp840600	134,743,431	DEM
kgp27492383	134,974,875	DEM
kgp8852139	135,033,545	DEM
kgp5352959	135,152,029	DEM
kgp1676062	135,231,526	DEM
kgp5859342	135,842,732	DEM
kgp1276914	135,962,024	DEM
rs7044834	136,040,899	DEM
rs8176704	136,135,552	DEM
rs505922	136,149,229	DEM
kgp7665392	136,365,210	DEM
kgp3389780	136,412,638	DEM
kgp5046126	136,637,867	DEM
rs9409863	136,643,239	DEM
kgp1498770	136,769,545	DEM
kgp7940722	136,783,505	DEM
kgp6740746	136,935,656	DEM
kgp170487	137,337,974	DEM
kgp10327190	135,312,230	GEM
rs10122574	135,365,135	GEM
rs11243956	135,840,945	GEM
rs509064	135,954,310	GEM
kgp11339176	135,974,100	GEM
kgp12147077	136,105,515	GEM
kgp9105822	136,162,255	GEM
kgp6406543	136,451,851	GEM
rs28404378	137,038,257	GEM
kgp11596945	137,167,572	GEM
rs877954	137,326,408	GEM
kgp6885766	137,369,131	GEM
rs11791555	137,408,867	GEM
kgp11976262	137,496,884	GEM

Table A.4: List of SNPs selected in the *iFunMed* fits with a posterior probability of inclusion threshold of 0.5. The annotation screening did not identify any enriched annotations for the listed candidate mediators; therefore, *iFunMed* results from fits without annotation (null model) are displayed.

Phenotype	Mediator Gene	Chrom	Start	End	# of SNPs	# of Subjects
Fasting Glucose (log)	*IL32*	chr16	1,616,201	4,617,621	1,948	1,501
Fasting Glucose (log)	*P2RX1*	chr17	2,300,929	5,319,808	2,581	1,501
HDL (log)	*CDA*	chr1	19,375,841	22,485,337	2,386	1,661
HDL (log)	*PSD4*	chr2	112,392,625	115,500,323	2,225	1,661
HDL (log)	*IL1RAP*	chr3	188,694,749	191,913,088	3,308	1,661
HDL (log)	*ASGR2*	chr17	5,467,424	8,557,608	3007	1,661
HDL (log)	*IGFLR1*	chr19	34,691,449	37,747,799	2,180	1,661
HDL (log)	*APMAP*	chr20	23,405,281	26,309,255	1,857	1,661

Table A.5: Details of loci considered for the mediation analysis of the FHS phenotypes fasting glucose and HDL.

Phenotype	Mediator Gene	# Annotations after Filtering	# of Identified Annotations DEM	GEM
Fasting Glucose (log)	*IL32*	139	3	0
Fasting Glucose (log)	*P2RX1*	109	0	5*
HDL (log)	*CDA*	91	0	0*
HDL (log)	*PSD4*	48	0	0
HDL (log)	*IL1RAP*	12	1	0
HDL (log)	*ASGR2*	115	0	0
HDL (log)	*IGFLR1*	124	0	0
HDL (log)	*APMAP*	54	1	0

Table A.6: Annotation strategy results for FHS phenotypes fasting glucose and HDL. Cases in asterisk denote loci with an elevated signal in either GWAS or eQTL ($-\log_{10}(\text{p-value}) > 20$) and low density from which 0.5% of the SNPs were trimmed to remove outliers.

Mediator Gene	Model	Tissue	Mark	Cell-type	Enrichment p-value
IL32	DEM	Immune/Hematopoietic	H3K27ac	CD3 primary	0.002
IL32	DEM	Immune/Hematopoietic	H3K27ac	Th0	0.001
IL32	DEM	Immune/Hematopoietic	H3K27ac	Th1	0.002
P2RX1	GEM	Cardiovascular	H3K4me1	Fetal heart	0.001
P2RX1	GEM	Skeletal Muscle	H3K4me1	Fetal trunk muscle	0.001
P2RX1	GEM	Cardiovascular	H3K9ac	Fetal heart	0.002
P2RX1	GEM	Other	H3K9ac	Penis foreskin keratinocyte primary	0.003
P2RX1	GEM	GI	H3K27ac	Duodenum mucosa	0.004
IL1RAP	DEM	Immune/Hematopoietic	H3K27ac	CD19	0.007
APMAP	DEM	CNS	H3K4me1	Fetal brain	0.000

Table A.7: Details of the annotations that were identified for the FHS phenotypes fasting glucose and HDL by the annotation screening strategy.

				Annotation				
IL32	Location	Model	Direction	CD3 Primary		Th0		Th1
rs1075581	4,135,898	DEM	0 to 1 (↑)	0		1		0
				Annotation				
P2RX1	Location	Model	Direction	Heart	Trunk Muscle	Heart	Keratinocyte	Duodenum
rs9652825	2,314,074	GEM	1 to 0 (↓)	0	0	0	0	0
rs7219019	2,633,324	GEM	1 to 0 (↓)	0	0	0	0	0
kgp23896402	2,908,975	GEM	1 to 1 (−)	1	1	0	0	0
kgp3755826	2,937,617	GEM	1 to 1 (−)	1	1	1	1	0
rs76056301	3,514,954	GEM	1 to 0 (↓)	0	0	0	0	0
rs224498	3,519,954	GEM	1 to 0 (↓)	0	0	0	0	0
kgp3692495	3,774,014	GEM	1 to 1 (−)	1	1	1	0	1
kgp2113525	3,790,498	GEM	1 to 1 (−)	0	1	0	1	1
kgp11900618	3,800,995	GEM	1 to 1 (−)	0	1	0	1	0
rs8076916	3,822,637	GEM	1 to 1 (−)	1	1	1	1	1
kgp10137990	3,822,926	GEM	1 to 1 (−)	1	1	1	1	1
kgp9641039	4,353,359	GEM	0 to 1 (−)	0	1	0	1	0
rs76395158	4,458,005	GEM	0 to 1 (↑)	1	1	1	1	0
rs117071988	4,502,386	GEM	0 to 1 (↑)	1	1	0	0	1
rs1050997	4,641,755	GEM	0 to 1 (↑)	0	1	0	0	1
				Annotation				
IL1RAP	Location	Model	Direction	CD19				
kgp9044897	188,819,337	DEM	1 to 1 (−)	0				
kgp5564150	188,952,988	DEM	1 to 1 (−)	0				
rs1515490	189,596,855	DEM	1 to 0 (−)	0				
rs2378570	190,154,740	DEM	1 to 1 (−)	1				
rs7641416	190,861,768	DEM	0 to 1 (↑)	1				
kgp7767169	191,001,699	DEM	1 to 1 (−)	1				
rs13059172	191,348,064	DEM	1 to 1 (−)	0				

Table A.8: List of SNPs selected in the analysis of FHS phenotypes fasting glucose and HDL. SNPs are labeled as 0 to 1 ((↑) direction) if they are selected only with the use of annotation and as 1 to 0 ((↓) direction) if they are excluded from the *iFunMed* fit with the use of annotation by thresholding poterior probability of inclusion at 0.5. SNPs selected with and without annotation are labeled as 1 to 1 (−). *APMAP* is not shown since there were no selected SNPs. Details of the annotations included for both models (DEM and GEM) at the individual SNP level are also displayed.

B APPENDIX B

B.1 Supplementary Text for "High dimensional sparse regression with auxiliary data on the features"

In this section, we provide the upper bound for the estimation error of $\boldsymbol{\beta}$. For simplicity of proof, we assume that the model does not have additional covariates X to adjust for. When there is a need to adjust for X, all the assumptions on G in the following proof should be modified into assumptions on $(I - X(X^\top X)^{-1}X^\top)G$ and all other parts of the proof remain the same. Our model setting is

$$Y = G\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{B.1}$$

$$\boldsymbol{\beta} = A\boldsymbol{\gamma} + \boldsymbol{\eta}, \tag{B.2}$$

where Y and G are centralized to have mean zero so no intercept is needed in the regression model. Denote $Z = (GA, G)$, $W = \mathrm{diag}(Z^\top Z)^{1/2}$ and $\widetilde{Z} = ZW^{-1}$ as the standardized Z so that the column l_2 norm of $\widetilde{Z}$ is one. Then we have

$$Y = GA\boldsymbol{\gamma} + G\boldsymbol{\eta} + \boldsymbol{\epsilon} = \widetilde{Z}\theta + \boldsymbol{\epsilon}$$

where $\theta = W(\gamma^\top, \eta^\top)^\top$. Our estimators for γ and η are obtained through the following optimization:

$$\begin{aligned} &\widehat{\theta} = \underset{\theta}{\operatorname{argmin}} \left\{ \frac{1}{2n} \|Y - \widetilde{Z}\theta\|_2^2 + \lambda \|\theta\|_1 \right\}, \\ &(\widehat{\gamma}^\top, \widehat{\eta}^\top)^\top = W^{-1}\widehat{\theta} \end{aligned} \tag{B.3}$$

Suppose $G = (G_1, \ldots, G_p)$ and $A = (A_1, \ldots, A_m)$. Define the following constants:

$$\mu_G = \max_{i \neq j} \frac{|G_i^\top G_j|}{\|G_i\|_2 \|G_j\|_2} \tag{B.4}$$

$$\tau_{max} = \max_j \frac{1}{\sqrt{n}} \|G_j\|_2 \tag{B.5}$$

$$\tau_{min} = \min_j \frac{1}{\sqrt{n}} \|G_j\|_2 \tag{B.6}$$

$$\mu_A = \max_{i \neq j} \frac{|A_i^\top A_j|}{\|A_i\|_2 \|A_j\|_2} \tag{B.7}$$

$$\kappa_1 = \max_j \|A_j\|_1 = \|A\|_1 = \sup_{x \neq 0} \frac{\|Ax\|_1}{\|x\|_1} \tag{B.8}$$

$$\kappa_{min} = \min_j \|A_j\|_2 \tag{B.9}$$

$$\kappa_{max} = \max_j \|A_j\|_2 \tag{B.10}$$

$$\kappa_\infty = \max_{i,j} |A_{ij}| \tag{B.11}$$

For any vector x, denote $S_x = \{i : x_i \neq 0\}$ as the support set of x, and $s_x = \#S_x$ as the number of nonzero elements of x.

Theorem B.1. *Suppose $\widehat{\gamma}$ and $\widehat{\eta}$ are obtained using* (B.3) *with*

$$\lambda = 2C\sigma\sqrt{\log(p+m)/n} \tag{B.12}$$

Then, under the condition of

$$\mu < 1/(4(s_\gamma + s_\eta) - 1) \tag{B.13}$$

with probability at least $1 - 2(p+m)^{1-C^2/2}$*, we have*

$$\|\hat{\beta} - \beta\|_2 \leqslant \frac{6CC_2}{1 - 4\mu(s_\gamma + s_\eta - 1)} \sqrt{\frac{\sigma^2(s_\gamma + s_\eta)\log(p+m)}{n}} \tag{B.14}$$

where

$$\mu = \max\left\{ \frac{(\tau_{max}/\tau_{min})^2(\mu_G \kappa_1^2 + \mu_A \kappa_{max}^2)}{\kappa_{min}^2 - \mu_G \kappa_1^2}, \frac{(\tau_{max}/\tau_{min})^2(\mu_G \kappa_1 + \kappa_\infty)}{\sqrt{\kappa_{min}^2 - \mu_G \kappa_1^2 (\tau_{max}/\tau_{min})^2}}, \mu_G \right\},$$

$$C_2 = \max\left\{ \sqrt{\frac{\kappa_{max}^2(1 + \mu_A(m-1))}{\tau_{max}^2 \kappa_{min}^2 - \mu_G \tau_{max}^2 \kappa_1^2}}, \frac{1}{\tau_{min}} \right\}.$$

Proof. Set $h = \widehat{\theta} - \theta$ and let S_h be the set of indices of the largest s values of h. By definition of $\widehat{\theta}$, we have

$$\begin{aligned} \frac{1}{2n}\|\epsilon - \widetilde{Z}h\|_2^2 + \lambda\|\widehat{\theta}\|_1 &= \frac{1}{2n}\|Y - \widetilde{Z}\widehat{\theta}\|_2^2 + \lambda\|\widehat{\theta}\|_1 \\ &\leqslant \frac{1}{2n}\|Y - \widetilde{Z}\theta\|_2^2 + \lambda\|\theta\|_1 = \frac{1}{2n}\|\epsilon\|_2^2 + \lambda\|\theta\|_1, \end{aligned}$$

which gives us

$$\begin{aligned}\frac{1}{2n}\|\boldsymbol{\epsilon}-\widetilde{Z}\mathbf{h}\|_2^2-\frac{1}{2n}\|\boldsymbol{\epsilon}\|_2^2 &\leqslant \lambda(\|\boldsymbol{\theta}\|_1-\|\widehat{\boldsymbol{\theta}}\|_1)\\ &= \lambda(\|\boldsymbol{\theta}_{\mathrm{supp}(\boldsymbol{\theta})}\|_1-\|\widehat{\boldsymbol{\theta}}_{\mathrm{supp}(\boldsymbol{\theta})}\|_1-\|\widehat{\boldsymbol{\theta}}_{\mathrm{supp}(\boldsymbol{\theta})^c}-\boldsymbol{\theta}_{\mathrm{supp}(\boldsymbol{\theta})^c}\|_1)\\ &\leqslant \lambda(\|\mathbf{h}_{\mathrm{supp}(\boldsymbol{\theta})}\|_1-\|\mathbf{h}_{\mathrm{supp}(\boldsymbol{\theta})^c}\|_1)\\ &\leqslant \lambda(\|\mathbf{h}_{S_h}\|_1-\|\mathbf{h}_{S_h^c}\|_1).\end{aligned} \tag{B.15}$$

Define event

$$\mathcal{A}=\left\{\|\boldsymbol{\epsilon}^\top\widetilde{Z}\|_\infty\leqslant\frac{n\lambda}{2}\right\}.$$

So under event $\mathcal{A}$,

$$\begin{aligned}\frac{1}{2n}\|\boldsymbol{\epsilon}-\widetilde{Z}\mathbf{h}\|_2^2-\frac{1}{2n}\|\boldsymbol{\epsilon}\|_2^2 &= \frac{1}{2n}\|\widetilde{Z}\mathbf{h}\|_2^2-\boldsymbol{\epsilon}^\top\widetilde{Z}\mathbf{h}/n\\ \geqslant -\frac{1}{n}\|\boldsymbol{\epsilon}^\top\widetilde{Z}\|_\infty\|\mathbf{h}\|_1 &\geqslant -\frac{\lambda}{2}\|\mathbf{h}\|_1=-\frac{\lambda}{2}(\|\mathbf{h}_{S_h}\|_1+\|\mathbf{h}_{S_h^c}\|_1).\end{aligned} \tag{B.16}$$

Combining inequalities (B.15) and (B.16), we have

$$\|\mathbf{h}_{S_h^c}\|_1\leqslant 3\|\mathbf{h}_{S_h}\|_1.$$

Therefore, under event $\mathcal{A}$,

$$\|\mathbf{h}_{S_h^c}\|_2^2\leqslant\|\mathbf{h}_{S_h^c}\|_1\|\mathbf{h}_{S_h^c}\|_\infty\leqslant 3\|\mathbf{h}_{S_h}\|_1\cdot\|\mathbf{h}_{S_h}\|_1/s\leqslant 3\|\mathbf{h}_{S_h}\|_2^2. \tag{B.17}$$

Next, we will prove that

$$\max_{i\neq j} |\widetilde{Z}_i^\top \widetilde{Z}_j| \leqslant \mu, \tag{B.18}$$

where

$$\mu = \max\left\{ \frac{\tau_{max}^2(\mu_G\kappa_1^2 + \mu_A\kappa_{max}^2)}{\tau_{min}^2\kappa_{min}^2 - \mu_G\tau_{max}^2\kappa_1^2}, \frac{\tau_{max}^2(\mu_G\kappa_1 + \kappa_\infty)}{\tau_{min}\sqrt{\tau_{min}^2\kappa_{min}^2 - \mu_G\tau_{max}^2\kappa_1^2}}, \mu_G \right\}.$$

In fact,

$$\begin{aligned}
\max_{i\neq j} |(A^\top G^\top GA)_{i,j}| \leqslant & \max_{i\neq j} \left| \sum_{k,l} A_{ki}(G^\top G)_{k,l} A_{lj} \right| \\
\leqslant & \max_{i\neq j} \sum_{k,l} |A_{ki}||A_{lj}| \frac{|G_k^\top G_l|}{\|G_k\|_2\|G_l\|_2} n\tau_{max}^2 \\
\leqslant & \max_{i\neq j} \left\{ n\tau_{max}^2\mu_G \sum_{k\neq l} |A_{ki}||A_{lj}| + n\tau_{max}^2 \sum_k |A_{ki}A_{kj}| \right\} \\
\leqslant & \max_{i\neq j} \left\{ n\tau_{max}^2\mu_G \|A_i\|_1\|A_j\|_1 + n\tau_{max}^2 \frac{|A_i^\top A_j|}{\|A_i\|_2\|A_j\|_2} \|A_i\|_2\|A_j\|_2 \right\} \\
\leqslant & \ n\tau_{max}^2(\mu_G\kappa_1^2 + \mu_A\kappa_{max}^2),
\end{aligned}$$

$$\begin{aligned}
\min_i \|GA_i\|_2^2 = & \|\sum_k G_k A_{ki}\|_2^2 = \sum_k \|G_k\|_2^2 A_{ki}^2 + \sum_{k\neq l} \langle G_k, G_l\rangle A_{ki}A_{li} \\
\geqslant & \min_k \|G_k\|_2^2 \cdot \sum_k A_{ki}^2 - \max_k \|G_k\|_2^2 \mu_G \cdot \sum_{k\neq l} A_{ki}A_{li} \\
\geqslant & n\tau_{min}^2\|A_i\|_2^2 - n\tau_{max}^2\mu_G\|A_i\|_1^2 \\
\geqslant & n\tau_{min}^2\kappa_{min}^2 - n\mu_G\tau_{max}^2\kappa_1^2.
\end{aligned}$$

Thus,

$$\max_{i\neq j}\frac{|(A^\top G^\top GA)_{i,j}|}{\|(GA)_i\|_2\|(GA)_j\|_2}\leqslant\frac{\tau_{max}^2(\mu_G\kappa_1^2+\mu_A\kappa_{max}^2)}{\tau_{min}^2\kappa_{min}^2-\mu_G\tau_{max}^2\kappa_1^2}.$$

$$\begin{aligned}\max_{i,j}|(G^\top GA)_{i,j}|\leqslant&\ \max_{i,j}\sum_k|(G^\top G)_{i,k}||A_{kj}|\\ \leqslant&\ \max_{i,j}\left\{n\tau_{max}^2\sum_{k\neq i}\frac{|G_i^\top G_k|}{\|G_i\|_2\|G_k\|_2}|A_{kj}|+n\tau_{max}^2|A_{ij}|\right\}\\ \leqslant&\ \max_{i,j}\left\{n\tau_{max}^2\mu_G\|A_j\|_1+n\tau_{max}^2|A_{ij}|\right\}\\ \leqslant&\ n\tau_{max}^2(\mu_G\kappa_1+\kappa_\infty),\end{aligned}$$

thus,

$$\max_{i,j}\frac{|(G^\top GA)_{i,j}|}{\|G_i\|_2\cdot\|(GA)_j\|_2}\leqslant\frac{\tau_{max}^2(\mu_G\kappa_1+\kappa_\infty)}{\tau_{min}\sqrt{\tau_{min}^2\kappa_{min}^2-\mu_G\tau_{max}^2\kappa_1^2}}$$

