THE UNIVERSITY OF CHICAGO

# METHODS TO DISSECT THE BIOLOGY OF COMPLEX PHENOTYPES USING GENOMIC, TRANSCRIPTOMIC, AND PHENOMIC DATA

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE BIOLOGICAL SCIENCES
AND THE PRITZKER SCHOOL OF MEDICINE
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

COMMITTEE ON GENETICS, GENOMICS, AND SYSTEMS BIOLOGY

BY
YANYU LIANG

CHICAGO, ILLINOIS
DECEMBER 2021

To my grandparents

# TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# ACKNOWLEDGMENTS

First and foremost, I would like to thank my PhD advisor, Hae Kyung Im. I cannot have gone this far without the strongest support and trust from Haky. She always inspired me with another perspective when I got stuck at problems. And, throughout my PhD, she always gave me freedom and flexibility to explore scientific questions. Her encouragement truly helped me build my confidence as a scientist.

I would like to express a special thank to my committee chair, Xin He. My journey at UChicago started at Xin's lab six years ago during which I was working on a Master's degree in computational biology at Carnegie Mellon. After a skype chat with him, luckily, I got the chance to spend a summer in his lab. During that summer, for the first time, I engaged in human genetics research and got fascinated with it.

I am grateful to my committee members, Matthew Stephens and Marcelo Nobrega, for taking time out of their busy schedules to engage in my committee meetings. I always got constructive feedbacks and additional insights from them. Their inputs really helped improving this dissertation and made my PhD journey stay on track.

I would like to thank other faculty members at the Human Genetics department and the Section of Genetic Medicine: Carole Ober, John Novembre, Mary Sara McPeek, Mengjie Chen, Andy Dahl, and many others for their inspirations from coursework/discussion and the creation of the stimulating, open, and collaborative environment for computational folks. Furthermore, I would like to thank Sue Levison who was always there whenever I had any administrative questions.

I would also like to thank my peers and friends: Yuwen Liu, Gao Wang, Min Qiao, Nicholas Knoblauch, Yifan Zhou, Alan Selewa, Joseph Marcus, Sahar Mozaffari, Arjun Biddanda, Yuxin Zou, Maryn Carlson, Yichen Hou, Joyce Shi, Bohou Wu, Rodrigo Bonazzola, Alvaro Barbeira, Milton Pividori, Owen Melia, Lili Wang, Natasha Santhanam, Festus Nyasimi, and many others. I learned tons of academic and non-academic knowledge and

insights from them. My PhD life could not have been colorful without them.

Finally, I want to express my deepest gratitude to my family. I would like to thank my parents for their unconditional love and support, which is far beyond something that words can express. I would like to thank my partner, He Ma, for his continuous encouragement which helped me get through all these tough moments in the journey. Besides, I would like to mention my furry friends, Gary and Roger who were born and grew up during my PhD journey for their invaluable companion.

# ABSTRACT

One of the big aims in human genetics is to understand the biological mechanism underlying the genetic associations. In the past decades, the rapid development of biotechnology has made tremendous progress to approach this aim. For instance, with advanced and specialized devices and data automation systems, more complex phenotypes can be measured at higher accuracy and in more individuals. And with the inventions in high-throughput sequencing, we can profile various types of biological molecules in organs, tissues, and cells. As a geneticist, we face a massive amount of biological data at different levels and of great diversity, creating unprecedented opportunities for making discoveries. However, making the best use of data and translating them into scientific insights remain challenging. In the current data-dominated era, statistical modeling has become a vital tool to fill the gap between biological data and scientific discoveries. My dissertation spans multiple topics in statistical genetics involving the handling of genomic, transcriptomic, and phenomic data. In Chapter 2, I propose a unified statistical framework, along with computationally efficient implementation, leveraging signals from both total counts and allele-specific counts to study the genetic effect of variants on cis-regulation. In Chapter 3, I show the utility of predicted transcriptome-based polygenic risk scores in terms of the prediction performance in the matched ancestry and cross ancestry. In Chapter 4, I propose a method to impute the parental origin of the haplotypes by exploiting the parental phenome and analyze the potential benefit of using these imputed haplotypes in a GWAS with parental phenotypes and offspring genotypes. In Chapter 5, I design and implement a data analysis pipeline studying the relation between magnetic resonance imaging-derived brain features and complex phenotypes by leveraging genetic evidence rather than purely observational data. Besides methodological advancements, I also involve in collaborative efforts on analyzing and integrating the state-of-the-art datasets to decipher the genetic basis of transcriptome in multi-tissue setting and how it relates to complex phenotype genetics, which is shown in Appendix A.

# CHAPTER 1

# INTRODUCTION

Life is complex. Meter-long whales, a hundred-meter high redwoods, and hundreds of micrometer long small creatures like *paramecium* are all lives and so are ourselves, human beings. But, in most cases, all these different types of lives do share basic fundamentals such as the central dogma [26] which points out how the genetic information flows from genetic materials to phenotypes at macro-scale. To understand how nature works has been rooted deeply in our heart. As part of the nature, our curiosity drives us to dig into ourselves. Generations of biologists have made tremendous progresses on the journey of understanding life and ourselves with countless groundbreaking technological advances, brilliant theories, and scientific discoveries.

Genetics, as a branch of biology, is still quite young, whose history is dated back to over one and a half century ago with the discovery of Mendelian inheritance [99]. Many branches in biology study life by making direct observations at different levels with different technologies. Whereas, geneticists focus on studying heredity and, more generally, linking genotypes to phenotypes. Diversity is the key. Naturally occurring and stabilized genetic variations make genetic findings possible. But, to identify such genotypes to phenotypes relation is not the end of the story and, instead, it shed light on the study of the mechanism behind a phenotype. Before the emerging of the sequencing technology, linkage studies have made great success on pinpointing the causal genes of Mendelian disorders which are heritable diseases caused by a mutation at a single genetic locus.

Whereas many phenotypes are complex in the sense that they are driven by a large number of genetic and environmental factors. Human height is a great example of complex phenotypes. Decades before the identification of the actual genetic materials, it has been realized that height is polygenic [39] in reconciling the Mendelian inheritance and continuous phenotypes, like height, whose distribution is bell-shape.

To study the genetic basis of complex phenotypes is quite challenging. By the nature of complex phenotypes, the effect of each genetic factor itself is small relative to the collective effect from all contributing factors. With the success of the Human Genome Project (HGP), we obtained the human genome for the first time in about 20 years ago [72, 143] which catalyzed a massive amount of advances in genome sequencing technology and bioinformatics. This stands as a milestone for human genetics research. If we think of HGP as going from 0 to 1 for the characterization of human genome, the journey from 1 to many went incredibly fast and, in a few years following, geneticists were able to characterize the genetic variations across the globe with genotyping arrays [48]. With these achievements, the genome-wide association study (GWAS), which associates each of the genome-wide genetic variations to a phenotype, became accomplishable. GWAS suits well in studying complex phenotypes. Nowadays, GWAS has achieved great success with hundreds of thousands of associations being identified via GWAS.

However, the challenges still remain. For instance, for many phenotypes, the GWAS sample size is still quite small due to the difficulty on collecting enough participants, e.g. late-onset dieases. And due to the extensive linkage disequilibrium (LD), it is very hard to pinpoint the causal variants within a genetic locus. Furthermore, even though the causal variants is known, the functional consequence of a genetic variant is usually unclear.

Researchers used both experimental and computational methods to approach these issues. Functional genomics has grown very rapidly in the past decade with a large number of high-throughput assays profiling a wide range of (epi)genomic markers, transcription process, and etc. These various omics data provide screenshots of genome status from different perspectives which has largely enlarged our knowledge of the genome function. Going beyond the scope of molecular biology, the endophenotypes, the potential intermediate phenotype of a complex phenotype such as metabolite levels, brain features, and etc, have also been systematically profiled these days. Moreover, large biobanks have been created providing

comprehensive phenome profile of millions of individuals along with genetic data. With all these data available, statistical models have been intensively applied to genetic data analysis to fill the gap between raw data and biological interpretations and generate or prioritize hypotheses for experimental follow-up.

Such a rich source of data enlarges the scope of human phenotypes greatly and enables us to study the genetics of molecular and intermediate phenotypes. It creates a lot of opportunities for statistical geneticists. With genomic, transcriptomic, phenomic data in hand, we can not only answer existing questions but also ask new questions which could not have been approachable before. Though the central question is still about the genetic basis of human phenotypes, now we can bring more aspects sitting in-between genotypes and phenotypes such as the role of molecular features and mediating phenotypes. To approach this goal, more sophisticated statistical models are needed to leverage evidences from multiple modalities and handle high-dimensional data.

In present days, the mission of human genetics is going beyond the basic scientific discoveries. Genetics collects knowledges on how the genetic effect flows from genotypes to phenotypes, i.e. from micro-scale to macro-scale. These knowledges also help us to understand the disease mechanisms, which is of great practical significance for improving the human health in the fields such as drug target identification, precision medicine, and etc.

As a statistical geneticist by training, I develop statistical and computational methods to leverage the richness of genomic, transcriptomic, and phenomic data for answering human genetics-related questions. My dissertation works span across multiple topics of human genetics. In Chapter 2, I propose a statistical framework to make the best use of RNA-sequencing based transcriptome data in the identification of the genetic basis of gene cis-regulation. It resolves two main issues in the previous approaches: i) prohibitive computational burden on large-scale transcriptome data; ii) lack of a systematic way to analyze genetic variants jointly. In Chapter 3, I focus on the polygenic risk scores (PRSs). PRS measures the genetically

determined disease risks or phenotype quantities (relative to a population) which could be potentially useful for doctors to access the health status of patients. I investigate the utility of a predicted transcriptome-based polygenic risk score in order to resolve two existing issues in current PRS approaches: i) lack of biological interpretability; ii) poor transferrability across different ancestry groups. Furthermore, I implement a computationally efficient approach to obtain the proposed scores from publicly available GWAS results. In Chapter 4, I propose and analyze a novel approach to perform GWAS with parental phenotypes, which can significantly increase the number of cases for late-onset diseases such as Alzheimer's disease. I perform theoretical and computational analysis to assess the feasibility of this approach under the current availability of parental phenotypes and outlook how much data is required in order to obtain good statistical power. In Chapter 5, with the availability of biobank-scale brain-related endophenotype data, I focus on understanding the relation between brain-related endophenotypes and complex phenotypes. I characterize the genetic architecture of brain endophenotypes and develop genetic predictors of these endophenotypes with genome-wide variants. And I implement a computational framework to look for associations between brain-related endophenotypes and complex phenotypes by leveraging the genome-wide genetic evidence from both sides. All these methodological advancements won't be meaningful without applications to real-world data. In Appendix A, I participate in cross-institution collaborative work to advance our understanding on the role of gene regulation in complex phenotypes using the state-of-the-art datasets and analytical tools, which shows the real journey to turn big data into biological insights.

In summary, the methodologies being proposed or developed in my dissertation can be classified into two categories: i) to develop novel methods to improve the statistical power of an existing problem (Chapter 2 and 4); ii) to propose a new perspective to study the biology of complex phenotypes (Chapter 3 and 5). Though these method advances focus on different topics, they all hold the same big aim which is to develop computational methods to make

the best use of the huge amount of genetic and phenomic data in fueling the life studying journey, for which Appendix A makes a case.

# CHAPTER 2

# A SCALABLE UNIFIED FRAMEWORK OF TOTAL AND ALLELE-SPECIFIC COUNTS FOR CIS-QTL, FINE-MAPPING, AND PREDICTION

Material from: Liang, Yanyu, François Aguet, Alvaro N. Barbeira, Kristin Ardlie, and Hae Kyung Im, "A scalable unified framework of total and allele-specific counts for cis-QTL, fine-mapping, and prediction.", Nature communications, published 2021, Nature Publishing Group [81]

## 2.1 Abstract

Genetic studies of the transcriptome help bridge the gap between genetic variation and phenotypes. To maximize the potential of such studies, efficient methods to identify expression quantitative trait loci (eQTLs) and perform fine-mapping and genetic prediction of gene expression traits are needed. Current methods that leverage both total read counts and allele-specific expression to identify eQTLs are generally computationally intractable for large transcriptomic studies. Here, we describe a unified framework that addresses these needs and is scalable to thousands of samples. Using simulations and data from GTEx, we demonstrate its calibration and performance. For example, mixQTL shows a power gain equivalent to a 29% increase in sample size for genes with sufficient allele-specific read coverage. To showcase the potential of mixQTL, we applied it to 49 GTEx tissues and found 20% additional eQTLs (FDR$<$ 0.05, per tissue) that are significantly more enriched among trait associated variants and candidate cis-regulatory elements comparing to the standard approach.

## 2.2 Introduction

Genome-wide association studies (GWAS) have identified tens of thousands of genomic loci associated with complex traits but most of these loci lie in non-coding regions of the genome, indicating transcriptome regulation as a potential key driver of disease biology. Multiple methods have been developed to integrate GWAS results with expression quantitative trait loci (eQTLs) and inform mechanisms underlying GWAS loci.

Two strategies are commonly employed: 1) association-based approaches including PrediXcan [43], fusion [51], and smr [168]; and 2) colocalization-based approaches including coloc [47], eCAVIAR [54], and enloc [153]. Association-based approaches correlate genetic predictors of gene expression with complex traits of interest. Colocalization-based approaches rely on high-quality eQTL mapping and fine-mapping results to identify potentially causal genes.

In addition to gene expression levels measured by total read counts, allele-specific expression (the relative expression difference between the two haplotypes) provides valuable additional information that can be leverage to improve eQTL mapping and fine-mapping. Several methods have been proposed to combine total and allele-specific read count for QTL mapping, such as TReCASE [133], WASP [141], and RASQUAL [70]). However, running these methods on sample sizes beyond a few hundred is generally computationally intractable, and as a result they have not been applied to large-scale studies like GTEx, which includes over 15,000 samples across 49 tissues. For fine-mapping, two approaches that combine both ASE and eQTL mapping via meta-analysis have been recently proposed [170, 148]. However, to our knowledge, no existing method provides a scalable unified framework combining total and allele-specific counts with explicit multi-SNP modeling for QTL mapping, fine-mapping, and prediction.

By assuming a log-linear model for transcript expression levels with independent reads from each haplotype and weak genetic effects, as proposed in [101], we derive two approxi-

mately independent equations for allelic imbalance (read count ratio between the two haplotypes) and total read count.

In this work, we develop a unified framework and computationally efficient algorithms combining total and allele-specific reads for QTL mapping, fine-mapping, and prediction. We demonstrate the resulting gain in performance with simulations under a range of different settings, applications to GTEx v8 data [139], and comparisons to a large-scale eQTL meta-analysis from eQTLGen [145]. We also generated mixQTL results for the full set of GTEx data and make this resource publicly available. The software, simulation, data preprocessing, and analysis pipelines can be found at `https://github.com/hakyimlab/mixqtl`, `https://github.com/liangyy/mixqtl-pipeline`, and `https://github.com/liangyy/mixqtl-gtex`. A computationally efficient GPU-based implementation of mixQTL has been embedded in tensorQTL `https://github.com/broadinstitute/tensorqtl`.

## 2.3 Results

#### 2.3.0.1 Overview of the statistical model

To develop a computationally efficient approach that integrates total and allele-specific count data, we assumed multiplicative cis-regulatory effects and noise, similarly to the model proposed in [101]. For a given gene, we modeled the haplotypic read count $Y_i^h$, which is the number of reads from haplotype $h$ of individual $i$ as

$$Y_i^h = L_i \cdot \theta_{0,i} \cdot \exp(\beta \cdot X_i^h) \cdot \exp(\epsilon_i^h), \tag{2.1}$$

where $L_i$ is the library size for individual $i$, $\theta_{0,i}$ is the baseline abundance (for a haplotype with the reference allele), $\exp(\beta)$ is the cis-regulatory effect (allelic fold change due to the presence of the alternative allele), $X_i^h$ indicates the dosage of the variant (0 if the individual has the reference allele, and 1 if they have the alternative one), and $\exp(\epsilon_i^h)$ is the

multiplicative noise.

Calculating the total read count as the sum of the two haplotypic counts and assuming weak cis-regulatory effects, we derived an approximately linear model for the logarithm of the haplotypic and total read counts (see details in Section 2.5.2, 2.5.3 and 2.9.1). In practice, we only observe the allele-specific reads that include a heterozygous site, which is a fraction of the total haplotypic count denoted as $Y_i^{(h)\text{obs}} = \alpha_i \cdot Y_i^h$. To take this partial readout into account, we modeled the observed total and allele-specific counts as

$$\begin{aligned}
\log Y_i^{(1)\text{obs}} &= \log L_i + \log \alpha_i + \log \theta_{0,i} + X_i^1 \beta + \epsilon_i^{(1)} \\
\log Y_i^{(2)\text{obs}} &= \log L_i + \log \alpha_i + \log \theta_{0,i} + X_i^2 \beta + \epsilon_i^{(2)} \\
\log \frac{Y_i^{\text{total}}}{2} &\approx \log L_i + \log \theta_{0,i} + \frac{X_i^1 + X_i^2}{2} \beta + \epsilon_i^{\text{trc}}
\end{aligned} \tag{2.2}$$

where the error terms are $\epsilon_i^{\text{trc}} \sim N(0, \frac{\sigma^2}{Y_i^{\text{total}}})$, $\epsilon_i^{(h)} \sim N(0, \frac{\sigma^2}{Y_i^{(h)\text{obs}}})$ and the errors of the two haplotypes are independent: $\epsilon^{(1)} \perp\!\!\!\perp \epsilon^{(2)}$. Here we let the $\epsilon$ terms have variance inversely proportional to the actual count and by doing so, we ensure that the variance of the count scales approximately linearly to the mean of the count as demonstrated in Section 2.9.1.2.

We further simplified the models by combining the two allele-specific counts and defining the baseline abundance variation as a random effect $z_i$ ($\log \theta_{0,i}$ = population mean + $z_i$). Then, we merge the total count term $\epsilon_i^{\text{trc}}$ and $z_i$ into one term $\tilde{z}_i$ (since $\epsilon_i^{\text{asc}}$ is approximately independent from both of them. See Section 2.5.4 and 2.9.4.1). The final model is

$$\log \frac{Y_i^{(1)\text{obs}}}{Y_i^{(2)\text{obs}}} = \quad (X_i^1 - X_i^2)\beta + \epsilon_i^{\text{asc}} \text{ (allelic imbalance eq.)} \tag{2.3}$$

$$\log \frac{Y_i^{\text{total}}}{2L_i} \approx \mu_0 + \frac{X_i^1 + X_i^2}{2}\beta \ + \tilde{z}_i \quad \text{(total read count eq.)} \tag{2.4}$$

where $\tilde{z}_i \sim N(0, \tilde{\sigma}_0^2)$ and $\epsilon_i^{\text{asc}} \sim N(0, \sigma^2 \cdot (\frac{1}{Y_i^{(1)\text{obs}}} + \frac{1}{Y_i^{(2)\text{obs}}}))$ and $\tilde{z}_i$ is approximately independent from $\epsilon^{\text{asc}}$.

This single SNP model extends to multiple SNPs in a straightforward manner by using a vector of allelic dosages $(X_{i1}, \cdots, X_{ip})$ and genetic effects $(\beta_1, \cdots, \beta_p)$ instead of the scalar values above. Here, $p$ represents the number of genetic variants in the cis-window of the gene under consideration (Section 2.9.3 and 2.9.5).

For cis-QTL mapping, we took advantage of the approximate independence of the allelic-imbalance and the total read counts in equations (2.3) and (2.4), solving them as separate linear regressions (for computational efficiency) and combining the results via inverse-variance weighted meta-analysis. We call this method mixQTL.

For the fine-mapping and prediction problems, we also leveraged the approximate independence of the allelic-imbalance and total read count equations. We used a two-step approach in which we first scale the two equations so that they become independent data points with equal variances. In the second step, we combined these data points into an augmented dataset and applied the existing algorithms SuSiE [149] and elastic net [40]. We term these methods mixFine and mixPred, for fine-mapping and prediction, respectively.

### 2.3.1 Simulation of total and allele-specific reads

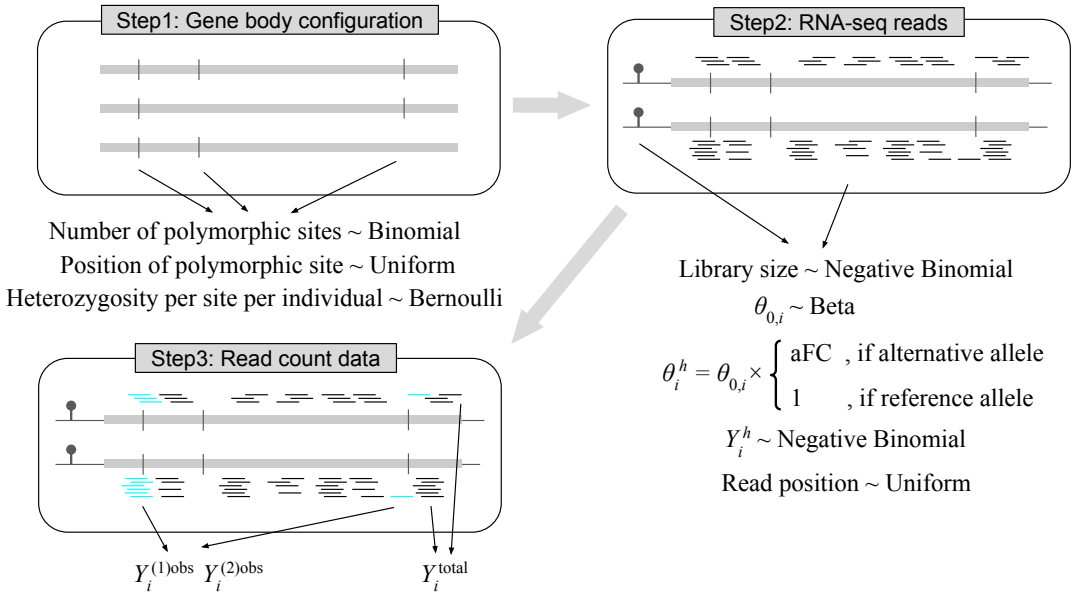



**Figure 2.1: Simulation scheme for total and allele-specific read counts.** Step 1 simulates a gene body configuration by first simulating the number of polymorphic sites of the gene followed by positioning these polymorphic sites uniformly across the gene body. For each individual, the heterozygosity of these polymorphic sites is drawn from a Bernoulli distribution. Step 2 simulates the haplotypic reads by first simulating Negative Binomial library size $L_i$, Beta baseline abundance $\theta_{0,i}$, and the genetic effect $\beta$. These parameters determine the abundance $\theta_i^h$ for each haplotypic transcript, in which aFC is the allelic fold change which equals $e^\beta$ in our parameterization. Then, the haplotypic read count $Y_i^h$ is generated using a Negative Binomial distribution given the expected count $L_i \times \theta_i^h$, where the reads are distributed uniformly across the gene body. In Step 3, the gene-level allele-specific counts $Y_i^{(h)\text{obs}}$ are determined by counting the reads that overlap heterozygous sites. $Y_i^{\text{total}}$ is calculated as the sum of the two haplotypic counts $Y_i^1$ and $Y_i^2$.

To assess the benefits of this unified framework over using only total read counts or allele-specific expression, we simulated haplotypic reads according to the framework illustrated in Figure 2.1, with additional details in Section 2.5.7 and 2.9.6. For mixQTL, we simulated data with a single causal variant and for mixPred and mixFine, we simulated data with 1-3 causal variants.

For all simulation settings, we set an average library size of 94 million reads (to approximately match GTEx v8 library sizes) and used a series of expression levels (expected value

of $\theta_{0,i}$ in Eq 2.1): from 50 to 1 read per million, corresponding to $\theta = 5 \times 10^{-5} \sim 10^{-6}$. The fraction of allele-specific reads was kept at consistent levels across simulations by using the same distribution of polymorphic sites per individual.

### *2.3.2 Combining total and allele-specific read counts improves cis-eQTL mapping*

To assess the gain in power of combining total and allele-specific read counts, we simulated 200 replicates with allelic fold change varying among 1, 1.01, 1.05, 1.1, 1.25, 1.5, 2, 3. We compared mixQTL with two methods: using either only allele-specific counts (ascQTL) or total counts (trcQTL). See details in Section 2.9.4.1.

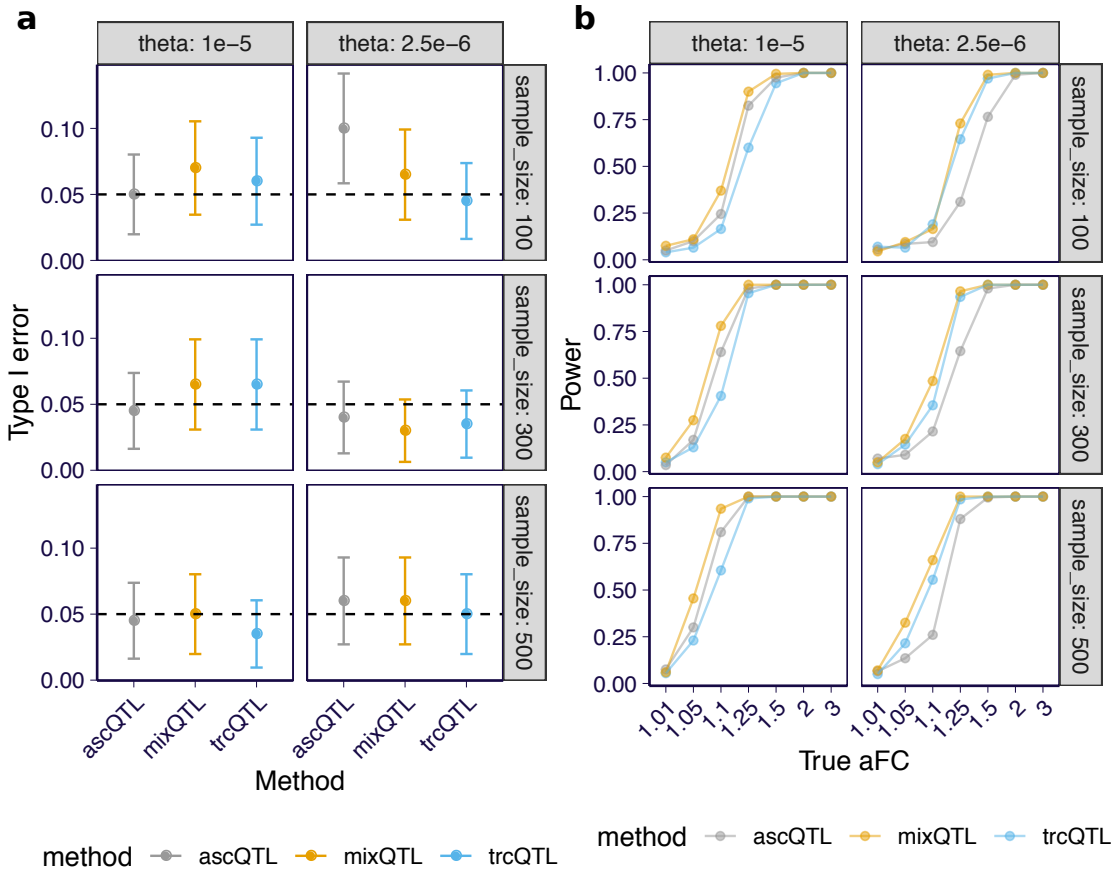



**Figure 2.2: QTL mapping performance for mixQTL and approaches based on either total reads (trcQTL) or allele-specific reads (ascQTL) on simulated data.** Each panel presents the results for two relative abundances of the gene, $\theta$, and three sample sizes. **(a)** Type I error (y-axis) at a 5% significance level across methods (x-axis) are shown. The dashed line represents the desired error rate under the null hypothesis. The error bar indicates the 95% confidence interval of the estimated error rate from 200 replicates. **(b)** Power (y-axis) at a 5% significance level across methods under a range of true aFC values (x-axis) are shown. Power is defined as the fraction of eQTLs passing the significance threshold.

All three methods had calibrated type I errors (Figure 2.2a and Figure 2.7). mixQTL outperformed both trcQTL and ascQTL in all simulation settings, demonstrating the benefits of combining total and allele-specific counts for cis-eQTL mapping (Figure 2.2b and Figure 2.8).

The power of ascQTL was sensitive to the number of allele-specific reads, as expected. As shown in Figure 2.2b, with $\theta$ controlling the expression level, ascQTL yielded much higher

power for higher expression levels. In contrast, trcQTL was less sensitive to the number of reads observed under the range of read counts in our simulation settings. Such sensitivity differences between ascQTL and trcQTL are consistent with the nature of count data, where the magnitude of the noise is inversely related to the count.