Therefore,

$$\begin{aligned}&\max_{i\neq j}|\widetilde{Z}_i^\top\widetilde{Z}_j|=\max_{i\neq j}\frac{|Z_i^\top Z_j|}{\|Z_i\|_2\|Z_j\|_2}\\ =&\max\left\{\max_{i\neq j}\frac{|(A^\top G^\top GA)_{i,j}|}{\|(GA)_i\|_2\|(GA)_j\|_2},\max_{i,j}\frac{|(G^\top GA)_{i,j}|}{\|G_i\|_2\cdot\|(GA)_j\|_2},\max_{i,j}\frac{|(G^\top G)_{i,j}|}{\|G_i\|_2\|G_j\|_2}\right\}\\ \leqslant&\max\left\{\frac{\tau_{max}^2(\mu_G\kappa_1^2+\mu_A\kappa_{max}^2)}{\tau_{min}^2\kappa_{min}^2-\mu_G\tau_{max}^2\kappa_1^2},\frac{\tau_{max}^2(\mu_G\kappa_1+\kappa_\infty)}{\tau_{min}\sqrt{\tau_{min}^2\kappa_{min}^2-\mu_G\tau_{max}^2\kappa_1^2}},\mu_G\right\}=\mu\end{aligned}$$

and we have (B.18). As a result of (B.18), we have

$$\begin{aligned}
\|\widetilde{Z}\mathbf{h}_{S_h}\|_2^2 &= \sum_{i\in S_h}\sum_{j\in S_h}(\widetilde{Z}^\top\widetilde{Z})_{i,j}h_ih_j \\
&\geqslant \sum_{i\in S_h}(\widetilde{Z}^\top\widetilde{Z})_{i,i}h_i^2 - \sum_{i,j\in S_h, i\neq j}\mu|h_i||h_j| \\
&\geqslant \sum_{i\in S_h}h_i^2 - \mu(\|\mathbf{h}_{S_h}\|_1^2 - \|\mathbf{h}_{S_h}\|_2^2) \\
&\geqslant \|\mathbf{h}_{S_h}\|_2^2 - \mu(s-1)\|\mathbf{h}_{S_h}\|_2^2 \\
&\geqslant (1-\mu(s-1))\|\mathbf{h}_{S_h}\|_2^2
\end{aligned}$$

and

$$\begin{aligned}
|\mathbf{h}_{S_h^c}^\top\widetilde{Z}^\top\widetilde{Z}\mathbf{h}_{S_h}| &= |\sum_{i\in S_h}\sum_{j\in S_h^c}(\widetilde{Z}^\top\widetilde{Z})_{i,j}h_ih_j| \\
&\leqslant \sum_{i\in S_h}\sum_{j\in S_h^c}\mu|h_i||h_j| = \mu\|\mathbf{h}_{S_h}\|_1\|\mathbf{h}_{S_h^c}\|_1 \\
&\leqslant 3\mu\|\mathbf{h}_{S_h}\|_1^2 \leqslant 3s\mu\|\mathbf{h}_{S_h}\|_2^2
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathbf{h}^\top\widetilde{Z}^\top\widetilde{Z}\mathbf{h}_{S_h} &\geqslant \|\widetilde{Z}\mathbf{h}_{S_h}\|_2^2 - |\mathbf{h}_{S_h^c}^\top\widetilde{Z}^\top\widetilde{Z}\mathbf{h}_{S_h}| \\
&\geqslant (1-\mu(4s-1))\,\|\mathbf{h}_{S_h}\|_2^2
\end{aligned} \tag{B.19}$$

On the other hand, according to the KKT condition of optimization (B.3), we have

$\|\widetilde{Z}^\top(y - \widetilde{Z}\widehat{\theta})\|_\infty \leqslant n\lambda$. Therefore, under event $\mathcal{A}$,

$$\begin{aligned} h^\top \widetilde{Z}^\top \widetilde{Z} h_{S_h} &\leqslant \|\widetilde{Z}^\top \widetilde{Z} h\|_\infty \|h_{S_h}\|_1 \\ &\leqslant \left(n\lambda + \|\widetilde{Z}^\top \epsilon\|_\infty\right) \|h_{S_h}\|_1 \\ &\leqslant (n\lambda + n\lambda/2) \|h_{S_h}\|_1 \\ &\leqslant \frac{3}{2} n\lambda\sqrt{s} \|h_{S_h}\|_2 \end{aligned} \tag{B.20}$$

Combining (B.19) and (B.20), we have

$$\|h_{S_h}\|_2 \leqslant \frac{3n\lambda\sqrt{s}/2}{1 - \mu(4s-1)}. \tag{B.21}$$

Together with (B.17), we have that under event $\mathcal{A}$,

$$\|h\|_2 = (\|h_{S_h}\|_2^2 + \|h_{S_h^c}\|_2^2)^{1/2} \leqslant 2\|h_{S_h}\|_2 \leqslant \frac{3n\lambda\sqrt{s}}{1 - \mu(4s-1)}. \tag{B.22}$$

Note that

$$\begin{aligned} \mathbb{P}(\mathcal{A}) = &\ 1 - \mathbb{P}\left(\|\epsilon^\top \widetilde{Z}\|_\infty \geqslant \frac{n\lambda}{2}\right) \\ &\geqslant 1 - 2(p+m)\exp\left\{-\frac{n^2\lambda^2}{8\sigma^2}\right\} = 1 - 2(p+m)^{1-C^2/2}. \end{aligned} \tag{B.23}$$

Therefore,

$$\mathbb{P}\left(\|h\|_2 \leqslant \frac{3n\lambda\sqrt{s}}{1 - \mu(4s-1)}\right) \geqslant 1 - 2(p+m)^{1-C^2/2} \tag{B.24}$$

Next, we develop the upper bound for $\hat{\beta} - \beta$ based on the upper bound of $\|h\|_2$.

Note that for any vector $\boldsymbol{\nu} \in \mathbb{R}^m$,

$$\begin{aligned}
\|A\boldsymbol{\nu}\|_2^2 &\leqslant \sum_i \nu_i^2 \|A_i\|_2^2 + \sum_{i\neq j} |\nu_i \nu_j| |A_i^\top A_j| \\
&\leqslant \kappa_{\max}^2 \|\boldsymbol{\nu}\|_2^2 + \mu_A \kappa_{\max}^2 \sum_{i\neq j} |\nu_i| |\nu_j| \\
&\leqslant \kappa_{\max}^2 \left(\|\boldsymbol{\nu}\|_2^2 + \mu_A (\|\boldsymbol{\nu}\|_1^2 - \|\boldsymbol{\nu}\|_2^2) \right) \\
&\leqslant \kappa_{\max}^2 (1 + \mu_A (m-1)) \|\boldsymbol{\nu}\|_2^2,
\end{aligned}$$

so that $\|A\| \leqslant \kappa_{\max} \sqrt{1 + \mu_A (m-1)}$, where $\|A\|$ is the spectrum norm of matrix A. Therefore, under event $\mathcal{A}$,

$$\begin{aligned}
&\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2 = \|A(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}) + \widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}\|_2 \overset{(B.3)}{=} \|[A\ I] W^{-1} (\hat{\theta} - \theta)\|_2 \\
= &\left\| [A\ I] \begin{bmatrix} \operatorname{diag}(\{\|(GA)_i\|_2^{-1}\}_{i=1}^m) & \\ & \operatorname{diag}(\{\|G_i\|_2^{-1}\}_{i=1}^p) \end{bmatrix} \right\| \cdot \|\mathbf{h}\|_2 \\
\leqslant &\max \left\{ \|A\| \cdot \max_i \{\|(GA)_i\|_2^{-1}\}, \max_i \|G_i\|_2^{-1} \right\} \cdot \|h\|_2 \\
\leqslant &\frac{3n\lambda\sqrt{s}}{1 - \mu(4s-1)} \max \left\{ \sqrt{\frac{\kappa_{\max}^2 (1 + \mu_A (m-1))}{n\tau_{\max}^2 \kappa_{\min}^2 - n\mu_G \tau_{\max}^2 \kappa_1^2}}, \frac{1}{\sqrt{n}\tau_{\min}} \right\} \\
\leqslant &\ 6C\sigma \sqrt{\frac{s \log(p+m)}{n}} \cdot \frac{1}{1 - \mu(4s-1)} \cdot \max \left\{ \sqrt{\frac{\kappa_{\max}^2 (1 + \mu_A (m-1))}{\tau_{\max}^2 \kappa_{\min}^2 - \mu_G \tau_{\max}^2 \kappa_1^2}}, \frac{1}{\tau_{\min}} \right\}.
\end{aligned}$$

□

The constants μ_G and μ_A characterize the orthogonality of columns of G and A respectively. If the columns of A are normalized to have ℓ_2 norm of one, then $\kappa_{\max} = \kappa_{\min} = 1$, and κ_1 characterizes the sparsity of the column vectors of A, and

κ_∞ characterizes the concentration of values within A. If the columns of G are normalized to have ℓ_2 norm of one, then $\tau_{\max} = \tau_{\min} = 1$. μ in Condition (B.13) becomes

$$\mu = \max\left\{\frac{(\mu_G\kappa_1^2 + \mu_A)}{1-\mu_G\kappa_1^2}, \frac{(\mu_G\kappa_1 + \kappa_\infty)}{\sqrt{1-\mu_G\kappa_1^2}}, \mu_G\right\}.$$

The requirement of small μ in Condition (B.13) indicates that aside from a well conditioned G, we also need a well conditioned A with sparse columns and values not concentrated on a few entries. Here are a few examples of A that would violate Condition (B.13):

1. A_i and A_j are colinear, so that $\mu_A \geqslant A_i^\top A_j$ is large.
2. $A_i = (1, 0, \ldots, 0)$, so that κ_∞ is large. In this case, GA_i and G_1 are colinear.
3. $A_i = (1/\sqrt{p}, \ldots, 1/\sqrt{p})$, so that κ_1 is large. In this case, the annotation A_i is not informative.

Denote $V = \text{diag}(G^\top G)^{1/2}$ and $\widetilde{G} = GV^{-1}$ as the standardized G so that the column ℓ_2 norm of $\widetilde{G}$ is one. Denote $\delta = V\beta$. Then we have

$$Y = \widetilde{G}\delta + \epsilon.$$

Theorem B.2. *Consider the estimation of* β *using* $\widehat{\beta} = V^{-1}\widehat{\delta}$ *where* $\widehat{\delta}$ *is obtained using*

$$\widehat{\delta} = \underset{\delta}{\text{argmin}}\left\{\frac{1}{2n}||Y - \widetilde{G}\delta||_2^2 + \lambda||\delta||_1\right\} \tag{B.25}$$

with

$$\lambda = 2C\sigma\sqrt{\log p}/n. \tag{B.26}$$

Then, under the condition of

$$\mu_G \leqslant \frac{1}{4s_\beta - 1}, \tag{B.27}$$

with probability at least $1 - 2p^{1-C^2/2}$*, we have*

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2 \leqslant \frac{6C/\tau_{min}}{1 - \mu_G(4s_\beta - 1)}\sqrt{\frac{\sigma^2 s_\beta \log p}{n}}. \tag{B.28}$$

Comparing Theorem 1 and Theorem 2, the rate of estimation error bound is much smaller using annotations when $\boldsymbol{\beta}$ is much larger than $s_\gamma + s_\eta \ll s_\beta$, which indicates that effect of genotype on phenotype is largely through the annotations.

B.2 Supplementary Figures for "High dimensional sparse regression with auxiliary data on the features"

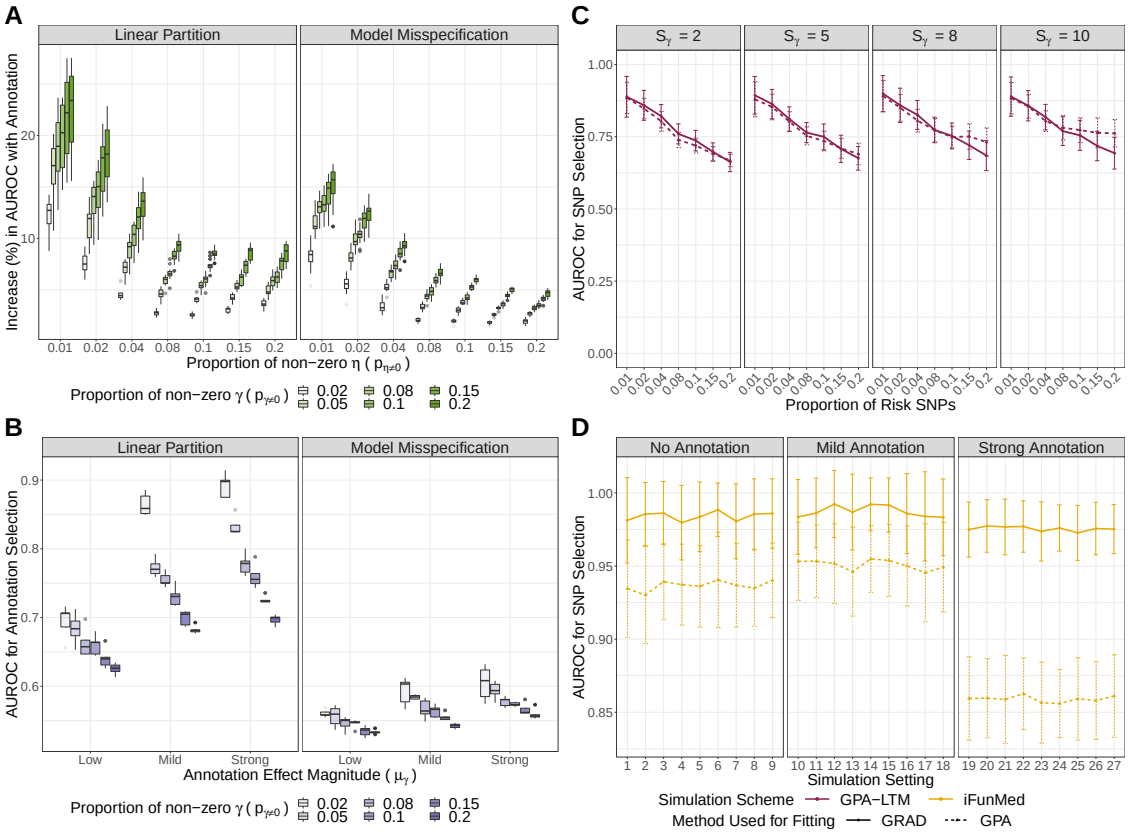

Figure B.1: **Simulation results comparing fits and without annotation in terms of area under the receiver operating characteristic curve (AUROC).**
(A) Percentage change in the AUROC for SNP selection across fits with the use of annotation comparing simulations generated by linear partition and model misspecification for different proportions of non-zero $\boldsymbol{\eta}$ and $\boldsymbol{\gamma}$ ($p_{\eta\neq0}$ and $p_{\gamma\neq0}$, respectively). **(B)** AUROC for annotation selection ($\boldsymbol{\gamma}$) for annotation selection ($\boldsymbol{\gamma}$) when the proportion of risk SNPs is 0.04 ($p_{\eta\neq0} = 0.04$) across fits comparing simulations generated by linear partition and model misspecification for different proportions of non-zero $\boldsymbol{\gamma}$ ($p_{\gamma\neq0}$). **(C, D)** GPA comparisons: average AUROC for SNP selection across 100 simulation replicates and their corresponding error bars (mean $\pm$ standard deviation) for GRAD and GPA. **(C)** Data generated using the GPA liability threshold model (GPA-LTM) . Results are divided by the number of risk annotations $S_\gamma \in \{2, 5, 8, 10\}$. **(D)** Data generated using the *iFunMed* model. Results are divided by their prior inclusion probabilities with the use of annotation (no annotation effect, mild annotation effect, and strong annotation effect).

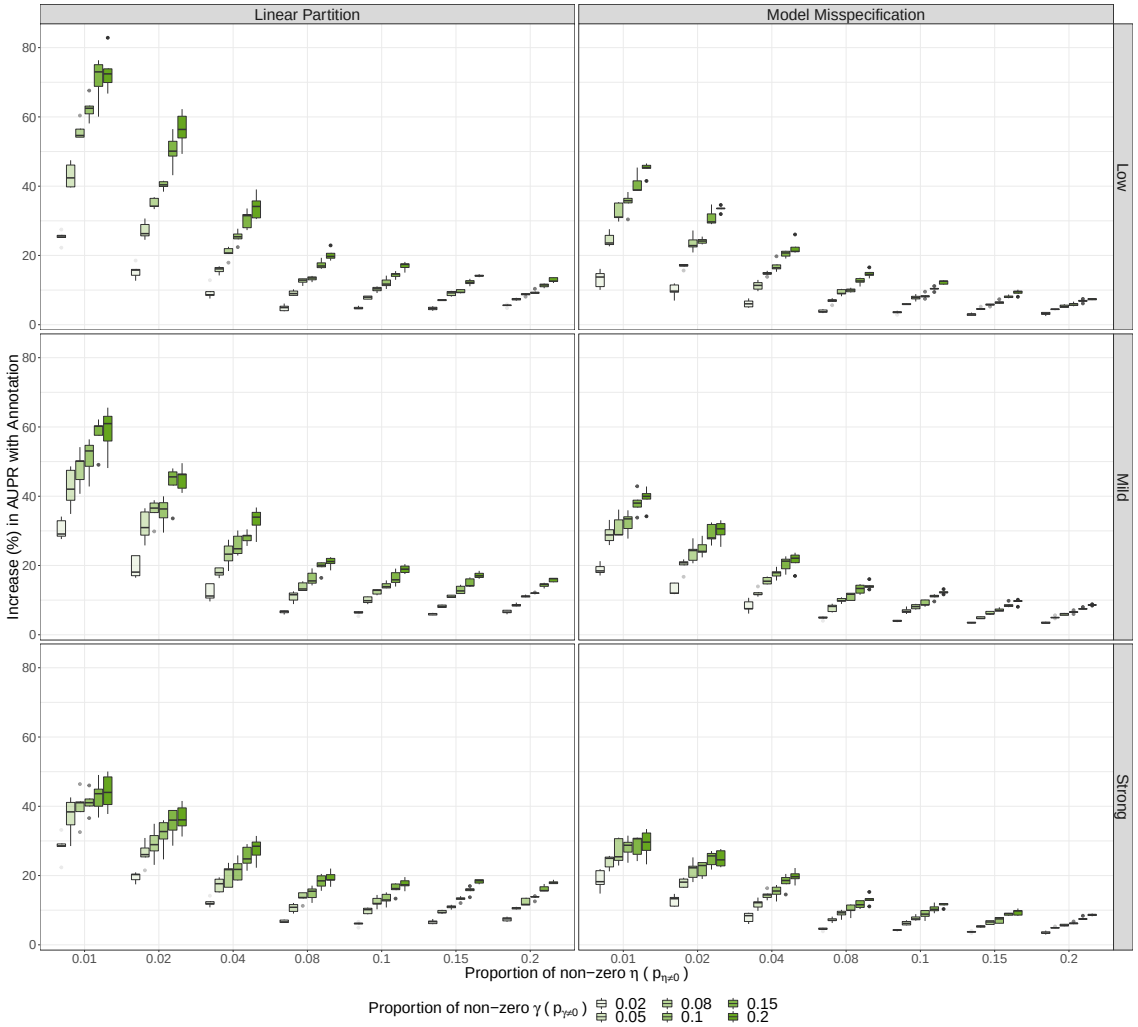

Figure B.2: **Percentage change in the area under the precision-recall curves (AUPR) for SNP selection across fits with the use of annotation comparing simulations generated by linear partition and model misspecification for different proportions of non-zero η and γ ($p_{\eta \neq 0}$ and $p_{\gamma \neq 0}$, respectively) divided by annotation effect magnitude (low, mild, and strong).**