### *2.3.3 Combining total and allele-specific read count improves fine-mapping*

To realistically simulate LD structure, we used the genotypes of European individuals from the 1000 Genomes projects phase 3 [1] within ±1MB cis-windows of 100 randomly selected genes. We applied mixFine and trcFine (which uses total read count only; Section 2.9.5.3) to the simulated data and characterized the fine-mapping results with two metrics: 1) power curve, defined as the proportion of detected variants among causal ones versus the number of detected variants, where detection was defined as the variant having posterior inclusion probability (PIP) > threshold (which is varied to get the desired number of detected SNPs); 2) the size of 95% credible set (CS) which contains the causal variant.

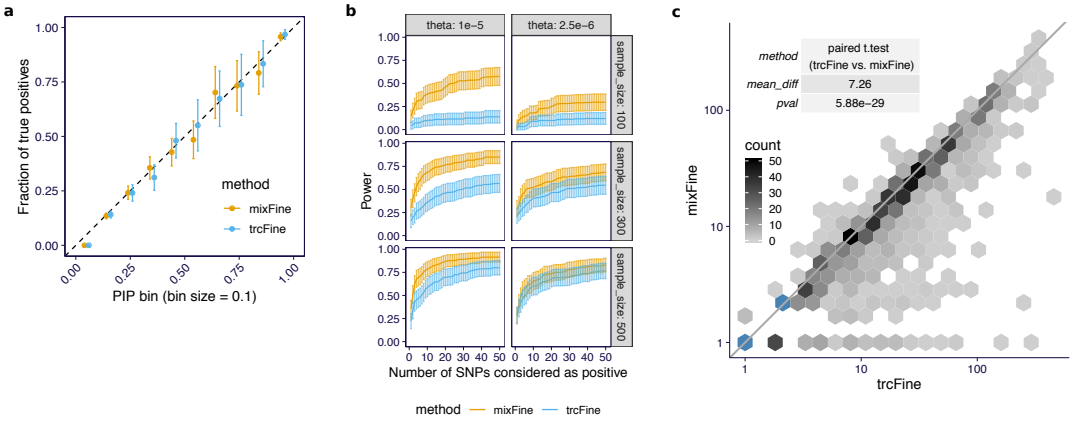



**Figure 2.3: Fine-mapping performance of the combined (mixFine) and total read-based (trcFine) approaches on simulated data. (a)** The observed true positive rate within SNPs binned by PIP are shown (aggregated across all simulation settings) for both mixFine (orange) and trcFine (blue). The plot is based on 10,211,200 simulations across the grid of simulation parameters. From left to right, the bin sizes for mixFine are 10,206,540, 2,554, 742, 335, 234, 128, 57, 56, 67, 487 and the bin sizes for trcFine are 10,208,066, 1,790, 495, 241, 152, 69, 52, 38, 48, 249. The error bars indicate the 95% confidence interval of the estimated fraction. **(b)** The power at a PIP cutoff (on y-axis) is plotted against the number of variants passing the PIP cutoff (on x-axis) for mixFine and trcFine. In each panel, the curve is based on 200 simulation replicates with 100 simulations having signals and 100 simulations being drawn from the null. The solid curves indicate the mean power (recall rate) among the 100 simulation replicates with signals and the error bars indicate the 95% confidence interval. **(c)** For the true signals captured in both mixFine and trcFine, the sizes of the 95% credible sets in the two methods are plotted (trcFine on x-axis and mixFine on y-axis). The table shows the average difference of the size (trcFine vs. mixFine) along with the p-value under paired t test (two-sided). The color of the hexagon bin indicates the count of data points in the bin. The blue bins have more than 50 counts.

The PIPs of both trcFine and mixFine were consistent with the proportion of true causal variants within each PIP bin (Figure 2.3a). By combining total and allele-specific reads, mixFine achieved higher power than trcFine (Figure 2.3b and Figure 2.10) across almost all simulation settings. mixFine achieved the highest improvement relative to trcFine at a high expression level ($\theta$), corresponding to high-quality allele-specific signals. The gain in power decreased with larger sample sizes.

The increased power was also reflected in the number and size of 95% CSs containing the true signals. As shown in Figure 2.3c and Figure 2.11, mixFine identified more true positive

95% CSs, and these 95% CSs were generally smaller than the ones of trcFine (paired t-test p=$5.88 \times 10^{-29}$) demonstrating that mixFine can pinpoint causal SNPs more accurately.

Overall, the combined method was more powerful for identifying causal variants, which is consistent with recent reports [170, 148].

### *2.3.4 Combining total and allele-specific read count improves prediction*

Using the data from the fine-mapping simulation, we tested the performance of mixPred and trcPred (Section 2.9.5.3) on held-out test data. Specifically, we split each simulation replicate into training (4/5) and test (1/5) sets. We trained prediction models using training data and evaluated the prediction performance on test data using Pearson correlation between predicted and true responses. For each data set, we repeated the splitting-training-evaluation procedure twice to reduce the stochasticity introduced by splitting.

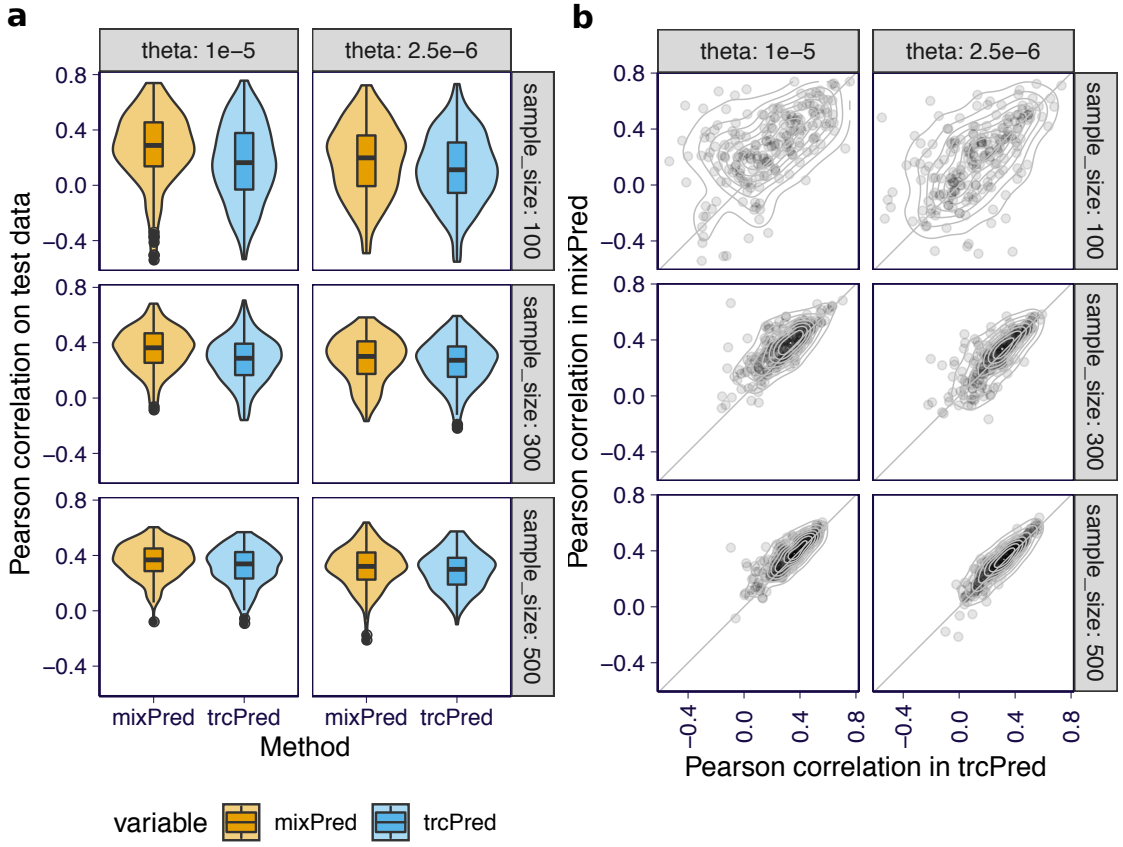



**Figure 2.4: Prediction performance of the combined (mixPred) and total read-based (trcPred) methods on simulated data. (a)** The overall distribution of Pearson correlations between predicted and observed total count abundance in log-scale, i.e., $\log(Y_i^{\text{total}}/L_i)$, for mixPred (orange) and trcPred (blue) across all data splits are shown. For each panel, the plot is based on 200 simulation replicates. In the boxplots, the lower and upper hinges show the first and third quartiles and the middle line shows the median. The whiskers extend from the hinge to the maximum and minimum at most 1.5x the inter-quartile range. All data points beyond the end of the whiskers are plotted individually. **(b)** For each split, the prediction performance of mixPred (y-axis) is plotted against the prediction performance of trcPred (x-axis).

Overall, mixPred achieved higher prediction accuracy than trcPred (Figure 2.4 and Figures 2.12 and 2.13). The gain in performance was more apparent when the expression level $\theta$ was higher and as a consequence the allele-specific count was larger.

### *2.3.5 mixQTL outperforms standard eQTL mapping in GTEx data*

Next, we compared mixQTL to the standard eQTL mapping approach (denoted here simply as eQTL) used by the GTEx consortium [139], using 670 whole blood RNA-seq samples from the v8 release (see Section 2.5.8). We included variants within a ±1 Mb cis-window around the transcription start site of each gene. Although mixQTL can be applied to all genes regardless of the number of allele-specific counts, we focus on examining the benefit of integrating allele-specific information and therefore limit these comparisons to genes with sufficient allele-specific counts, based on the following criteria: 1) at least 15 samples having at least 50 allele-specific counts for each haplotype; and 2) at least 500 samples having a total read count of at least 100. 5,734 (28%) of genes passed these filters. We then stratified these genes by their median expression level (read counts) into low, medium, and high expression tertiles. For genes with below-threshold allele-specific counts, the calculation can be performed using total read counts only, such that all genes considered using the standard approach are also tested in mixQTL. Performance for these genes was similar to the standard eQTL approach (Figure 2.14).

All three approaches mixQTL, aseQTL, and trcQTL were relatively well-calibrated when permuting data in four randomly selected genes (Figure 2.15). The estimated effect sizes were consistent with allelic fold change estimates from the main GTEx v8 analysis (Figure 2.16).

To further compare the performance of the methods, we used eQTLGen [145], a large-scale meta-analysis of over 30,000 blood samples, as our "ground truth" eQTL discovery reference (Section 2.9.8). We selected a random subset of 100,000 variant/gene pairs tested by eQTLGen with FDR $< 0.05$ as the set of "ground truth" eQTLs. We also selected a random set of 100,000 variant/gene pairs with $p > 0.50$ as a background set of "non-significant" eQTLs. Among these pairs, 96,660 and 78,691 of the "ground truth" and "non-significant" pairs had matching data in GTEx.

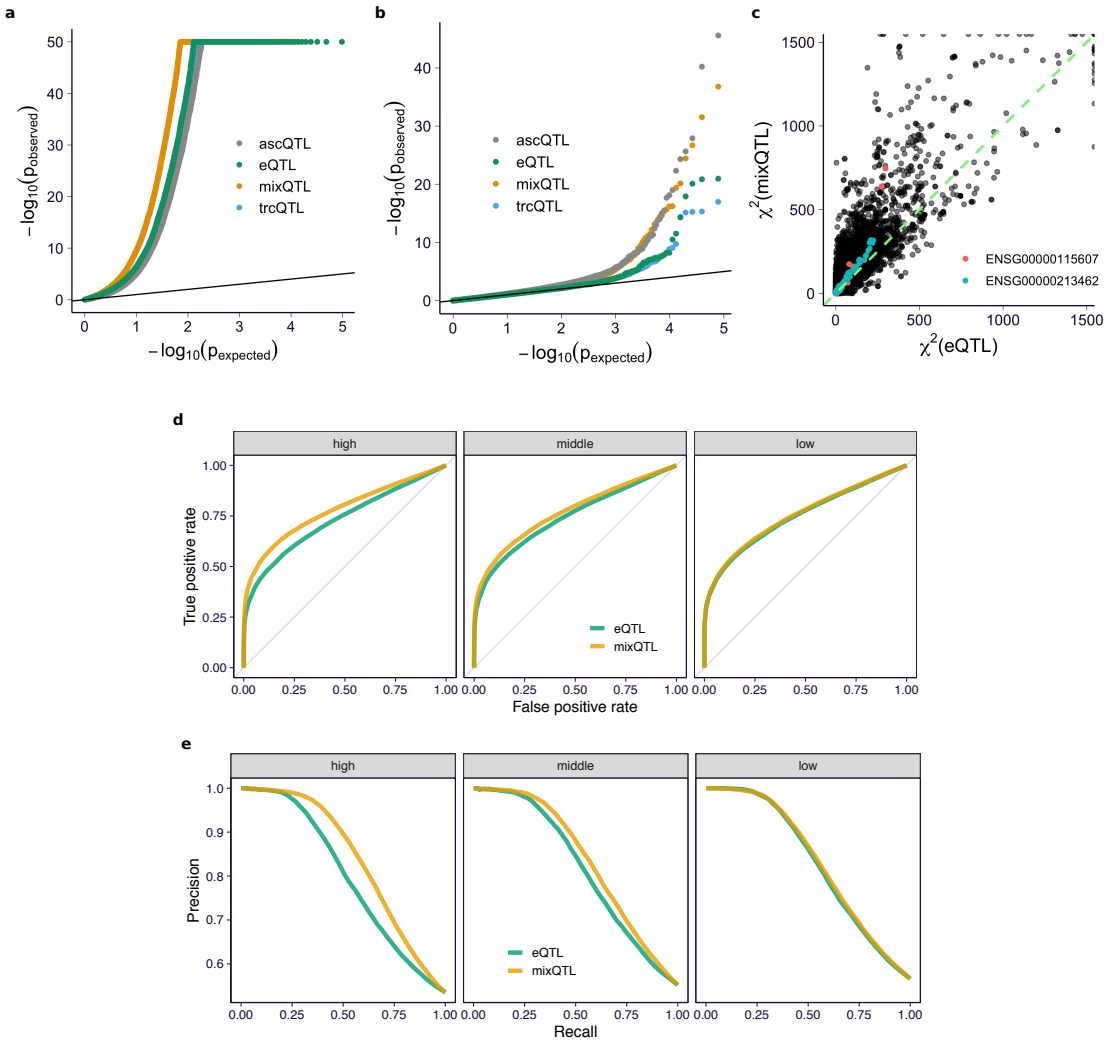



**Figure 2.5: Performance of mixQTL on GTEx v8 whole blood RNA-seq. (a)** QQ-plot of nominal p-values for a random subset (size = 96,660) of cis-eQTLs (FDR $< 0.05$) reported in eQTLGen. **(b)** QQ-plot of nominal p-values for a random subset (size = 78,691) of variant/gene pairs with p-value $> 0.5$ in eQTLGen. **(c)** $\chi^2$ statistics from eQTL analysis (x-axis) and mixQTL analysis (y-axis) among a random subset (size = 96,660) of cis-eQTLs (FDR $< 0.05$) reported in eQTLGen. Two randomly selected genes (ENSG00000115607 and ENSG00000213462) are highlighted in red and green, respectively. **(d, e)** ROC and PR curves for mixQTL and the standard eQTL method measured in eQTLGen. Each panel shows the results of genes stratified by expression level tertiles.

For the "ground truth" eQTLs, mixQTL yielded more significant p-values compared to the standard eQTL, ascQTL, and trcQTL approaches (Figure 2.5). The "non-significant" variant/gene pairs showed moderate enrichment for small p-values for all methods (Figure

2.5b), likely reflecting a combination of false negatives in eQTLGen and potential false positives in our analysis. Overall, we found that mixQTL achieves increased power compared to standard eQTL mapping on real data for the set of genes with sufficient total and allele-specific read counts.

As an intuitive measure of improved performance, we estimated the effective sample size gain of mixQTL compared to standard eQTL mapping as the median of the ratio between mixQTL $\chi^2$ statistics and eQTL $\chi^2$ statistics. mixQTL showed a 29% increase in effective sample size compared to the standard eQTL mapping approach (Figure 2.5c).

To account for the trade-off between true and false positive rates, as well as between precision and power, we used receiver operating characteristic (ROC) and precision-recall (PR) curves to compare the performance of mixQTL and standard eQTL approaches using the eQTLGen "ground truth" and "non-significant" eQTLs. We found that mixQTL achieves higher performance in both ROC (Figure 2.5d) and PR curves (Figure 2.5e). Consistent with simulation results, this gain is more significant for genes with higher expression levels.

To determine whether the eQTLGen-based analysis above depended on the selected random subset of cis-eQTLs, we repeated the analysis for multiple samplings of eQTLGen results and found no substantive differences in the results.

### *2.3.6 mixQTL is scalable to full GTEx eQTL analysis*

To compare the performance and computational cost of mixQTL and the existing QTL mapping approaches which can leverage both total and allele-specific counts, we ran RASQUAL on two of the GTEx tissues, kidney cortex (sample size = 73; a subset of 4,596 genes) and whole blood (a subset of 192 genes; Section 2.9.9). We observed concordant effect size estimates (Figure 2.17A). As expected, because RASQUAL models counts directly instead of approximating them with a log linear model, it yielded more significant results than mixQTL (Figure 2.17B). On average, RASQUAL took 47 seconds per gene in kidney cortex and 826

seconds per gene in whole blood whereas mixQTL took 0.065 seconds (723 times faster) and 0.33 seconds (2,480 times faster), respectively.

Given this computational efficiency, we decided to run mixQTL on the 49 tissues from the GTEx v8 release. This corresponded to 15,201 samples in total, and took approximately 54 CPU hours in total (without permutations).

mixQTL's runtime scaled linearly as a function of sample size (Figure 2.18A), with the tissue with the largest sample size (skeletal muscle, n = 706) taking 0.34 seconds per gene on average.

At FDR cutoff 0.05, on average, mixQTL identified 1440 more genes and about 618,000 more eQTLs than the standard eQTL approach (Figures 2.18B and 2.18C).

### *2.3.7 Fine-mapping and prediction model building in GTEx data*

We applied mixFine to the GTEx v8 whole blood RNA-seq data, using the same subset of genes with high expression and allelic counts that were used in the comparison of mixQTL vs. standard eQTL approach above. We compared mixFine to the SuSiE fine-mapping approach [149], applied to inverse normal transformed expression values in the standard eQTL mapping pipeline [139]. We corrected for sex, 5 genetic principal components, WGS platform, WGS library prep protocol (PCR), and 60 PEER factors. We refer to the latter as the "standard approach" below for simplicity.

To compare the power of causal variant detection, we performed a subsampling analysis on a random subset of 1,000 genes. First, we defined "consensus SNPs" as the variants with PIP $> 0.5$ in both mixFine and the "standard approach" using all samples. Similarly, a variant was defined as "top SNP" if it was the most significant variant within the 95% CS for both mixFine and the "standard approach". Then, we compared how well the "consensus SNPs" and "top SNPs" were detected by mixFine and the standard fine-mapping approach using only a subset of samples. We subsampled to 90%, 80%, $\cdots$, 30% of samples, and

repeated each random subsampling step 10 times.

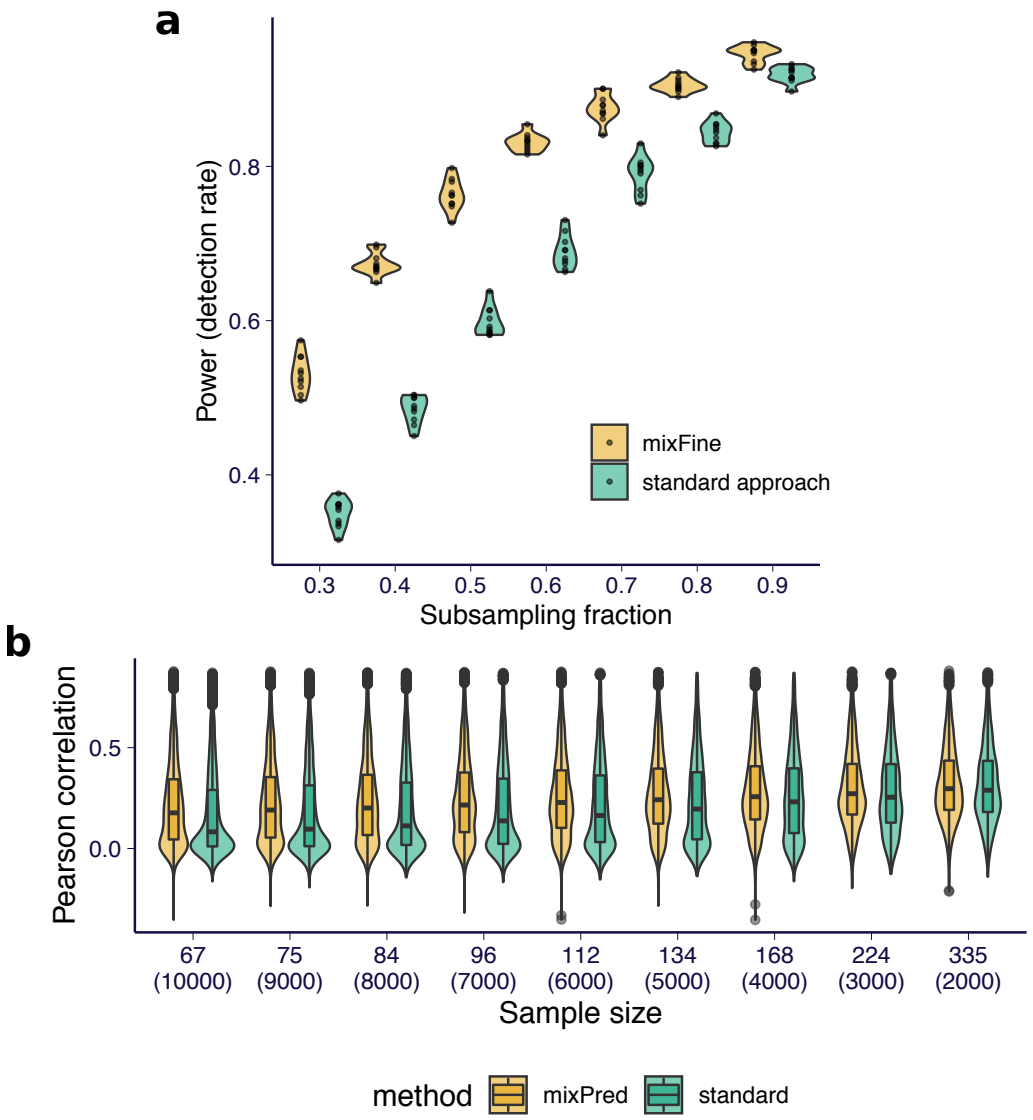



**Figure 2.6: Performance of mixFine and mixPred on GTEx v8 whole blood RNA-seq.** **(a)** The fraction of detected "consensus SNPs" among all 272 "consensus SNPs" in full data as a function of subsampling level, for mixFine and the standard approach, are shown. The subsampling analysis are repeated 10 times. The plot shows the results of all the ten replications. **(b)** The Pearson correlation between observed and predicted expression across all models trained from 1,000 genes are shown. "Standard" corresponds to the elastic net model as implemented in [43]. The results are stratified by sample size used for training. For each sample size, the distribution of the Pearson correlation across all cross-validation folds and genes are shown (the corresponding total number of observations is shown in the parentheses). In the boxplots, the lower and upper hinges show the first and third quartiles and the middle line shows the median. The whiskers extend from the hinge to the maximum and minimum at most 1.5x the inter-quartile range. All data points beyond the end of the whiskers are plotted individually.

Among the 1,000 genes, there were 272 "consensus SNPs" being identified in the full data. At each subsampling level, mixFine, on average, detected more "consensus SNPs" than the standard approach (Figure 2.6a) and performance improved most on the more highly expressed genes (top tertile) (Figure 2.19). Moreover, mixFine detected "top SNPs" in 95% CSs with an average of 9.5 variants, whereas the corresponding 95% CS from the standard approach had 14.6 variants on average (Figure 2.20). Furthermore, since the power gain would be more apparent in small sample sizes, we ran mixFine and standard eQTL approach in 26 GTEx v8 tissues with sample size $< 260$. We examined the enrichment of the top QTL and PIP in different functional annotations, including regulatory element annotations, candidate cis-regulatory elements (cCREs) [104], and GWAS catalog (Section 2.9.10). We found that the variants with the most significant mixQTL p-value or the highest mixFine PIP were more enriched in GWAS catalog variants and cCREs than the standard approach. We found enrichment of enhancer, promoter, and transcription factor binding sites but the difference in enrichment between mixQTL and standard QTL methods was not significant (Figure 2.22). The reduced enrichment compared to cCREs are likely due to the fact that we used tissue specific annotations for cCREs and cross tissue annotations for enhancers, promoters, and TF. These results indicate that, when sufficient counts are available, mixFine, the multi-SNP model combining total and allele-specific counts, can better pinpoint causal cis-eQTLs than the standard approach on real data.

To compare the performance of mixPred and the standard method on real data, we implemented a cross-validated evaluation pipeline where we split the GTEx v8 whole blood data into $k$ folds. At each fold, we trained the prediction model using one fold of the data and evaluated the performance (by Pearson correlation between predicted and observed $\log(Y_i^{\text{total}}/L_i)$) on the remaining $(k-1)$ folds. We applied this evaluation pipeline to mixPred and the standard approach (elastic net as in [43]) on the same 1000 genes as the subsampling analysis with $k = 10, 9, \ldots, 2$ (corresponding to sample size $= 67, 75, \cdots, 335$). At the

same sample size, we observed, on average, significantly higher performance in mixPred as compared to the standard approach, and the performance gain was greater for smaller sample sizes (Figure 2.6b and Table 2.2).

## 2.4 Discussion

We proposed a unified framework that integrates both allele-specific and total read counts to estimate genetic cis-regulatory effects, resulting in improved eQTL mapping, fine-mapping, and prediction of gene expression traits. Our suite of tools (mixQTL, mixFine, and mixPred) can be scaled to much larger sample sizes (thousands) due to the underlying log-linear approximation. By assuming weak multiplicative genetic effects consistent with observations (most estimated log allelic fold changes of cis-eQTLs have a median absolute value of 0.153 and a 95th percentile of 0.845 (Figure 2.21)), we transform the observed read counts into two approximately independent quantities: allelic imbalance and total read count. Leveraging this independence, we developed computationally efficient approaches that integrate both allele-specific and total reads.

Specifically, mixQTL estimates the genetic effect separately for allelic imbalance and total read counts, and combines the resulting statistics via meta-analysis. These calculations have computationally efficient closed-form solutions, enabling their use in permutation schemes applied to compute FDR in eQTL mapping [124, 110, 136].

Furthermore, the simple multi-SNP extension and the approximate independence of the terms enable use of a two-step inference procedure. In a first step, the allelic imbalance and total read count are scaled such that the error terms have the same variance. And in a second step, given their approximate independence, the pair of equations (from allelic imbalance and total counts) can simply be input into existing fine-mapping and prediction algorithms.

We showed through simulations and applications to GTEx v8 data that our suite of

methods outperforms methods that rely on total read counts alone. Compared to existing QTL mapping methods that integrate total and allele-specific reads, such as RASQUAL [70], mixQTL has slightly lower power (Figure 2.17B). This is expected since RASQUAL models count data directly and mixQTL relies on approximations. However, the computational burden of RASQUAL is prohibitive for large datasets. In practice, the most suitable approach will depend on computational capacity and sample sizes. For datasets with small sample sizes (e.g., fewer than 100 samples), RASQUAL or WASP remain preferable. The computational efficiency of mixQTL makes it applicable to large sample sizes, and, moreover, enables using the mixQTL model in place of the standard eQTL mapping approach that relies on inverse normal transformed counts.

Comparing to another recent fine-mapping method PLASMA [148] which also combines the total and allele-specific reads, our proposed method took a different approach on unifying the signals. In PLASMA, QTLs are called from total and allele-specific counts separately and then these results, in terms of z-scores, are combined via a Bayesian model assuming that QTLs from total and allele-specific counts share exactly the same causal status. Due to handling z-scores, the scale of the signal becomes entangled with the sample size and how exactly the QTLs are called (since different methods differ in statistical power). To handle this complication, PLASMA introduced additional hyperparameter, the jointness hyperparameter. Although the results were shown to be robust to this jointness parameter in a simulated dataset, it still introduces uncertainty in real-world application. Whereas, our proposed method handles total and allele-specific signals in a unified framework, which allows us to combine the effect sizes (instead of z-scores) directly in mixQTL and to use one linear model leveraging both total and allele-specific signals in fine-mapping and prediction problems. With this setup, we avoid handling extra hyperparameters and the resulting estimates are clear in terms of the scale. Regarding the statistical power, if the cis-regulation is the underlying molecular mechanism, our proposed method should be more powerful than

PLASMA but, at the same time, our methods are prone to model misspecification if factors other than the cis-regulation drive the signals.