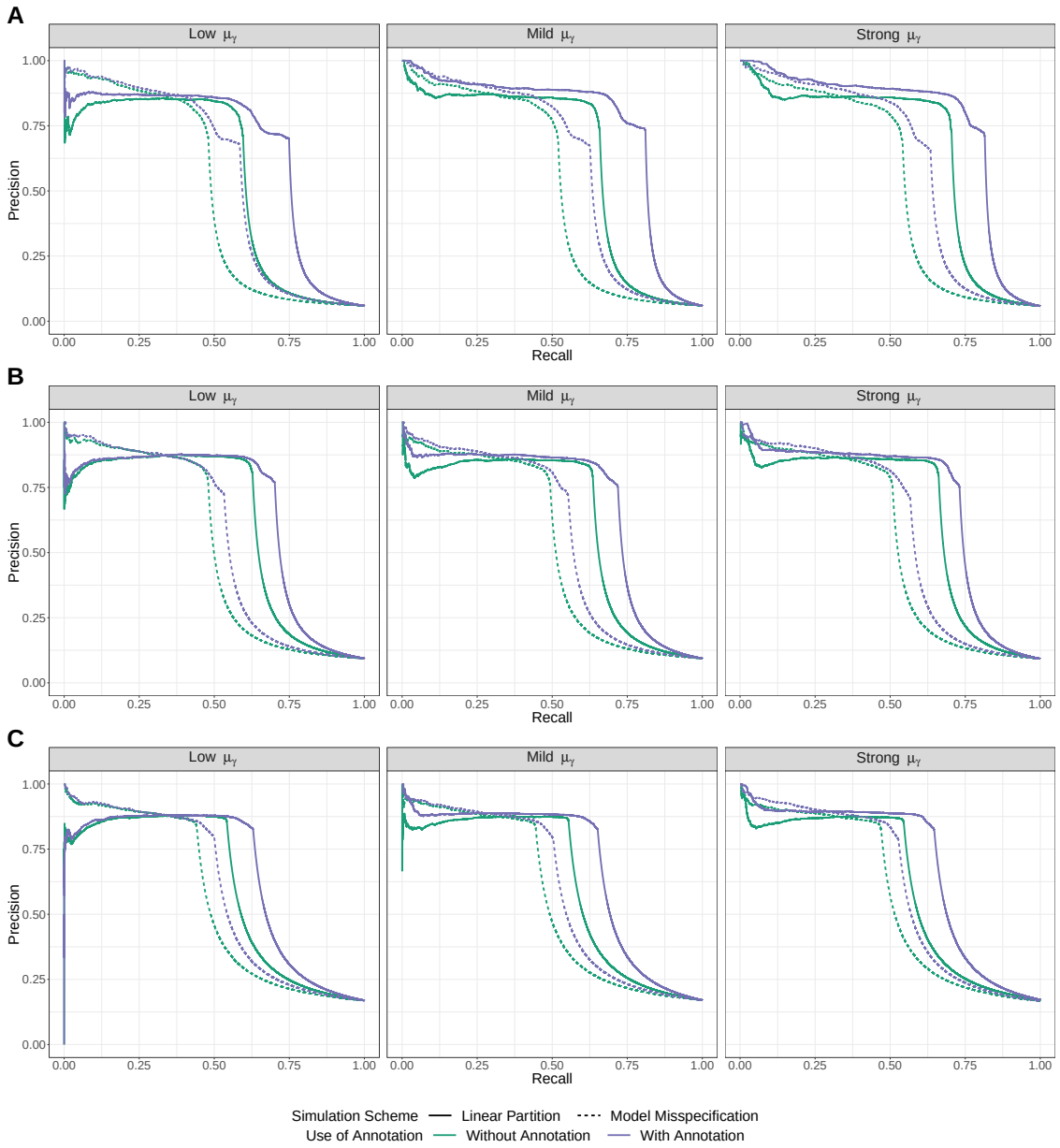

Figure B.3: **SNP selection precision-recall curves for fits with and without annotation comparing simulations generated by linear partition and model misspecification schemes for different simulation scenarios for low, mild, and strong annotation effect magnitude μ_γ.**
(A) $p_{\eta\neq 0} = 0.04$, $p_{\gamma\neq 0} = 0.08$, and $\sigma^2 = 150$. **(B)** $p_{\eta\neq 0} = 0.08$, $p_{\gamma\neq 0} = 0.05$, and $\sigma^2 = 200$. **(C)** $p_{\eta\neq 0} = 0.15$, $p_{\gamma\neq 0} = 0.1$, and $\sigma^2 = 150$.

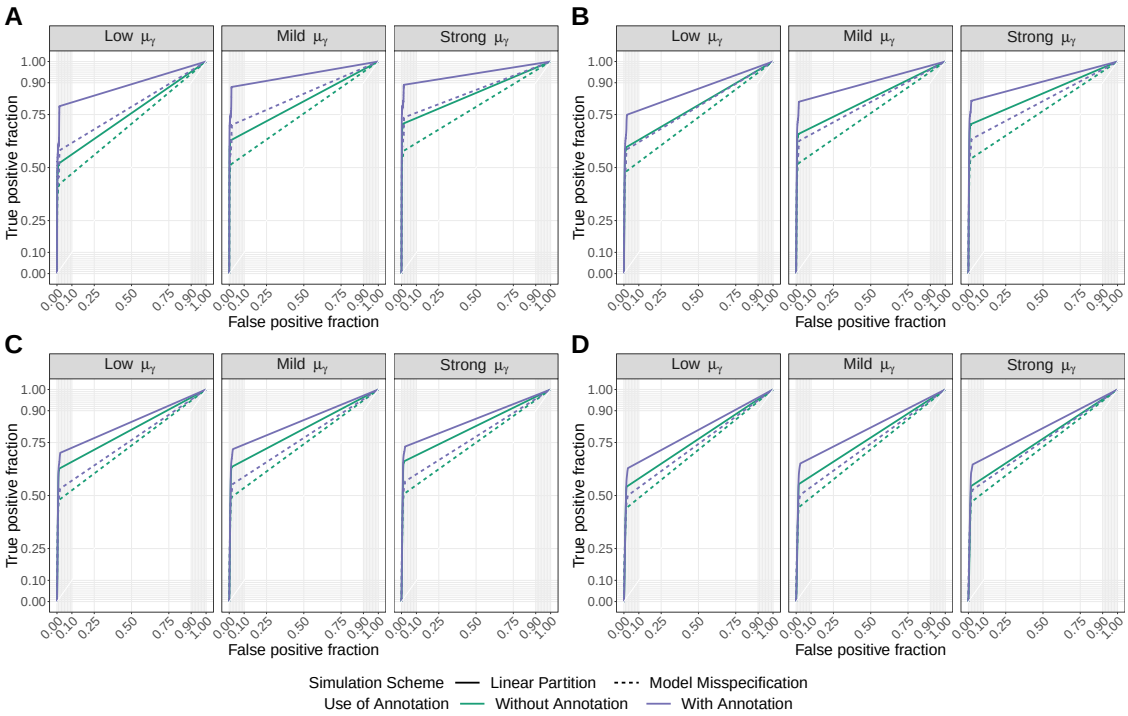

Figure B.4: **SNP selection receiver operating characteristic curves (ROC) for fits with and without annotation comparing simulations generated by linear partition and model misspecification schemes for different simulation scenarios for low, mild, and strong annotation effect magnitude μ_γ.**
(A) $p_{\eta\neq 0} = 0.01$, $p_{\gamma\neq 0} = 0.05$, and $\sigma^2 = 100$. **(B)** $p_{\eta\neq 0} = 0.04$, $p_{\gamma\neq 0} = 0.08$, and $\sigma^2 = 150$. **(C)** $p_{\eta\neq 0} = 0.08$, $p_{\gamma\neq 0} = 0.05$, and $\sigma^2 = 200$. **(D)** $p_{\eta\neq 0} = 0.15$, $p_{\gamma\neq 0} = 0.1$, and $\sigma^2 = 150$.

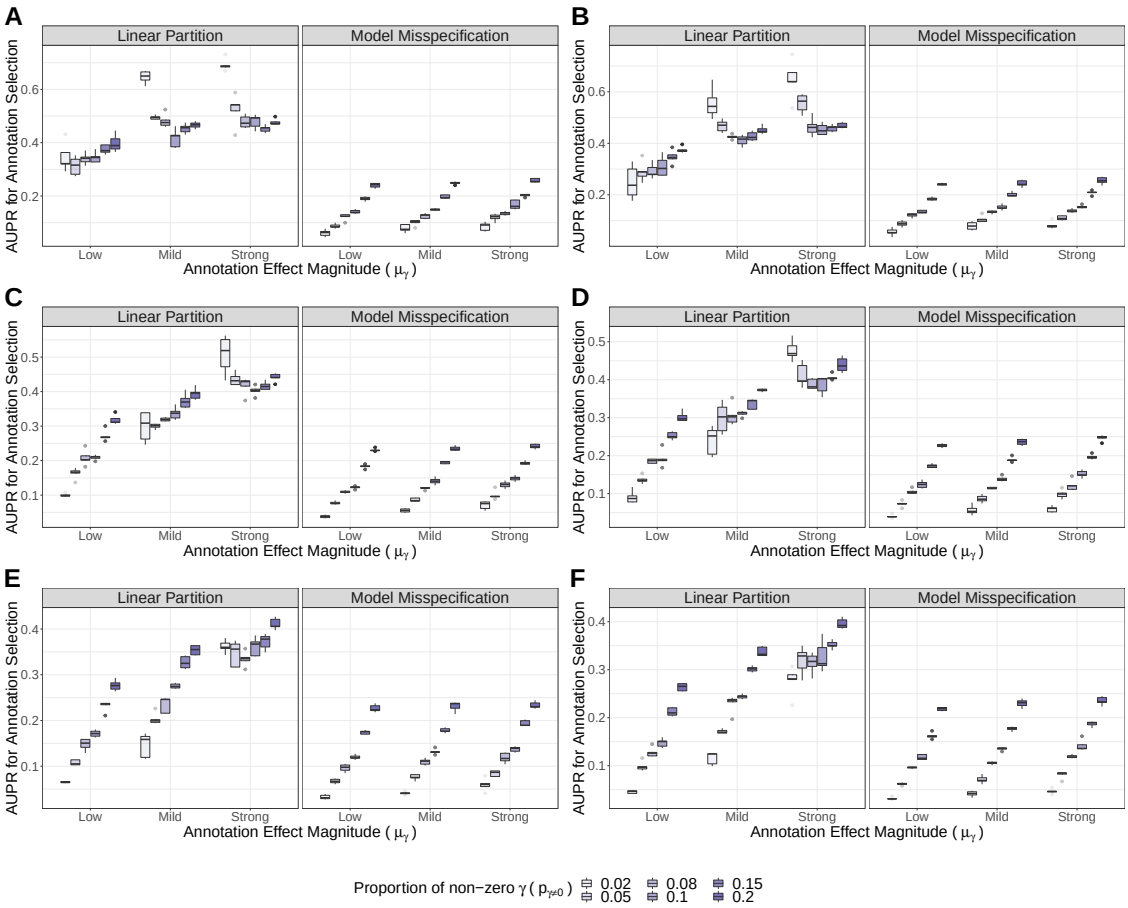

Figure B.5: **Area under the precision-recall curve (AUPR) for annotation selection (γ) for varying proportion of risk SNPs $p_{\eta\neq0}$ across fits comparing simulations generated by linear partition and model misspecification for different proportions of non-zero γ ($p_{\gamma\neq0}$).**
(A) $p_{\eta\neq0} = 0.01$. **(B)** $p_{\eta\neq0} = 0.02$. **(C)** $p_{\eta\neq0} = 0.08$. **(D)** $p_{\eta\neq0} = 0.1$. **(E)** $p_{\eta\neq0} = 0.15$. **(F)** $p_{\eta\neq0} = 0.2$.

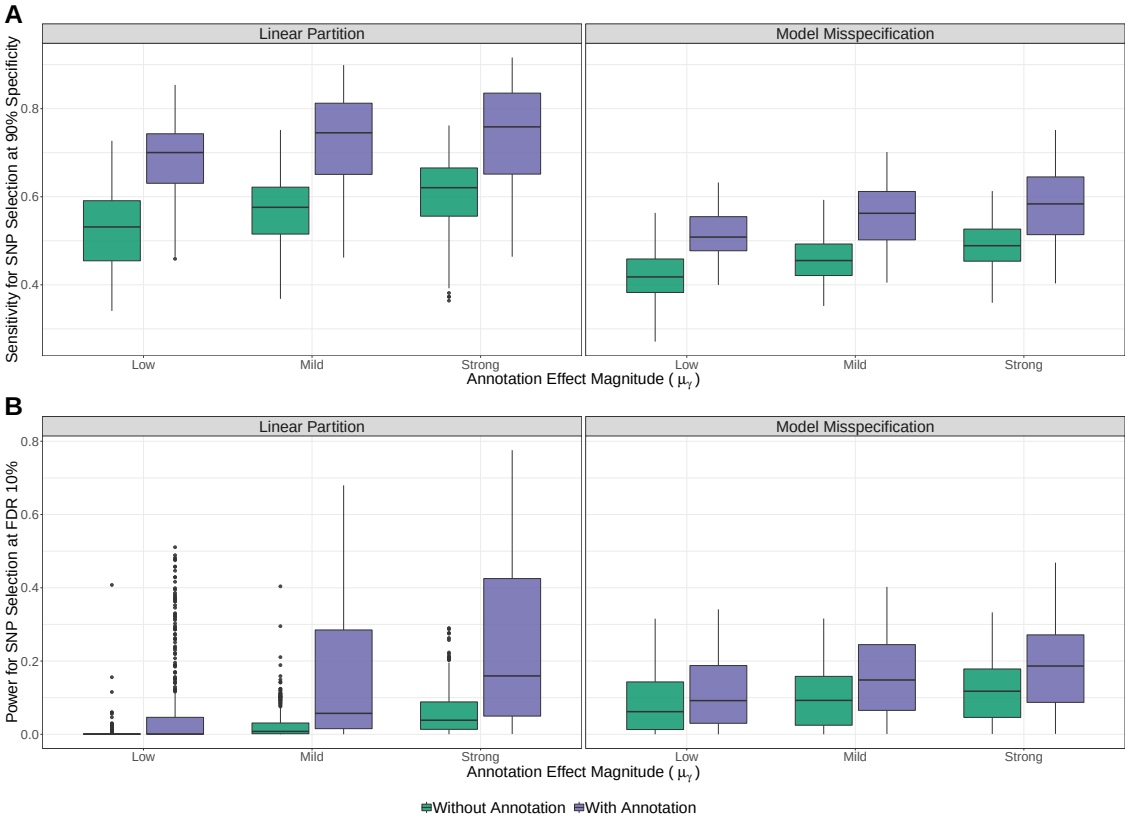

Figure B.6: **SNP selection sensitivity and power calculations for fits with and without annotation comparing simulations generated by linear partition and model misspecification schemes for different simulation scenarios for low, mild, and strong annotation effect magnitude μ_γ.**
(A) Sensitivity at 90% specificity. **(B)** Power at FDR 10%.

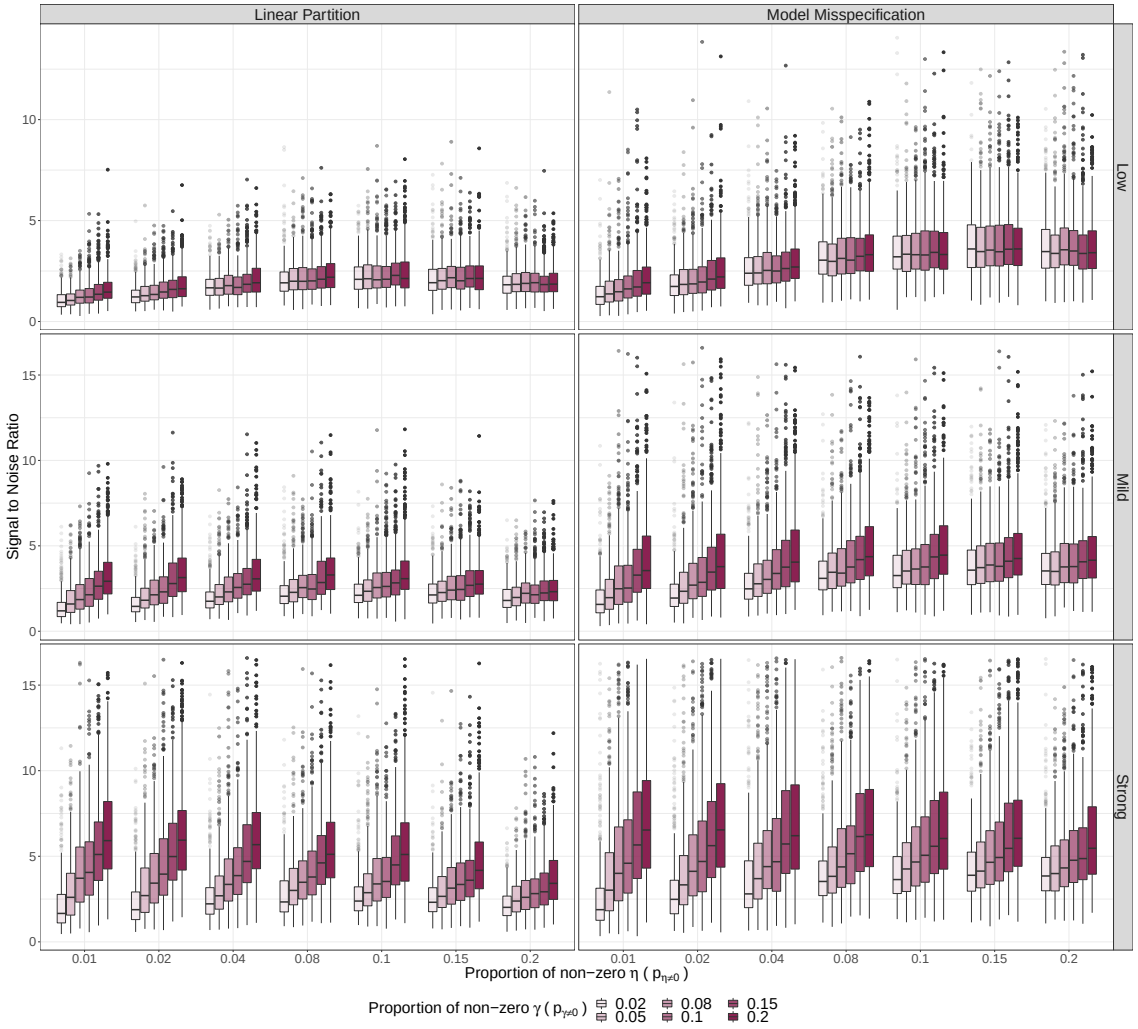

Figure B.7: **Signal to noise ratio $\frac{\|G\hat{\beta}\|_2}{\|e\|_2}$ for fits with annotation comparing simulations generated by linear partition and model misspecification schemes for different proportions of non-zero η and γ ($p_{\eta\neq0}$ and $p_{\gamma\neq0}$, respectively) divided by annotation effect magnitude (low, mild, and strong). For visualization purposes, values on the top 0.5% were removed.**