Given the unified modeling framework and computationally scalable tools proposed here, we anticipate that combining total and allele-specific read counts will find widespread use for cis-QTL mapping, fine-mapping, and prediction of gene expression.

## 2.5 Methods

### 2.5.1 Notation and terminology

| Notation | Description | Synonym in text | Observable |
|---|---|---|---|
| $i$ | Individual index. | - | - |
| $h$ | Haplotype index, with $h = 1, 2$ for diploid. | - | - |
| $X_i^h$ | Alternative allele count (0 or 1) of the variant linking to the gene haplotype $h$. | allelic dosage | Yes |
| $L_i$ | The total number of reads in the RNA-seq library. | library size | Yes |
| $Y_i^h$ | Count of reads originated from gene haplotype $h$. | haplotypic (read) count | No |
| $Y_i^{(h)\text{obs}}$ | Allele-specific read count that gets aligned to the gene haplotype $h$. | allele-specific (read) count | Yes |
| $Y_i^{\text{total}}$ | Total count of reads originated from any of the two gene haplotypes (sum). | total (read) count | Yes |
| $\theta_{0,i}$ | The abundance of the gene haplotype relative to the total transcriptome when the linked causal variants are all in reference alleles | baseline (relative) abundance | No |
| $\theta_i^h$ | The abundance of the gene haplotype $h$ relative to the total transcriptome in individual $i$ | (relative) abundance; expression level$^\dagger$ | No |
| $\beta$ | The log fold change of gene haplotype abundance when linking to alternative allele relative the reference allele | allelic fold change (aFC) in natural log scale | No |
| $\frac{Y_i^{(1)\text{obs}}}{Y_i^{(2)\text{obs}}}$ | The ratio of the allele-specific counts between two haplotypes | allelic imbalance | Yes |
| $Y_i^{\text{trc}}$ | Shorthand of the term $\log \frac{Y_i^{\text{total}}}{2L_i}$. | - | - |
| $Y_i^{\text{asc}}$ | Shorthand of the term $\log \frac{Y_i^{(1)\text{obs}}}{Y_i^{(2)\text{obs}}}$ | - | - |
| $\theta$ | Only used in simulation where $\theta = \text{E}(\theta_{0,i})$ | expression level$^\star$ | - |

**Table 2.1: Summary of notation and terminology used in the paper.** The "Description" column contains a brief definition of each "Notation", and the "Synonym in text" column contains the corresponding terminology used in the text. The "Observable" column indicates whether the entity is an observable variable or not. (†, ⋆: expression level does not strictly refer to $\theta_i^h$ or $\text{E}(\theta_{0,i})$, but more generally to the abundance of the gene transcripts relative to the transcriptome.)

### 2.5.2 Statistical model of cis-regulation

For individual $i$, let $X_i^1$ and $X_i^2$ be the number of alternative alleles in each of the two haplotypes at the variant of interest. Let $Y_i^1$ and $Y_i^2$ be the number of reads mapped to each of the two haplotypes (i.e., haplotypic counts; in practice, these quantities are unobserved)

and $L_i$ the library size for individual $i$. As proposed in [101], we use the concept of allelic fold change (aFC) to represent the genetic effect on cis-expression. We denote $\theta_{0,i}$ as the baseline abundance of the transcripts originating from each of the gene haplotype without considering genetic effect. Let $\beta$ be the genetic effect of a variant of interest, which is defined as the log fold change relative to the reference allele. Then, the transcript abundance of each haplotype $h$ after accounting for the genetic effect is $\theta_i^h = \theta_{0,i} \times g(\beta, X_i^h)$ where $g(\beta, X_i^h)$ is $e^\beta$ if $X_i^h$ is the alternative allele; otherwise $g(\beta, X_i^h) = 1$. We model read count $Y_i^h$ as

$$\log Y_i^h | L_i, \theta_i^h \sim N(\log(L_i \theta_i^h), \tau_i^h). \tag{2.5}$$

In an RNA-seq experiment, a fraction of reads contribute to allele-specific read counts. Let $\alpha_i$ denote the fraction of allele-specific reads in individual $i$, which depends on the number of heterozygous sites within the transcript. Instead of observing haplotypic counts $Y_i^1$ and $Y_i^2$, we observe total read count $Y_i^{\text{total}}$ and gene-level allele-specific read counts $Y_i^{(1)\text{obs}}$ and $Y_i^{(2)\text{obs}}$. Similarly, we further assume that the baseline abundance of allele-specific reads per haplotype is $\theta_{0,i} \times \alpha_i$, so we have

$$\log Y_i^{(1)\text{obs}} | L_i, \theta_i^1, \alpha_i \sim N(\log(\alpha_i L_i \theta_i^1), \tau_i^{(1)}) \tag{2.6}$$

$$\log Y_i^{(2)\text{obs}} | L_i, \theta_i^2, \alpha_i \sim N(\log(\alpha_i L_i \theta_i^2), \tau_i^{(2)})$$

$$\log Y_i^{\text{total}} | L_i, \theta_i^1, \theta_i^2 = \log(Y_i^1 + Y_i^2) | L_i, \theta_i^1, \theta_i^2 \tag{2.7}$$

$$\sim N(\log[L_i(\theta_i^1 + \theta_i^2)], \tau_i) \tag{2.8}$$

### *2.5.3 Linearizing the model by approximation*

Based on the model described above along with approximations under weak effect assumptions, we propose the following linear mixed effects model (see Section 2.9.2 for derivation):

$$\underbrace{\log \frac{Y_i^{\text{total}}}{2L_i}}_{Y_i^{\text{trc}}} = \mu_0 + + \underbrace{\frac{X_i^1 + X_i^2}{2}}_{X_i^{\text{trc}}} \beta + \epsilon_i^{\text{trc}} \tag{2.9}$$

$$\underbrace{\log \frac{Y_i^{(1)\text{obs}}}{Y_i^{(2)\text{obs}}}}_{Y_i^{\text{asc}}} = \underbrace{(X_i^1 - X_i^2)}_{X_i^{\text{asc}}} \beta + \epsilon_i^{\text{asc}} \tag{2.10}$$

$$z_i \sim N(0, \sigma_0^2),\ \epsilon_i^{\text{trc}} \sim N(0, \frac{\sigma^2}{Y_i}),\ \epsilon_i^{\text{asc}} \sim N(0, \underbrace{\frac{\sigma^2 Y_i^{(1)} Y_i^{(2)}}{Y_i^{(1)} + Y_i^{(2)}}}_{\sigma^2/w_i}), \tag{2.11}$$

where $z_i$ is the individual-level random effect capturing the between-individual variation of $\theta_{i,0}$. Notice that the individual-level random effect cancels out when we take the difference between the two log-scale allele-specific read counts (allelic imbalance in log-scale). The scaling of $\epsilon^{\text{trc}}$ and $\epsilon^{\text{asc}}$ in Eq 2.11 is to ensure that variance of read count scales linearly with the magnitude of read count (see Section 2.9.1.2). In other words, this model ensures $\text{Var}(Y) \approx \text{constant} \times \text{E}(Y)$, such that over-dispersion is implicitly taken into account.

Since $\epsilon_i^{\text{asc}}$ is approximately independent to $\epsilon_i^{\text{trc}}$ (see Section 2.9.4), $\epsilon_i^{\text{trc}}$ and $z_i$ can be merged into one term $\tilde{z}_i$. So, we can further simplify Eq 2.9, 2.10 as

$$Y_i^{\text{trc}} = \mu_0 + X_i^{\text{trc}} \beta^{\text{trc}} + \tilde{z}_i \quad , \tilde{z}_i \quad \sim N(0, \tilde{\sigma}_0^2) \tag{2.12}$$

$$Y_i^{\text{asc}} = \quad X_i^{\text{asc}} \beta^{\text{asc}} + \epsilon_i^{\text{asc}}, \epsilon^{\text{asc}} \sim N(0, \sigma^2/w_i) \tag{2.13}$$

Eqs 2.12, 2.13 are applicable to both single-SNP and multi-SNP scenarios. In the single-SNP

case, $X_i$ and $\beta$ are scalars, and in the multi-SNP case, $X_i$ and $\beta$ are replaced by vectors including all SNPs within the cis-window (see Section 2.9.3).

### *2.5.4 Numerically efficient QTL mapping leveraging approximate independence of allelic imbalance and total read count*

The likelihood function corresponding to the proposed model in Eqs 2.12, 2.13 approximately takes the form

$$\prod_i \Pr(Y_i^{\text{total}}|\mu_0, \sigma_0^2, \sigma^2, \beta) \cdot \Pr(\frac{Y_i^{(1)\text{obs}}}{Y_i^{(2)\text{obs}}}|\sigma^2, \beta),$$

factoring into total read count and allelic imbalance components. (see Section 2.9.2.2). This means that the likelihood for total read count and the ratio of allele-specific read counts provide approximately independent information on $\beta$, and enables us to solve each component separately and combine the results via meta-analysis (standard approach with independent studies [38]). Specifically, we fit $\beta^{\text{trc}}$ and $\beta^{\text{asc}}$ using total and allele-specific observations as two separate linear regression problems, and meta-analyze the results using inverse-variance weighting (see details in Section 2.9.4.2).

### *2.5.5 Two-step inference procedure for multi-SNP model*

The prediction and fine-mapping problems both rely on the linearized model Eq 2.12, 2.13, but with different objectives. For prediction, the objective is to find the best predictor, whereas for fine-mapping, the objective is to infer whether $\beta_k$ is non-zero. Existing solvers for both prediction and fine-mapping use total read information only and assume that data $(X, y)$ follows the model $y = X\beta + \epsilon$, where the noise term $\epsilon$ is independent across the rows of the data matrix. We will refer to this model as the 'canonical' linear model. We propose a two-step inference procedure that first processes the data such that it approximates $y = X\beta+\epsilon$, and then uses existing solvers for prediction and fine-mapping problems, respectively.

For the first step, we process total and allele-specific reads separately to fit the 'canonical' linear model. Specifically, we estimate $\sigma^2$ from $(Y^{\text{asc}}, X^{\text{asc}})$ based on Eq 2.13 by further assuming the genetic effect as random effect and estimating $\sigma^2$ using R package EMMA [62]. And similarly, based on Eq 2.12 and the random effect assumption, we estimate $\widetilde{\sigma}_0^2$ from $(Y^{\text{trc}}, X^{\text{trc}})$. To account for the intercept term $\mu_0$ in Eq 2.12, we center $Y^{\text{trc}}$ and $X^{\text{trc}}$ by subtracting the mean values across all samples and then scale the centered $(Y^{\text{trc}}, X^{\text{trc}})$ by $1/\widehat{\widetilde{\sigma}}_0$. And similarly, we scale $(Y^{\text{asc}}, X^{\text{asc}})$ by $w/\hat{\sigma}$. These linear transformations ensure that the transformed $(\tilde{Y}^{\text{trc}}, \tilde{X}^{\text{trc}})$ and $(\tilde{Y}^{\text{asc}}, \tilde{X}^{\text{asc}})$ both approximately follow $Y = X\beta + \epsilon$. The implementation details are described in Section 2.9.5. At the second step, we concatenate the transformed data from both total and allele-specific read counts as $(\tilde{Y}, \tilde{X})$, which is compatible with existing solvers for prediction and fine-mapping problems.

### *2.5.6 Adjusting for covariates*

When analyzing real data, we need to take covariates such as sex, batch effect, population stratification into account. Here, we adapt the procedure which has been proposed previously [101]. We regress out the effect of covariates beforehand and use the residual as the response in both QTL mapping and fitting multi-SNP model. Specifically, let $c_1, \cdots, c_K$ denote the $K$ covariates to be considered. We first regress $Y^{\text{trc}}$ against $c_1, \cdots, c_K$ jointly and select the covariates with nominally significant coefficients ($\text{p} < 0.05$). Then we regress $Y^{\text{trc}}$ against the selected covariates jointly and set the residuals as the adjusted $Y^{\text{trc}}$ for QTL mapping and multi-SNP inference downstream.

### *2.5.7 Simulation scheme*

We simulate RNA-seq reads with total and allele-specific readouts as sketched in three steps in Figure 2.1. In step 1, we specify, for each individual $i$, the position of heterozygous sites within the gene body. The expected read count from each haplotype transcripts, $\text{E}(Y_i^h)$, is

determined by the RNA-seq library size $L_i$, the baseline abundance of the transcript $\theta_{0,i}$, and the genetic effect $\beta$. In step 2, given the expected haplotypic count, we draw $Y_i^h$ from Negative Binomial to model the variation among count data. In step 3, we position the reads randomly along the gene body and readout observed allele-specific count $Y_i^{(h)\text{obs}}$ by counting the number of reads overlapping heterozygous sites simulated in step 1. The total read count readout is $Y_i = Y_i^1 + Y_i^2$, which is independent of the number of heterozygous sites.

To survey a wide range of parameters, we simulate data with a grid of parameters. We vary sample size among 100, 200, ..., 500. At library size around 90 million, we vary the level of $\theta_{0,i}$ to cover the gene with different expression levels, among $5 \times 10^{-5}, 2.5 \times 10^{-5}, 1 \times 10^{-5}, 2.5 \times 10^{-6}, 1 \times 10^{-6}$. The genetic effect, aFC, is set to 1 (null), 1.01, 1.05, 1.1, 1.25, 1.5, 2, 3 in the single SNP model. For the multi-SNP scenario, we set the number of causal SNPs between 1 and 3 with heritability from 0.2 to 0.55. The number of polymorphic sites within the gene body is centered around 10 with minor allele frequency from 0.05 to 0.3. A detailed description and parameter settings are provided in the Section 2.9.6.

### *2.5.8 Analysis of GTEx v8 data*

We downloaded the phased genotypes, total read count matrix, and variant-level allele-specific read counts for whole blood from GTEx release 8 [139] via dbGaP (accession number phs000424.v8.p1). To obtain gene-level read counts, we summed over allele-specific counts at all the heterozygous sites for each gene haplotype. We also obtained library size, sex, and genotype PCs from GTEx v8. For comparisons with the inverse normalization-based approach, we also downloaded normalized expression matrices.

Similarly to the GTEx v8 analyses [139], we restricted the analysis to the cis-regulatory window defined as 1Mbp up/downstream of the transcription start site of each gene.

To obtain the PEER factors for mixQTL analysis, we ran `peertool` [130] on a matrix

with value $\log(\frac{Y_{i,g}}{2L_i})$ for individual $i$ and gene $g$ (imputed by k-nearest neighbors if $Y_{i,g}$ is zero using `impute::impute.knn` in R).

We considered very large allele-specific counts to be likely alignment artifacts and removed individuals with allele-specific read counts greater than 1000. To further limit the influence of large count outliers on the estimated log fold-change, $\hat{\beta}^{\text{asc}}$, we set the largest weight $\left(\frac{1}{Y^{(1)\text{obs}}} + \frac{1}{Y^{(2)\text{obs}}}\right)^{-1}$ to be at most $K$ fold to the smallest one, where $K = \min(10, \text{sample size}/10)$.

Specific analyses focused on high or low expression were performed with different gene filtering criteria as stated in the Results section.

For analyses of the full GTEx v8 dataset, we built a data analysis pipeline at `https://github.com/liangyy/mixqtl-gtex/tree/master/mixqtl` which relied on the tensorQTL implementation of mixQTL. We included all genes regardless of expression level and analyzed the 22 autosomes for each of the 49 tissues. Specifically, since mixQTL can only work with non-zero total read count, we imputed the samples with missing total read count as 1. And in the mixQTL call, all total read counts were included and all allele-specific counts with more than $\geq 15$ reads (on both haplotypes) were included.

## 2.6 Data Availability

Genotype-Tissue Expression (GTEx) project's raw whole transcriptome and genome sequencing data are available via dbGaP accession number phs000424.v8.p2. All processed GTEx data are available via GTEx portal (`http://gtexportal.org/`). The mixQTL full summary statistics for 49 GTEx tissues are listed in [81] Supplementary Data 1.

## 2.7 Code Availability

Softwares mixQTL, mixFine, and mixPred in R `https://github.com/hakyimlab/mixqtl`.

A reproducible pipeline for the simulated data and some GTEx data analysis `https://github.com/liangyy/mixqtl-pipeline`.

A reproducible pipeline for the massive GTEx data analysis `https://github.com/liangyy/mixqtl-gtex`.

A GPU-based implementation embedded in tensorQTL `https://github.com/broadinstitute/tensorqtl`.

## 2.8 Supplementary Figures and Tables

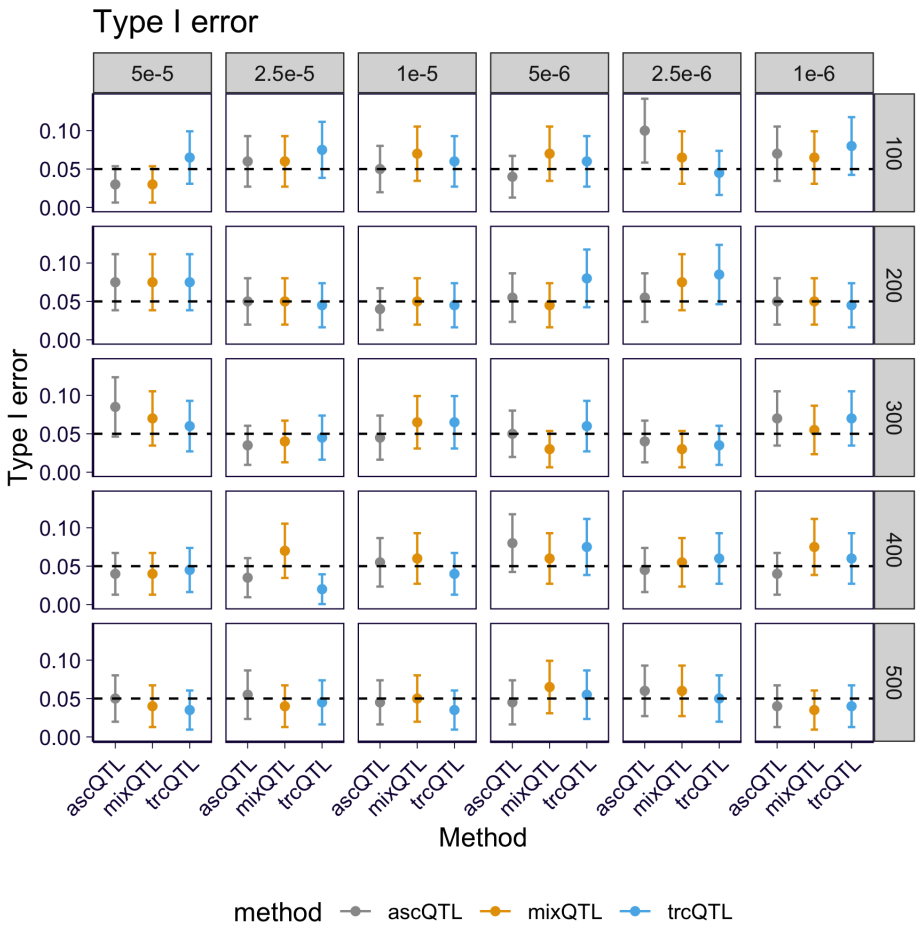



**Figure 2.7: Type I error of mixQTL, ascQTL, and trcQTL on the full grid of simulations.** Each panel shows results on data simulated under a pair of $\theta$ (relative abundance in the simulation, by column) and sample size (by row). The error rate under significance level $\alpha = 0.05$ from 200 replicates is shown. The error bar indicates the 95% confidence interval of the estimated error rate.

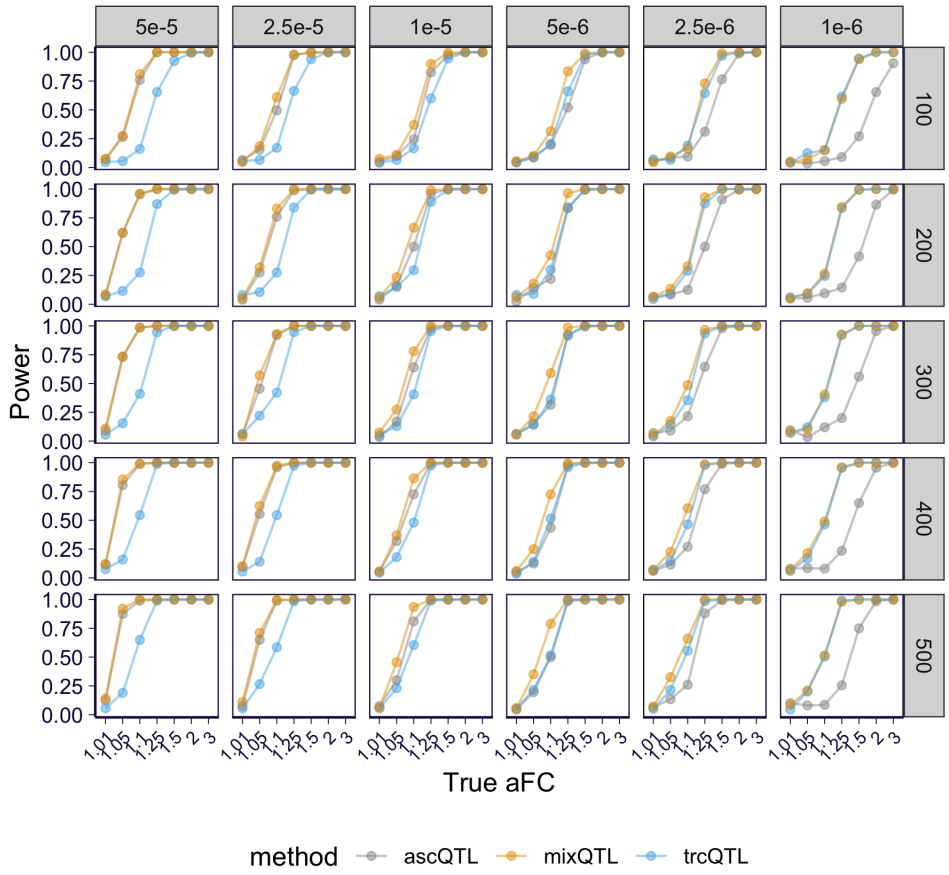



**Figure 2.8: Power of mixQTL, ascQTL, and trcQTL on the full grid of simulations.** Each panel shows results on data simulated under a pair of $\theta$ (relative abundance in the simulation, by column) and sample size (by row). The power is calculated under significance level $\alpha = 0.05$.

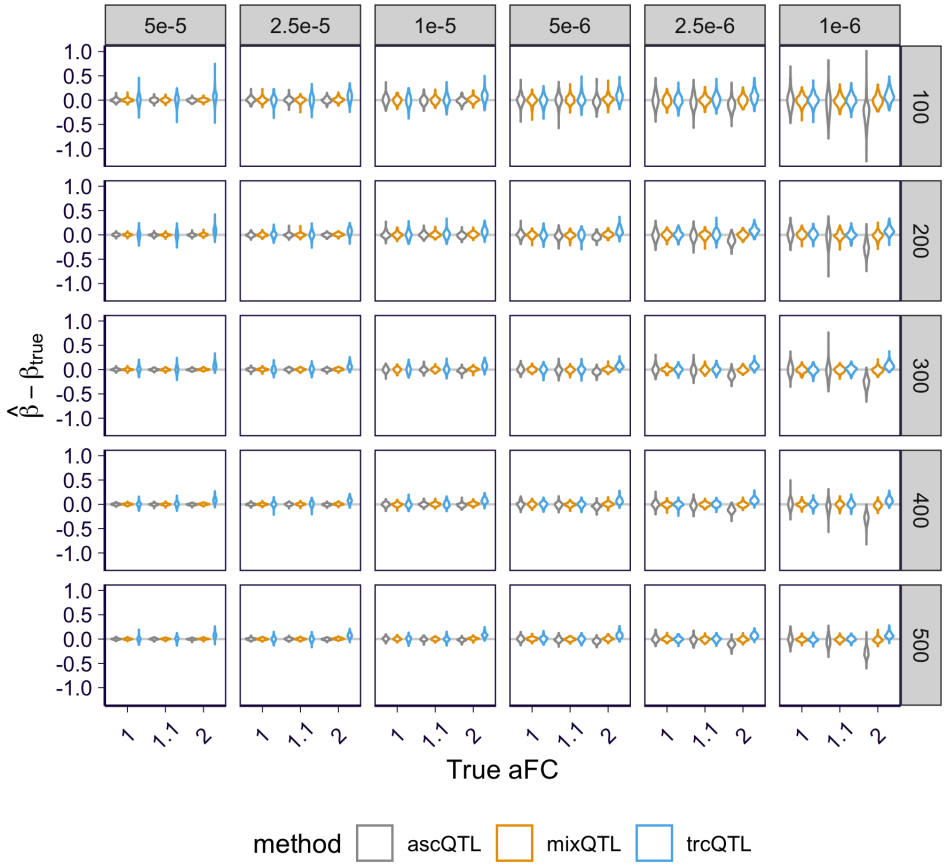



**Figure 2.9: Difference between $\hat{\beta}$ and true $\beta$ of mixQTL, ascQTL, and trcQTL on the full grid of simulations.** Each panel shows results on data simulated under a pair of $\theta$ (relative abundance in the simulation, by column) and sample size (by row). The difference between the estimated log allelic fold change and the true log allelic fold change is shown on y-axis.

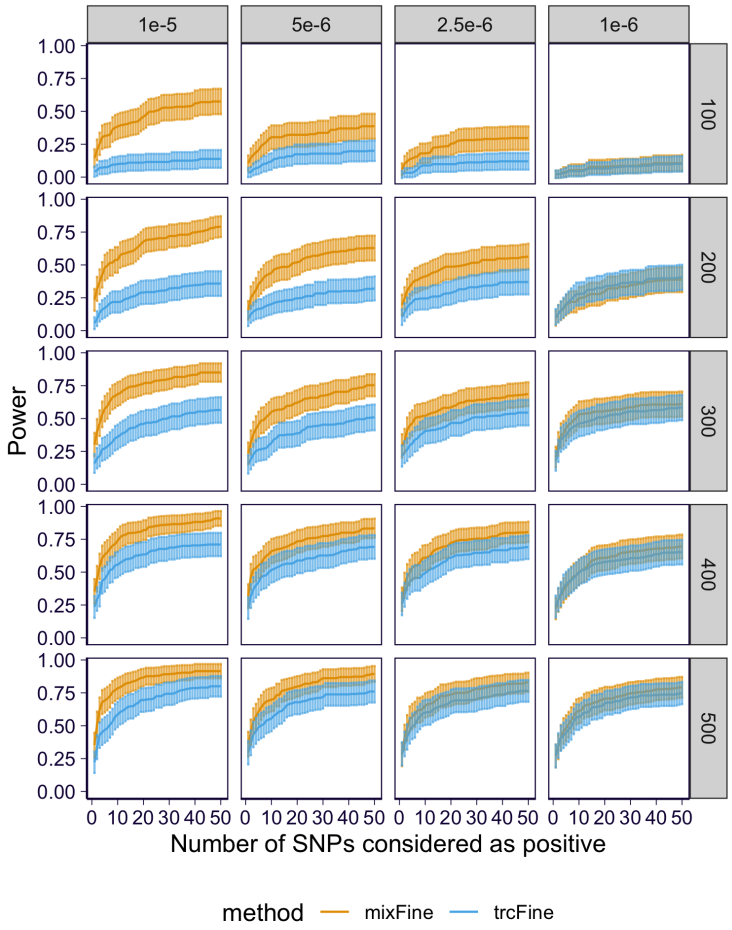



**Figure 2.10: Power curves of mixFine and trcFine on the full grid of simulations.** Each panel shows results on data simulated under a pair of $\theta$ (relative abundance in the simulation, by column) and sample size (by row). In each panel, the curve is based on 200 simulation replicates with 100 simulations having signals and 100 simulations being drawn from the null. The solid curves indicate the mean power (recall rate) among 100 simulation replicates and the error bars indicate the 95% confidence interval.