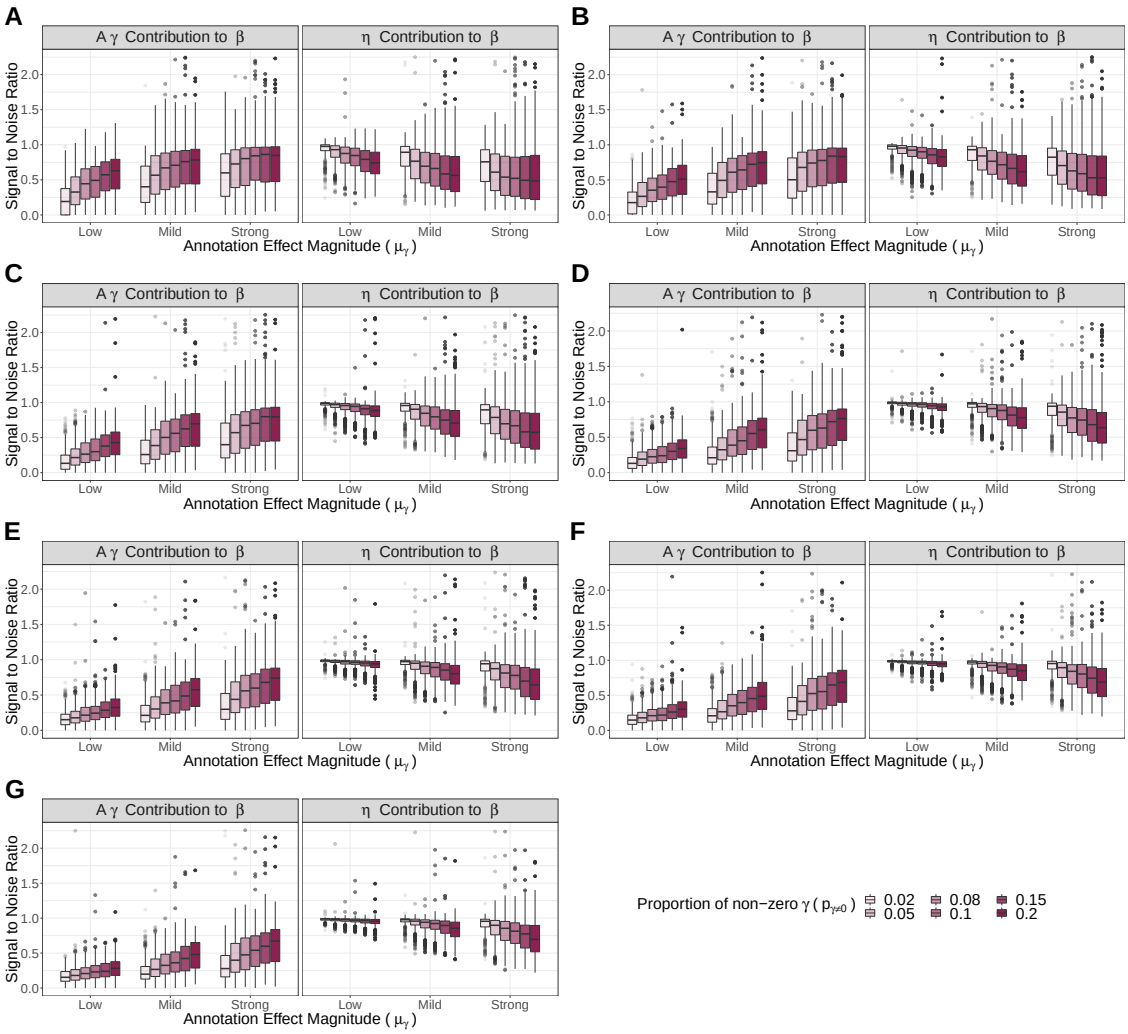

Figure B.8: **Signal of each component ($A\gamma$ and η) to β ratio for fits with annotation comparing simulations generated by linear partition and model misspecification schemes for different proportions of non-zero η and γ ($p_{\eta\neq0}$ and $p_{\gamma\neq0}$, respectively) divided by annotation effect magnitude μ_γ (low, mild, and strong). The $A\gamma$ contribution to β corresponds to $\frac{\|A\hat{\gamma}\|_2}{\|\hat{\beta}\|_2}$ and the η contribution to β corresponds to $\frac{\|\hat{\eta}\|_2}{\|\hat{\beta}\|_2}$. For visualization purposes, values on the top 0.5% were removed.**

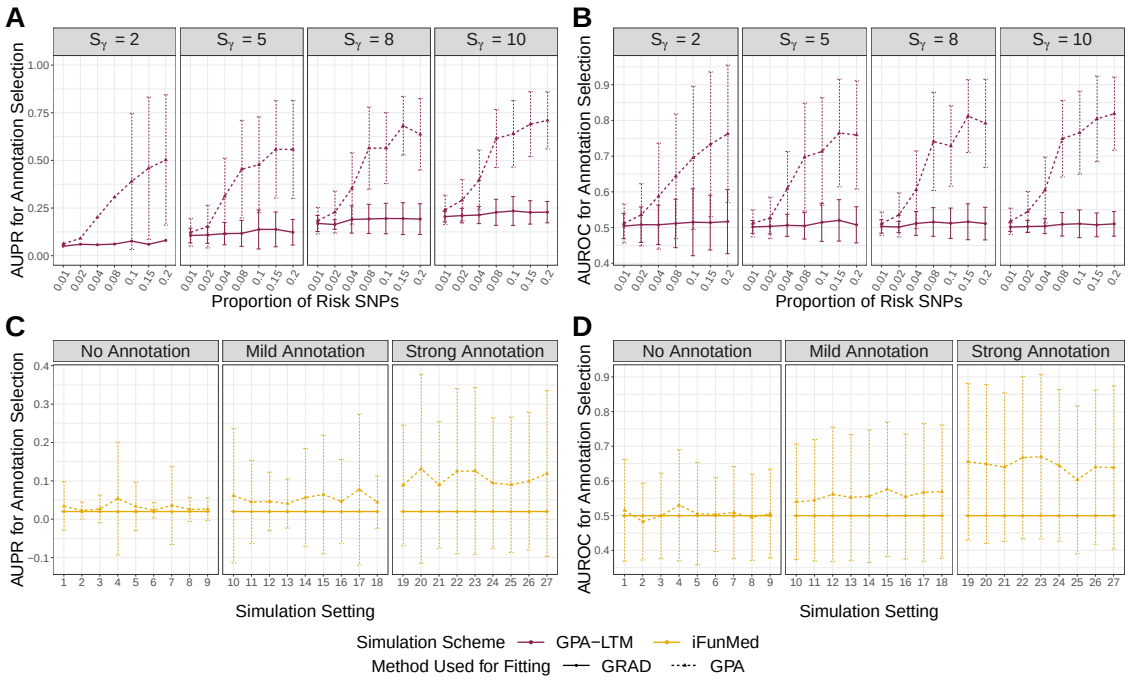

Figure B.9: **Annotation selection simulations results for GRAD and GPA comparisons.**
For the GPA liability threshold model (GPA-LTM) used for simulations, results are divided by the number of risk annotations $S_\gamma \in \{2, 5, 8, 10\}$, and for the *iFunMed* simulations results are divided by their prior inclusion probabilities with the use of annotation (no annotation effect, mild annotation effect, and strong annotation effect). Results are summarized by area under the precision-recall curves (AUPR) and area under the receiver operating characteristic curves (AUROC). **(A)** Average AUPR for annotation selection across 100 simulation replicates and their corresponding error bars (mean $\pm$ standard deviation) for GRAD and GPA. **(B)** Average AUROC for annotation selection across 100 simulation replicates and their corresponding error bars (mean $\pm$ standard deviation) for GRAD and GPA.

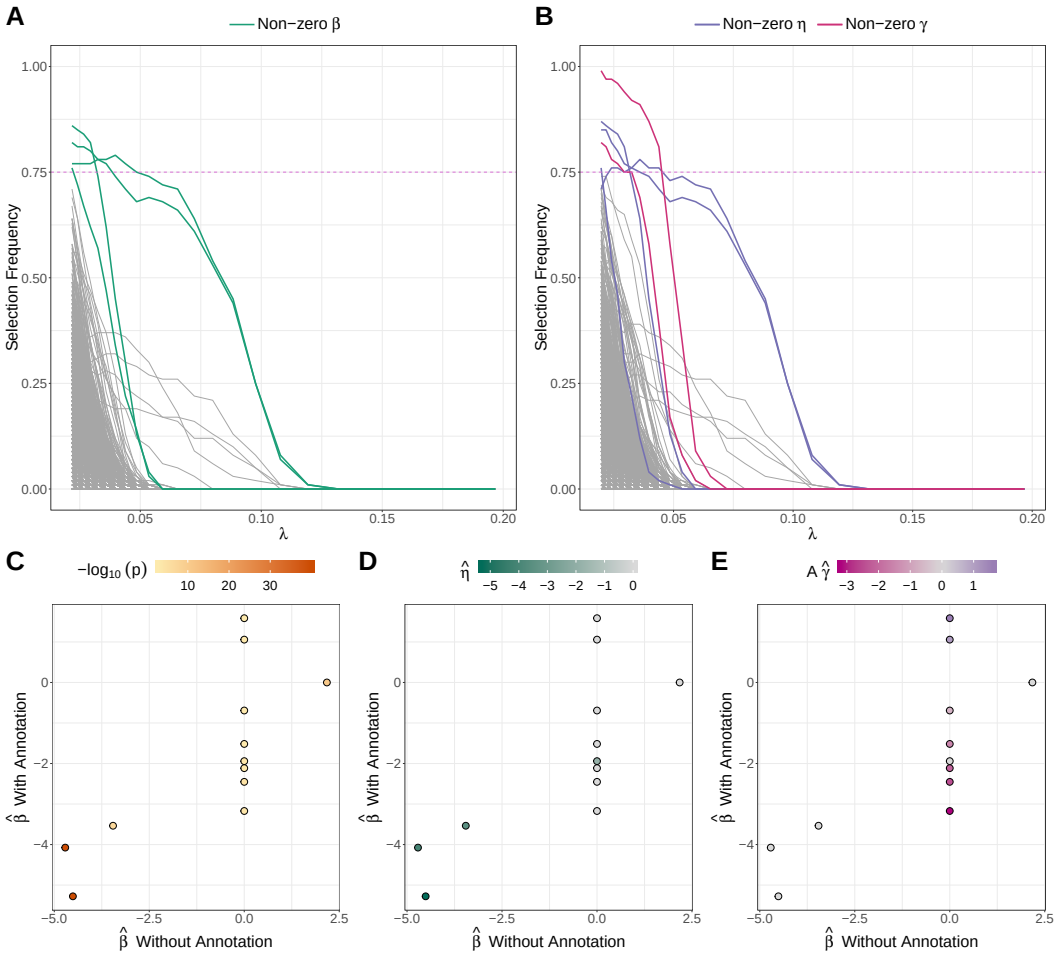

Figure B.10: **Stability selection results for factor VII.**
(A, B) Stability paths for each parameter included in the model. Colored paths indicate non-zero estimated parameters. Dashed line represents selection frequency cutoff of 0.75. **(A)** Without annotation and **(B)** with annotation. **(C, E)** Estimated SNP effect sizes across fits with and without annotation. SNPs with effect sizes exactly equal to zero with and without annotation are omitted. SNPs are colored by **(C)** $-\log_{10}$ transformed p-values from univariate GWAS associations, **(D)** strength of the annotation free contribution $\hat{\boldsymbol{\eta}}$ from the model with annotation, and **(E)** strength of the annotation contribution $A\hat{\boldsymbol{\gamma}}$ from the model with annotation.

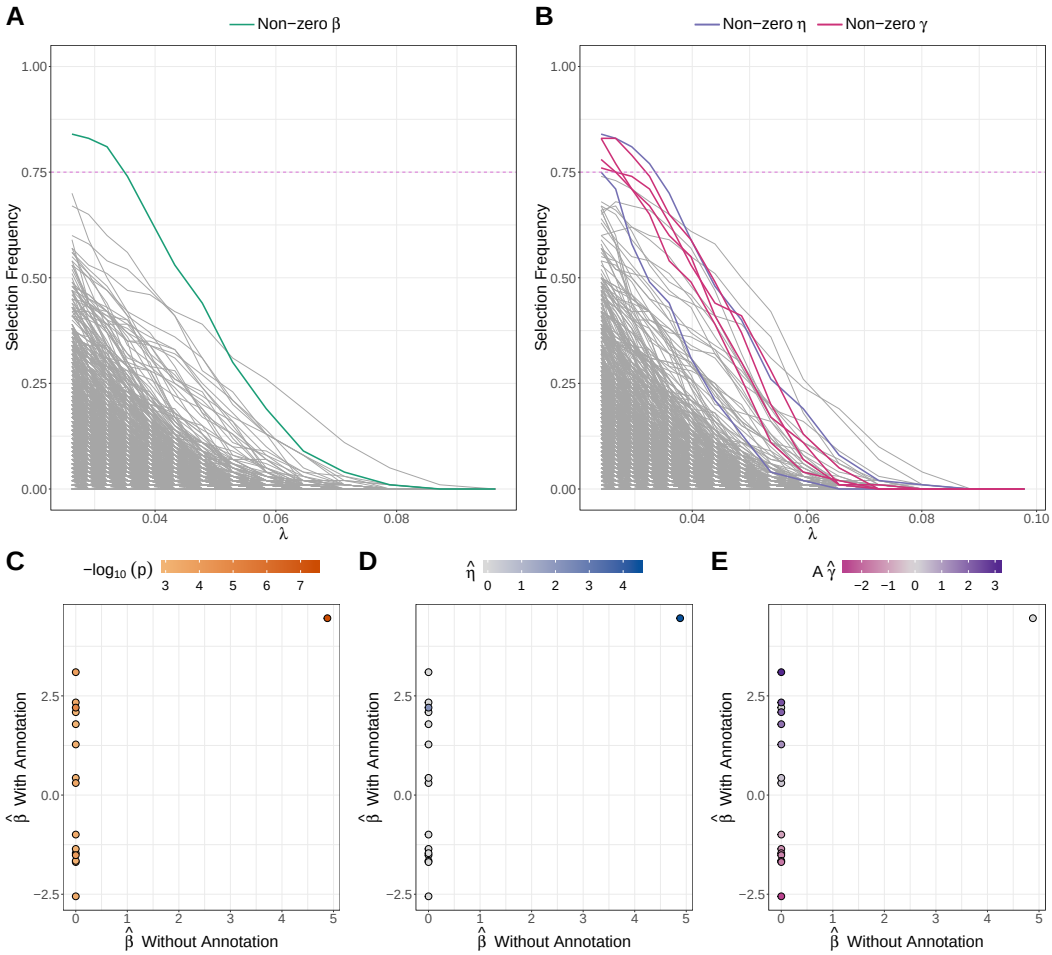

Figure B.11: **Stability selection results for fasting glucose** (log)**.**
(A, B) Stability paths for each parameter included in the model. Colored paths indicate non-zero estimated parameters. Dashed line represents selection frequency cutoff of 0.75. **(A)** Without annotation and **(B)** with annotation. **(C, E)** Estimated SNP effect sizes across fits with and without annotation. SNPs with effect sizes exactly equal to zero with and without annotation are omitted. SNPs are colored by **(C)** $-\log_{10}$ transformed p-values from univariate GWAS associations, **(D)** strength of the annotation free contribution $\hat{\boldsymbol{\eta}}$ from the model with annotation, and **(E)** strength of the annotation contribution $A\hat{\boldsymbol{\gamma}}$ from the model with annotation.

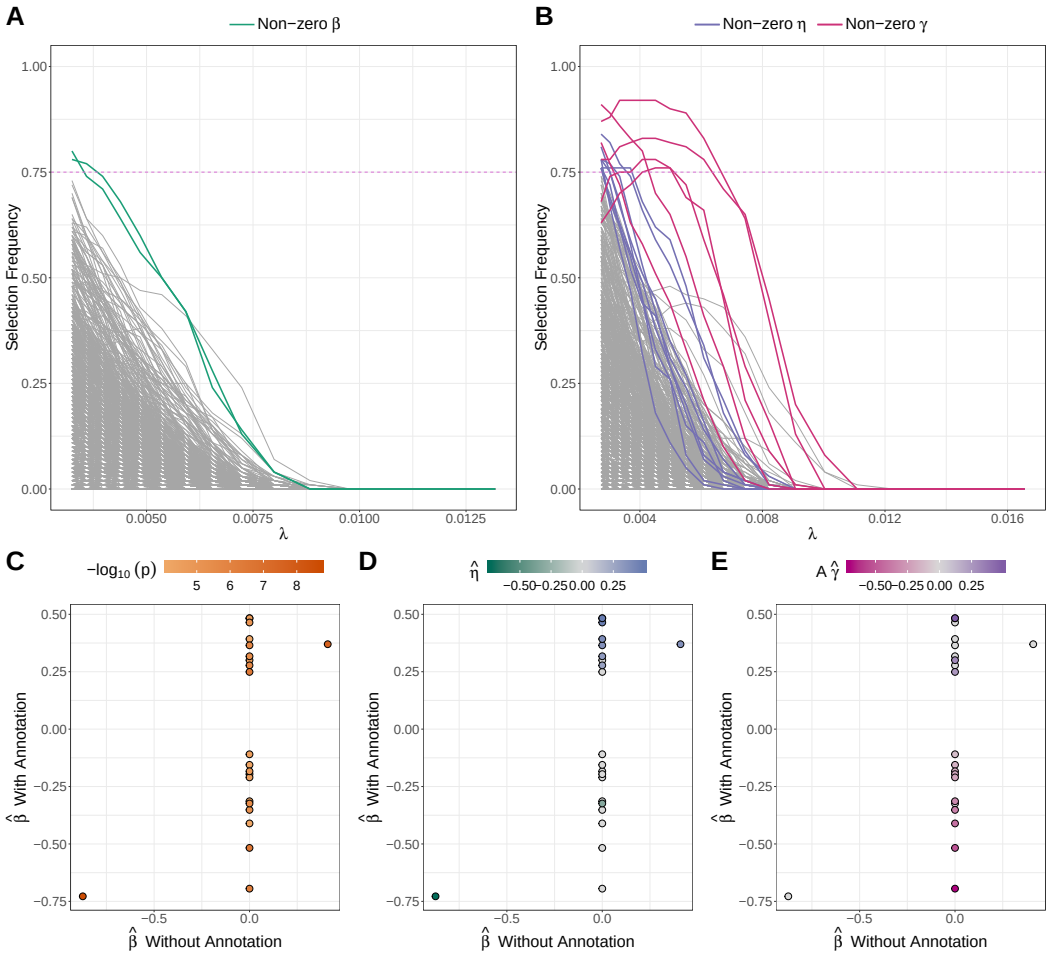

Figure B.12: **Stability selection results for Height.**
(A, B) Stability paths for each parameter included in the model. Colored paths indicate non-zero estimated parameters. Dashed line represents selection frequency cutoff of 0.75. **(A)** Without annotation and **(B)** with annotation. **(C, E)** Estimated SNP effect sizes across fits with and without annotation. SNPs with effect sizes exactly equal to zero with and without annotation are omitted. SNPs are colored by **(C)** $-\log_{10}$ transformed p-values from univariate GWAS associations, **(D)** strength of the annotation free contribution $\hat{\boldsymbol{\eta}}$ from the model with annotation, and **(E)** strength of the annotation contribution $A\hat{\boldsymbol{\gamma}}$ from the model with annotation.