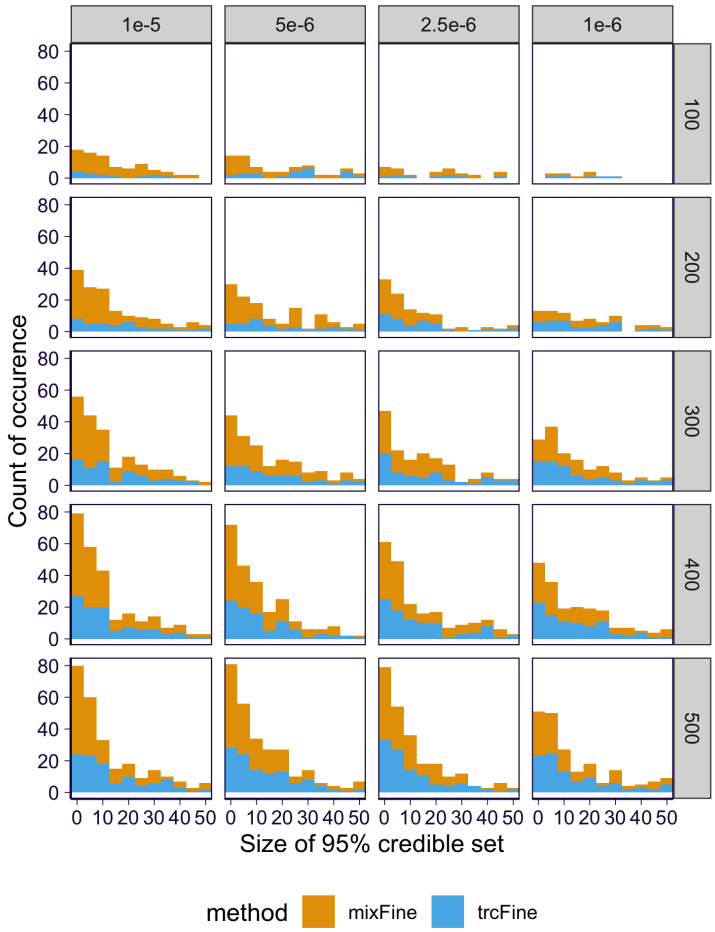



**Figure 2.11: Distribution of the positive 95% CS's which contain causal variants in mixFine and trcFine on the full grid of simulations.** Each panel shows results on data simulated under a pair of $\theta$ (relative abundance in the simulation, by column) and sample size (by row).

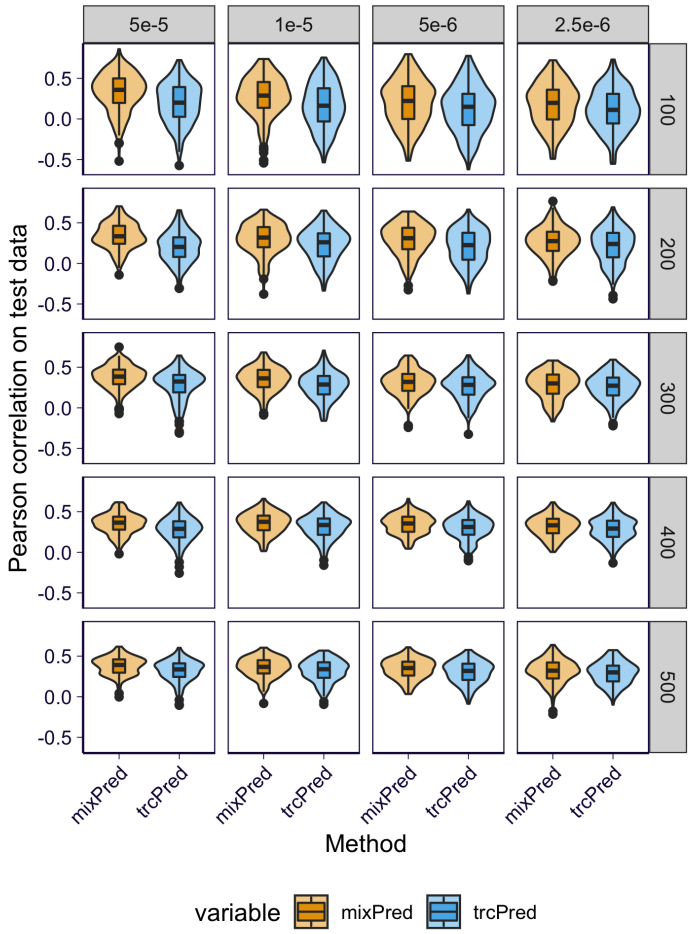



**Figure 2.12: Distribution of Pearson correlations between predicted and observed expression level (in the scale $\log(Y_i^{\mathbf{total}}/L_i)$) for mixPred and trcPred on the full grid of simulations.** Correlation is calculated on held-out test data. Each panel shows results on data simulated under a pair of $\theta$ (relative abundance in the simulation, by column) and sample size (by row). For each panel, the plot is based on 200 simulation replicates. In the boxplots, the lower and upper hinges show the first and third quartiles and the middle line shows the median. The whiskers extend from the hinge to the maximum and minimum at most 1.5x the inter-quartile range. All data points beyond the end of the whiskers are plotted individually.

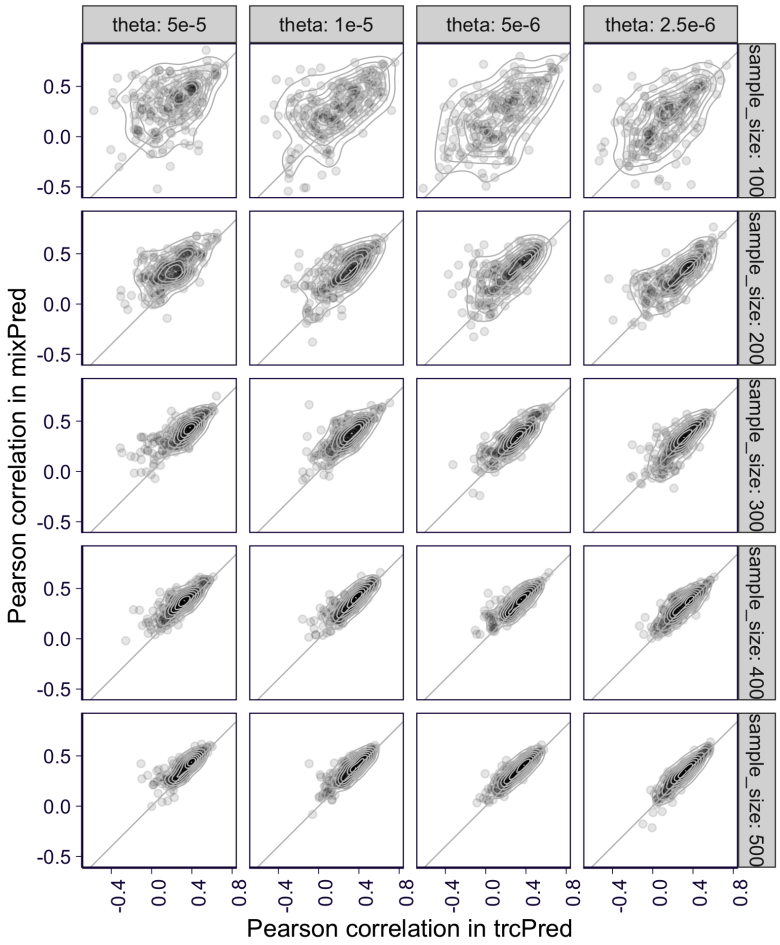



**Figure 2.13: Pairwise comparison of prediction performance of mixPred and trcPred on the full grid of simulations.** Correlation of predicted versus observed expression level (in the scale $\log(Y_i^{\text{total}}/L_i)$) is calculated on held-out test data. The prediction performance of mixPred (y-axis) is plotted against the prediction performance of trcPred (x-axis) for each split. Each panel shows results on data simulated under a pair of $\theta$ (relative abundance in the simulation, by column) and sample size (by row).

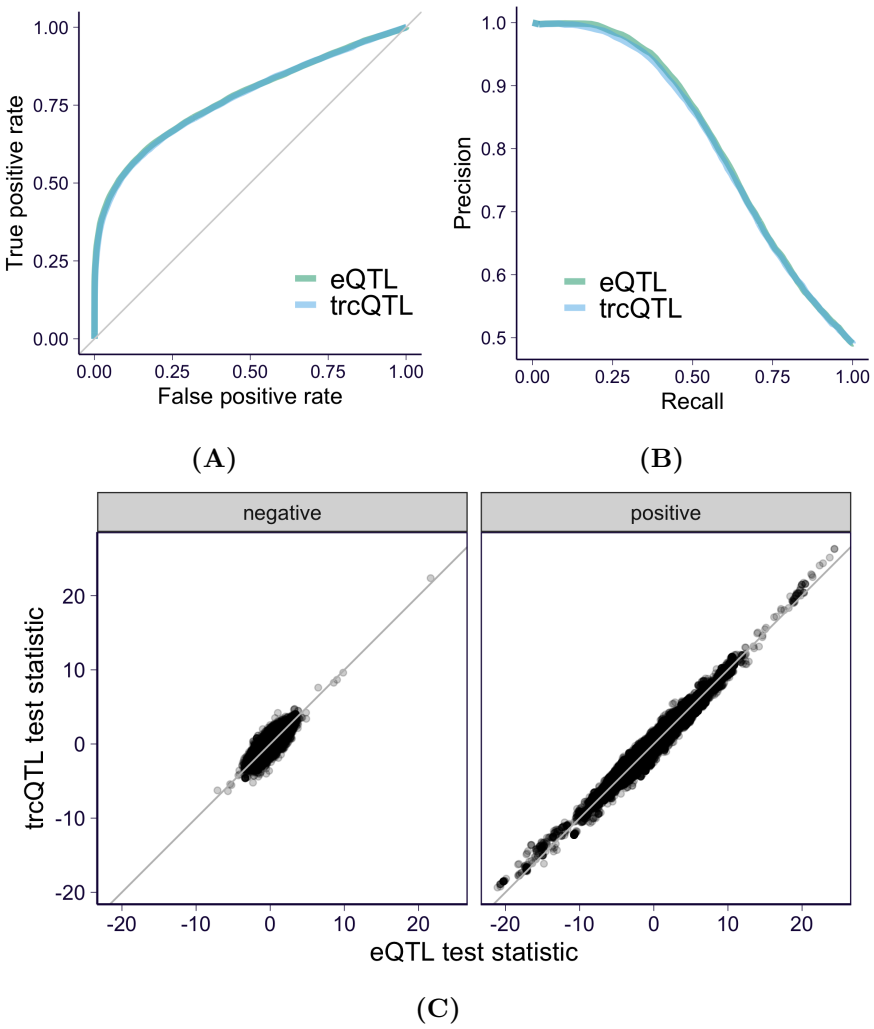



**Figure 2.14: The performance of trcQTL and the standard eQTL approach on genes with low total read counts.** Genes with low total counts are defined as having no more than 50 total read counts in any one sample. In GTEx v8 whole blood samples, we extracted 912 genes with low total counts and calculated trcQTL estimates for variants in the corresponding cis-windows. To compare the power of trcQTL and eQTL, we used the 85,129 variant/gene pairs with FDR $<$ 0.05 in eQTLGen as a "ground truth" set. We also randomly selected 88,242 variant/gene pairs from the pairs with p-value $>$ 0.5 in eQTLGen as a negative set. **(A,B)** ROC and PR curves for trcQTL and the standard eQTL method. **(C)** Test statistics for the standard eQTL method (x-axis) and trcQTL (y-axis). The variant/gene pairs in the eQTLGen negative set are shown in the left panel, and pairs in the "ground truth" set in the right panel.

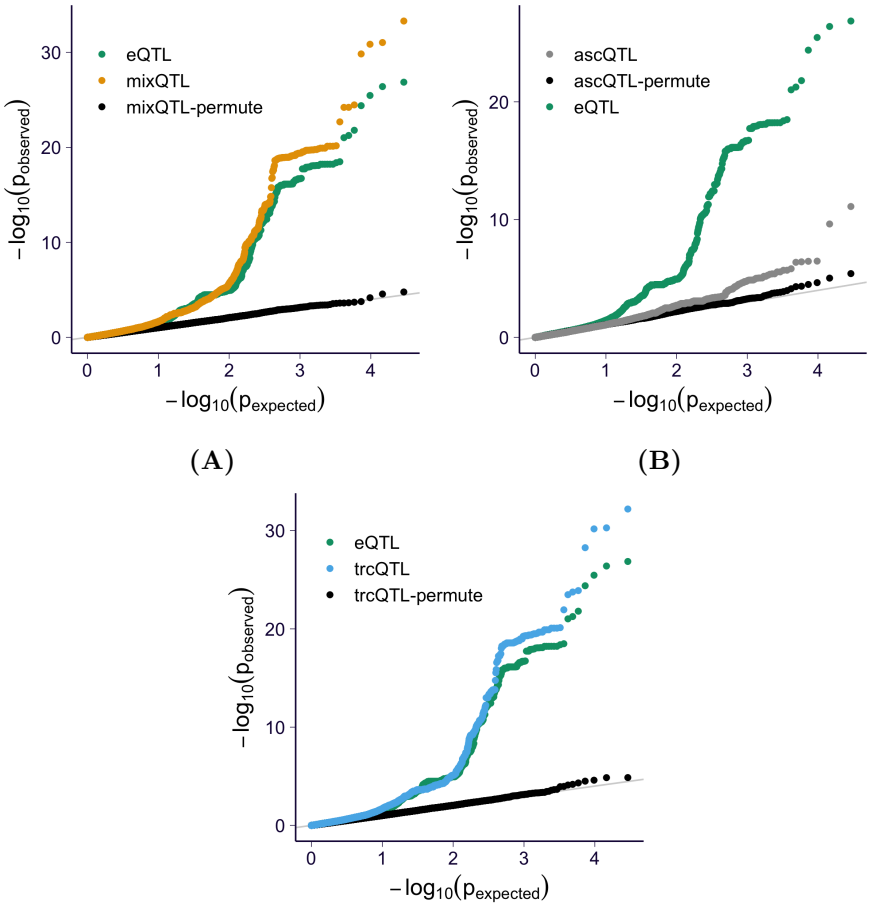



**Figure 2.15: QQ-plot of nominal p-values from ascQTL and trcQTL on four randomly selected genes in GTEx v8 whole blood RNA-seq.** The nominal p-values of trcQTL and ascQTL are compared against the standard eQTL method for four randomly selected genes ENSG00000000457, ENSG00000001461, ENSG00000002834, and ENSG00000277734. The results of ascQTL and trcQTL on permuted genotypes are shown in black. **(A)** Results from mixQTL. **(B)** Results from ascQTL. **(C)** Results from trcQTL.

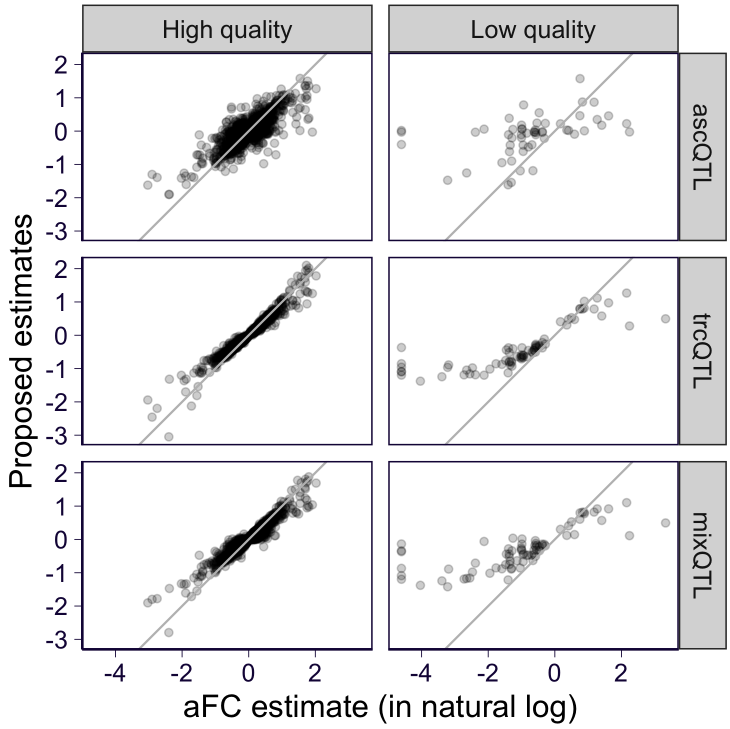
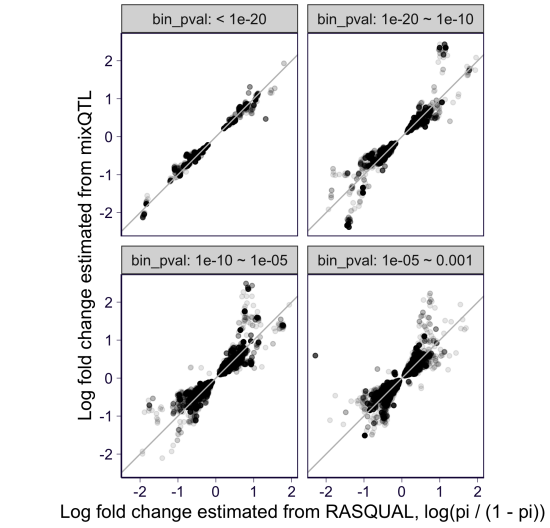
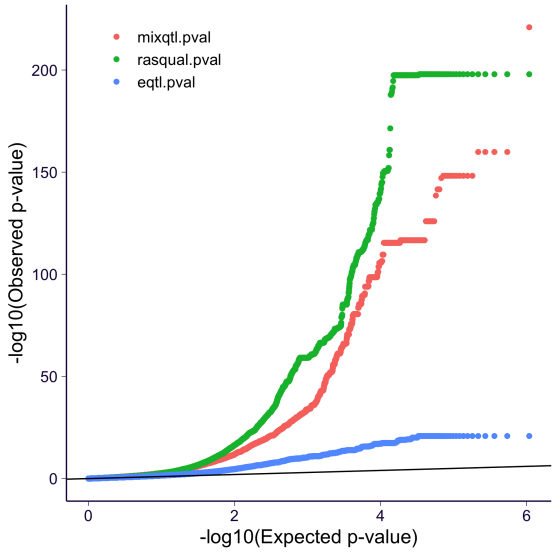



**Figure 2.16: Comparison of aFC estimates from GTEx v8 and the estimated allelic fold change of ascQTL, trcQTL, and mixQTL.** The estimates of the top variants in the eGenes of GTEx v8 whole blood are shown (based on eQTL results). On the x-axis, the aFC estimate reported by GTEx v8 is shown (the reported value is in $\log_2$ and, for visualization, we rescale it to natural log scale by multiplying the value with log(2)). On the y-axis, the estimated allelic fold changes (in natural log scale) of ascQTL, trcQTL, and mixQTL are shown. The variant/gene pairs are stratified on the basis of the quality of aFC estimate, which is defined as 'high quality' if the 95% confidence interval of $\log_2$ aFC is smaller than 1 and the low and high boundaries of the 95% confidence interval are not more extreme than $-\log_2(50)$ and $\log_2(50)$, and as 'low quality' otherwise.




**(A)**




**(B)**

**Figure 2.17: The performance of RASQUAL in GTEx v8 kidney cortex RNA-seq.** Here we show the results on kidney cortex for the gene/variants pairs within ±50kbp of the transcription start side. We tested the gene with enough allele-specific counts. Specifically, we include genes that have more than 100 reads (total count) in at least 80% of the samples and 50 allele-specific reads per haplotypes (both haplotypes should meet the criterion) in at least 15 samples. With these criteria, 4,596 genes are included. **(A)** The estimated effect sizes (in terms of log fold change) of both RASQUAL (on x-axis) and mixQTL (on y-axis). For RASQUAL, the log fold change is calculated from RASQUAL parameter $\pi$ using the relation that log fold change $= \log \frac{\pi}{1-\pi}$. The plot includes variant/gene pairs that both RASQUAL and mixQTL p-values pass some cutoffs (as stratified in the different panels). The concordance is similar across different minor allele frequencies. **(B)** QQ-plot of all the variant/gene pairs being tested.

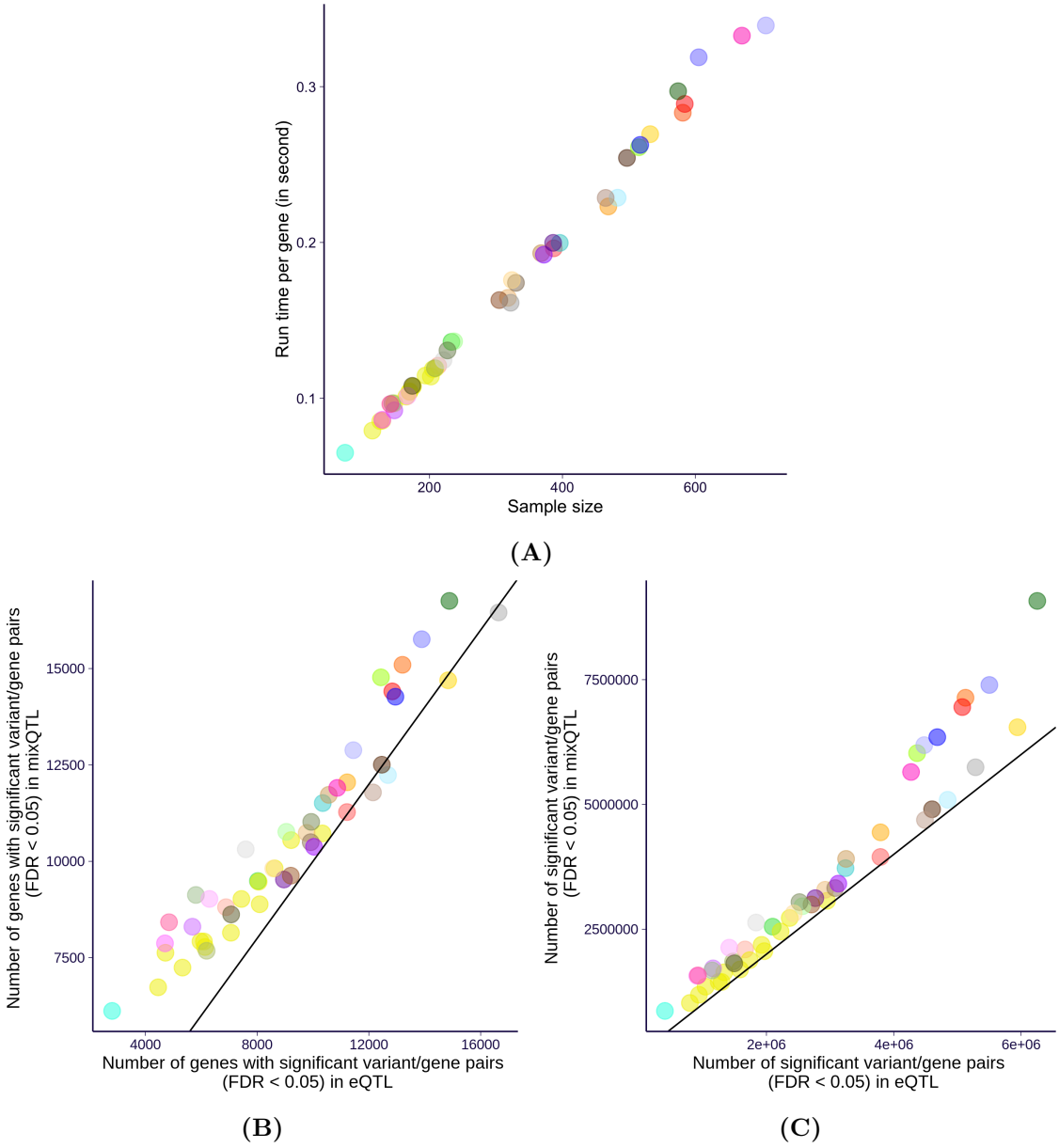



**Figure 2.18: Running mixQTL on the full GTEx v8 data. (A)** The average runtime (clock time under 8 CPU cores) per gene is shown for each of the 49 tissues (y-axis) against the corresponding sample size (x-axis). **(B)** The number of genes that have at least one variant passing FDR control at 0.05 is shown for both mixQTL (y-axis) and the standard approach (x-axis). In the GTEx v8 main eQTL analysis, "eGene" was defined based on permutation-based analysis. Here we do not perform permutation so, to avoid confusion, we do not use the term "eGene". **(C)** The number of variant/gene pairs that pass FDR control at 0.05 is shown for both mixQTL (y-axis) and the standard approach (x-axis).

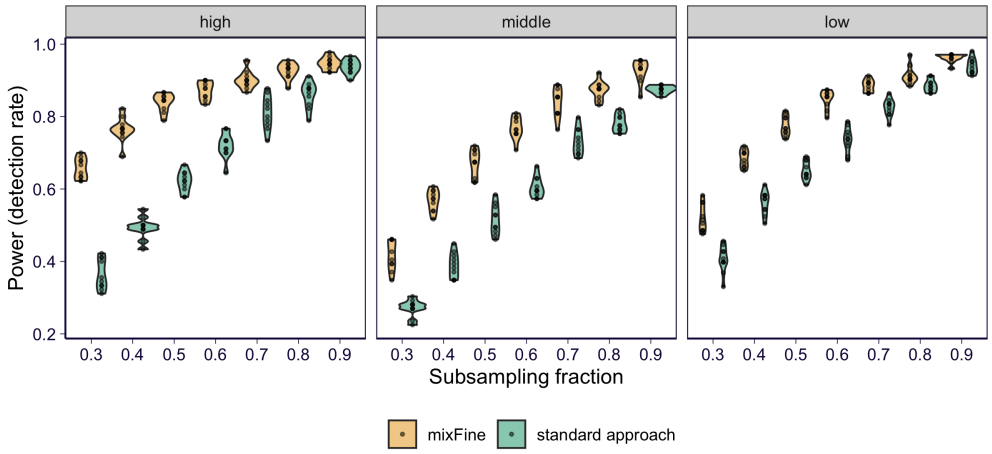



**Figure 2.19: The performance of mixFine on GTEx v8 whole blood RNA-seq stratified by expression level.** At each subsampling level (x-axis), the fraction of "consensus SNPs" being detected is shown on the y-axis. Each panel shows the results of genes stratified by expression level tertiles in which the fraction is calculated within each expression level category. Among the 272 "consensus SNPs", 90 belong to "high" expression level, 89 belong to "middle" level, and 103 belong to "low" level. The subsampling analysis are repeated 10 times. The plot of each panel shows the results of all the ten replications.

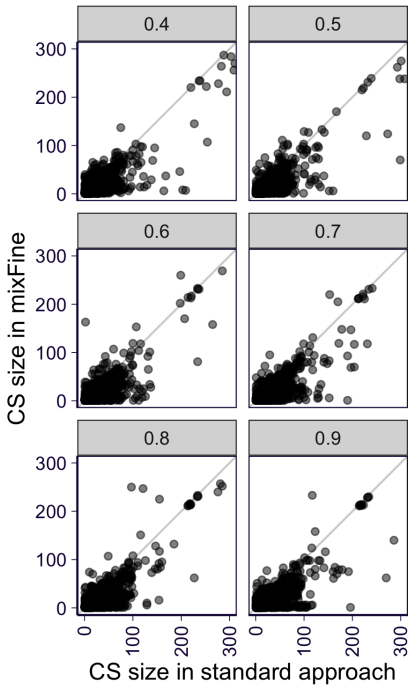



**Figure 2.20: The performance of mixFine on GTEx v8 whole blood RNA-seq on pinpointing the "top" SNPs.** At each subsampling level (shown in each panel), we compare mixFine (y-axis) and the standard method (x-axis) on the size of 95% CS's which are paired by sharing the same "top SNP".

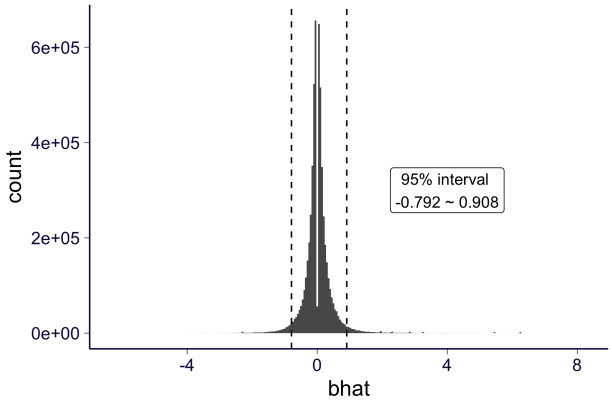



**Figure 2.21: The estimated cis-eQTL effect size in GTEx v8 whole blood.** We examined the estimated effect sizes by mixQTL (in GTEx v8 whole blood) among the variant/gene pairs with FDR $< 0.05$. The 95% intervals (2.5% quantile to 97.5% quantile) of the estimated effect size are shown. Note that the estimated effect size (x-axis) is defined as allelic fold change in log-scale.