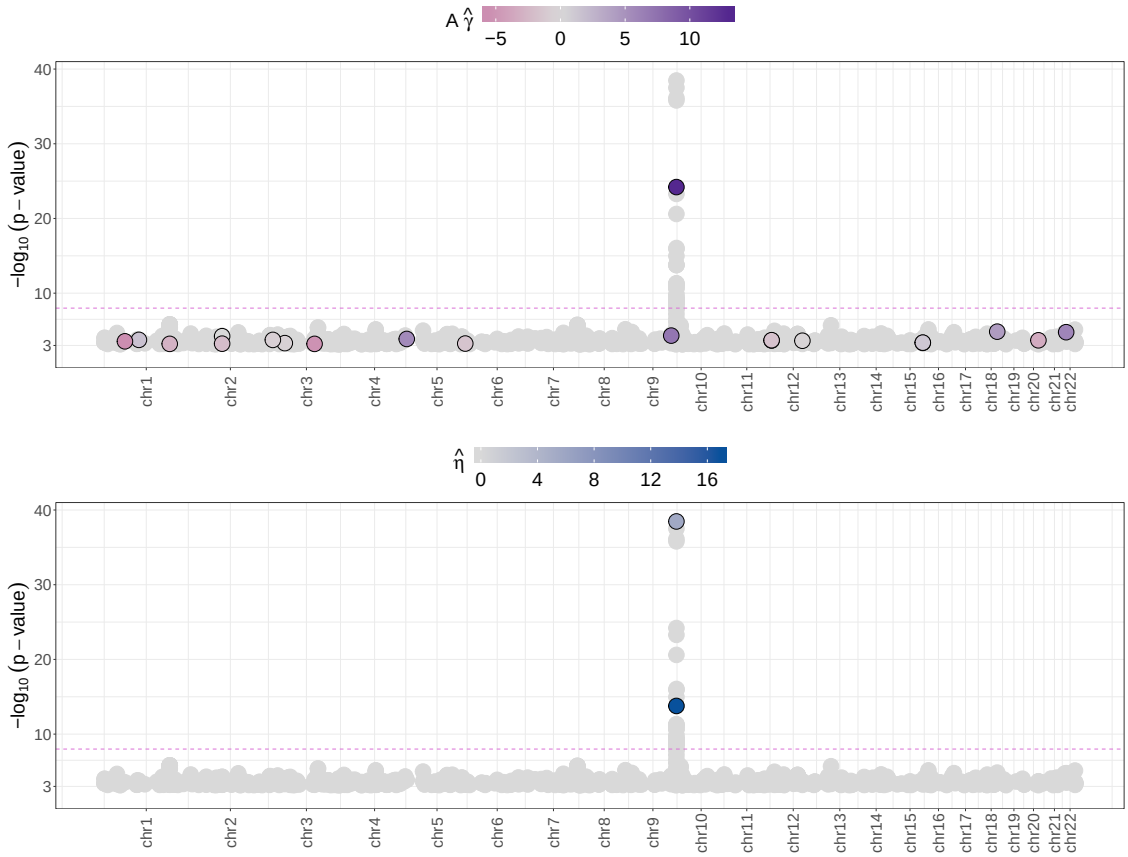

Figure B.13: **Manhattan plots of von Willebrand factor for the 1,000 SNPs considered in the analysis.**
SNPs are colored by their effect sizes and dashed line represents GWAS traditional cutoff at $p - value = 10^{-8}$. The upper panel highlights non-zero SNPs that are captured by the annotation and lower panel represents annotation free non-zero SNPs. In both cases, gray represents zero estimated effects.

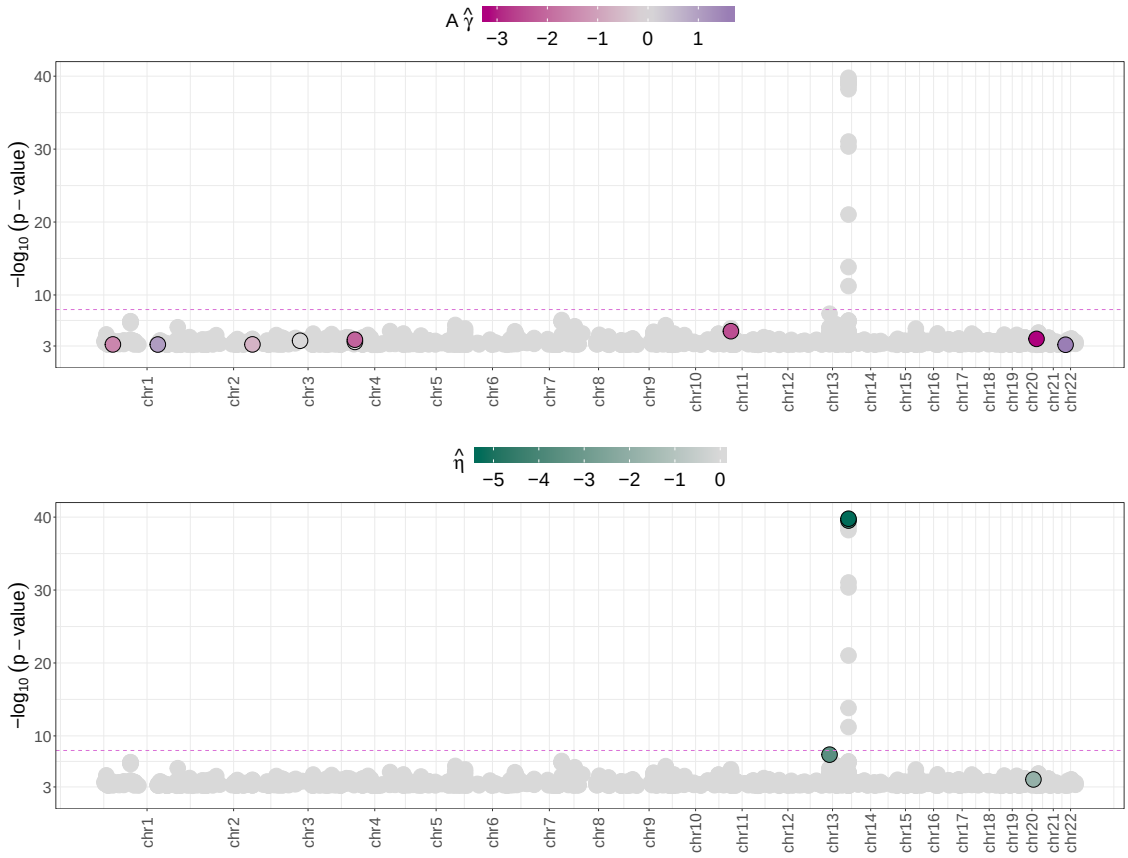

Figure B.14: **Manhattan plots of factor VII for the 1,000 SNPs considered in the analysis.**
SNPs are colored by their effect sizes and dashed line represents GWAS traditional cutoff at $p - value = 10^{-8}$. The upper panel highlights non-zero SNPs that are captured by the annotation and lower panel represents annotation free non-zero SNPs. In both cases, gray represents zero estimated effects.

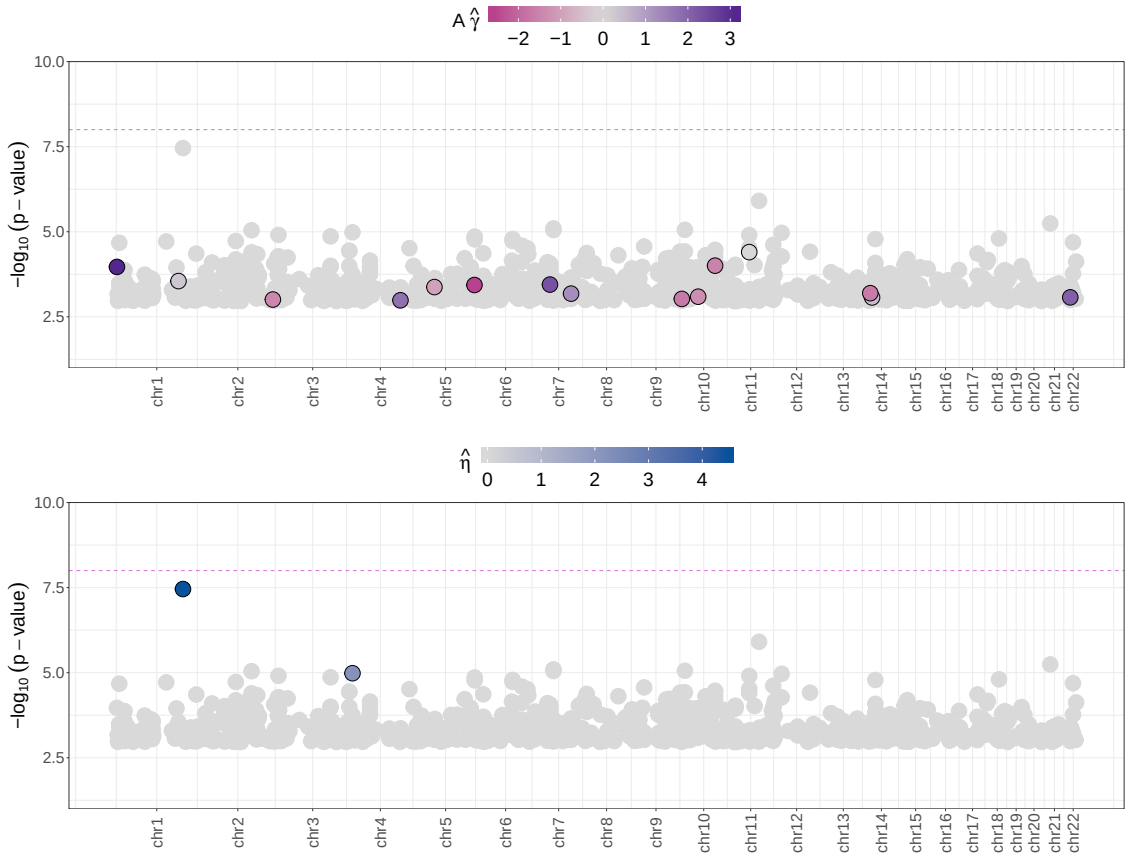

Figure B.15: **Manhattan plots of fasting glucose (**log**) for the 1,000 SNPs considered in the analysis.**
SNPs are colored by their effect sizes and dashed line represents GWAS traditional cutoff at $p-value = 10^{-8}$. The upper panel highlights non-zero SNPs that are captured by the annotation and lower panel represents annotation free non-zero SNPs. In both cases, gray represents zero estimated effects.

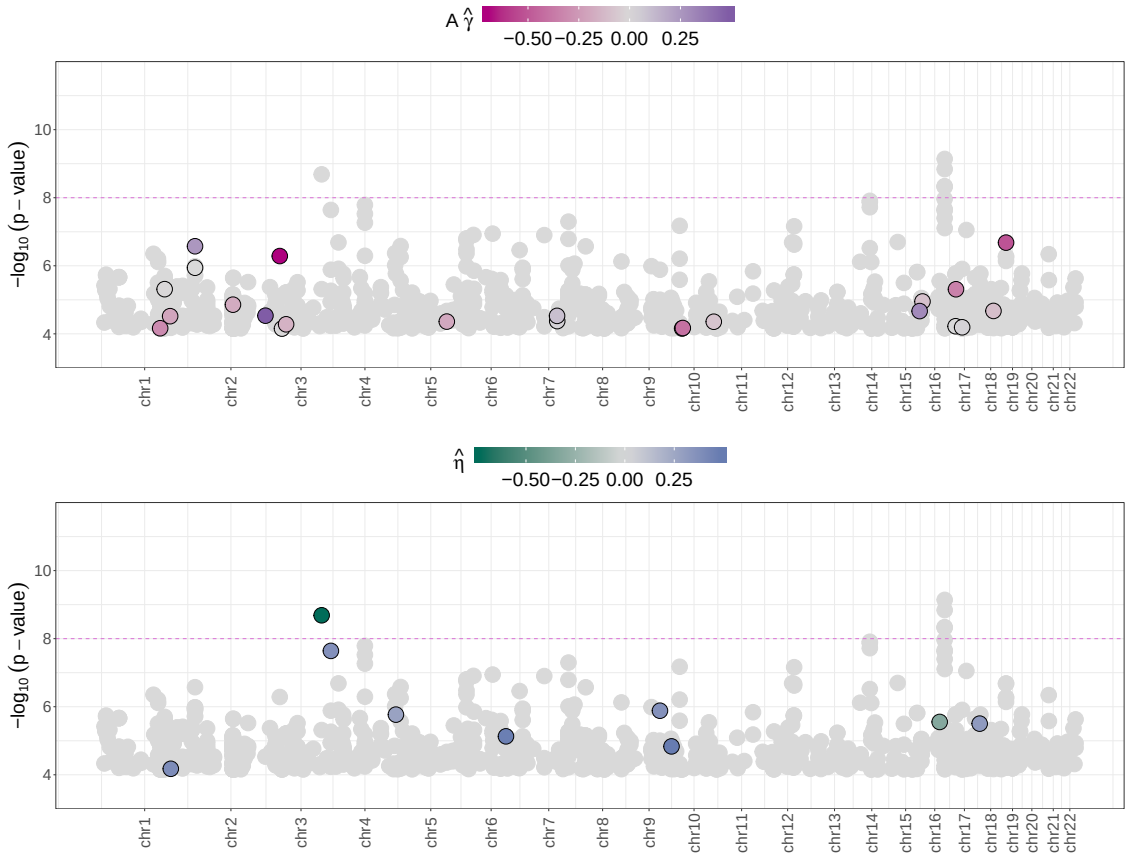

Figure B.16: **Manhattan plots of height for the 1,000 SNPs considered in the analysis.**
SNPs are colored by their effect sizes and dashed line represents GWAS traditional cutoff at $p - value = 10^{-8}$. The upper panel highlights non-zero SNPs that are captured by the annotation and lower panel represents annotation free non-zero SNPs. In both cases, gray represents zero estimated effects.

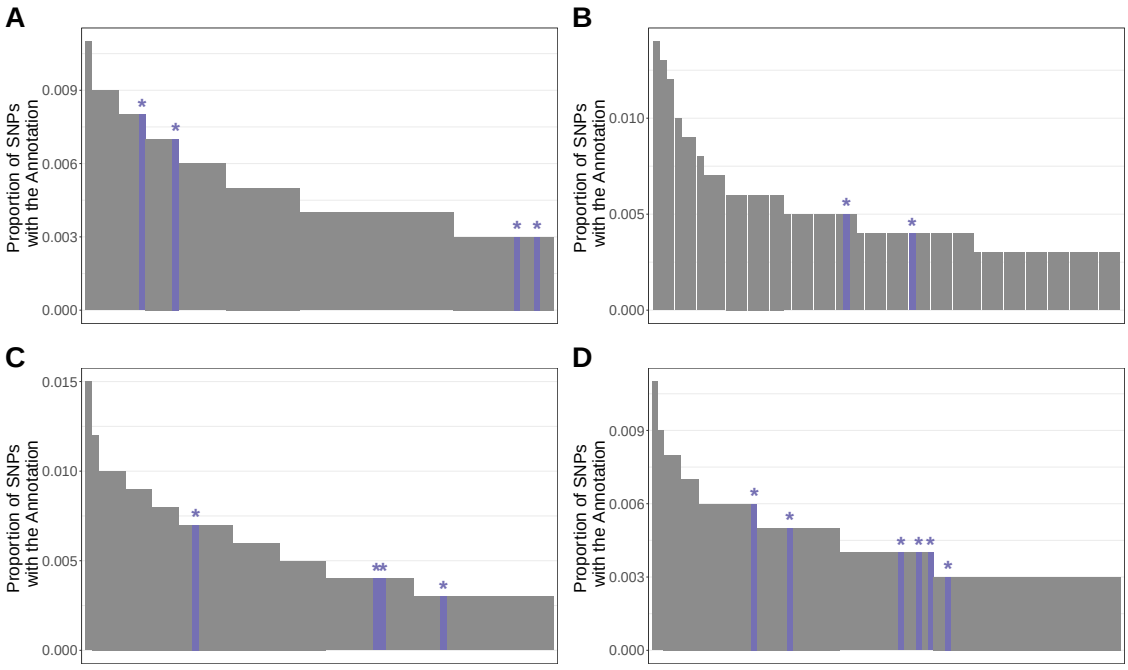

Figure B.17: **Proportion of SNPs with corresponding non-zero entries across the m annotations for the four phenotypes considered. Selected annotations by GRAD are highlighted in purple and with asterisks.**
(A) von Willebrand factor with $m = 70$ annotations. **(B)** Factor VII with $m = 64$ annotations. **(C)** Fasting glucose (log) with $m = 69$ annotations. **(D)** Height with $m = 78$ annotations.

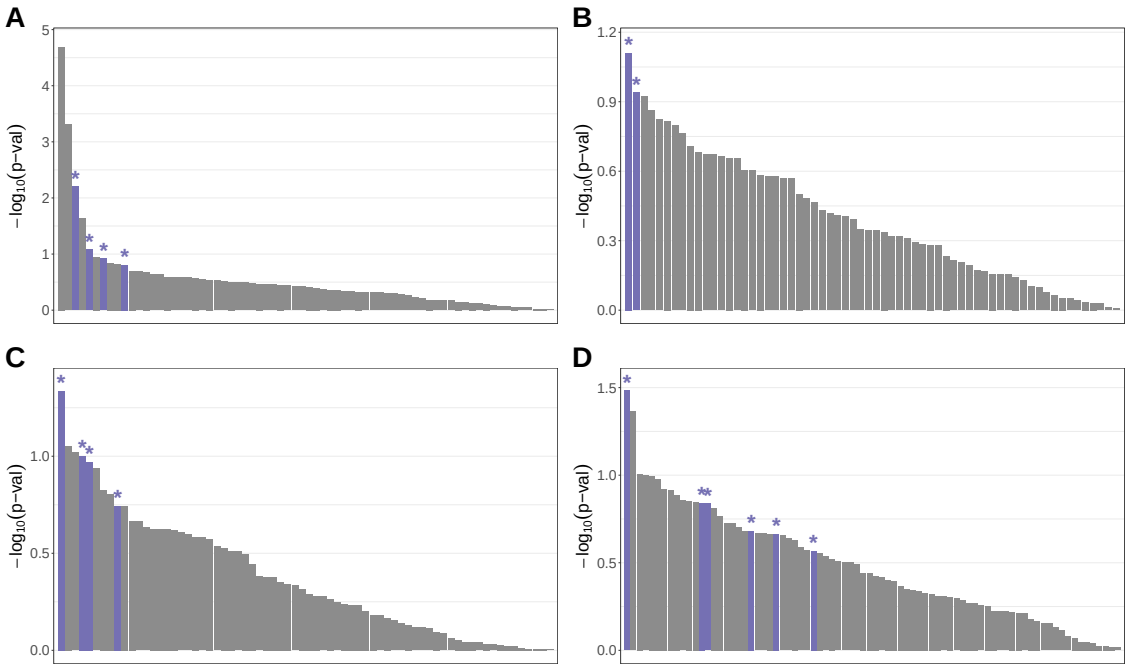

Figure B.18: $-\log_{10}$ **transformed p-values from the univariate association analysis of GWAS summary statistics (Z_Y) with individual annotations, i.e. $Z_Y \sim A_j$ for $j = 1, \ldots, m$, across the four phenotypes considered. Selected annotations by GRAD are highlighted in purple and with asterisks.**
(A) Von Willebrand factor with $m = 70$ annotations. **(B)** Factor VII with $m = 64$ annotations. **(C)** Fasting glucose (log) with $m = 69$ annotations. **(D)** Height with $m = 78$ annotations.