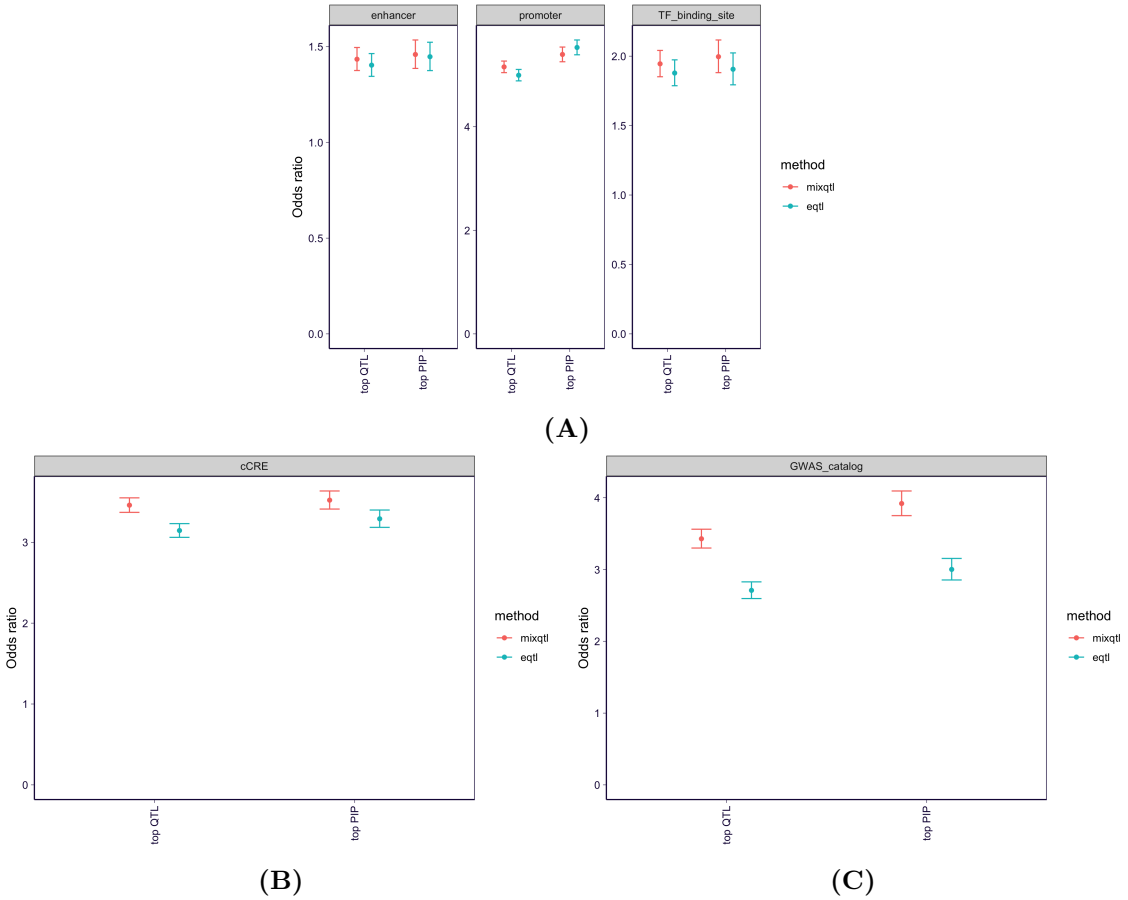



**Figure 2.22: Enrichment in functional annotation for GTEx v8 tissues.** The enrichment is measured by odds ratio which is based on the 2-by-2 table indicating if the variant is in the annotation and if the variant is the top signal within a gene according to mixQTL or mixFine. The result is calculated by aggregating across 26 GTEx v8 tissues which have sample size $< 260$ and 221,920,351 tissue-gene-variant tuples are considered in total. The error bar indicates the 95% confidence interval. The enrichment is examined among all genes with enough allele-specific counts. **(A)** The enrichment of top mixQTL and mixFine signal in regulatory element annotations curated by GTEx v8 paper [139]. **(B)** The enrichment of top mixQTL and mixFine signal in candidate cis-regulatory elements (cCREs) [104] where only 10 of the 26 tissues are included due to the lack of matched tissue in cCRE data. In total, 85,170,905 tissue-gene-variant tuples are considered. **(C)** The enrichment of top mixQTL and mixFine signal in GWAS catalog variants.

| nfold | sample_size | pairwise_diff | diff_ci95_low | diff_ci95_high | pval | median_mixpred | median_standard |
|---|---|---|---|---|---|---|---|
| 10 | 67 | 0.052 | 0.047 | 0.057 | 1.318e-72 | 0.175 | 0.070 |
| 9 | 75 | 0.050 | 0.044 | 0.055 | 4.828e-58 | 0.185 | 0.079 |
| 8 | 84 | 0.049 | 0.044 | 0.054 | 4.569e-63 | 0.198 | 0.100 |
| 7 | 96 | 0.047 | 0.042 | 0.053 | 1.350e-57 | 0.214 | 0.119 |
| 6 | 112 | 0.043 | 0.038 | 0.049 | 4.884e-53 | 0.228 | 0.152 |
| 5 | 134 | 0.036 | 0.031 | 0.041 | 1.483e-39 | 0.241 | 0.195 |
| 4 | 168 | 0.028 | 0.023 | 0.032 | 1.791e-27 | 0.251 | 0.219 |
| 3 | 224 | 0.017 | 0.012 | 0.021 | 2.535e-12 | 0.266 | 0.254 |
| 2 | 335 | 0.007 | 0.002 | 0.011 | 3.354e-03 | 0.292 | 0.287 |

**Table 2.2: The pairwise comparison of the prediction performance between mixPred and the standard approach based on the cross-validated evaluation.** The GTEx v8 whole blood data (sample size = 670) is split into $k$ folds. To evaluate the prediction performance, we train a model using one fold of the data and measure the performance on the held-out $(k-1)$ folds. This routine is applied to 1,000 genes and, for each gene, it is repeatedly $k$ times going through each of the $k$ folds. The prediction performance is measured by Pearson correlation. The **nfold** columns shows the number of folds, and, correspondingly, the **sample_size** column shows the number of samples used for training. The **pairwise_diff** column shows the average pairwise difference (mixPred vs. the standard approach) of the prediction performance among all folds and genes. And the **diff_ci95_low** and **diff_ci95_high** columns show the lower and upper bounds of the 95% confidence interval of the pairwise difference. The **pval** shows the p-value of the pairwise difference under paired t test (two-sided). The median of the prediction performance among all folds and genes are shown in the **median_mixpred** and **median_standard** columns for mixPred and the standard approach respectively.

## 2.9 Supplementary Notes

### *2.9.1 Statistical model for read count*

Here we introduce the statistical model of read count in this paper. For completeness, we opt for keeping some text that overlaps with main text. Recall that $i$ indexes individual and $h$ indexes haplotypes. $X_i^h$ is the phased genotype of the corresponding individual $i$ haplotype $h$. $Y_i^{\text{total}}$ is the total read count within the gene body and $L_i$ is the library size. $Y_i^{(h)\text{obs}}$ is the allele-specific read count of the corresponding haplotype transcript $h$ and $Y_i^h$ is the actual (though unobserved) read count of the haplotype transcript $h$. $\alpha_i$ is the expected fraction of allele-specific reads in individual $i$. Additionally, the cis-genetic effect of a single

SNP on haplotype $h$ is represented as $g(\beta, X_i^h)$ where

$$g(\beta, X_i^h) = \begin{cases} 1 & \text{, if } X_i^h = 0 \\ e^\beta & \text{, if } X_i^h = 1 \end{cases} \tag{2.14}$$

$$= e^{X_i^h \beta} \tag{2.15}$$

We assume multiplicative effect when there are multiple causal SNPs. And the effect of multiple SNPs $j = 1, \cdots, p$ is

$$\prod_{j=1}^{p} g(\beta_j, X_{ij}^h) = e^{\sum_j X_{ij}^h \beta_j} \tag{2.16}$$

$$= e^{\mathbf{X}_i^h \boldsymbol{\beta}} \tag{2.17}$$

$$:= g(\boldsymbol{\beta}, \mathbf{X}_i^h) \tag{2.18}$$

#### 2.9.1.1 Overview

We model haplotypic count $Y_i^h$ as lognormal distribution as follow.

$$\log Y_i^h \sim N(\log(L_i \theta_i^h), \tau_i^h) \tag{2.19}$$

$$\theta_i^h = \theta_{0,i} \times g(\boldsymbol{\beta}, \mathbf{X}_i^h), \tag{2.20}$$

$\theta_{0,i}$ is the baseline abundance of haplotype transcript without considering genetic effect (*i.e.* it represents the abundance when the affecting SNP is reference allele).

In practice, we do not observe $Y_i^h$ but allele-specific read count $Y_i^{(h)\text{obs}}$. So, we further assume that the baseline abundance of corresponding allele-specific reads are $\theta_{0,i}^{(1)} = \theta_{0,i}^{(2)} = \alpha_i \theta_{0,i}$. And by definition, total read count $Y_i^{\text{total}} = Y_i^1 + Y_i^2$. So, similar to Eq 2.19, 2.20,

$Y_i^{(h)\text{obs}}$ and $Y_i^{\text{total}}$ follow

$$\log Y_i^{(h)\text{obs}} \sim N(\log(L_i\theta_i^{(h)}), \tau_i^{(h)}) \tag{2.21}$$

$$\log Y_i^{\text{total}} \sim N(\log(L_i\theta_i), \tau_i) \tag{2.22}$$

$$\theta_i^{(h)} = \alpha_i\theta_{0,i} \times g(\beta, \mathbf{X}_i^h) \tag{2.23}$$

$$\theta_i = \theta_{0,i} \times [g(\beta, \mathbf{X}_i^1) + g(\beta, \mathbf{X}_i^2)] \tag{2.24}$$

#### 2.9.1.2 Parameterizing $\tau$ to weight total and AS count properly

Note that lognormal distribution has the following property.

$$\log X \sim N(\mu, \tau) \tag{2.25}$$

$$X \sim \text{lognormal}(\mu, \tau) \text{ , by definition of lognormal} \tag{2.26}$$

$$\text{E}(X) = e^{\mu+\frac{\tau}{2}} \tag{2.27}$$

$$\text{Var}(X) = (e^{\tau} - 1)(e^{2\mu+\tau}) \tag{2.28}$$

When modeling read count, given the mean, we would like the variance to scale linearly with the mean (as assumed in RASQUAL [70]). In other word, we want to ensure that $\text{Var}(X)/\text{E}(X)$, also known as over-dispersion parameter, is roughly a constant. From Eq 2.27, 2.28 we have $\text{Var}(X) = (e^{\tau} - 1)\text{E}(X)^2$. For count data, since $\tau$ is capturing the variation of count in log-scale, $\tau$ is typically close to 0. So $e^{\tau} - 1 \approx \tau$ and $\text{Var}(X) \approx \tau\text{E}(X)^2$. This result suggests that to ensure $\text{Var}(X)/\text{E}(X) =$ constant, $\tau$ should be approximately proportional to $1/\text{E}(X)$. So, for the distribution of $Y \sim \text{lognormal}(\log(L\theta), \tau)$, we impose the constraint on $\tau$ such that $\tau \approx \sigma^2/\text{E}(Y)$. In practice, $\text{E}(Y)$ is unknown so that we plug-in $Y$ in replace of $\text{E}(Y)$.

### *2.9.2 Single-SNP model*

On the basis of the model described in Section 2.9.1.1, we propose the single-SNP model where we focus on one "test SNP" $X_i^h$ instead of the whole phased haplotype $\mathbf{X}_i^h$. Hence, the cis-genetic effect of interest is $g(\beta, X_i^h)$.

#### 2.9.2.1 From likelihood to linear mixed model

Here, we model cis-genetic effect of test SNP as allelic fold change (aFC) [101]. So $\beta$ is log-scale aFC in $g(\beta, X_i^{(h)}) = e^{X_i^{(h)}\beta}$. From Eq 2.21, 2.23, we have (for $h = 1, 2$)

$$\log Y_i^{(h)\text{obs}} = \log L_i + \log \theta_i^{(h)} + \epsilon_i^{(h)} \tag{2.29}$$

$$= \log L_i + \log \alpha_i + \log \theta_i^h + \epsilon_i^{(h)} \tag{2.30}$$

$$= \log L_i + \log \alpha_i + \log \theta_{0,i} + \log(e^{X_i^h\beta}) + \epsilon_i^{(h)} \tag{2.31}$$

$$= \log L_i + \log \alpha_i + \log \theta_{0,i} + X_i^h\beta + \epsilon_i^{(h)} \tag{2.32}$$

$$\epsilon_i^{(h)} \sim N(0, \frac{\sigma^2}{Y_i^{(h)}}), \tag{2.33}$$

where the error term scaling in Eq 2.33 follows from the discussion in Section 2.9.1.2. To further simplify the term $\log \theta_{0,i}$, as the variation of baseline abundance among individuals, we assume $\log \theta_{0,i} \sim N(\mu_0, \sigma_0^2)$. So that Eq 2.32, 2.33 can be further written as

$$\log Y_i^{(h)\text{obs}} = \mu_0 + \log L_i + \log \alpha_i + z_i + X_i^h\beta + \epsilon_i^{(h)} \tag{2.34}$$

$$\epsilon_i^{(h)} \sim N(0, \frac{\sigma^2}{Y_i^{(h)\text{obs}}}),\ z_i \sim N(0, \sigma_0^2), \tag{2.35}$$

which is the approximated likelihood function for allele-specific counts $Y_i^{(1)\text{obs}}$ and $Y_i^{(2)\text{obs}}$. Such likelihood function is equivalent to linear mixed effects model.

Furthermore, we can linearize the likelihood of total read count $Y_i^{\text{total}}$ in similar fashion. From Eq 2.22, 2.24 , we have

$$\log Y_i^{\text{total}} = \mu_0 + \log L_i + z_i + \log(\theta_i^1 + \theta_i^2) + \epsilon_i \tag{2.36}$$

$$= \mu_0 + \log L_i + z_i + \log(e^{X_i^1\beta} + e^{X_i^2\beta}) + \epsilon_i \tag{2.37}$$

$$\epsilon_i \sim N(0, \frac{\sigma^2}{Y_i^{\text{total}}}),\ z_i \sim N(0, \sigma_0^2) \tag{2.38}$$

Here we linearize $\log(e^{X_i^1\beta} + e^{X_i^2\beta})$ under the weak-effect assumption as follow

$$\log(e^{X_i^1\beta} + e^{X_i^2\beta}) = \log[(X_i^1 e^{\beta} + 1 - X_i^1) + (X_i^2 e^{\beta} + 1 - X_i^2)] \tag{2.39}$$

$$= \log(2 + X_i e^{\beta} - X_i) \quad \text{, let } X_i = X_i^1 + X_i^2 \tag{2.40}$$

$$= \log[2 + X_i(e^{\beta} - 1)] \tag{2.41}$$

$$= \log 2 + \frac{1}{2}(e^{\beta} - 1)X_i + o(X_i(e^{\beta} - 1)) \tag{2.42}$$

$$\approx \log 2 + \frac{1}{2}X_i\beta \quad \text{, when } \beta \text{ is close to 0} \tag{2.43}$$

So that Eq 2.37 can be approximated as

$$\log \frac{Y_i^{\text{total}}}{2} \approx \mu_0 + \log L_i + z_i + \frac{X_i^1 + X_i^2}{2}\beta + \epsilon_i \tag{2.44}$$

In summary, combining Eq 2.34 ,2.38, 2.35, 2.44, we have a linear mixed effects model unifying total and allele-specific read counts after linearization along with other approximations. And it also serves as an approximated likelihood for total and allele-specific reads, in which we can see that these read counts are not independent since they share the same random effect $z_i$.

#### 2.9.2.2 Simplifying the model

Note that $\alpha_i$ is not observed so that we are unable to solve the model proposed in Section 2.9.2.1 in a computationally efficient manner. Here we address this problem by re-parameterizing the model. In principle, conditioning on genetic effect $\beta$, the ratio of allele-specific reads should be independent to the observations on the total read counts. This intuition motivates us to model the ratio of $Y_i^{(1)\text{obs}}$ and $Y_i^{(2)\text{obs}}$ rather than each of them separately. Mathematically, we subtract $\log Y_i^{(2)\text{obs}}$ from $\log Y_i^{(1)\text{obs}}$, which gives

$$\log \frac{Y_i^{(1)\text{obs}}}{Y_i^{(2)\text{obs}}} = (X_i^1 - X_i^2)\beta + \epsilon_i^{\text{asc}} \tag{2.45}$$

$$\epsilon_i^{\text{asc}} \sim N(0, \sigma^2(\frac{1}{Y_i^{(1)\text{obs}}} + \frac{1}{Y_i^{(2)\text{obs}}})), \tag{2.46}$$

where both $z_i$ and $\alpha_i$ cancel out. This result naturally shows that the likelihood function of $Y_i^{\text{total}}$ and $\frac{Y_i^{(1)\text{obs}}}{Y_i^{(2)\text{obs}}}$ takes the form:

$$\mathcal{L}(\mathbf{Y}^{\text{total}}, \frac{\mathbf{Y}^{(1)\text{obs}}}{\mathbf{Y}^{(2)\text{obs}}}; \mu_0, \sigma_0^2, \sigma^2, \beta) = \prod_i \Pr(Y_i^{\text{total}}|\mu_0, \sigma_0^2, \sigma^2, \beta) \Pr(\frac{Y_i^{(1)\text{obs}}}{Y_i^{(2)\text{obs}}}|\sigma^2, \beta) \tag{2.47}$$

$$= \underbrace{\prod_i \Pr(Y_i^{\text{total}}|\mu_0, \sigma_0^2, \sigma^2, \beta)}_{\text{total read count likelihood}} \underbrace{\prod_i \Pr(\frac{Y_i^{(1)\text{obs}}}{Y_i^{(2)\text{obs}}}|\sigma^2, \beta)}_{\text{allele-specific read count likelihood}} \tag{2.48}$$

$$:= \mathcal{L}^{\text{trc}}(\mathbf{Y}^{\text{total}}) \times \mathcal{L}^{\text{asc}}(\frac{\mathbf{Y}^{(1)\text{obs}}}{\mathbf{Y}^{(2)\text{obs}}}) \tag{2.49}$$

With the simplification shown in Eq 2.45, the model used for inference can be summarized as follow

$$\log \frac{Y_i^{\text{total}}}{2L_i} = \mu_0 + z_i + \frac{X_i^1 + X_i^2}{2}\beta \quad + \epsilon_i^{\text{trc}} \tag{2.50}$$

$$\log \frac{Y_i^{(1)\text{obs}}}{Y_i^{(2)\text{obs}}} = \qquad\qquad (X_i^1 - X_i^2)\beta + \epsilon_i^{\text{asc}} \tag{2.51}$$

$$z_i \sim N(0, \sigma_0^2),\ \epsilon_i^{\text{trc}} \sim N(0, \frac{\sigma^2}{Y_i^{\text{total}}}),\ \epsilon_i^{\text{asc}} \sim N(0, \frac{\sigma^2 Y_i^{(1)\text{obs}} + Y_i^{(2)\text{obs}}}{Y_i^{(1)\text{obs}} Y_i^{(2)\text{obs}}}) \tag{2.52}$$

### *2.9.3 Generalizing to multi-SNP model*

The linearized model described in Eq 2.50, 2.51, 2.52 is easily extensible to multi-SNP scenario since we assume multiplicative genetic effect, as described in Section 2.18. To see the extension, all we need to examine is how $\log \theta_i^h$ and $\log(\theta_i^1 + \theta_i^2)$ as compared to the single

SNP case since the rest of the terms stay the same.

$$\log \theta_i^h = \log \theta_{0,i} + \log g(\beta, \mathbf{X}_i^h) \tag{2.53}$$

$$= \log \theta_{0,i} + \log e^{\mathbf{X}_i^h \beta} \tag{2.54}$$

$$= \log \theta_{0,i} + \mathbf{X}_i^h \beta \tag{2.55}$$

$$\log(\theta_i^1 + \theta_i^2) = \log \theta_{0,i} + \log\{\prod_j [1 + (e^{\beta_j} - 1)X_{ij}^1] + \prod_j [1 + (e^{\beta_j} - 1)X_{ij}^2]\}, \tag{2.56}$$

$$\text{similar to Eq 2.39} \tag{2.57}$$

$$\approx \log \theta_{0,i} + \log[1 + \sum_j (e^{\beta_j} - 1)X_{ij}^1 + 1 + \sum_j (e^{\beta_j} - 1)X_{ij}^2], \tag{2.58}$$

$$\text{high orders term like } (e^{\beta_j} - 1)X_{ij}^1(e^{\beta_{j'}} - 1)X_{ij'}^1 \text{ are ignored} \tag{2.59}$$

$$= \log \theta_{0,i} + \log(2 + \sum_j (e^{\beta_j} - 1)X_{ij}) \ , X_{ij} := X_{ij}^1 + X_{ij}^2 \tag{2.60}$$

$$\approx \log \theta_{0,i} + \log 2 + \frac{1}{2}\mathbf{X}_i \beta \ , \text{follows similarly as Eq 2.42, 2.43} \tag{2.61}$$

So, we can simply plug-in the multi-SNP version of $\log \theta_i^h$ and $\log(\theta_i^1 + \theta_i^2)$ to Eq 2.30 and 2.36 respectively and the similar conclusion follows with $\mathbf{X}$ and $\beta$ in replace of $X$ and $\beta$.

### *2.9.4 QTL mapping procedure*

In the following, we describe the mixQTL procedure to map cis-eQTLs under the model proposed in Eq 2.50, 2.51, 2.52.

#### 2.9.4.1 Converting the problems into two linear regressions

Instead of solving the proposed mixed effects model using numerical solver, we propose a meta-analysis procedure. In this procedure, we solve Eq 2.50 and 2.51 separately and meta-analyze the estimates afterwards.

Here we recognize that $\epsilon_i^{\text{trc}}$ in Eq 2.50 is approximate independent to $\epsilon_i^{\text{asc}}$ in Eq 2.51. The reason is that, under the model assumption, the read counts from each of the two haplotypes are independent (conditioning on $z_i$ and library size), which is also true in log-scale, *i.e.* $\epsilon^{(1)} \perp\!\!\!\perp \epsilon^{(2)}$. So, $\epsilon^{(1)} + \epsilon^{(2)} \perp\!\!\!\perp \epsilon^{(1)} - \epsilon^{(2)}$, which means that the sum of logarithm of the haplotypic counts, $\log Y_i^1 + \log Y_i^2$, is independent to the haplotypic imbalance signal, $\log Y_i^1/Y_i^2$. Furthermore, under the weak effect size assumption, $\log Y_i^1 + \log Y_i^2 \approx \log Y_i^{\text{total}}$ so that $\epsilon_i^{\text{trc}}$ is approximately independent to $\epsilon_i^{\text{asc}}$. Besides, $z_i$ represents baseline abundance, which is independent of the multiplicative errors $\epsilon_i^{\text{trc}}$ and $\epsilon_i^{\text{asc}}$. So, we can further simplify Eq 2.50 by merging the noise term $\epsilon_i^{\text{trc}}$ and $z_i$ as a new term $\tilde{z}_i$. Such simplification results in the following linear model

$$Y_i^{\text{trc}} = \mu_0 + X_i^{\text{trc}}\beta^{\text{trc}} + \tilde{z}_i,\ \tilde{z}_i \sim N(0, \tilde{\sigma}_0^2)\ , \tag{2.62}$$

where $X^{\text{trc}} := \frac{X^1+X^2}{2},\ Y^{\text{trc}} = \log \frac{Y_i^{\text{total}}}{2L_i}$. Eq 2.62 itself can be used for QTL mapping and we call this approach trcQTL in the paper.

For solving Eq 2.51, notice that it is weighted simple linear regression with the form

$$Y_i^{\text{asc}} = X_i^{\text{asc}}\beta^{\text{asc}} + \epsilon_i^{\text{asc}},\ \epsilon_i^{\text{asc}} \sim N(0, \sigma^2/w_i)\ , \tag{2.63}$$

where $Y_i^{\text{asc}} = \log \frac{Y_i^{(1)\text{obs}}}{Y_i^{(2)\text{obs}}},\ X_i^{\text{asc}} = X_i^1 - X_i^2,\ w_i = \frac{Y_i^{(1)\text{obs}} Y_i^{(2)\text{obs}}}{Y_i^{(1)\text{obs}} + Y_i^{(2)\text{obs}}}$. We call QTL mapped by Eq 2.63 ascQTL.

Note that we can combine Eq 2.62 and 2.63 and solve them jointly in close form. But here we still prefer meta-analysis for two reasons: 1) it allows combining summary statistics across studies; and 2) it allows the over-dispersion in total and allele-specific read counts to be different which is more realistic in practice since total and allele-specific read counts may go through different pre-processing steps.

Since the inference of linear regression has analytical solution which only involves $X^T X$

and $X^T Y$, we can solve it quickly and in a parallel way as proposed by Matrix eQTL [124]. We sketch the pseudocode on calculating trcQTL and ascQTL estimates in matrix form in Section 2.9.7.

#### 2.9.4.2 Meta-analysis for QTL mapping

Once we obtain estimated $\widehat{\beta}^{\text{trc}}$ and $\widehat{\beta}^{\text{asc}}$, we can use these estimates to approximate $\mathcal{L}^{\text{trc}}$ and $\mathcal{L}^{\text{asc}}$ in Eq 2.49. Specifically, when sample size is large,

$$\mathcal{L}^{\text{trc}}(Y_i^{\text{total}}|\beta) \approx N(\beta; \widehat{\beta}^{\text{trc}}, \text{se}(\widehat{\beta}^{\text{trc}})) \tag{2.64}$$

$$\mathcal{L}^{\text{asc}}(\frac{Y_i^{(1)\text{obs}}}{Y_i^{(2)\text{obs}}}|\beta) \approx N(\beta; \widehat{\beta}^{\text{asc}}, \text{se}(\widehat{\beta}^{\text{asc}})) \tag{2.65}$$

So that the joint likelihood, as factorized in Eq 2.48, is simply $N(\beta; \widehat{\beta}^{\text{trc}}, \text{se}(\widehat{\beta}^{\text{trc}})) \times N(\beta; \widehat{\beta}^{\text{asc}}, \text{se}(\widehat{\beta}^{\text{asc}}))$. As shown previously [73], maximizing the approximate joint likelihood is equivalent to inverse-variance meta-analysis, which takes the form

$$\widehat{\beta}^{\text{mix}} = \frac{w^{\text{trc}}\widehat{\beta}^{\text{trc}} + w^{\text{asc}}\widehat{\beta}^{\text{asc}}}{w^{\text{trc}} + w^{\text{asc}}} \tag{2.66}$$

$$\text{se}(\widehat{\beta}^{\text{mix}}) = \sqrt{\frac{1}{w^{\text{trc}} + w^{\text{asc}}}}\ , \tag{2.67}$$

where $w^{\text{trc}} = 1/\text{se}(\widehat{\beta}^{\text{trc}})^2$ and $w^{\text{asc}} = 1/\text{se}(\widehat{\beta}^{\text{asc}})^2$.

### *2.9.5 Inference procedure for multi-SNP model*

With the simplification made in Section 2.9.4.1, the multi-SNP model can be written as

$$Y_i^{\text{trc}} = \mu_0 + \mathbf{X}_i^{\text{trc}}\boldsymbol{\beta} + \tilde{z}_i \quad , \tilde{z}_i \quad \sim N(0, \tilde{\sigma}_0^2) \tag{2.68}$$

$$Y_i^{\text{asc}} = \qquad \mathbf{X}_i^{\text{asc}}\boldsymbol{\beta} + \epsilon_i^{\text{asc}}, \epsilon^{\text{asc}} \sim N(0, \sigma^2/w_i)\ . \tag{2.69}$$

#### 2.9.5.1 Motivating two-step inference procedure

Here we focus on two inference problems under the multi-SNP model: 1) construct genetic predictor of expression; and 2) infer whether $\beta_k$ is non-zero, *i.e.* causal SNP. Problem 1) is prediction problem in machine learning context and in terms of building genetic predictor, elastic net has been used for this task as implemented in the PrediXcan method[43]. For problem 2), the inference problem is formulated into a Bayesian variable selection problem and efficient solvers such as susieR [149] and DAP-G [76] have been developed in the context of eQTL analysis.

However, the existing methods only use total read information (typically inverse normalized expression) and they assume the inversely normalized expression $Y$ and genotype vector $\mathbf{X}$ follow $Y \sim N(\mathbf{X}\beta, \nu)$. The modeling assumption is very close to Eq 2.68, 2.69 but it requires equal variance in error term and shared intercept across all observations. To apply the existing tools, we need to bypass the gap between our model and their modeling assumption. For this reason, we propose a two-step inference procedure to perform inference for multi-SNP model. In step 1, we infer $\widetilde{\sigma}_0^2$ and $\sigma^2$ and transform the data such that they approximately follow $Y \sim N(\mathbf{X}\beta, \nu)$. And in step 2, we apply the transformed data to existing solvers for both prediction and fine-mapping problems.