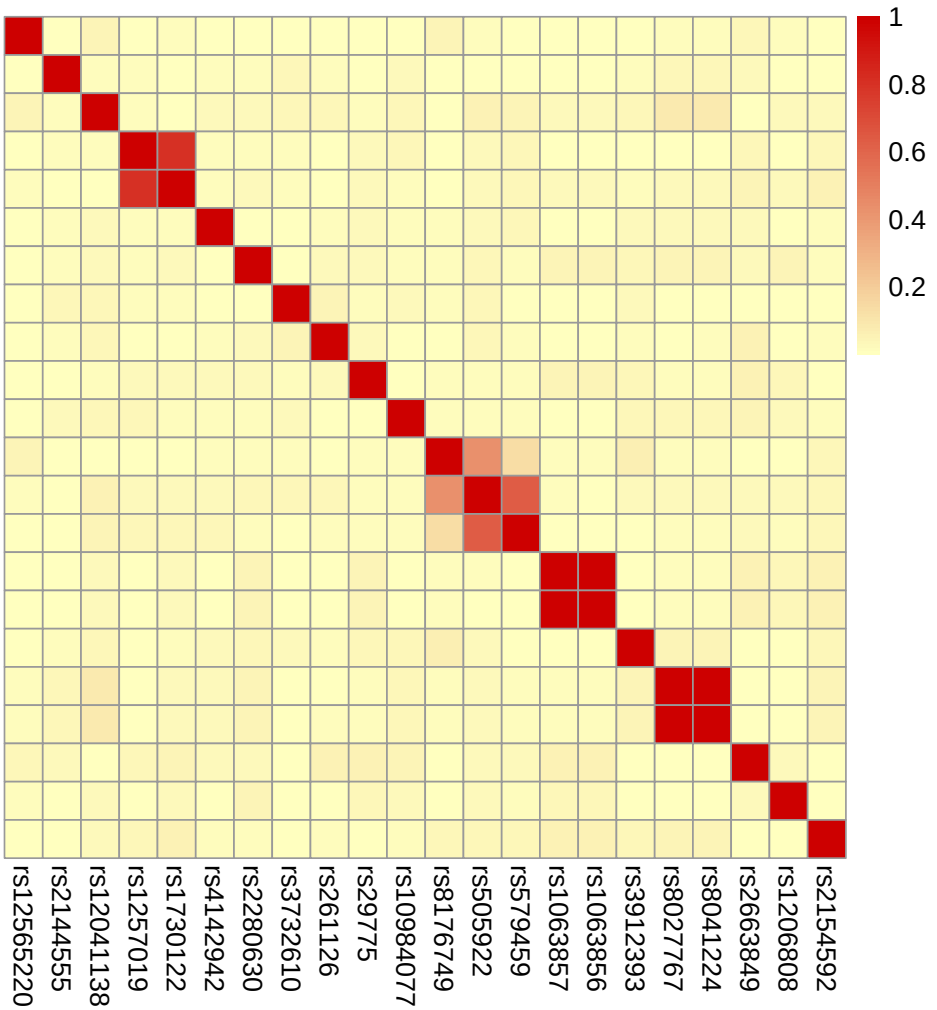

Figure B.19: **Pearson's correlation magnitude, i.e. absolute value, of non-zero SNPs from the fit with annotation for von Willebrand factor.**

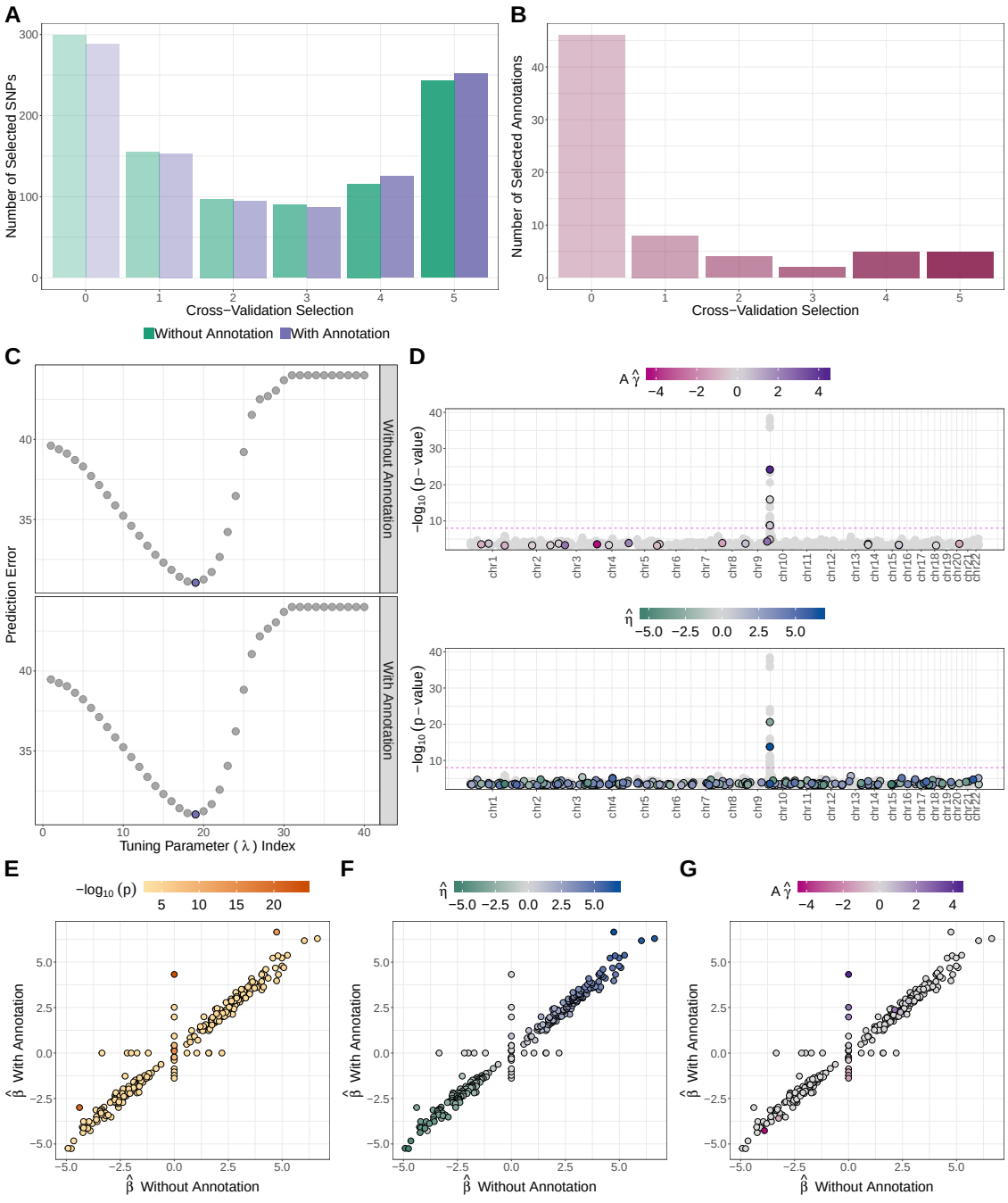
A
Number of Selected SNPs
Cross−Validation Selection
Without Annotation
With Annotation
B
Number of Selected Annotations
Cross−Validation Selection
C
Prediction Error
Without Annotation
With Annotation
Tuning Parameter (λ) Index
D
$A\hat{\gamma}$
$-\log_{10}$ (p − value)
$\hat{\eta}$
E
$-\log_{10}$ (ρ)
$\hat{\beta}$ With Annotation
$\hat{\beta}$ Without Annotation
F
G

Figure B.20: **5-fold cross-validation results for von Willebrand factor.**
(A) Number of selected SNPs across the five cross-validation sets for fits without and with annotation. **(B)** Number of selected annotations across the five cross-validation sets. **(C)** Prediction error across 40 values of the tuning parameter λ for fits without and with annotation. Colored point represents the selected λ value that minimizes the prediction error. **(D)** Manhattan plots from univariate GWAS results. SNPs are colored by their effect sizes and dashed line represents GWAS traditional cutoff at $p - \text{value} = 10^{-8}$. The upper panel highlights non-zero SNPs that are captured by the annotation and lower panel represents annotation free non-zero SNPs. In both cases, gray represents zero estimated effects. **(E, G)** Estimated SNP effect sizes across fits with and without annotation. SNPs with effect sizes exactly equal to zero with and without annotation are omitted. SNPs are colored by **(E)** $-\log_{10}$ transformed p-values from univariate GWAS associations, **(F)** strength of the annotation free contribution $\hat{\boldsymbol{\eta}}$ from the model with annotation, and **(G)** strength of the annotation contribution $A\hat{\boldsymbol{\gamma}}$ from the model with annotation.

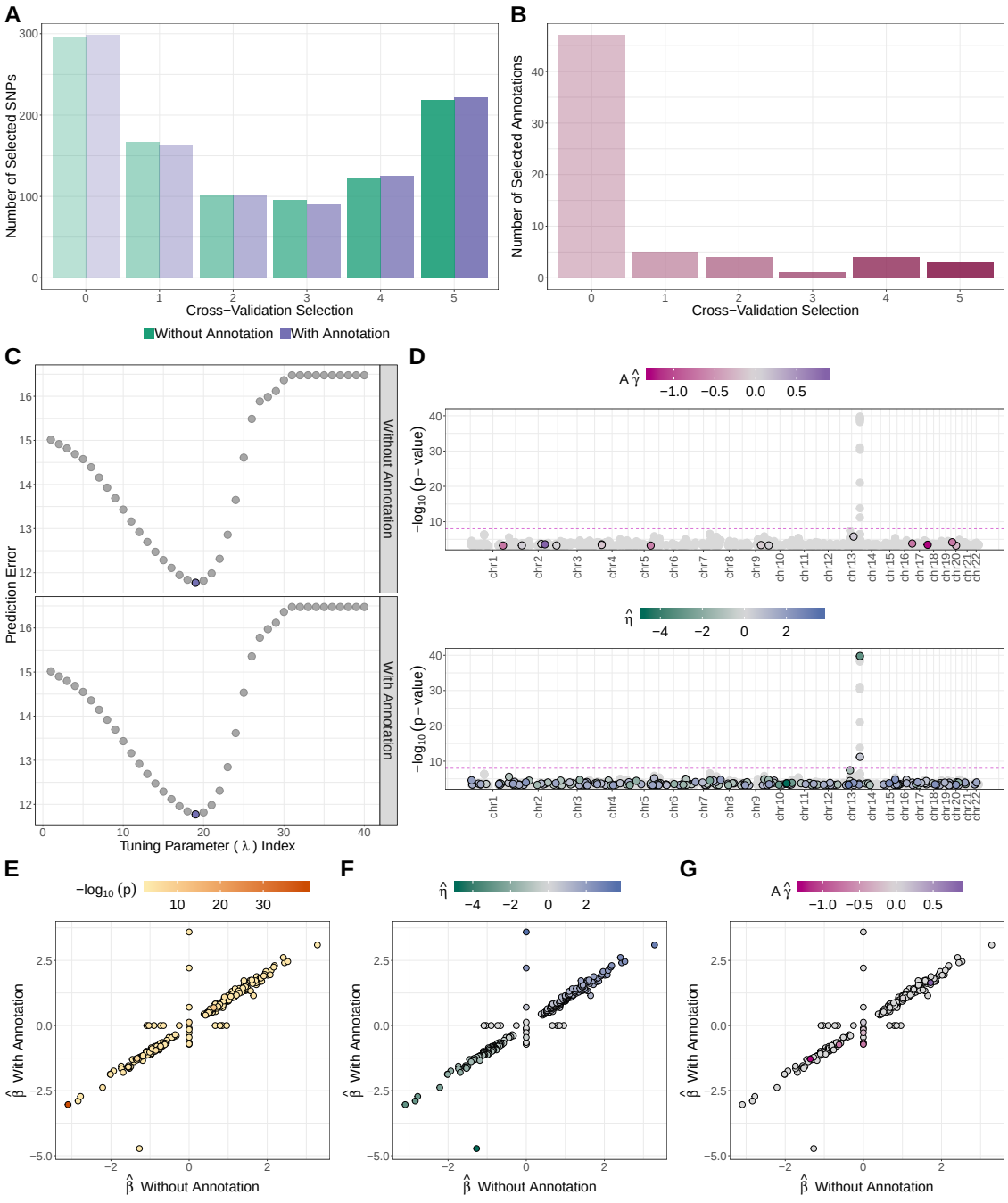
A
Number of Selected SNPs
Cross−Validation Selection
Without Annotation
With Annotation
B
Number of Selected Annotations
Cross−Validation Selection
C
Prediction Error
Without Annotation
With Annotation
Tuning Parameter (λ) Index
D
−log10 (p − value)
E
−log10 (ρ)
β̂ With Annotation
β̂ Without Annotation
F
η̂
G

Figure B.21: **5-fold cross-validation results for factor VII.**
(A) Number of selected SNPs across the five cross-validation sets for fits without and with annotation. **(B)** Number of selected annotations across the five cross-validation sets. **(C)** Prediction error across 40 values of the tuning parameter λ for fits without and with annotation. Colored point represents the selected λ value that minimizes the prediction error. **(D)** Manhattan plots from univariate GWAS results. SNPs are colored by their effect sizes and dashed line represents GWAS traditional cutoff at $p - value = 10^{-8}$. The upper panel highlights non-zero SNPs that are captured by the annotation and lower panel represents annotation free non-zero SNPs. In both cases, gray represents zero estimated effects. **(E, G)** Estimated SNP effect sizes across fits with and without annotation. SNPs with effect sizes exactly equal to zero with and without annotation are omitted. SNPs are colored by **(E)** $-\log_{10}$ transformed p-values from univariate GWAS associations, **(F)** strength of the annotation free contribution $\hat{\boldsymbol{\eta}}$ from the model with annotation, and **(G)** strength of the annotation contribution $A\hat{\boldsymbol{\gamma}}$ from the model with annotation.

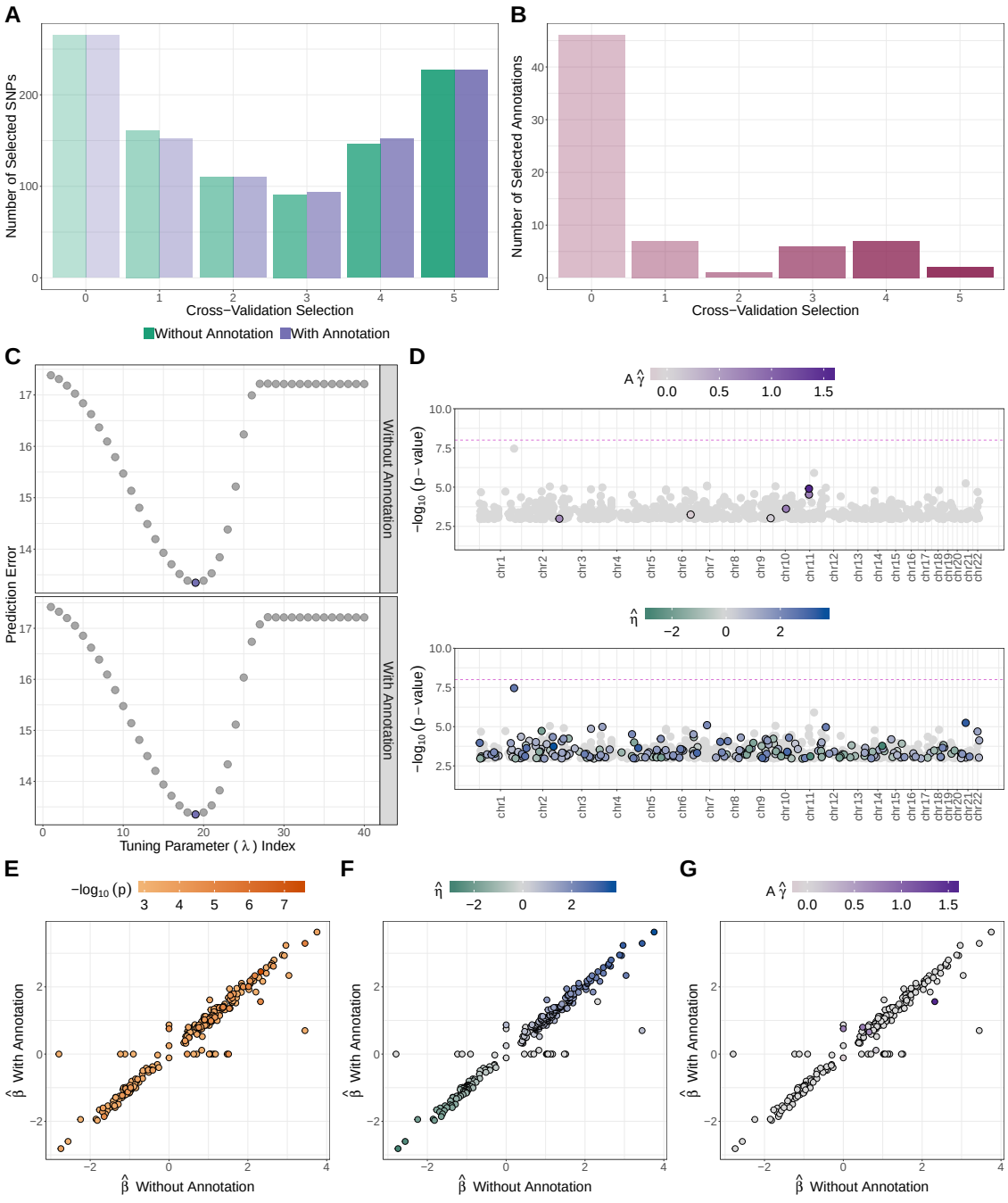
A
Number of Selected SNPs
Cross−Validation Selection
Without Annotation
With Annotation
B
Number of Selected Annotations
Cross−Validation Selection
C
Prediction Error
Without Annotation
With Annotation
Tuning Parameter (λ) Index
D
$\hat{\gamma}_A$
$-\log_{10}$ (p − value)
$\hat{\eta}$
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chr20
chr21
chr22
E
$-\log_{10}$ (p)
$\hat{\beta}$ With Annotation
$\hat{\beta}$ Without Annotation
F
G

Figure B.22: **5-fold cross-validation results for fasting glucose (**$l\log$**).**
(A) Number of selected SNPs across the five cross-validation sets for fits without and with annotation. **(B)** Number of selected annotations across the five cross-validation sets. **(C)** Prediction error across 40 values of the tuning parameter λ for fits without and with annotation. Colored point represents the selected λ value that minimizes the prediction error. **(D)** Manhattan plots from univariate GWAS results. SNPs are colored by their effect sizes and dashed line represents GWAS traditional cutoff at $p-\text{value} = 10^{-8}$. The upper panel highlights non-zero SNPs that are captured by the annotation and lower panel represents annotation free non-zero SNPs. In both cases, gray represents zero estimated effects. **(E, G)** Estimated SNP effect sizes across fits with and without annotation. SNPs with effect sizes exactly equal to zero with and without annotation are omitted. SNPs are colored by **(E)** $-\log_{10}$ transformed p-values from univariate GWAS associations, **(F)** strength of the annotation free contribution $\hat{\boldsymbol{\eta}}$ from the model with annotation, and **(G)** strength of the annotation contribution $A\hat{\boldsymbol{\gamma}}$ from the model with annotation.