#### 2.9.5.2 Inferring $\widetilde{\sigma}_0^2$ and $\sigma^2$

To estimate $\widetilde{\sigma}_0^2$ and $\sigma^2$ from Eq 2.68 and Eq 2.69, we further assume that the genetic effects $\beta_1, \cdots, \beta_P$ (for all the SNPs within the cis-window) follow $\beta_p \sim_{iid} N(0, V_g)$. Or equivalently, we assume

$$Y^{\text{trc}} \sim N(\mu_0, \widetilde{\sigma}_0^2 I_N + V_g \mathbf{X}_i^{\text{trc}} (\mathbf{X}_i^{\text{trc}})') \tag{2.70}$$

$$Y^{\text{asc}} \sim N(0, \sigma^2 I_N + V_g \mathbf{X}_i^{\text{asc}} (\mathbf{X}_i^{\text{asc}})') \tag{2.71}$$

Under the mixed effect model Eq 2.70, we solve for $\tilde{\sigma}_0^2$ using total read count data. And similarly, under the random effect model Eq 2.71, we solve for $\sigma^2$ using allele-specific count data. The actual computation is done using R package EMMA [62].

#### 2.9.5.3 Data transformation and inference

Once we obtain $\hat{\tilde{\sigma}}_0^2$ and $\hat{\sigma}^2$, we shift and re-scale the total and allelic imbalance observations by

$$\widetilde{Y}_i^{\text{trc}} = \frac{\text{center}(Y_i^{\text{trc}})}{\hat{\tilde{\sigma}}_0}, \ \widetilde{\mathbf{X}}_i^{\text{trc}} = \frac{\text{center}(\mathbf{X}_i^{\text{trc}})}{\hat{\tilde{\sigma}}_0} \tag{2.72}$$

$$\widetilde{Y}_i^{\text{asc}} = \frac{Y_i^{\text{asc}}}{\hat{\sigma}}, \qquad \widetilde{\mathbf{X}}_i^{\text{asc}} = \frac{\mathbf{X}_i^{\text{asc}}}{\hat{\sigma}} \ , \tag{2.73}$$

where the function center$(\cdot)$ centers the input by subtracting the population mean (mean across all samples). By centering $Y_i^{\text{trc}}$ and $\mathbf{X}_i^{\text{trc}}$, effectively, we account for the term $\mu_0$ in Eq 2.68, which has been deployed previously by [124, 91]. And the transformed data (on the left-hand side) is used for downstream analysis on performing prediction and fine-mapping.

Specifically, we concatenate $\widetilde{\mathbf{Y}}^{\text{trc}}$ and $\widetilde{\mathbf{Y}}^{\text{asc}}$ into one vector $\mathbf{Y} \in \mathbb{R}^{(N^{\text{trc}}+N^{\text{asc}})\times 1}$ and similarly we concatenate $\widetilde{\mathbf{X}}^{\text{trc}}$ and $\widetilde{\mathbf{X}}^{\text{asc}}$ into one matrix $\mathbf{X} \in \mathbb{R}^{(N^{\text{trc}}+N^{\text{asc}})\times p}$ where $p$ is the number of SNPs. To perform fine-mapping, we run `susieR::susie(X = X, Y = Y, intercept = FALSE, standardize = FALSE)` with `X` equal to $\mathbf{X}$ and `Y` equal to $\mathbf{Y}$. To build prediction model, we run `glmnet::glmnet(x = X, y = Y, lambda = lambda, alpha = 0.5)` with `x` equal to $\mathbf{X}$ and `y` equal to $\mathbf{Y}$. The hyperparamter `lambda` is selected by 5-fold nested cross-validation where at each `lambda` the 5-fold cross-validation are repeated three times and `lambda` that has lowest cross-validated mean squared error (averaged across three runs) is used. For comparison, we feed the part of total read count data $(\mathbf{X}^{\text{trc}}, \mathbf{Y}^{\text{trc}})$ directly into: 1) susieR for fine-mapping; and 2) elastic net for prediction. The procedure is the same but $\mathbf{X}, \mathbf{Y}$ are replaced by $\mathbf{X}^{\text{trc}}, \mathbf{Y}^{\text{trc}}$. And we call this total read count-only approach for

fine-mapping and prediction as trcFine and trcPred.

### *2.9.6 Simulating RNA-seq reads*

To examine the performance of the methods, we propose and implement a simulation scheme which generates total and allele-specific read counts. The simulation procedure includes three parts: 1) simulate gene body which will be aligned by reads; 2) randomly draw the causal variants; 3) simulate the number of reads for each haplotype transcript and place these reads to the gene body obtained in step 1). The total and allele-specific read counts can be directly read out from step 3) where the total read count is the sum of two haplotypic read counts and the allele-specific read count is the number of reads overlapping with heterozygous sites within gene body.

In step 1), we fix the length of gene body to be 10kbp. To simulate the heterozygous sites within gene body for each individual, we start with determining the position of polymorphic sites along gene body. We first sample the number of polymorphic sites from Binomial distribution, and then draw their positions and minor allele frequencies (MAFs). And finally, whether a polymorphic site is heterozygous in an individual is determined by Bernoulli distribution with MAF. The procedure is sketched as follow.

1. Number of polymorphic site within gene body $N_h \sim \text{Binomial}(L_{\text{gene}}, f^h)$, where $L_{\text{gene}} = 10^4, f^h = 0.001$.

2. Position $P_m$ $(m = 1, \cdots, N_h)$ of these polymorphic sites are sampled by $P_m \sim \text{Sample}(\{1, \cdots, L_{\text{gene}}\})$. And the corresponding MAF $f_m$ are drawn from $f_m \sim \text{Uniform}(\text{maf}^l, \text{maf}^h)$, where $\text{maf}^l = 0.05, \text{maf}^h = 0.3$.

3. For each individual $i$, whether the $m$th polymorphic site is heterozygous (denote as $Z_{im}$) is determined by $Z_{im} \sim \text{Bernoulli}(2f_m(1 - f_m))$.

In step 2), the genetic effect equals to $e^{X_i^h \beta}$ (in single-SNP model) and $e^{\mathbf{X}_i^h \boldsymbol{\beta}}$ (in multi-SNP model). To do so, we need to obtain haplotype and effect size. For single-SNP model, we first sample MAF of the causal variants and obtain the two haplotypes of each individual by drawing from Bernoulli. For multi-SNP model, we use the 1000G phase3 genotypes of European individuals. In brief, we randomly select 200 genes on chromosome 22 and extract phased genotypes of 1Mbp cis-window surrounding the transcription start site of them (excluding variants with allele frequency $< 0.01$ or $> 0.99$). The genetic effect size, $e^{\beta}$, ranges among 1, 1.01, 1.05, 1.1, 1.25, 1.5, 2, 3 for single-SNP case. In multi-SNP case, the number of causal SNPs is sampled from 1, 2, 3 and the genetic effect ranges from 0.015 to 0.075 such that the heritability ranges approximately from 19.4% to 54.5%. The detailed procedure for sampling $e^{X^h \beta}$ and $e^{\mathbf{X}_i^h \boldsymbol{\beta}}$ is as follow.

- **Single-SNP scenario**:

    1. Sampling $X_i^h$: MAF of causal SNP $f^c \sim \text{Uniform}(\text{maf}^l, \text{maf}^h)$ and $X_i^h \sim \text{Bernoulli}(f^c)$ where $\text{maf}^l = 0.05, \text{maf}^h = 0.3$.
    2. Setting up $\beta$: fixed to 1, 1.01, ..., 2, 3.

- **Multi-SNP scenario**:

    1. Sampling $\mathbf{X}_i^h$: obtained from 1000G phased genotypes.
    2. Setting up $\boldsymbol{\beta}$: number of causal SNPs $\sim \text{Sample}(\{1, 2, 3\})$ and the genetic variation $v_g \sim \text{Uniform}(0.015, 0.075)$. The genetic effect of causal variants are determined by randomly partition the genetic variation and convert per-SNP genetic variation into effect size by $\beta_k = \sqrt{v_{g,k}/(2f_k(1-f_k))}$ where $f_k$ is MAF of $k$th causal SNP.

In the step 3), the last step, we sample the reads coming from each of the haplotype transcripts. The procedure is as follow.

1. For individual $i$, sample library size $L_i \sim$ NegativeBinomial(size, prob) where size $=$ $15, \text{prob} = 1.6 \times 10^{-7}$ (Negative Binomial follows parameterization in `rnbinom` in R).

2. And then, sample individual-specific baseline abundance $\theta_{0,i} \sim$ Beta where $\mathrm{E}(\theta_{0,i})$ ranges among $5\times10^{-5}, 2.5\times10^{-5}, 1\times10^{-5}, 5\times10^{-6}, 2.5\times10^{-6}, 1\times10^{-6}$ and $\mathrm{sd}(\theta_{0,i}) = \mathrm{E}(\theta_{0,i})/4$ (so that the non-genetic variation is roughly $1/4^2 = 1/16$).

3. The actual relative abundance of haplotype $h$ in individual $i$ is $\theta_i^h = \theta_{0,i} e^{X_i^h \beta}$ or $\theta_i^h = \theta_{0,i} e^{\mathbf{X}_i^h \beta}$

4. Sample actual read count for each haplotype: $Y_i^h \sim$ NegativeBinomial(size, prob) where size $= 2L_i\theta_i^h, \text{prob} = \frac{2}{3}$. This corresponds to $\mathrm{E}(Y_i^h) = L_i\theta_i^h$ and $\mathrm{Var}(Y_i^h) = \frac{3}{2}\mathrm{E}(Y_i^h)$.

5. Randomly place reads, $Y_i^h$ in total, onto the corresponding gene body simulated in step 1) where the read is aligned to each position of gene body with equal probability.

6. Total count is $Y_i^{\text{total}} = Y_i^1 + Y_i^2$ and allele-specific count $Y_i^{(h)\text{obs}}$ is the number of reads (as part of $Y_i^h$) that overlaps with the heterozygous sites of the individual (indicated by $Z_{i\cdot}$).

### *2.9.7 Pseudocode on solving trcQTL and ascQTL in matrix form*

We sketch the matrix operations for solving a grid of least squares problems $\mathbf{y}_k \sim \mathbf{x}_j$ for each pair of $j, k$ where we let $Y = [\mathbf{y}_1, \cdots, \mathbf{y}_K]$ and $X = [\mathbf{x}_1, \cdots, \mathbf{x}_n]$. To obtain nominal p-value, $K = 1$. For permutation procedure proposed in fastQTL [110], $K$ equals to the number of permutation and $\mathbf{y}_k$ is the $k$th permuted $\mathbf{y}$.

To ensure trcQTL and ascQTL ran on the same permuted $\mathbf{y}$, we perform permutation before removing low count observations. So that in each permutation, different individuals

are removed by low-count filter. To account for this fact, we introduce mask $M \in \{0,1\}^{n \times K}$ where $M_{ik}$ indicating if the $i$th individual is included in $k$th permutation.

For trcQTL, the corresponding least squares problem has intercept, as mentioned in Eq 2.62. The pseudocode to solve the grid of trcQTL problems for all cis-SNP of a gene is sketched in Algorithm 1 where $Y = \mathbf{Y}^{\text{trc}}$ for nominal pass and $Y_{\cdot k} = P_k \mathbf{Y}^{\text{trc}}$ with permutation matrix $P_k$ for permutation pass.

Note that the pseudocode only requires basic matrix operation. The matrix operation is element-wise if not notice explicitly. The Einstein summation is represented by `einsum` with similar arguments as `numpy.einsum` in Python. For instance, `einsum('ij,jk→ik', A, B)` means that to take the inner product of the $i$ row in A and $k$ column in B as the element at $i$th row and $j$th column in the output matrix.

Similar to trcQTL, the corresponding least squares problem of ascQTL is weighted without intercept, as mentioned in Eq 2.63. The pseudocode to solve the grid of ascQTL problems for all cis-SNP of a gene is sketched in Algorithm 2 where $Y = \mathbf{Y}^{\text{asc}}$ for nominal pass and $Y_{\cdot k} = P_k \mathbf{Y}^{\text{asc}}$ with permutation matrix $P_k$ for permutation pass. And $W$ as the weight matrix should be permutate accordingly, *i.e.* $W_{\cdot k} = P_k \mathbf{w}$. And to obtain valid mixQTL estimates under permutation, $P_k$ is required to be shared by trcQTL and ascQTL in permutation pass.

Note that both Algorithm 1 and Algorithm 2 are iteration free. And throughout the computation, only two-way tensors are involved explicitly so that the memory usage does not blow up.

**Algorithm 1:** Solve multiple least squares problems $y = a + bx + e$ in matrix form

**Input :** $Y \in \mathbb{R}^{n \times K},\ X \in \mathbb{R}^{n \times p},\ M \in \{0,1\}^{n \times K}$.

**Output:** $\widehat{A}, \widehat{B}, \mathrm{se}(\widehat{A}), \mathrm{se}(\widehat{B}) \in \mathbb{R}^{K \times p}$ where $\widehat{A}_{kj}, \widehat{B}_{kj}, \mathrm{se}(\widehat{A}_{kj}), \mathrm{se}(\widehat{B}_{kj})$ are estimates of $Y_{\cdot k} = A_{kj} + B_{kj} X_{\cdot j} + \epsilon$ where data is masked by $M_{\cdot k}$.

**Function** `SolveMatrixLSwithIntercept`$(Y, X, M)$**:**

- $U = \mathrm{matrix}(1,\ \mathrm{dim} = \mathrm{dim}(X))$;
- $n = \mathrm{einsum}(\text{`ik}\rightarrow\text{k'},\ M)$;
- $Y = YM$;
- $T_1 = \mathrm{einsum}(\text{`ij,ik}\rightarrow\text{jk'},\ X,\ Y)$;
- $T_2 = \mathrm{einsum}(\text{`ij,ik}\rightarrow\text{jk'},\ U,\ Y)$;
- $S_{11} = X^2$;
- $S_{11} = \mathrm{einsum}(\text{`ij,ik}\rightarrow\text{jk'},\ S_{11},\ M)$;
- $S_{22} = U^2$;
- $S_{22} = \mathrm{einsum}(\text{`ij,ik}\rightarrow\text{jk'},\ S_{22},\ M)$;
- $S_{12} = XU$;
- $S_{12} = \mathrm{einsum}(\text{`ij,ik}\rightarrow\text{jk'},\ S_{12},\ M)$;
- $\Delta = |S_{11}S_{22} - S_{12}S_{12}|$;
- $\widehat{B} = (S_{22}T_1 - S_{12}T_2)/\Delta$;
- $\widehat{A} = (S_{11}T_2 - S_{12}T_1)/\Delta$;
- $Y_{sq} = \mathrm{einsum}(\text{`ik,ik}\rightarrow\text{k'},\ Y,\ Y)$;
- $R_{sq} = Y_{sq} - 2\widehat{B}T_1 - 2\widehat{A}T_2 + 2\widehat{B}\widehat{A}S_{12} + \widehat{B}^2 S_{11} + \widehat{A}^2 S_{22}$;
- $\widehat{\sigma} = \sqrt{R_{sq}/(n-2)}$;
- $\mathrm{se}(\widehat{B}) = \widehat{\sigma}\sqrt{S_{22}/\Delta}$;
- $\mathrm{se}(\widehat{A}) = \widehat{\sigma}\sqrt{S_{11}/\Delta}$;
- **return** $\widehat{A}, \widehat{B}, \mathrm{se}(\widehat{A}), \mathrm{se}(\widehat{B})$

**End**

**Algorithm 2:** Solve multiple least squares problems $y = bx + e$ with weight $w$ in matrix form

**Input :** $Y \in \mathbb{R}^{n\times K},\ X \in \mathbb{R}^{n\times p},\ M \in \{0,1\}^{n\times K},\ W \in \mathbb{R}_{+}^{n\times K}$.

**Output:** $\widehat{B} \in \mathbb{R}^{K\times p}$ and $\mathrm{se}(\widehat{B}) \in \mathbb{R}^{K\times p}$ where $\widehat{B}_{kj}, \mathrm{se}(\widehat{B}_{kj})$ are estimates of $Y_{\cdot k} = B_{kj}X_{\cdot j} + \epsilon$ where data is weighted by $W_{\cdot k}$ and masked by $M_{\cdot k}$.

**Function** `SolveMatrixLSwithWeight`$(Y, X, M, W)$**:**

- $n = \mathrm{einsum}(\text{'ik}\rightarrow\text{k'},\ M)$;
- $W = WM$;
- $Y_{sqW} = Y\sqrt{W}$;
- $Y = YW$;
- $T = \mathrm{einsum}(\text{'ij,ik}\rightarrow\text{jk'},\ X,\ Y)$;
- $S = X^2$;
- $S = \mathrm{einsum}(\text{'ij,ik}\rightarrow\text{jk'},\ S,\ W)$;
- $\widehat{B} = T/S$;
- $Y_{sq} = \mathrm{einsum}(\text{'ik,ik}\rightarrow\text{k'},\ Y_{sqW},\ Y_{sqW})$;
- $R_{sq} = Y_{sq} - 2\widehat{B}T + \widehat{B}^2 S_{11}$;
- $\widehat{\sigma} = \sqrt{R_{sq}/(n-1)}$;
- $\mathrm{se}(\widehat{B}) = \widehat{\sigma}/\sqrt{S}$;
- **return** $\widehat{B}, \mathrm{se}(\widehat{B})$

**End**

### *2.9.8 Evaluating QTL mapping performance using eQTLGen results*

To evaluate the performance of QTL mapping method, we treat eQTLGen [145] as a silver standard, in the sense that eQTLs identified as positive in eQTLGen are treated as the true associations and the non-significant variant/gene pairs in eQTLGen are treated as true non-associations. Although 336 GTEx samples are included in eQTLGen analysis, they make up of only around 1.5% of total samples. So, eQTLGen results are unlikely driven by

GTEx samples. And besides, GTEx v8 includes additional samples that are not included in eQTLGen. Therefore, eQTLGen is an approximately independent eQTL study with much larger sample size (50-fold relative to GTEx v8) and diverse populations (predominantly Europeans along with other populations).

To simplify the analysis, we randomly select 100,000 eQTLGen cis-eQTLs (FDR $< 0.05$) as the true associations in the silver standard. And we randomly collect 100,000 variant/gene pairs in eQTLGen with p-value $> 0.5$ as the true non-associations. Among those variant/gene pairs in silver standard, 96,660 true associations and 78,691 true non-associations are included in both our mixQTL mapping pipeline and GTEx v8 analysis. So that we keep only these variant/gene pairs for downstream analysis.

#### 2.9.8.1 Comparing the effective sample size

To compare the effective sample size between mixQTL and eQTL approaches, we performed analysis similar to [90]. Here, we utilize the fact that $\chi^2$ statistic scales proportionally with the sample size, among those true associations. So, we can calculate the ratio $\chi^2_{\text{mixQTL}}$ over $\chi^2_{\text{eQTL}}$ for each truly associated variant/gene pair as the measure of effective sample size of mixQTL relative to eQTL approach. Specifically, we calculate the relative effective sample size using the true associations in the silver standard constructed above (as the proxy of true associations based on independent evidence). Note that the gain of power in mixQTL depends on the amount of allele-specific observations so we measured the average relative effective sample size as the median of the $\chi^2$ ratio. Among the 96,660 variant/gene pairs collected as true associations in silver standard, we measured the median of $\chi^2_{\text{eQTL}}$ as 2.59 and the median of $\chi^2_{\text{mixQTL}}$ as 3.56. And the median of the ratio $\chi^2_{\text{mixQTL}}$ over $\chi^2_{\text{eQTL}}$ is 1.29. In other word, it suggests that the mixQTL approach (with 670 individuals) is equivalent to the eQTL approach with 863 individuals.

#### 2.9.8.2 Drawing receiver operating characteristic and precision-recall curves

The ROC and PR curves are constructed using $-\log(p)$ as prediction score (higher means more likely to be causal). To simplify the calculation, we evaluate the performance measures at a grid of score cutoffs: 0.1, 0.2, ..., 1.9, 2, 2.2, ..., 2.8, 3, 4, ..., 50. For ROC curve, we calculate true positive rate and false positive rate at these cutoffs. And similarly, for PR curve, we calculate precision and power at these cutoffs.

### *2.9.9 Running RASQUAL on GTEx data*

We implemented the RASQUAL analysis pipeline for GTEx v8 data at `https://github.com/liangyy/run-rasqual` and ran RASQUAL on kidney cortex and whole blood data in GTEx v8. We focused on the genes with enough allele-specific reads. To ensure this, we required the genes to pass the following two criteria: 1. The gene should have more than 100 reads (total count) in at least 80% of the samples; 2. The gene should have $\geq 50$ allele-specific reads (per haplotypes and both haplotypes should meet the criteria) in at least 15 samples. With these criteria, we tested 4,596 genes in kidney cortex (sample size = 73) among 22 autosomes and 192 genes in whole blood (sample size = 670) on chromosome 22. Instead of using RASQUAL default parameters, we fixed two of the hyperparameters, $\delta$ (=0.5) and $\phi$ (=0.01), controlling mapping error rate and mapping bias. We made this choice for two reasons: 1. These two parameters are not considered in mixQTL analysis; 2. To estimate these parameters take time and by fixing these the running time for RASQUAL reduced substantially. RASQUAL was run with 8 CPU cores and 16gb RAM.

### *2.9.10 Examining the enrichment in functional annotations*

We focused the analysis on 26 GTEx v8 tissues which have sample size $< 260$. Furthermore, we focused on the genes with sufficient amount of allele-specific counts. Specifically, for each tissue, we selected the genes passing the criteria described in Section 2.9.9.

Regarding the functional annotation, we included the functional annotation constructed by GTEx v8 working group (see more details in [139] Section section 9). We also looked at the candidate Cis-Regulatory Elements (cCREs) in ENCODE [104] where we manually selected ENCODE tissue/cell line that matches with the GTEx tissue. With this restrictive matching, we included 10 of the 26 tissues for the cCRE enrichment analysis. Moreover, to ensure the quality of the annotation, we excluded the cCREs that are labelled as "Unclassified". Lastly, we also considered GWAS catalog where we label GWAS catalog variant as 1 and the rest of the genome as 0.

Since all these annotations are binary, for each functional annotation, we formed a 2-by-2 table (functional annotation against whether the variant is top signal in mixQTL or mixFine) aggregating across all tissues. The enrichment in functional annotation was measured as the odds ratio calculated on the basis of the 2-by-2 table.

# CHAPTER 3

# DEVELOPING AND EXAMINING THE PERFORMANCE OF POLYGENIC TRANSCRIPTOME RISK SCORES

In Section 3.1, I implemented a predicted transcriptome-based polygenic risk score and examined the performance and portability across different ancestry groups. In Section 3.2, I developed computer codes calculating the proposed score in Section 3.1 on the basis of GWAS summary statistics instead of individual-level data.

## 3.1 Polygenic transcriptome risk scores (PTRS) can improve portability of polygenic risk scores across ancestries

Material from: Liang, Yanyu, Milton Pividori, Ani Manichaikul, Abraham A. Palmer, Nancy J. Cox, Heather E. Wheeler, and Hae Kyung Im, "Polygenic transcriptome risk scores improve portability of polygenic risk scores across ancestries", Biorxiv, preprint 2020, Cold Spring Harbor Laboratory Press [83]

### *3.1.1 Abstract*

Polygenic risk scores (PRS) are a valuable tool to translate the results of genome-wide association studies (GWAS) into clinical practice. To date, most GWAS have been based on individuals of European-ancestry leading to poor performance in non-European populations. We introduce the Polygenic Transcriptome Risk Score (PTRS), which is based on predicted transcript levels (rather than SNPs), and explore the portability of PTRS across populations using UK Biobank data. We show that PTRS has a significantly higher portability (wilcoxon p=0.013) in the African samples where the loss of performance is most acute.

### *3.1.2 Introduction*

Polygenic risk scores (PRS) for a variety of traits are becoming accurate enough to be useful for clinical practice, realizing the longstanding goal of personalized medicine. PRS for coronary artery disease (CAD) have been shown to provide prediction that has been compared to monogenic mutations of hypercholesterolemia [66]. In practice, PRS may impact a larger proportion of patients compared to monogenic mutations; for example, PRS for CAD provide potentially actionable information for 8% of the population (for whom the risk increases by three-fold) whereas known monogenic mutations are only informative for about 0.4% of patients. However, a major limitation of this approach is that PRS developed in one human ancestry group do not perform well in other ancestry groups, limiting their utility and exacerbating already severe health disparities [27, 97]. This problem is being addressed by large efforts such as Human Heredity and Health in Africa (H3Africa) [24], Million Veterans Project [45], AllofUs [108] and TOPMED [135] that are recruiting individuals from diverse ancestry groups.

However, these efforts are time consuming, enormously expensive and will have to be repeated at scale, for numerous traits, across numerous ancestry groups. Therefore, methods that can use GWAS results from one population for prediction in other ancestry groups are highly desirable. Analysis of GWAS conducted in different populations suggested that a considerable fraction of causal SNPs are shared across populations [127]. Hence, efforts to develop methods that transfers knowledge about the influence of genes on traits across populations could help improve prediction in underrepresented ancestry groups in a cost-effective manner.

It is now widely understood that many association signals are driven by their effects on the transcriptome. PrediXcan [43] and other TWAS methods [51, 55] leverage reference transcriptome datasets to train prediction models of gene expression levels and correlate the genetically predicted gene expression levels with complex traits to identify causal genes.

Assuming that the role of genes on traits is conserved across populations, we hypothesized that prediction at the level of estimated transcript abundance rather than SNPs might help the portability across populations.

Therefore, we propose the polygenic transcriptomic risk score (PTRS) as a gene-based complement to the PRS that can help improve portability across human ancestry groups. We recognize that PTRS does not outperform the state of the art PRS [111]. However, integrating PTRS to PRS construction has many desirable properties. One advantage of PTRS is that the smaller number of features (tens of thousands of genes rather than millions of SNPs), means that optimizing the parameters to build PTRS is more manageable than PRS. Another advantage of PTRS is that training transcriptome prediction models requires much smaller samples than training PRS, and can then be used for prediction of many different traits. Furthermore, training data for non-European individuals are becoming increasingly available. Finally, because PTRS is gene-based, it is inherently more biologically interpretable than PRS.

In this paper, we explore the properties of PTRS using the UK Biobank (UKB), which provides genotype and phenotype data in half a million individuals [20]. Although the majority of participants in UKB are of European-descent, several thousand individuals of non-European descent are also available, and can be used to compare prediction by PRS and PTRS across ancestries. We start by quantifying how much of the trait heritability can be explained by the genetically predicted transcriptome. We then build PRS and PTRS for a range of complex traits and compared their prediction accuracy and portability across populations.

### *3.1.3 Results*

Before describing the results we define and clarify some terminology. In this paper, there are two types of prediction: 1) gene expression level prediction from genotype data and 2) com-

plex trait prediction using PRS or PTRS. PRS uses genotype data directly and PTRS uses linear combinations of genotypes representing predicted gene expression levels. To simplify exposition, we will only use the term *training* for the calculation of weights for predicting gene expression levels using genotype data. The *training* of transcriptome (gene expression levels) prediction weights had been performed previously and we simply downloaded them from predictdb.org. When we estimate optimal weights for PRS and PTRS, we will use the terms *building* or *constructing*. We performed the *building* of PRS and PTRS using the *discovery* set. The testing of the risk scores, PRS and PTRS, were performed in what we call the *target* sets. For the remainder of the paper, we will refer to individuals by their ancestry and drop the -descent suffix. Unless otherwise clarified, we will use the term transcriptome to mean the set of predicted expression levels of genes. GTEx EUR transcriptome should be interpreted as the set of predicted gene expression levels using weights trained in European samples from GTEx. Similarly, MESA EUR transcriptome, will refer to the predicted transcriptome using weights trained with the MESA European samples. MESA AFHI transcriptome will refer to the predicted transcriptome using weights trained with a combination of African American and Hispanic individuals from the MESA study.

#### 3.1.3.1 Experimental setup

An overview of the experimental setup describing the discovery, training, and target sets used in the paper is shown in Figure 3.1. We randomly selected 356,476 unrelated Europeans in the UK Biobank for the discovery set. For testing the performance of risk scores, we constructed 5 target sets with participants of various ancestries in the UK Biobank. We used 6,413 African, 1,326 East Asian, and 6,479 South Asian individuals for the non European target sets. We also reserved two randomly selected sets of 5,000 Europeans as additional target sets. One was selected as the EUR reference set and the second European target was used as a test set to assess the variability of the results within the same ancestry.

For predicting the transcriptome, we downloaded prediction weights from multiple ancestries collected in predictdb.org. The first set of models had been trained in European individuals from the GTEx v8 release [6] in whole blood. The second set of models had been trained using array-based expression in monocyte samples of Europeans, African Americans, and Hispanics from the MESA cohort [100].

For our tests, we focused on the 17 anthropomorphic and blood phenotypes used by Martin et al. [97]. We sampled our discovery and target sets randomly so that there is no exact match with Martin et al's discovery set.