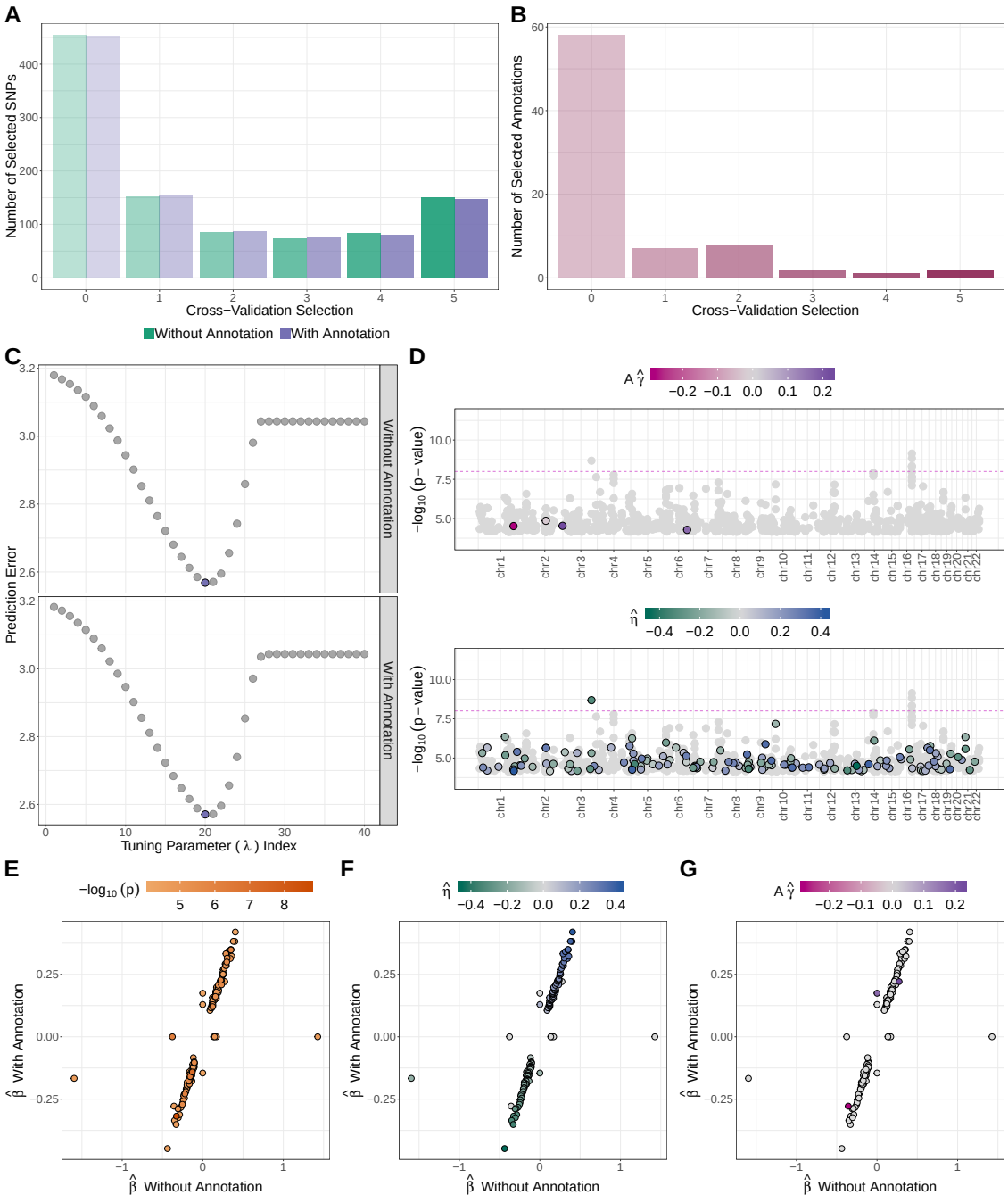
A
Number of Selected SNPs
Cross−Validation Selection
Without Annotation
With Annotation
B
Number of Selected Annotations
Cross−Validation Selection
C
Prediction Error
Without Annotation
With Annotation
Tuning Parameter (λ) Index
D
−log10 (p − value)
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chr20
chr21
chr22
E
−log10 (p)
β̂ With Annotation
β̂ Without Annotation
F
η̂
G
γ̂

Figure B.23: **5-fold cross-validation results for height.**
(A) Number of selected SNPs across the five cross-validation sets for fits without and with annotation. **(B)** Number of selected annotations across the five cross-validation sets. **(C)** Prediction error across 40 values of the tuning parameter λ for fits without and with annotation. Colored point represents the selected λ value that minimizes the prediction error. **(D)** Manhattan plots from univariate GWAS results. SNPs are colored by their effect sizes and dashed line represents GWAS traditional cutoff at $p - \text{value} = 10^{-8}$. The upper panel highlights non-zero SNPs that are captured by the annotation and lower panel represents annotation free non-zero SNPs. In both cases, gray represents zero estimated effects. **(E, G)** Estimated SNP effect sizes across fits with and without annotation. SNPs with effect sizes exactly equal to zero with and without annotation are omitted. SNPs are colored by **(E)** $-\log_{10}$ transformed p-values from univariate GWAS associations, **(F)** strength of the annotation free contribution $\hat{\boldsymbol{\eta}}$ from the model with annotation, and **(G)** strength of the annotation contribution $A\hat{\boldsymbol{\gamma}}$ from the model with annotation.

B.3 Supplementary Tables for "High dimensional sparse regression with auxiliary data on the features"

Table B.1: Details on the simulation settings for the iFunMed scheme. The SNP effect error variance component corresponds to the ν parameter and the error variance to σ^2 in the iFunMed model. The annotation effects correspond to the γ parameter on iFunMed. The prior probabilities of being non-zero with the use of annotation changes as 0.018 for with and without annotation for no effect, from 0.011 to 0.076 for a mild effect, and 0.047 to 0.269 for strong effect changes.

	No Annotation Effect									Mild Annotation Effect									Strong Annotation Effect								
Set Number	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27
SNP effect variance	10	20	100	10	20	100	10	20	100	10	20	100	10	20	100	10	20	100	10	20	100	10	20	100	10	20	100
Error variance	0.1	0.1	0.1	1	1	1	5	5	5	0.1	0.1	0.1	1	1	1	5	5	5	0.1	0.1	0.1	1	1	1	5	5	5

Table B.2: Power calculations with FDR at 10% from precision-recall curves (AUPR) stratified by annotation effect magnitude μ_γ (low, mild, and strong) for simulation scenarios displayed in Figure 1C ($p_{\eta \neq 0} = 0.01$, $p_{\gamma \neq 0} = 0.05$, and $\sigma^2 = 100$) with and without annotation.

	Linear Partition		Model Misspecification	
Annotation Effect	Without Annotation	With Annotation	Without Annotation	With Annotation
Low	0.004	0.303	0.153	0.197
Mild	0.295	0.662	0.209	0.372
Strong	0.157	0.729	0.231	0.352

Table B.3: List of SNPs selected in GRAD for factor VII. SNP signals refers to the direction of the estimated SNP effect sizes for fits without and with annotation. Details of the annotations included at the individual SNP level are also displayed. Bold SNPs have evidence of association in the GWAS Catalog.

		Signal		Annotation	
SNP ID	Location	wo. Ann	w. Ann	Mxi1	CTCF
rs4970519	chr1:27659500	0	−	0	+
rs11264339	chr1:155140648	0	+	+	0
rs2706126	chr2:178026558	0	−	0	+
rs9878609	chr3:72305546	0	+	0	−
rs13148961	chr4:40625135	0	−	0	+
rs11732608	chr4:40629682	0	−	−	0
rs2553808	chr11:35164108	0	−	0	+
rs9597985	chr13:59912839	−	−	0	0
rs555212	**chr13:113756540**	+	**0**	**0**	**0**
rs1755685	**chr13:113757192**	−	−	**0**	**0**
rs488703	**chr13:113770876**	−	−	**0**	**0**
rs11696570	chr20:37253950	0	−	0	0
rs16992555	chr20:46044044	0	−	−	0
rs2239961	chr22:21363960	0	+	+	0

Table B.4: List of SNPs selected in GRAD for fasting glucose (log). SNP signals refers to the direction of the estimated SNP effect sizes for fits without and with annotation. Details of the annotations included at the individual SNP level are also displayed.

		Signal		Annotation			
SNP ID	Location	wo. Ann	w. Ann	Elf-1	MafK	Ccn-T2	C/EBPβ
rs12408116	chr1:12533468	0	+	−	0	0	0
rs12026202	chr1:188187805	0	+	0	0	0	+
rs3850625	chr1:201016296	+	+	0	0	0	0
rs715049	chr2:217657102	0	−	+	0	0	−
rs4580644	chr4:15785201	0	+	0	0	0	0
rs11556167	chr4:152682046	0	+	0	+	0	0
rs13176438	chr5:59759107	0	−	0	0	0	−
rs11134864	chr5:174035919	0	−	0	−	0	0
rs1004558	chr7:44240407	0	+	−	0	−	0
rs13238018	chr7:104430466	0	+	0	+	0	0
rs1993181	chr10:4891168	0	−	0	0	0	−
rs7350420	chr10:51594462	0	−	0	0	+	0
rs11189479	chr10:99836031	0	−	0	0	0	−
rs17146413	chr11:64638041	0	+	0	0	−	−
rs12886379	chr14:34638094	0	−	0	−	0	0
rs9322996	chr14:39693018	0	+	0	0	0	+
rs8140067	chr22:32871442	0	+	0	0	−	0

Table B.5: List of SNPs selected in GRAD for height. SNP signals refers to the direction of the estimated SNP effect sizes for fits without and with annotation. Details of the annotations included at the individual SNP level are also displayed.

		Signal		Annotation					
SNP ID	Location	wo. Ann	w. Ann	EBF1	CHD2	C/EBPβ	Max	TAL1	Stat3
rs3791020	chr1:173813197	0	−	0	0	0	0	−	0
rs1569879	chr1:186547952	0	−	0	0	0	0	0	−
rs4363479	chr1:202112285	0	−	0	0	+	0	0	0
rs7519922	chr1:203991273	0	+	0	0	0	0	0	0
rs925255	chr2:28614794	0	+	0	−	0	0	0	0
rs1396733	chr2:28642747	0	+	0	0	−	0	0	0
rs3806502	chr2:136288273	0	−	0	+	+	0	0	0
rs17369895	chr2:228483942	0	+	0	0	−	0	0	0
rs17685252	chr3:27719152	0	−	0	0	0	0	0	−
rs9850318	chr3:34018833	0	−	0	0	+	0	0	0
rs893566	chr3:45673062	0	−	0	0	0	+	0	0
rs11717486	chr3:146536741	−	−	0	0	0	0	0	0
rs1114277	chr3:172783783	+	+	0	0	0	0	0	0
rs4554118	chr4:184576088	0	+	0	0	0	0	0	0
rs260718	chr5:139132796	0	−	0	0	0	+	0	0
rs2068981	chr6:127697992	0	+	0	0	0	0	0	0
rs6466121	chr7:106183083	0	+	0	0	0	−	+	0
rs849299	chr7:106666157	0	+	0	0	0	0	+	0
rs1329393	chr9:98318926	0	+	0	0	0	0	0	0
rs7081523	chr10:1240519	0	+	0	0	0	0	0	0
rs793088	chr10:31364621	0	+	0	0	0	0	0	+
rs17296289	chr10:33260699	0	−	0	0	0	0	−	0
rs2275044	chr10:121201626	0	−	−	0	0	0	0	0
rs15564	chr16:677854	0	+	0	−	0	0	0	0
rs1620139	chr16:8755121	0	−	−	0	0	0	0	0
rs4784817	chr16:57565774	0	−	0	0	0	0	0	0
rs872300	chr17:16277776	0	+	0	0	0	0	0	+
rs1242482	chr17:17352341	0	−	0	0	0	+	0	0
rs4796224	chr17:34842521	0	+	0	−	0	+	0	0
rs2016639	chr18:6943264	0	+	0	0	0	0	0	0
rs12607412	chr18:46072320	0	−	0	0	0	+	0	0
rs8108874	chr19:11797112	0	−	−	0	0	0	0	0

REFERENCES

Aleva, Floor E, Frank L van de Veerdonk, Yang Li, Rahajeng N Tunjungputri, Sami Simons, Philip G De Groot, Mihai M Netea, Yvonne F Heijdra, Quirijn de Mast, and André JAM van der Ven. 2018. The effects of signal transducer and activator of transcription three mutations on human platelets. *Platelets* 29(6):602–609.

Alexander, David H, and Kenneth Lange. 2011. Stability selection for genome-wide association. *Genetic Epidemiology* 35(7):722–728.

Alipanahi, Babak, Andrew Delong, Matthew T Weirauch, and Brendan J Frey. 2015. Predicting the sequence specificities of dna-and rna-binding proteins by deep learning. *Nature Biotechnology* 33(8):831.

Araki, Toshiyuki, and Jeffrey Milbrandt. 2000. Ninjurin2, a novel homophilic adhesion molecule, is expressed in mature sensory and enteric neurons and promotes neurite outgrowth. *Journal of Neuroscience* 20(1):187–195.

Barbeira, Alvaro N., Scott P. Dickinson, Rodrigo Bonazzola, Jiamao Zheng, Heather E. Wheeler, Jason M. Torres, Eric S. Torstenson, Kaanan P. Shah, Tzintzuni Garcia, Todd L. Edwards, Eli A. Stahl, Laura M. Huckins, , Dan L. GTEx Consortium, Nicolae, Nancy J. Cox, and Hae Kyung Im. 2018. Exploring the phenotypic consequences of tissue specific gene expression variation inferred from gwas summary statistics. *Nature Communications* 9(1):1825.

Baron, Reuben M, and David A Kenny. 1986. The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology* 51(6):1173.

Benjamini, Yoav, and Yosef Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)* 57(1):289–300.

Bogdan, Małgorzata, Ewout Van Den Berg, Chiara Sabatti, Weijie Su, and Emmanuel J Candès. 2015. Slope - adaptive variable selection via convex optimization. *Annals of Applied Statistics* 9(3):1103.

Bolado-Carrancio, A, José A Riancho, Jesús Sainz, and José C Rodríguez-Rey. 2014. Activation of nuclear receptor nr5a2 increases glut4 expression and glucose metabolism in muscle cells. *Biochemical and Biophysical Research Communications* 446(2):614–619.

Boyle, Evan A, Yang I Li, and Jonathan K Pritchard. 2017. An expanded view of complex traits: from polygenic to omnigenic. *Cell* 169(7):1177–1186.

Broholm, Christa, Anders H Olsson, Alexander Perfilyev, Ninna S Hansen, Maren Schrölkamp, Klaudia S Strasko, Camilla Scheele, Rasmus Ribel-Madsen, Brynjulf Mortensen, Sine W Jørgensen, et al. 2016. Epigenetic programming of adipose-derived stem cells in low birthweight individuals. *Diabetologia* 59(12):2664–2673.

Brzyski, Damian, Christine B Peterson, Piotr Sobczyk, Emmanuel J Candès, Malgorzata Bogdan, and Chiara Sabatti. 2017. Controlling the rate of gwas false discoveries. *Genetics* 205(1):61–75.

Buniello, Annalisa, Jacqueline A L MacArthur, Maria Cerezo, Laura W Harris, James Hayhurst, Cinzia Malangone, Aoife McMahon, Joannella Morales, Edward Mountjoy, Elliot Sollis, et al. 2018. The nhgri-ebi gwas catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research* 47(D1):D1005–D1012.

Carbonetto, Peter, and Matthew Stephens. 2012. Scalable variational inference for bayesian variable selection in regression, and its accuracy in genetic association studies. *Bayesian Analysis* 7(1):73–108.

Carithers, Latarsha J., Kristin Ardlie, Mary Barcus, Philip A. Branton, Angela Britton, Stephen A. Buia, Carolyn C. Compton, David S. DeLuca, Joanne Peter-Demchok, Ellen T. Gelfand, Ping Guan, Greg E. Korzeniewski, Nicole C. Lockhart,

Chana A. Rabiner, Abhi K. Rao, Karna L. Robinson, Nancy V. Roche, Sherilyn J. Sawyer, Ayellet V. Segrè, Charles E. Shive, Anna M. Smith, Leslie H. Sobin, Anita H. Undale, Kimberly M. Valentino, Jim Vaught, Taylor R. Young, and Helen M. Moore. 2015. A novel approach to high-quality postmortem tissue procurement: the gtex project. *Biopreservation and Biobanking* 13(5):311–319.

Chaibub-Neto, E., M. P. Keller, A. D. Attie, and B. S. Yandell. 2010. Causal graphical models in systems genetics: A unified framework for joint inference of causal network and genetic architecture for correlated phenotypes. *Annals of Applied Statistics* 4(1):320–339.

Chang, Christopher C, Carson C Chow, Laurent CAM Tellier, Shashaank Vattikuti, Shaun M Purcell, and James J Lee. 2015. Second-generation plink: rising to the challenge of larger and richer datasets. *Gigascience* 4(1):7.

Chatterjee, N., J. Shi, and M. García-Closas. 2016. Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nature Reviews Genetics* 14210(2014):14205–14210.

Chen, Wenan, Shannon K McDonnell, Stephen N Thibodeau, Lori S Tillmans, and Daniel J Schaid. 2016. Incorporating functional annotations for fine-mapping causal variants in a bayesian framework using summary statistics. *Genetics* 204(3): 933–958.

Choi, Hak-Jong, Yanbiao Geng, Hoonsik Cho, Sha Li, Pramod Kumar Giri, Kyrie Felio, and Chyung-Ru Wang. 2011. Differential requirements for the ets transcription factor elf-1 in the development of nkt cells and nk cells. *Blood* 117(6): 1880–1887.

Chung, Dongjun, Can Yang, Cong Li, Joel Gelernter, and Hongyu Zhao. 2014. Gpa: a statistical approach to prioritizing gwas results by integrating pleiotropy and annotation. *PLoS Genetics* 10(11):e1004787.

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

Consortium, ENCODE Project. 2012. An integrated encyclopedia of dna elements in the human genome. *Nature* 489(7414):57.

Corn, Paul G, M Stacey Ricci, Kimberly A Scata, Andrew M Arsham, M Celeste Simon, David T Dicker, and Wafik S El-Deiry. 2005. Mxi1 is induced by hypoxia in a hif-1–dependent manner and protects cells from c-myc-induced apoptosis. *Cancer Biology & Therapy* 4(11):1285–1294.

Cui, Ying, Wen Chen, Jinfeng Chi, and Lei Wang. 2016. Comparison of transcriptome between type 2 diabetes mellitus and impaired fasting glucose. *Medical science monitor: international medical journal of experimental and clinical research* 22: 4699.

Davey Smith, George, and Shah Ebrahim. 2003. 'mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *International Journal of Epidemiology* 32(1):1–22.

Desch, Karl C., Ayse B. Ozel, David Siemieniak, Yossi Kalish, Jordan A. Shavit, Courtney D. Thornburg, Anjali A. Sharathkumar, Caitlin P. McHugh, Cathy C. Laurie, Andrew Crenshaw, Daniel B. Mirel, Yoonhee Kim, Cheryl D. Cropp, Anne M. Molloy, Peadar N. Kirke, Joan E. Bailey-Wilson, Alexander F. Wilson, James L. Mills, John M. Scott, Lawrence C. Brody, Jun Z. Li, and David Ginsburg. 2013. Linkage analysis identifies a locus for plasma von willebrand factor undetected by genome-wide association. *Proceedings of the National Academy of Sciences* 110(2): 588–593.

Edgar, Ron, Yaron Mazor, Ariel Rinon, Jacob Blumenthal, Yaron Golan, Ella Buzhor, Idit Livnat, Shani Ben-Ari, Iris Lieder, Alina Shitrit, Yaron Gilboa, Ahmi Ben-Yehudah, Osnat Edri, Netta Shraga, Yoel Bogoch, Lucy Leshansky, Shlomi Aharoni, Michael D. West, David Warshawsky, and Ronit Shtrichman. 2013. Lifemap discovery^{tm}: the embryonic development, stem cells, and regenerative medicine research portal. *PloS One* 8(7):e66629.

Emilsson, Valur, Marjan Ilkov, John R Lamb, Nancy Finkel, Elias F Gudmundsson, Rebecca Pitts, Heather Hoover, Valborg Gudmundsdottir, Shane R Horman, Thor Aspelund, et al. 2018. Co-regulatory networks of human serum proteins link genetics to disease. *Science* 361(6404):769–773.