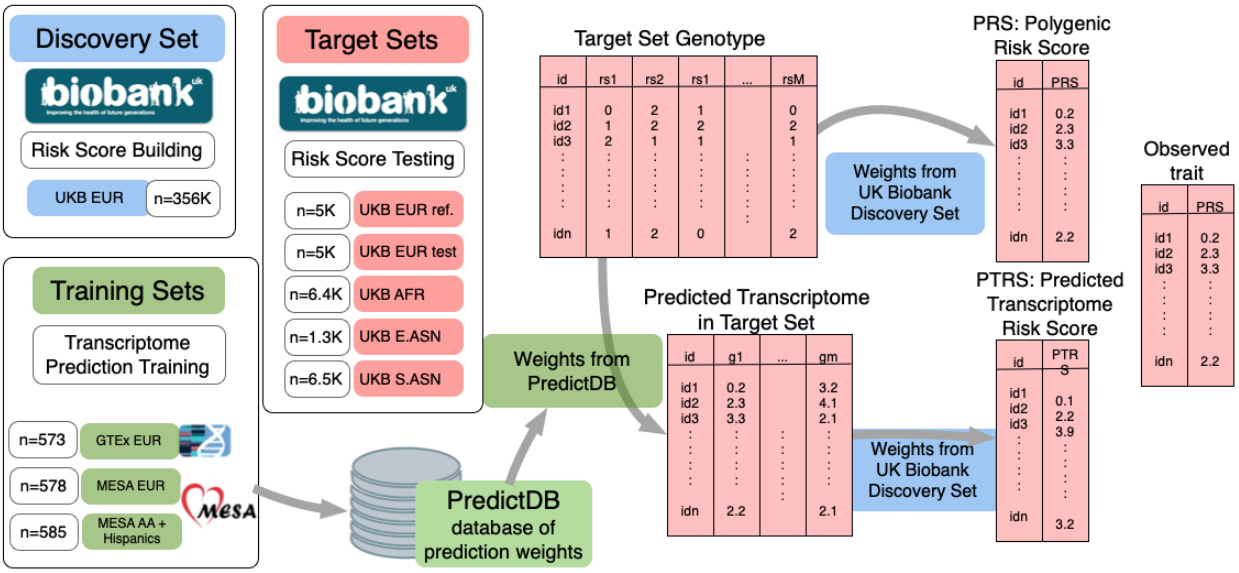



**Figure 3.1: Experiment setup for examining portability.** This figure summarizes the experimental set up used for testing the portability of PRS and PTRS across populations. The weights for calculating PRS and PTRS were estimated in the *discovery set*, which consisted of 345K randomly sampled individuals of European-descent from the UK Biobank. The *training* sets where the weights for the prediction of transcriptomes were computed are shown in green. We downloaded the weights trained previously from predictdb.org. We sampled 5 *target sets* from the UK Biobank for testing the risk scores: two randomly sampled sets of European-, one African-, one East Asian-, and one South Asian-descent individuals. For each of the 5 *target* sets, predicted transcriptomes were calculated using the weights trained in each of the three *training* sets: GTEx EUR, MESA-EUR, MESA-AFHI.

#### 3.1.3.2 Predicted transcriptome captures a higher portion of chip heritability than expected

Before building PTRS, we started by determining whether prediction of traits was possible using just predicted transcriptomes. If feasible, we also wanted to know how much of the trait variation can be captured by the predicted transcriptome. For this purpose, we calculated the proportion of variance explained (PVE) by the predicted transcriptome assuming random effects of gene expression levels. The approach is analogous to standard SNP-heritability estimation [160]. For heritability estimation one would use the genetic relatedness matrix. Here, we use the "predicted expression relatedness matrix".

In this section, we calculated the predicted transcriptome using the GTEx EUR weights using the European target set genotype data. Using these predicted expression levels, we calculated the "predicted expression relatedness matrix" (instead of the genetic relatedness matrix) and applied the standard restricted maximum likelihood estimation to calculate the proportion of variance explained by the predicted transcriptome.

Since the PVE for each trait will depend on the heritability of the trait, we calculated the proportion of PVE divided by the heritability of the trait. This quantity represents the proportion of heritability explained by the predicted transcriptome. We also estimated the SNP-heritability using the restricted maximum likelihood approach in the same cohort.

Figure 3.2A shows the distribution of the proportion of heritability explained by the GTEx EUR transcriptome calculated in the EUR target set. We found that the GTEx EUR whole blood based predicted transcriptome captured on average 22.9% (s.e.=2.9%) of the trait heritability. This result is largely consistent to the estimates reported previously by [162] for the subset of traits used here which were mostly blood related. Notice that the predicted transcriptome had fewer than 1% of the number of features used in the calculation of the chip heritability (fewer than 10K genes predicted in whole blood compared to roughly 1 Million independent SNPs). Therefore, this result constitutes a 20-fold increase in the per

feature proportion of heritability explained.

#### 3.1.3.3 Aggregating predicted transcriptomes in multiple tissues increases the PVE

To explore ways to increase the proportion of variance explained (PVE), we calculated the proportion explained collectively by the transcriptome predicted in 10 tissues selected among the ones with largest sample sizes in GTEx, including muscle, adipose, tibial artery, breast, lung, fibroblast, lung, tibial nerve, and skin, with sample sizes ranging from 337 to 602 (Table 3.3). As anticipated, we found that, collectively, the predicted transcriptomes in 10 tissues explained a larger portion of heritability: on average 35.5% (s.e. = 4.7%) of the heritability corresponding to a 48% increase relative to whole blood alone. This result suggests that adding transcriptomes from multiple tissues will improve predictions in general.

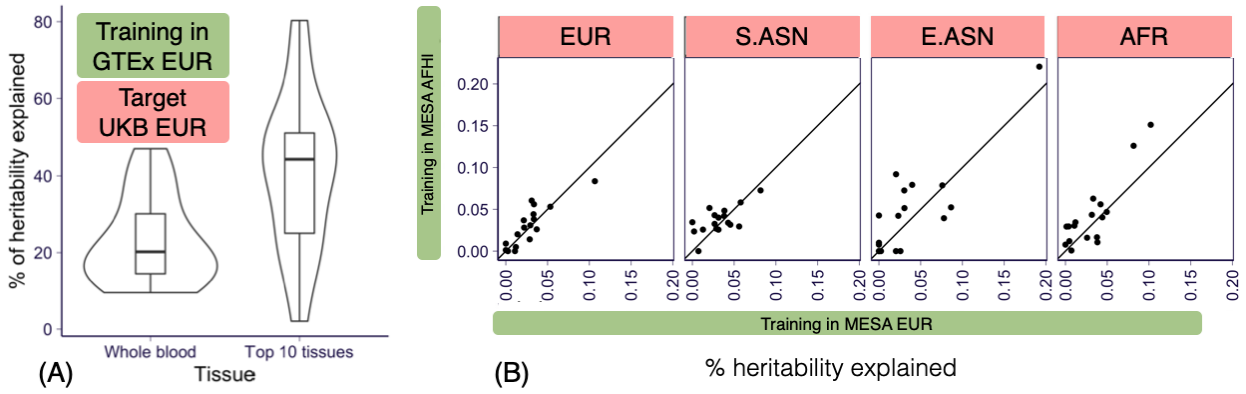



**Figure 3.2: Proportion of variance explained (PVE) by the predicted transcriptome.** **(A)** shows the ratio of PVE (the proportion of phenotypic variation explained by the predicted transcriptome) of GTEx EUR transcriptome model over the chip heritability using whole blood on the left and using 10 tissues on the right. The 10 tissues with highest sample sizes were selected from GTEx (muscle, adipose, tibial artery, breast, lung, fibroblast, lung, tibial nerve, and skin, with sample sizes ranging from 337 to 602). Notice that in general we used GTEx whole blood and MESA monocyte-based predictors except for this panel where we used 10 tissues from GTEx. **(B)** The PVE of MESA EUR-based predicted transcriptome (x-axis) and MESA AFHI-based predicted transcriptome (y-axis) are shown. Each panel presents the results calculated in one ancestry group and each point presents one trait.

#### 3.1.3.4 Matching training and target ancestries may increase the proportion of variance explained

So far, we have used GTEx EUR transcriptomes to calculate the PVE. It is reasonable to assume that matching the training and target populations, i.e. using transcriptomes trained in the same population as the target sets should be beneficial. We tested this hypothesis here.

We took advantage of the availability of trans-ancestry transcriptome prediction models from the MESA cohort [100]. One of them (MESA-EUR) was trained in a European population and the other one (MESA AFHI) was trained in a combination of African American and Hispanic populations (Table 3.3). We decided to use the combined (African American and Hispanic) transcriptome prediction since the similarity of the sample sizes (578 vs 585) would make the comparison with the European trained models more fair.

We found that (Figure 3.2B) in the African target set, using the ancestry matched MESA AFHI transcriptome yielded a higher PVE albeit not significant (1.1%, p=0.065) proportion of variance explained than when using the MESA EUR transcriptome. For the European target set, the difference between using the MESA AFHI or the EUR transcriptomes close to 0 (0.3%, p=0.50). This lack of significant difference could be attributed to the limitations of the array-based MESA transcriptome data prompting the need to generate improved non-European transcriptome predictors.

#### 3.1.3.5 Building PRS and PTRS

After having determined that it is possible to capture a significant portion of trait variability using predicted transcriptome, we proceeded to build the PRS and PTRS in our discovery set (356K Europeans from the UK biobank).

We built PTRS weights using elastic net, a regularized linear regression approach, which selects a sparse set of predicted expression features to make up the PTRS. For PRS weights,

we used the standard LD clumping and p-value thresholding approach (see details Section 3.1.5.3). To rule out that using elastic net for PTRS instead of clumping and thresholding as done for PRS was driving our conclusions, we also ran the comparison using clumping and thresholding for PTRS and found no change in the substance of our results. See (Figure 3.6 and 3.7).

We quantified the prediction accuracy in each target set using the partial $R^2$ ($\widetilde{R}^2$), which provides a measure of correlation between predicted and observed outcomes with the added benefit of taking covariates into account (see details in Section 3.1.5.12). We split the target set into validation and test set and determined the hyperparameters (penalty for elastic net PTRS and p-value threshold for PRS/PTRS) in the validation set and calculated the $\widetilde{R}^2$ under the selected hyperparameter in the test set. To reduce the stochasticity of the estimated performance we repeated the random splitting 10 times and report the average $\widetilde{R}^2$ across the ten splitting schemes. The weights were calculated in the discovery set for the different hyperparameters.

#### 3.1.3.6 PTRS prediction accuracy achieves the expectation given by their PVE explained

The prediction accuracy of PTRS (GTEx EUR based) was on average lower than the accuracy of PRS (paired t-test p = 0.03) as shown for the 17 traits in Figure 3.3A for the European target set and in Figure 3.8 for the other ancestries. It is worth noting that since predicted transcriptomes were explaining about a fifth of the heritability (i.e. what common genetic variants could explain), we would expect that the prediction performance of PTRS would be about a fifth of the PRS performance. However, PTRS performance was around half of the PRS performance. This better than expected performance suggests that integrating predicted transcriptomes and other omics is a promising avenue to improve PRS performance in general.

We found that PTRS was much closer to the optimal performance upper bound (PVE) than PRS was to its upper bound (heritability), as shown in Figure 3.3B, where each score is compared to its upper bound. This is consistent with the reported monotonically increasing relationship between a measure of prediction performance (correlation between the genetic component and the predicted values, slightly different from our definition, using lowercase $r$ for distinction) and the per-feature heritability/PVE: $r^2 = \frac{n \cdot h^2/n_{\text{features}}}{1+n \cdot h^2/n_{\text{features}}}$ [28]. The close-to-optimum prediction performance of PTRS indicates that the predicted expression of each gene carries a higher per-gene PVE than genetic variant's per-variant heritability.

#### 3.1.3.7 Matching training and target ancestries may improve prediction accuracy

We have suggestive evidence that matching training and target ancestries can improve the PVE in the African population. To test whether matching the training and target ancestries would improve the PTRS prediction accuracy, we examined the difference between using the African transcriptome (MESA AFHI) vs the European transcriptome (MESA EUR).

For the European target set, the European transcriptome based PTRS had better accuracy than the AFHI transcriptome based one, with an gain of 0.74% (s.e.=0.096%) when using European vs AFHI, as hypothesized. For the African target set, however, the difference was not significant indicating that improvements in prediction across ancestries are probably needed to detect the potential gain.

To avoid differences due to having different number of predicted genes, PTRS were built using only the genes that were present in both training sets, EUR and AFHI. See details in Section 3.1.5.9.

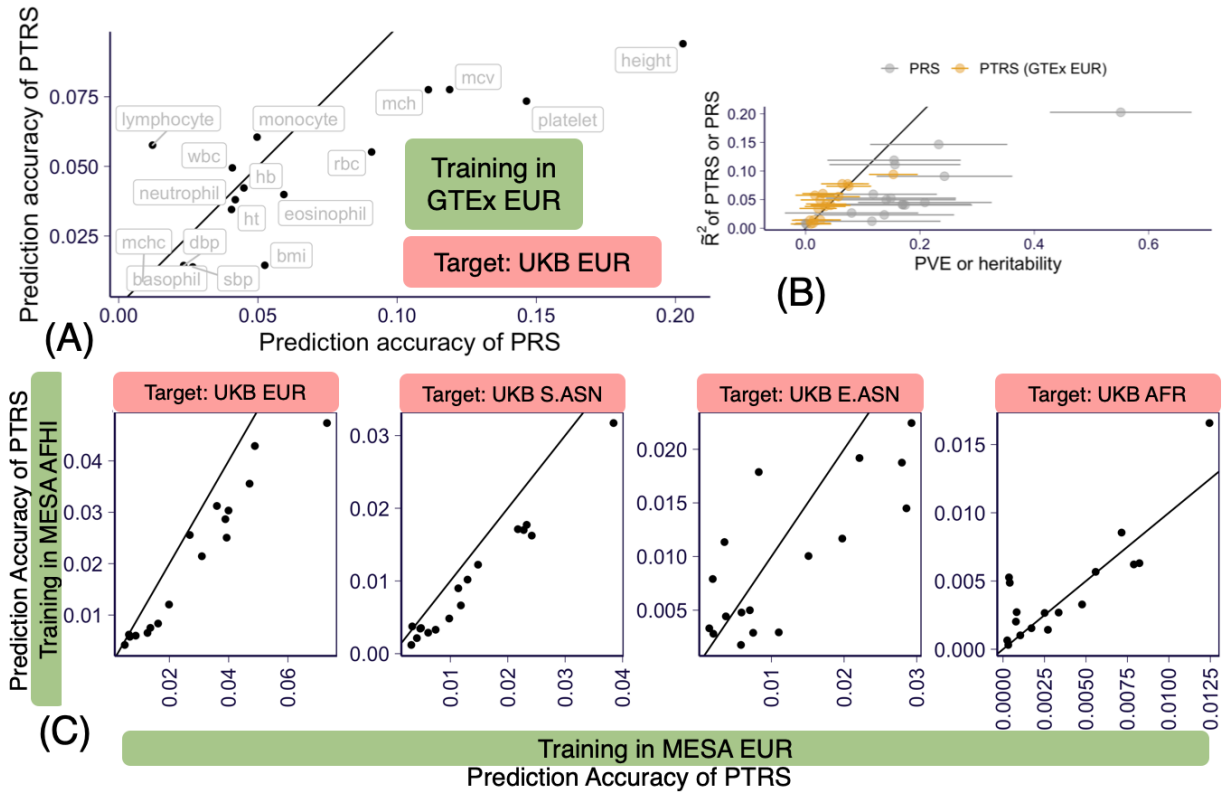



**Figure 3.3: Prediction accuracy of predicted transcriptome risk scores (PTRS). (A)** Prediction accuracy, measured by partial $\widetilde{R}^2$, of PTRS (on y-axis) compared to the accuracy of PRS (on x-axis). **(B)** Prediction performance is shown on x-axis and heritability (for PRS) and proportion of variance explained (for PTRS) are shown on y-axis. **(C)** Prediction accuracy of MESA AFHI PTRS vs MESA EUR PTRS in the 4 target populations.

### 3.1.3.8 PTRS improves portability into the African population

To test our hypothesis that PTRS can generalize more robustly across populations than the standard PRS, we defined '*portability*' as the predictive accuracy in each population relative to the European reference target set (EUR ref.). This is calculated as the ratio of the $\widetilde{R}^2$ in the target population divided by the $\widetilde{R}^2$ in the European reference target set. Thus, by definition the portability in the European reference set is 1.

Consistent with reports by [97], the portability of PRS degrades with the genetic distance to the European discovery set as shown in gray in Figure 3.4A. The portability of PTRS

(shown in orange) also decreases with genetic distance to the discovery set, with the African target sets showing the largest loss of accuracy, as expected. However, we also observed that the portability of PTRS in the African target set was significantly higher than the portability of PRS (wilcoxon p=0.013). These results provide strong proof of principle that integrating predicted transcriptome as done with PTRS has the potential to improve portability of risk scores across populations despite the limitations of currently available AFR training data. In the European test set, we observed quite a bit of variability in the portability, ranging from 0.47 to 2.28, despite the fact that both European target sets were randomly sampled from the same European UK Biobank participant set. As expected, the median portability in the second EUR target set is centered around 1.

We verified that the increase in portability of PTRS in the African target set was not driven by the use of elastic net instead of clumping and thresholding for building the PTRS as shown in Figure 3.7 where portability is shown for the clumping and thresholding approach. By looking at the comparison between prediction performance and portability (Figure 3.9), we found no evidence that the increase in portability in the African target set could be driven by low prediction performance in the set.

Furthermore, we examined whether the population-matched transcriptome models could improve portability using the MESA models. However, we did not observe significant improvement relative to PRS (Figure 3.10).

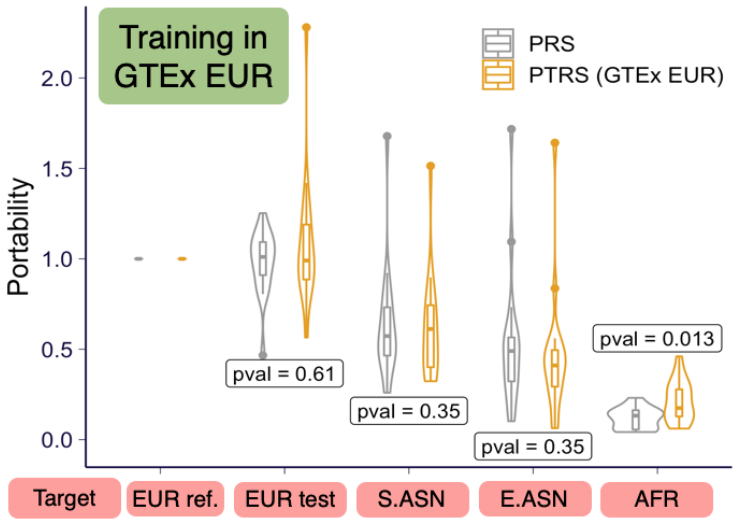
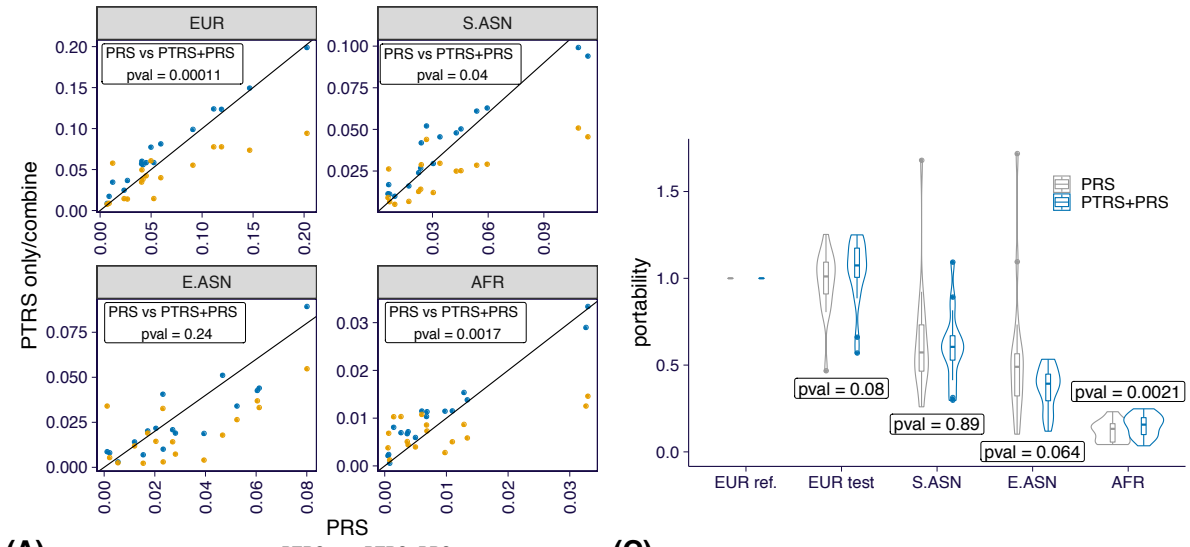



**Figure 3.4: Portability of PTRS for 17 quantitative phenotypes in UK Biobank.** The portability of PTRS trained and calculated using GTEx EUR whole blood samples are shown in yellow with the PRS shown in gray. 'EUR ref.' set is used as the reference population in the calculation of portability (3.1.5.14) so that the portability is always 1. Recall that the absolute performance of PTRS is, on average, lower than PRS as shown in Figure 2.3A and Figure 3.8.

Taken together, our results provide support to our hypothesis that PTRS can improve the portability of PRS in general. Also they suggests that adding transcriptomes predicted in other tissues and other omics data are promising avenues to further improve PRS portability.

### 3.1.3.9 Combining PTRS and PRS achieves comparable performance as PRS alone




(B)

| sample | R2 (PRS – PTRS) | wilcox_pval |
|---|---|---|
| AFR | −0.00218 | 0.001680 |
| EUR test | −0.01090 | 0.000107 |
| EUR ref. | −0.00635 | 0.034800 |
| E.ASN | 0.00369 | 0.243000 |
| S.ASN | −0.00407 | 0.039500 |

(D)

| sample | portability (PRS – PTRS) | wilcox_pval |
|---|---|---|
| AFR | −0.0339 | 0.00209 |
| EUR test | −0.0563 | 0.07970 |
| E.ASN | 0.1650 | 0.06380 |
| S.ASN | 0.0240 | 0.89000 |

**Figure 3.5: Prediction performance and portability of the score combining PTRS and PRS.** Combing the elastic net-based PTRS and PRS, the prediction accuracy and portability are shown below. **(A)** The prediction accuracy of the PRS is shown on x-axis and it is compared against the prediction accuracy of the PTRS (yellow) or the combined score on y-axis. The results on all of the 17 quantitative traits are shown. Each panel corresponds to one ancestry group. The p-values are for comparing PRS accuracy versus the combined score accuracy via the paired Wilcoxon signed rank test. **(B)** A summary of the difference between the prediction accuracy of PRS and the combined score is shown for each of the ancestry group. The second column shows the mean difference and the third column shows the results of the paired Wilcoxon signed rank test comparing the accuracy of PRS versus the combined score. **(C)** The portability of PRS (gray) and the combined score (blue) is shown for each ancestry group. The violin and box plots summarize the results from the 17 quantitative traits. The p-values are for comparing PRS portability versus the combined score portability via the paired Wilcoxon signed rank test. **(D)** A summary of the difference between the portability of PRS and the combined score is shown for each of the ancestry group. The second column shows the mean difference and the third column shows the results of the paired Wilcoxon signed rank test comparing the portability of PRS versus the combined score.

To leverage the better prediction performance of PRS and the higher portabililty of PTRS in African ancestry, we combined PRS and PTRS via a weighted sum (Section 3.1.5.13). Combining the elastic net-based PTRS and the clumping and thresholding-based PRS, the prediction performance of the combined score is significantly higher than PRS alone in all but E.ASN ancestry (Figure 3.5A and 3.5B). Though the improvement is not significant when combining the clumping and thresholding-based PTRS and PRS, the performance is still comparable to PRS alone (Figure 3.11). Moreover, the combined score remained more portable than PRS (Figure 3.5C, 3.5D, and 3.12).

These results suggest that we can take the advantages of both PTRS and PRS by combining PTRS and PRS together. The combined scores achieve similar performance as PRS in the training ancestry, European ancestry, while transfer better to African ancestry.

### *3.1.4 Discussion*

In this paper we introduced the polygenic transcriptomic risk score and used it to address a major problem in human genetics, namely the poor ability to use genotype to predict phenotype using PRS that are trained in one ancestry group but applied to another ancestry group. We started by establishing that prediction of complex traits using the predicted transcriptome is possible by showing that the total trait variation explained via predicted transcriptomes ranges from 22.9% (using whole blood) to 35.5% (with a broader sets of tissues) of the SNP heritability, i.e. the total variation that can be explained using common SNPs. Comparing to a recent work, MESC, which quantifies the genetic effects mediated through gene regulation by utilizing GWAS summary statistics and eQTL datasets [162], our estimated fractions are slightly higher. This difference could be attributed to the difference in the eQTL datasets being used. The eQTL datasets used in our analysis are larger in sample sizes than the ones being used in MESC. This enabled our analysis to capture more transcriptomic regulation which was missed due to lack of power in MESC. Despite the

slight difference discussed above, consistent with our findings, MESC also suggested that by leveraging more tissue types, the gene regulation can explain larger proportion of genetic effects on complex phenotypes.

Promisingly, the actual predictors built on predicted transcriptomes had performances that were more than double the expected 22.9% of the PRS performance. We found that the portability of PTRS was significantly higher than the portability of PRS in the African target set. African populations are the most affected by the Eurocentric bias in GWAS studies. Our study results suggest that investing in multi-omic studies of diverse populations may be an effective way to reduce current genomic disparities by taking better advantage of existing GWAS studies.

In this paper, we have explored the use of PTRS to address the limited portability of PRS across ancestry groups. One intriguing application of our PTRS method is to transfer the polygenic knowledge derived in humans to other model systems. Currently, most attempts to follow up on human GWAS findings in model organisms focus on individual genes, despite the overwhelming evidence that individual genes seldom contribute even 1% of trait variability. As we improve our PTRS, it would be possible to build models based on human discovery sets and then use them to predict those traits in model organisms based on either measured or predicted transcript levels. Whether this approach could successfully predict traits that can be measured in other species (e.g. body size, blood pressure) is currently unknown. The success of PTRS for cross-species translation is theoretically dependent on the extent to which gene expression differences have similar impacts across species. However, if we could build a transferable PTRS, we could envision to run experiments on model organisms, measure their transcriptomes (in the right context and tissue), and predict complex traits (e.g. behavioral traits) that are hard or impossible to measure. This platform could be used to test various interventions and obtain effects on phenotypes that are more relevant to humans, opening up completely novel lines of investigation.

Our study points to promising strategies to improve risk prediction in general but it also has several limitations. First, PTRS are based on prediction models of gene expression traits which we estimated to account for less than a third of the chip heritability of the complex traits considered here. We expect this limitation to be mitigated as additional transcriptome reference sets in different contexts as well as other omics data covering mediating mechanisms missed in current models. The increase in the proportion of variance explained from a fifth when using whole blood predictors alone compared to over 35.5% when using 10 tissues indicates that much can be gained by increasing the breadth of reference omic data. Second, we used single tissue prediction models for most of the analysis in this paper, which captured a fifth of the variation in the complex traits here. This can be improved by developing approaches to integrate multiple tissue models. Third, weights for PRS were calculated using GWAS summary results (thresholding and pruning method) whereas PTRS weights were calculated using individual level data due to computational considerations. Individual-level based PRS would likely perform better, although whether they would be more portable is not obvious. For reassurance, we have ruled out the possibility that the portability improvement in the African target set was due to this different treatment of PTRS and PRS by re-calculating the portability of PTRS using the same approach as PRS (summary statistics based clumping followed by p-value thresholding). Developments of methods to optimally combine PRS and PTRS should be encouraged. Fourth, higher quality prediction models of the transcriptome in non-European ancestries are limited. Here we used predictors trained in monocyte samples assayed with older array technology. Multiple ancestry models are currently being generated by us and other groups. For example, the MESA TOPMED project has assayed RNAseq, protein, methylation, and metabolomics data in African Americans, Hispanics, and Asian ancestry individuals which will allow the development of improved prediction models. Currently, our methods are based on transcriptomic data from bulk tissues, however, just as the choice of tissue is important, the choice of tissue and cell type

has the potential to provide even better performance, however the availability of such data, as well as the methods for implementing it, are currently limited. Finally, other approaches such as prioritizing functional annotations for variant selection [2], more diverse set of variants [23], and better leveraging fine-mapping approaches should be combined with the one shown here to optimize the portability across populations.