Fan, Jianqing, and Runze Li. 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association* 96(456):1348–1360.

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

Franchini, Massimo, Franco Capra, Giovanni Targher, Martina Montagnana, and Giuseppe Lippi. 2007. Relationship between abo blood group and von willebrand factor levels: from biology to clinical implications. *Thrombosis Journal* 5(1):14.

Franchini, Massimo, Silvia Crestani, Francesco Frattini, Cinzia Sissa, and Carlo Bonfanti. 2014. Abo blood group and von willebrand factor: biological implications. *Clinical Chemistry and Laboratory Medicine (CCLM)* 52(9):1273–1276.

Gagliano, Sarah A, Michael R Barnes, Michael E Weale, and Jo Knight. 2014. A bayesian method to incorporate hundreds of functional characteristics with association evidence to improve variant prioritization. *PLoS One* 9(5):e98122.

Gamazon, Eric R, Heather E Wheeler, Kaanan P Shah, Sahar V Mozaffari, Keston Aquino-Michaels, Robert J Carroll, Anne E Eyler, Joshua C Denny, GTEx Consortium, Dan L Nicolae, Nancy J Cox, and Hae Kyung Im. 2015. A gene-based association method for mapping traits using reference transcriptome data. *Nature Genetics* 47(9):1091.

Guan, Yongtao, Matthew Stephens, et al. 2011. Bayesian variable selection regression for genome-wide association studies and other large-scale problems. *Annals of Applied Statistics* 5(3):1780–1815.

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

Heideman, Marinus R, Cesare Lancini, Natalie Proost, Eva Yanover, Heinz Jacobs, and Jan-Hermen Dannenberg. 2014. Sin3a associated Hdac1 and Hdac2 are essential for hematopoietic stem cell homeostasis and contribute differentially to hematopoiesis. *Haematologica* 99(8):1292–303.

Hirschhorn, Joel N, et al. 2009. Genomewide association studies–illuminating biologic pathways. *New England Journal of Medicine* 360(17):1699.

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

Huang, B., S. Sivaganesan, P. Succop, and E. Goodman. 2004. Statistical assessment of mediational effects for logistic mediational models. *Statistics in Medicine* 23(17): 2713–2728.

Huang, Y-T., T. J. VanderWeele, and X. Lin. 2014. Joint analysis of SNP and gene expression data in genetic association studies of complex diseases. *Annals of Applied Statistics* 22(1):352–376.

Huang, Yen-Tsung, Liming Liang, Miriam F Moffatt, William OCM Cookson, and Xihong Lin. 2015. igwas: Integrative genome-wide association studies of genetic and genomic data for disease susceptibility using mediation analysis. *Genetic Epidemiology* 39(5):347–356.

Jansen, Rick, Jouke-Jan Hottenga, Michel G Nivard, Abdel Abdellaoui, Bram Laport, Eco J de Geus, Fred A Wright, Brenda WJH Penninx, and Dorret I Boomsma. 2017. Conditional eqtl analysis reveals allelic heterogeneity of gene expression. *Human Molecular Genetics* 26(8):1444–1451.

Jin, Wanzhu, Allison B. Goldfine, Tanner Boes, Robert R. Henry, Theodore P. Ciaraldi, Eun-Young Kim, Merve Emecan, Connor Fitzpatrick, Anish Sen, Ankit Shah, Edward Mun, Martha Vokes, Joshua Schroeder, Elizabeth Tatro, Jose Jimenez-Chillaron, and Mary-Elizabeth Patti. 2011. Increased srf transcriptional activity in human and mouse skeletal muscle is a signature of insulin resistance. *The Journal of Clinical Investigation* 121(3):918–929.

Kaji, Hidesuke. 2013. High-density lipoproteins and the immune system. *Journal of Lipids* 2013:684903.

Kannel, William B, Manning Feinleib, Patricia M McNamara, Robert J Garrison, and William P Castelli. 1979. An investigation of coronary heart disease in families: the framingham offspring study. *American Journal of Epidemiology* 110(3):281–290.

Kelley, David R, Jasper Snoek, and John L Rinn. 2016. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Research* 26(7):990–999.

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

Koizume, Shiro, and Yohei Miyagi. 2015. Tissue factor–factor vii complex as a key regulator of ovarian cancer phenotypes. *Biomarkers in cancer* 7:BIC–S29318.

Lettre, Guillaume, and John D Rioux. 2008. Autoimmune diseases: insights from genome-wide association studies. *Human Molecular Genetics* 17(R2):R116–R121.

Li, Yue, and Manolis Kellis. 2016. Joint bayesian inference of risk variants and tissue-specific epigenomic enrichments across multiple complex human diseases. *Nucleic Acids Research* 44(18):e144–e144.

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

Lim, Kihong, and Hyo-Ihl Chang. 2010. Elevated o-linked n-acetylglucosamine correlated with reduced sp1 cooperative dna binding with its collaborating factors in vivo. *Bioscience, Biotechnology, and Biochemistry* 1007022039–1007022039.

MacIver, Nancie J, Sarah R Jacobs, Heather L Wieman, Jessica A Wofford, Jonathan L Coloff, and Jeffrey C Rathmell. 2008. Glucose metabolism in lymphocytes is a regulated process with significant effects on immune cell function and survival. *Journal of Leukocyte Biology* 84(4):949–957.

Mailman, Matthew D, Michael Feolo, Yumi Jin, Masato Kimura, Kimberly Tryka, Rinat Bagoutdinov, Luning Hao, Anne Kiang, Justin Paschall, Lon Phan, and Stephen T. Sherry. 2007. The ncbi dbgap database of genotypes and phenotypes. *Nature Genetics* 39(10):1181–1186.

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

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

Maurano, M. T., R. Humbert, E. Rynes, R. E. Thurman, E. Haugen, H. Wang, A. P. Reynolds, R. Sandstrom, H. Qu, J. Brody, A. Shafer, F. Neri, K. Lee, T. Kutyavin, S. Stehling-Sun, A. K. Johnson, T. K. Canfield, E. Giste, M. Diegel, D. Bates, R. S. Hansen, S. Neph, P. J. Sabo, S. Heimfeld, A. Raubitschek, S. Ziegler, C. Cotsapas, N. Sotoodehnia, I. Glass, S. R. Sunyaev, R. Kaul, and J. A. Stamatoyannopoulos. 2012. Systematic localization of common disease-associated variation in regulatory DNA. *Science* 337(6099):1190–1195.

Meinshausen, Nicolai, and Peter Bühlmann. 2010. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 72(4):417–473.

Minelli, Cosetta, Alessandro De Grandi, Christian X Weichenberger, Martin Gögele, Mirko Modenese, John Attia, Jennifer H Barrett, Michael Boehnke, Giuseppe Borsani, Giorgio Casari, and John R. Thompson. 2013. Importance of different types of prior knowledge in selecting genome-wide findings for follow-up. *Genetic Epidemiology* 37(2):205–213.

Newton, Michael A, Amine Noueiry, Deepayan Sarkar, and Paul Ahlquist. 2004. Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics* 5(2):155–176.

Nicolae, Dan L, Eric Gamazon, Wei Zhang, Shiwei Duan, M Eileen Dolan, and Nancy J Cox. 2010. Trait-associated snps are more likely to be eqtls: annotation to enhance discovery from gwas. *PLoS Genetics* 6(4):e1000888.

Nikpay, Majid, Anuj Goel, Hong-Hee Won, Leanne M Hall, Christina Willenborg, Stavroula Kanoni, Danish Saleheen, Theodosios Kyriakou, Christopher P Nelson, Jemma C Hopewell, et al. 2015. A comprehensive 1000 genomes–based genome-wide association meta-analysis of coronary artery disease. *Nature Genetics* 47(10): 1121.

Nomoto, Hiroshi, Takuma Kondo, Hideaki Miyoshi, Akinobu Nakamura, Yoko Hida, Ken-ichiro Yamashita, Arun J Sharma, and Tatsuya Atsumi. 2015. Inhibition of small maf function in pancreatic β-cells improves glucose tolerance through the enhancement of insulin gene transcription and insulin secretion. *Endocrinology* 156(10):3570–3580.

Ormerod, John T, and Matt P Wand. 2010. Explaining variational approximations. *The American Statistician* 2(64):140–153.

Ormerod, John T, Chong You, and Samuel Müller. 2017. A variational bayes approach to variable selection. *Electronic Journal of Statistics* 11(2):3549–3594.

Palasca, Oana, Alberto Santos, Christian Stolte, Jan Gorodkin, and Lars Juhl Jensen. 2018. Tissues 2.0: an integrative web resource on mammalian tissue expression. *Database* 2018.

Peyvandi, Flora, Isabella Garagiola, and Luciano Baronciani. 2011. Role of von willebrand factor in the haemostasis. *Blood Transfusion* 9(Suppl 2):s3.

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

Prasad, Rashmi B, and Leif Groop. 2015. Genetics of type 2 diabetes - pitfalls and possibilities. *Genes* 6(1):87–123.

Pruim, Randall J, Ryan P Welch, Serena Sanna, Tanya M Teslovich, Peter S Chines, Terry P Gliedt, Michael Boehnke, Gonçalo R Abecasis, and Cristen J Willer. 2010. Locuszoom: regional visualization of genome-wide association scan results. *Bioinformatics* 26(18):2336–2337.

Purcell, Shaun, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel AR Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul IW De Bakker, Mark J Daly, and P. C. Sham. 2007. Plink: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics* 81(3): 559–575.

Rahman, Shaikh M, Rachel C Janssen, Mahua Choudhury, Karalee C Baquero, Rebecca M Aikens, A Becky, and Jacob E Friedman. 2012. Ccaat/enhancer-binding protein β (c/ebpβ) expression regulates dietary-induced inflammation in macrophages and adipose tissue in mice. *Journal of Biological Chemistry* 287(41): 34349–34360.

Reshef, Yakir A, Hilary K Finucane, David R Kelley, Alexander Gusev, Dylan Kotliar, Jacob C Ulirsch, Farhad Hormozdiari, Joseph Nasser, Luke O'Connor, Bryce Van De Geijn, et al. 2018. Detecting genome-wide directional effects of transcription factor binding on polygenic disease risk. *Nature Genetics* 50(10):1483.

Roadmap Epigenomics Consortium. 2015. Integrative analysis of 111 reference human epigenomes. *Nature* 518(7539):317–330.

Rojo, Constanza, Pixu Shi, Ming Yuan, and Sündüz Keleş. 2020. High dimensional sparse regression with auxiliary data on the features. *(Under preparation)*.

Rojo, Constanza, Qi Zhang, and Sündüz Keleş. 2019. ifunmed: Integrative functional mediation analysis of gwas and eqtl studies. *Genetic Epidemiology* 43(7): 742–760.

Roshyara, Nab Raj, Holger Kirsten, Katrin Horn, Peter Ahnert, and Markus Scholz. 2014. Impact of pre-imputation snp-filtering on genotype imputation results. *BMC Genetics* 15(1):88.

Schelldorfer, Jürg, Peter Bühlmann, and Sara van de Geer. 2011. Estimation for high-dimensional linear mixed-effects models using l1-penalization. *Scandinavian Journal of Statistics* 38(2):197–214.

Shin, Sunyoung, Rebecca Hudson, Christopher Harrison, Mark Craven, and Sunduz Keleş. 2018. atSNP Search: a web resource for statistically evaluating influence of human genetic variation on transcription factor binding. *Bioinformatics* bty1010.

Slattery, Matthew, Tianyin Zhou, Lin Yang, Ana Carolina Dantas Machado, Raluca Gordân, and Remo Rohs. 2014. Absence of a simple code: how transcription factors read the genome. *Trends in Biochemical Sciences* 39(9):381–399.

Smith, Nicholas L, Ming-Huei Chen, Abbas Dehghan, David P Strachan, Saonli Basu, Nicole Soranzo, Caroline Hayward, Igor Rudan, Maria Sabater-Lleal, Joshua C Bis, Moniek P M de Maat, Ann Rumley, Xiaoxiao Kong, Qiong Yang, Frances M K Williams, Veronique Vitart, Harry Campbell, Anders Mälarstig, Kerri L Wiggins, Cornelia M Van Duijn, Wendy L McArdle, James S Pankow, Andrew D Johnson, Angela Silveira, Barbara McKnight, Andre G Uitterlinden, Wellcome Trust Case Control Consortium;, Nena Aleksic, James B Meigs, Annette Peters, Wolfgang Koenig, Mary Cushman, Sekar Kathiresan, Jerome I Rotter, Edwin G Bovill, Albert Hofman, Eric Boerwinkle, Geoffrey H Tofler, John F Peden, Bruce M Psaty, Frank Leebeek, Aaron R Folsom, Martin G Larson, Timothy D

Spector, Alan F Wright, James F Wilson, Anders Hamsten, Thomas Lumley, Jacqueline C M Witteman, Weihong Tang, and Christopher J O'Donnell. 2010. Novel associations of multiple genetic loci with plasma levels of factor vii, factor viii, and von willebrand factor: The charge consortium. *Circulation* 121(12):1382–1392.

Sun, Tingni, and Cun-Hui Zhang. 2012. Scaled sparse linear regression. *Biometrika* 99(4):879–898.

Teslovich, Tanya M, Kiran Musunuru, Albert V Smith, Andrew C Edmondson, Ioannis M Stylianou, Masahiro Koseki, James P Pirruccello, Samuli Ripatti, Daniel I Chasman, Cristen J Willer, et al. 2010. Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* 466(7307):707.

Thompson, John R., Martin Gögele, Christian X. Weichenberger, Mirko Modenese, John Attia, Jennifer H. Barrett, Michael Boehnke, Alessandro De Grandi, Francisco S. Domingues, Andrew A. Hicks, Fabio Marroni, Cristian Pattaro, Fabrizio Ruggeri, Giuseppe Borsani, Giorgio Casari, Giovanni Parmigiani, Andrea Pastore, Arne Pfeufer, Christine Schwienbacher, Daniel Taliun, CKDGen Consortium, Caroline S. Fox, Peter P. Pramstaller, and Cosetta Minelli. 2013. SNP prioritization using a Bayesian probability of association. *Genetic Epidemiology* 37(2):214–221.

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

Tzikas, Dimitris G., Aristidis C. Likas, and Nikolaos P. Galatsanos. 2008. The variational approximation for Bayesian inference. *IEEE Signal Processing Magazine* 25(6):131–146.

Wen, Xiaoquan. 2016. Molecular qtl discovery incorporating genomic annotations using bayesian false discovery rate control. *Annals of Applied Statistics* 10(3):1619–1638.

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

Wen, Xiaoquan, Francesca Luca, and Roger Pique-Regi. 2015. Cross-population joint analysis of eqtls: fine mapping and functional annotation. *PLoS genetics* 11(4): e1005176.

Williams, Frances MK, Angela M Carter, Pirro G Hysi, Gabriela Surdulescu, Dylan Hodgkiss, Nicole Soranzo, Matthew Traylor, Steve Bevan, Martin Dichgans, Peter MW Rothwell, et al. 2013. Ischemic stroke is associated with the abo locus: the euroclot study. *Annals of Neurology* 73(1):16–31.

Williams, Stephen R., Fang-Chi Hsu, Keith L. Keene, Wei-Min Chen, Godfrey Dzhivhuho, Joe L. Rowles, Andrew M. Southerland, Karen L. Furie, Stephen S. Rich, Bradford B. Worrall, and Michèle M. Sale. 2017. Genetic drivers of von willebrand factor levels in an ischemic stroke population and association with risk for recurrent stroke. *Stroke* 48(6):1444–1450.

Xiang, Yaozu, and John Hwa. 2016. Regulation of vwf expression, and secretion in health and disease. *Current Opinion in Hematology* 23(3):288.

Xiong, Q., N. Ancona, E. R. Hauser, S. Mukherjee, and T. S. Furey. 2012. Integrating genetic and gene expression evidence into genome-wide association analysis of gene sets. *Genome Research* 22(2):386–397.

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

Yao, C., B. H. Chen, R. Joehanes, B. Otlu, X. Zhang, C. Liu, T. Huan, O. Tastan, L. A. Cupples, J. B. Meigs, C. S. Fox, J. E. Freedman, P. Courchesne, C. J. O'Donnell, P.J. Munson, S. Keleş, and D. Levy. 2015. Integromic analysis of genetic variation and gene expression identifies networks for cardiovascular disease phenotypes. *Circulation* 131(6):536–549.

Zhang, Cun-Hui, et al. 2010. Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics* 38(2):894–942.

Zhang, Mingfeng, Liming Liang, Nilesh Morar, Anna L. Dixon, G. Mark Lathrop, Jun Ding, Miriam F. Moffatt, William O. C. Cookson, Peter Kraft, Abrar A. Qureshi, and Jiali Han. 2012. Integrating pathway analysis and genetics of gene expression for genome-wide association study of basal cell carcinoma. *Human Genetics* 131(4): 615–623.

Zhang, Qi, and Sündüz Keleş. 2017. An empirical bayes test for allelic-imbalance detection in chip-seq. *Biostatistics* 19(4):546–561.

Zhao, Qingyuan, Jingshu Wang, Wes Spiller, Jack Bowden, and Dylan S Small. 2019. Two-sample instrumental variable analyses using heterogeneous samples. *Statistical Science* 34(2):317–333.

Zheng, Hongzhi, Jingqi Fu, Peng Xue, Rui Zhao, Jian Dong, Dianxin Liu, Masayuki Yamamoto, Qingchun Tong, Weiping Teng, Weidong Qu, Qiang Zhang, Melvin E. Andersen, and Jingbo Pi. 2015. Cnc-bzip protein nrf1-dependent regulation of glucose-stimulated insulin secretion. *Antioxidants & Redox Signaling* 22(10):819–831.

Zhong, Hua, John Beaulaurier, Pek Yee Lum, Cliona Molony, Xia Yang, Douglas J. MacNeil, Drew T. Weingarth, Bin Zhang, Danielle Greenawalt, Radu Dobrin, Ke Hao, Sangsoon Woo, Christine Fabre-Suver, Su Qian, Michael R. Tota, Mark P. Keller, Christina M. Kendziorski, Brian S. Yandell, Victor Castro, Alan D. Attie, Lee M. Kaplan, and Eric E. Schadt. 2010. Liver and adipose expression associated snps are enriched for association to type 2 diabetes. *PLoS Genetics* 6(5):e1000932.

Zhou, Jian, and Olga G Troyanskaya. 2015. Predicting effects of noncoding variants with deep learning–based sequence model. *Nature Methods* 12(10):931.

Zhou, Zhou, Francisca C Gushiken, Doug Bolgiano, Breia J Salsbery, Niloufar Aghakasiri, Naijie Jing, Xiaoping Wu, K Vinod Vijayan, Rolando E Rumbaut, Roberto Adachi, et al. 2013. Stat3 regulates collagen-induced platelet aggregation independent of its transcription factor activity. *Circulation* 127(4):476.

Zhu, Xiang, and Matthew Stephens. 2017. Bayesian large-scale multiple regression with summary statistics from genome-wide association studies. *Annals of Applied Statistics* 11(3):1561.

Zhu, Zhihong, Futao Zhang, Han Hu, Andrew Bakshi, Matthew R Robinson, Joseph E Powell, Grant W Montgomery, Michael E Goddard, Naomi R Wray, Peter M Visscher, and Jian Yang. 2016. Integration of summary data from gwas and eqtl studies predicts complex trait gene targets. *Nature Genetics* 48(5):481–487.

Zou, Hui, and Trevor Hastie. 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67(2):301–320.

Zuo, C., S. Shin, and S. Keleş. 2015. atsnp: transcription factor binding affinity testing for regulatory snp detection. *Bioinformatics* 31:3353–3355.