### *3.1.5 Methods*

#### 3.1.5.1 Obtaining individuals and phenotypes from UK Biobank

We used data from UK Biobank downloaded on July 19 2017. We excluded related individuals and the ones with high missing rate or other sequencing quality issues. As covariates, we extracted age at recruitment (Data-Field 21022), sex (Data-Field 31), and the first 20 genetic PCs. The ancestry information of individuals was obtained from Data-Field 21000 and we kept individuals labelled as 'British', 'Indian', 'Chinese', or 'African' (according to Data-Coding 1001: `http://biobank.ctsu.ox.ac.uk/crystal/coding.cgi?id=1001`). Throughout the paper, we labeled 'British' individuals as EUR, 'Indian', 'Bangladeshi', and 'Pakistani' individuals as S.ASN, 'Chinese' individuals as E.ASN, and 'African' and 'Caribbean' individuals as AFR. The measurements of the 17 quantitative phenotypes (as shown in Table 3.1) across all available instances and arrays were retrieved. The data retrieval described above was performed using ukbREST [115] with the query YAML file available at `https://github.com/liangyy/ptrs-ukb/blob/master/output/query_phenotypes.yaml`.

If one individual has multiple measurements for the same phenotype (in more than one instances and/or more than one arrays), we collapsed multiple arrays by taking the average and we aggregated measurements across multiple instances by taking the first non-missing value. Individuals with missing phenotype in any of the 17 quantitative phenotypes or covariates were excluded.

#### 3.1.5.2 Quality control on self-reported ancestry

To ensure the quality of ancestry label, we removed individuals who deviate substantially from the population that they were assigned to. Specifically, for population $k$ among the 4 populations (EUR, S.ASN, E.ASN, and AFR), we treated the distribution of the individuals, in the space of the first 10 PCs, as multivariate normal. And we calculated the observed population mean $\widehat{\mu}_k$ and covariance $\widehat{\Sigma}_k$ accordingly. Then, for each individual $i$ in population $k$, we evaluated the "similarity" $S_{ik}$ to the population $k$ as $S_{ik} = \log \Pr(\mathrm{PC}_i^1, \cdots, \mathrm{PC}_i^{10}; \widehat{\mu}_k, \widehat{\Sigma}_k)$. Intuitively, if an individual has genetic background differing from is the assigned population, the corresponding $S_{ik}$ will be much larger than others. So, we filtered out individuals with $S_{ik} \leq -50$ in the assigned population $k$. This cutoff was picked such that $S_{ik'}$ for any un-assigned population $k'$ has $S_{ik'} \leq -50$ for all individuals.

The number of individuals remained after data retrieval and ancestry quality control is listed Table 3.2.

#### 3.1.5.3 Performing GWAS and building LD clumping and p-value thresholding based PRS models

We built PRS using the genotypes and phenotypes of the individuals in the discovery data set (the details of data splitting is described in Section 3.1.5.14). We performed GWAS (linear regression) using `linear_regression_rows` in hail v0.2 where we included covariates: first 20 genetic PCs, age, sex, age$^2$, sex $\times$ age, and sex $\times$ age$^2$. In the GWAS run, we excluded variants with minor allele frequency $< 0.001$ and variants that significantly deviate from Hardy-Weinberg equilibrium (p-value $< 10^{-10}$). And the phenotype in their original scales were used.

To obtain relatively independent associations for PRS construction, we ran LD clumping using `plink1.9` with option `--clump --clump-p1 1 --clump-r2 0.1 --clump-kb 250`. This command extracted genetic variants in the order of their GWAS significances and

excluded all variants having $R^2 > 0.1$ to or 250 kb within any variants that have already been included. The PRS was constructed on the basis of the set of variants obtained from the LD clumping along with the marginal effect size estimated in GWAS run. Specifically, we calculated PRSs at a series of GWAS p-value thresholds: $5 \times 10^{-8}$, $10^{-7}$, $10^{-6}$, $10^{-5}$, $10^{-4}$, $10^{-3}$, 0.01, 0.05, 0.1, 0.5, and 1. In other word, at threshold $t$, the PRS for individual $i$ was calculated as

$$\text{PRS}_i^t = \sum_{j:p_j \leq t} X_{ij} \widehat{b}_j, \tag{3.1}$$

where $X_{ij}$ is the effect allele dosage of variant $j$ in individual $i$ and $\widehat{b}_j$ is the estimated effect size of variant $j$ from GWAS run.

At the testing stage, given the genotype of an individual, we calculated the PRS of the individual using Eq 3.1.

#### 3.1.5.4 Computing the predicted transcriptome

We computed predicted gene expression for all individuals passing filtering steps and quality control. We utilized two sets of prediction models: 1) CTIMP models (proposed in [55]) trained on GTEx v8 EUR individuals [8]; and 2) elastic net models which were trained on Europeans (EUR) or African Americans in combination with Hispanics (AFHI) [100]. The sample size and tissue informations of the prediction models are listed in Table 3.3.

#### 3.1.5.5 Estimating PVE by predicted transcriptome of a single tissue

To assess the potential predictive power of predicted transcriptome on the phenotypes of interest, we estimated the proportion of phenotypic variation that could be explained by the predicted transcriptome in aggregate. Specifically, we assume the following mixed effect

model (for individual $i$).

$$Y_i = \mu + \sum_l C_{il} a_l + \sum_g \widetilde{T}_{ig} \beta_g + \epsilon_i \tag{3.2}$$

$$\epsilon_i \sim_{iid} N(0, \sigma_e^2) \tag{3.3}$$

$$\beta_g \sim_{iid} N(0, \frac{\sigma_g^2}{M}), \tag{3.4}$$

where $M$ denotes the number of genes, $C_{il}$ is the $l$th covariate, $\widetilde{T}_{ig}$ is the inverse normalized predicted expression for gene $g$, and $Y_i$ is the observed phenotype. By inverse normalization, we converted the predicted expression $\widehat{T}_{ig}$ to $\widetilde{T}_{ig}$ by $\widetilde{T}_{ig} = \Phi^{-1}(\frac{\text{rank}(\widehat{T}_{ig})}{N+1})$ within each gene $g$ where $N$ is the number of individuals and 'rank' is in increasing order. So that we have $\widetilde{T}_{ig} \sim N(0,1)$. The parameters of the model were estimated using hail v0.2 `stats.LinearMixedModel.from_kinship` with $K$ matrix being set as $\widetilde{T}\widetilde{T}^t/M$. And PVE is calculated as $\frac{\hat{\sigma}_g^2}{\hat{\sigma}_e^2+\hat{\sigma}_g^2}$. The same set of covariates as Section 3.1.5.3 were used.

The PVE estimation was performed for each transcriptome model and population pairs. For non-European populations, all individuals were included in the analysis. We randomly selected 5,000 EUR individuals for the analysis.

#### 3.1.5.6 Estimating PVE by predicted transcriptome of multiple tissues

The genetic effects on the complex trait can be mediated through the regulation of expression in different tissues so that including predicted transcriptomes in multiple tissues could potentially improve the prediction performance. To quantify the gain, we performed the PVE analysis as described in Section 3.1.5.5 using predicted expression in 10 GTEx tissues (listed in Table 3.3) instead of just one. To avoid colinearity issues caused by the high correlation of predicted expression among tissues, we used eigenvectors of the predicted transcriptome matrix instead of the actual predicted expression. More specifically, for each gene g, we formed the matrix of predicted expression across the 10 tissues and performed singular value

decomposition. We only retained eigenvectors with singular values that were at least 1/30 of the maximum singular value of the gene's expression matrix across tissue. This approach is similar to the one used for combining PrediXcan association in multiple tissues [9].

#### 3.1.5.7 Estimating chip heritability

The chip heritability was estimated using the mixed effect model which is similar to the one for PTRS PVE estimation. But here we replace predicted transcriptome with genome-wide variant genotypes. The same cohorts and covariates were used as the ones in PTRS PVE estimation.

#### 3.1.5.8 Transcriptome prediction models for PTRS construction

The predicted transcriptome in the discovery set (UKB EUR) was calculated using models from GTEx [8] and MESA EUR based models [100] listed in Table 3.3). The GTEx EUR whole blood transcriptome consisted of 7,041 genes. For the MESA transcriptomes, we restricted the prediction to the 4,041 genes that were present in both the MESA EUR models and the MESA AFHI models (to ensure that PTRS built in the discovery set with the EUR models could be computed without missing genes in the target sets using the AFHI models).

#### 3.1.5.9 Building PTRS models using elastic net

For each of the 17 quantitative phenotypes, we trained elastic net model to predict the phenotype of interest using the predicted transcriptome (in a single tissue) as features. The same set of covariates as described in Section 3.1.5.3 were used. Let $\widehat{T}_g \in \mathbb{R}^{N\times 1}$ denote the standardized predicted expression level of gene $g$ across $N$ individuals. Similarly, let $C_l \in \mathbb{R}^{N\times 1}$ denote the observed value of the $l$th standardized covariate. We fit the following

elastic net model.

$$\beta^{\mathrm{EN}} = \arg\min_{\beta} \overbrace{\frac{1}{N}\|Y - X\beta - \beta_0\|_2^2}^{\text{loss: } l(\beta)} + \lambda\alpha\|\beta\|_1 + \lambda(1-\alpha)\|\beta\|_2^2 \tag{3.5}$$

$$X := [\widehat{T}_1, \cdots, \widehat{T}_M, C_1, \cdots, C_L] \ , \tag{3.6}$$

where $\beta_0$ is the intercept, $M$ is the number of genes, $L$ is the number of covariates, $\|\beta\|_2^2$ is the $l_2$ norm and $\|\beta\|_1$ is the $l_1$ norm of the effect size vector. Here, $\alpha$ controls the relative contribution of the $l_1$ penalization term and $\lambda$ controls the overall strength of regularization.

The model fitting procedure was implemented using tensorflow v2 with mini-batch proximal gradient method and the code is available at `https://github.com/liangyy/ptrs-tf`. We trained models at $\alpha = 0.1$ ($\alpha = 0.5$ and 0.9 show similar performance). And fixing the $\alpha$ value, as suggested in [40], we trained a series of models for a sequence of $\lambda$'s starting from the highest. The maximum $\lambda$ value, $\lambda_{\max}$, was determined as the smallest $\lambda$ such that Eq 3.7 is satisfied.

$$\mid \nabla l(\beta) \mid \leq \alpha\lambda, \tag{3.7}$$

where the gradient is evaluated at

$$\beta_0 = \overline{Y}, \beta_{\text{covariate}} = 0, \beta_{\text{transcriptome}} = 0 \tag{3.8}$$

So, at $\lambda = \lambda_{\max}$, Eq 3.8 is the solution to Eq 3.5, which could serve as the initial points for the subsequent fittings of $\lambda$'s. We estimated $\lambda_{\max}$ using the first 1000 individuals of the data. And the sequence of $\lambda$ was set to be 20 equally spaced points in log scale with the maximum value being $1.5\lambda_{\max}$ and the minimum value being $\lambda_{\max}/10^4$. Among these PTRS models generated at different $\lambda$ values, we only kept the first 11 non-degenerate PTRS models so that we have the same number of candidate models for both PRS and PTRS.

#### 3.1.5.10 Building PTRS models using the LD clumping and p-value thresholding based approach

To implement clumping and thresholding based PTRS, we first obtained the gene/phenotype associations (PrediXcan association [43]) by associating the predicted gene expression with the phenotypes with the same set of covariates as PRS models. Next, we performed gene-based LD clumping to extract the roughly independent genes for PTRS models. The LD clumping procedure is decribed as follow. First, we prioritized genes according to their PrediXcan p-values. Second, we traversed the prioritized gene list from the most significant ones to the least ones, in which we included a gene into the returned gene list if the squared correlation (in terms of the predicted expression) of this gene and any other genes that have been included in the current returned gene list is smaller than 0.1.

Among these genes being selected in LD clumping, we built PTRS models by p-value thresholding. Specifically, at p-value thresholds $10^{-6}$,$5\times10^{-6}$, $10^{-5}$, $5\times10^{-5}$, $10^{-4}$, $5\times10^{-4}$, $10^{-3}$, $5 \times 10^{-3}$, 0.01, 0.05, 0.1, 0.5, 1, we built a PTRS model by keeping the genes with p-values smaller than the cutoff. For these genes being included in the final PTRS model, we used the effect sizes estimated in the PrediXcan association as the PTRS weights.

#### 3.1.5.11 Calculating PTRS in target sets

At the testing stage, given the standardized (within the population) predicted transcriptome of an individual, we calculated the PTRS of the individual using the following:

$$\text{PTRS}_i^{\lambda} = \sum_g \widehat{T}_{ig}\beta_g^{\lambda}, \tag{3.9}$$

where $\beta^{\lambda}$ denotes the $\beta^{\text{EN}}$ obtained at hyperparamter equal to $\lambda$. For the PTRS built upon from GTEx EUR predicted transcriptome, the target PTRS was calculated with the GTEx EUR transcriptome (transcriptome predicted with GTEx EUR gene expression prediction

models). To examine the utility of population-matched prediction model, the PTRS on the target set were calculated with of both MESA EUR and MESA AFHI transcriptomes.

#### 3.1.5.12 Quantifying the prediction accuracy of PRS and PTRS with partial $R^2$ ($\widetilde{R^2}$)

To measure the predictive performance of PRS and PTRS, we calculated the partial $R^2$ of PRS/PTRS against the observed phenotype accounting for the set of covariates listed in Section 3.1.5.3. Specifically, for individual $i$, let $\hat{y}_i$ denote the predicted phenotype which could be either PRS or PTRS and $y_i$ denote the observed phenotype. Partial $R^2$ (denoted as $\widetilde{R}^2$ below) is defined as the relative difference in sum of squared error (SSE) between two linear models: 1) $y \sim 1 + \text{covariates}$ (null model); and 2) $y \sim 1 + \text{covariates} + \hat{y}$ (full model), *i.e.* $\widetilde{R}^2 = 1 - \frac{\text{SSE}_{\text{full}}}{\text{SSE}_{\text{null}}}$. To enable fast computation, we calculated $\widetilde{R}^2$ using an equivalent formula shown in Eq 3.10 which relies on the projection matrix of the null model.

$$\widetilde{R}^2 = \frac{\mathcal{C}^2(y, \hat{y})}{\mathcal{C}(y, y)\mathcal{C}(\hat{y}, \hat{y})} \tag{3.10}$$

$$\mathcal{C}(u, v) := u^t v - u^t H v, \tag{3.11}$$

where $H$ is the projection matrix of the null model, *i.e.* $H = \widetilde{C}(\widetilde{C}^t\widetilde{C})^{-1}\widetilde{C}^t$ where $\widetilde{C} = [1, C_1, \cdots, C_L]$ with $C_l$ being the $l$th covariate.

As stated in the results section, PTRS weights were computed in the discovery set (UKB EUR) and tested in the 5 target sets. To determine the hyperparameter (p-value cutoff in the clumping and thresholding based approach or $\lambda$ value in elastic net), for each target set, we further split the target set into two equal-size parts, a validation set and a test set. First, we calculated $\widetilde{R}^2$ for all hyperparameters in the validation set and we selected the hyperparameter that maximized the $\widetilde{R}^2$ in the validation set. And then, we calculated the $\widetilde{R}^2$ under the selected hyperparamter in the test set. This procedure was repeated 10 times

and we reported the average $\widetilde{R}^2$ in the test set as the prediction accuracy.

#### 3.1.5.13 Combining PTRS and PRS

We implemented a procedure to combine PTRS and PRS via a weighted sum, i.e. combined score = $c_1$PRS + $c_2$PTRS with $c_1$ and $c_2$ being the extra coefficients to be measured. Given a sequence of PTRSs and PRSs at different hyperparameters (trained with the discovery set, UKB EUR), we used the following procedure to obtain the prediction accuracy in each of the ancestry groups. Similar to the prediction accuracy quantification of PTRS and PRS alone, we first split the target set into validation and test sets. We used the validation set to: i) determine the coefficients ($c_1$ and $c_2$) for each PTRS and PRS combination, and ii) select the best combined score among all PTRS and PRS combinations. And the test set was used to obtain an unbiased measure of the prediction accuracy of the selected score in the corresponding ancestry. To realize the two goals above using the validation set, we further split the validation set into two equal-size parts. Using the first half of the validation set, we determined the coefficients for each of the PTRS and PRS combinations with least squares. Next, using the other half of the validation set, we calculated the prediction accuracy of each PTRS and PRS combination with the coefficients obtained above. We note that this accuracy measure is unbiased since neither PTRS and PRS weights nor the coefficients for combining PTRS and PRS were obtained using this data. Based on these prediction accuracy measures, for each ancestry, we selected the best-performing PRS and PTRS combination along with the corresponding coefficients. The actual prediction performance being reported was calculated using the test set and the combined score being selected above.

#### 3.1.5.14 Quantifying the portability of PRS and PTRS

Portability was defined as the ratio of the prediction accuracy in each target set divided by the prediction accuracy in the European reference set. Therefore, by definition, portability

in the EUR ref. set was 1.

When calculating the portability of PTRS using MESA AFHI transcriptome, we used the MESA EUR model $\widetilde{R}^2_{\text{EUR ref.}}$ as the reference. This is a conservative choice since MESA EUR model is expected to perform better than MESA AFHI model among EUR individuals.

### *3.1.6 Supplementary Tables*

| **UKB Field Description** | **UKB Field ID** | **Tag** | **Phenotype Category** |
|---|---|---|---|
| Standing height | 50 | Height | Height |
| Diastolic blood pressure, automated reading | 4079 | DBP | Blood pressures |
| Systolic blood pressure, automated reading | 4080 | SBP | Blood pressures |
| Body mass index (BMI) | 21001 | BMI | BMI |
| White blood cell (leukocyte) count | 30000 | WBC | Blood cell counts |
| Red blood cell (erythrocyte) count | 30010 | RBC | Blood cell counts |
| Haemoglobin concentration | 30020 | Hb | Haemoglobin related |
| Haematocrit percentage | 30030 | Ht | Haemoglobin related |
| Mean corpuscular volume | 30040 | MCV | Haemoglobin related |
| Mean corpuscular haemoglobin | 30050 | MCH | Haemoglobin related |
| Mean corpuscular haemoglobin concentration | 30060 | MCHC | Haemoglobin related |
| Platelet count | 30080 | Platelet | Blood cell counts |
| Lymphocyte count | 30120 | Lymphocyte | Blood cell counts |
| Monocyte count | 30130 | Monocyte | Blood cell counts |
| Neutrophill count | 30140 | Neutrophil | Blood cell counts |
| Eosinophill count | 30150 | Eosinophil | Blood cell counts |
| Basophill count | 30160 | Basophil | Blood cell counts |

**Table 3.1: Meta information of the phenotypes retrieved from UK Biobank which were used in the analysis.** The "Tag" column shows the short name of the phenotyes used in this paper. And phenotypes are assigned into five categories which are shown in "Phenotype Category" column

| **Ancestry** | **Number of individuals** |
|---|---|
| AFR | 6413 |
| EUR | 356476 |
| E.ASN | 1326 |
| S.ASN | 6479 |

**Table 3.2: Number of individuals included in the analysis stratified by ancestry.**

| Method | Data source | Population | Tissue | Number of genes | Sample size | Tag |
|---|---|---|---|---|---|---|
| CTIMP | GTEx V8 | European | Adipose_Subcutaneous | 9228 | 491 | |
| CTIMP | GTEx V8 | European | Artery_Tibial | 9027 | 489 | |
| CTIMP | GTEx V8 | European | Breast_Mammary_Tissue | 8127 | 337 | |
| CTIMP | GTEx V8 | European | Cells_Cultured_fibroblasts | 8731 | 417 | |
| CTIMP | GTEx V8 | European | Lung | 8954 | 444 | |
| CTIMP | GTEx V8 | European | Muscle_Skeletal | 7671 | 602 | |
| CTIMP | GTEx V8 | European | Nerve_Tibial | 10184 | 449 | |
| CTIMP | GTEx V8 | European | Skin_Sun_Exposed_Lower_leg | 9474 | 517 | |
| CTIMP | GTEx V8 | European | Thyroid | 9827 | 494 | |
| CTIMP | GTEx V8 | European | Whole_Blood | 7041 | 573 | GTEx EUR |
| Elastic Net | MESA | | Monocyte | 4670 | 578 | MESA EUR |
| Elastic Net | MESA | African American or Hispanic | Monocyte | 5554 | 585 | MESA AFHI |

**Table 3.3: Meta information of the prediction models used in the analysis.** The highlighted prediction models were used to build PTRS. The "Tag" column shows the short name of the models used in this paper.

### *3.1.7 Supplementary Figures*

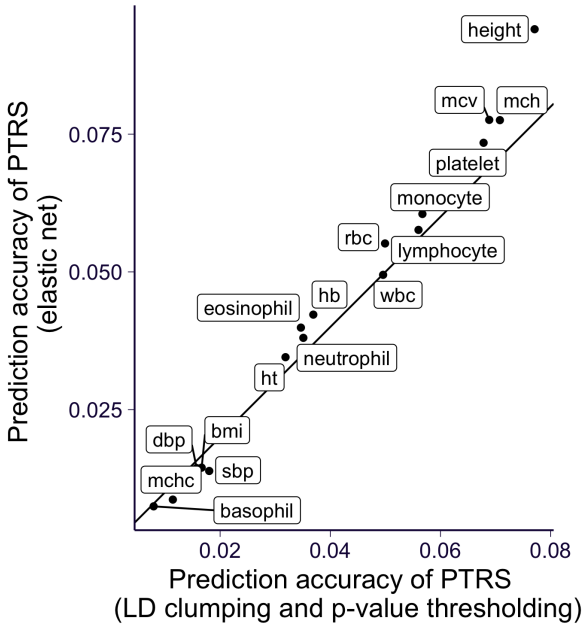



**Figure 3.6: Prediction accuracy of PTRS built with the elastic nets vs the LD clumping and p-value thresholding approach.** The prediction accuracy of PTRS built with the LD clumping and p-value thresholding approach was shown on x-axis. And the accuracy of PTRS built with the elastic net was shown on y-axis. The PTRS construction was based on the transcriptome models from GTEx EUR whole blood samples.

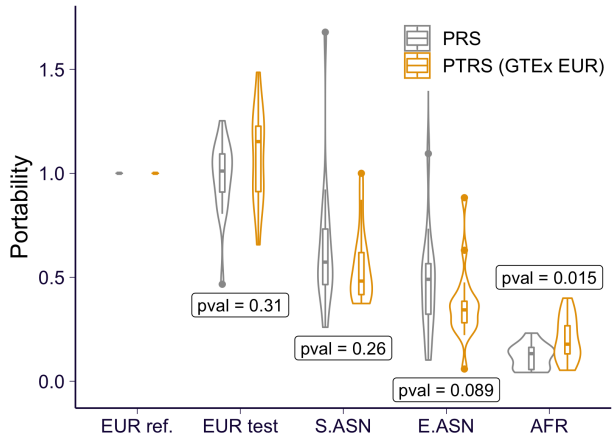



**Figure 3.7: Portability of LD clumping and p-value thresholding based PTRS for 17 quantitative phenotypes in UK Biobank.** The portability of clumping and thresholding based PTRS trained and calculated using GTEx EUR whole blood samples are shown in yellow with the PRS shown in gray. 'EUR ref.' set is used as the reference population in the calculation of portability so that the portability is always 1.

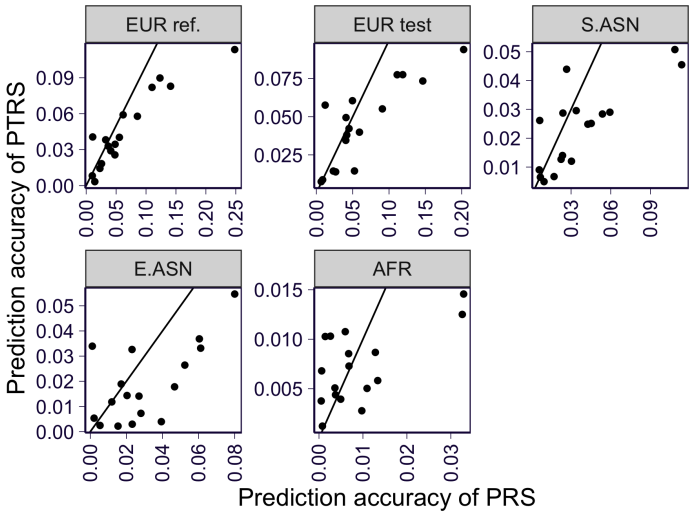



**Figure 3.8: Prediction accuracy of PTRS vs PRS in all ancestral groups.** Prediction accuracy, measured by partial $\tilde{R}^2$, of PTRS (on y-axis) was compared to the accuracy of PRS (on x-axis). Each panel corresponds to each target set. The PTRS construction was based on the transcriptome models from GTEx EUR whole blood samples.

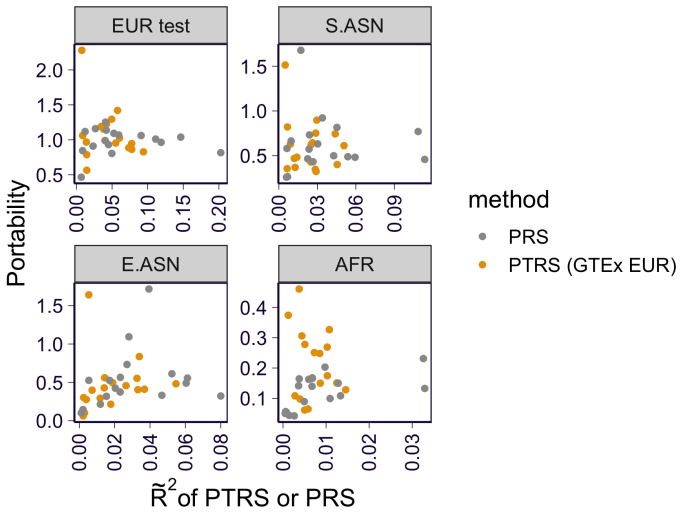
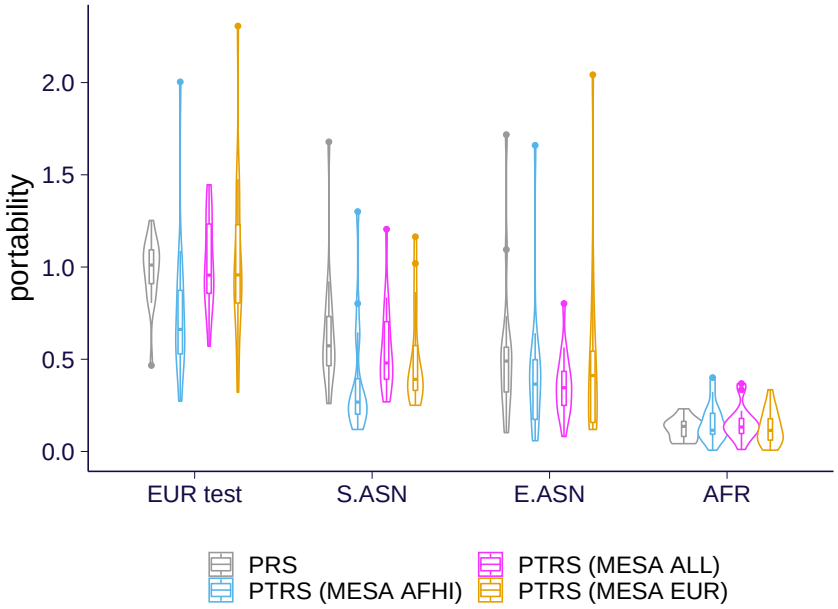
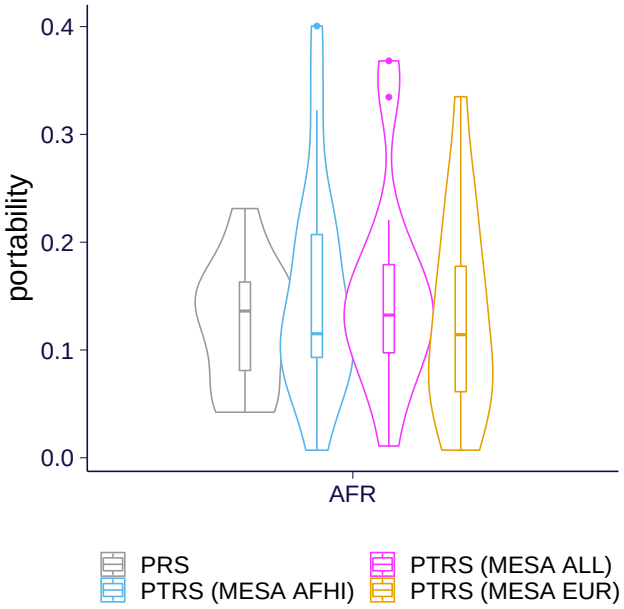



**Figure 3.9: Prediction accuracy vs portability of PTRS in all ancestral groups.** Portability of PTRS (y-axis) was compared to the prediction accuracy, measured by partial $\widetilde{R}^2$, of PTRS (on x-axis). Each panel corresponds to each target set. The PTRS construction was based on the transcriptome models from GTEx EUR whole blood samples.

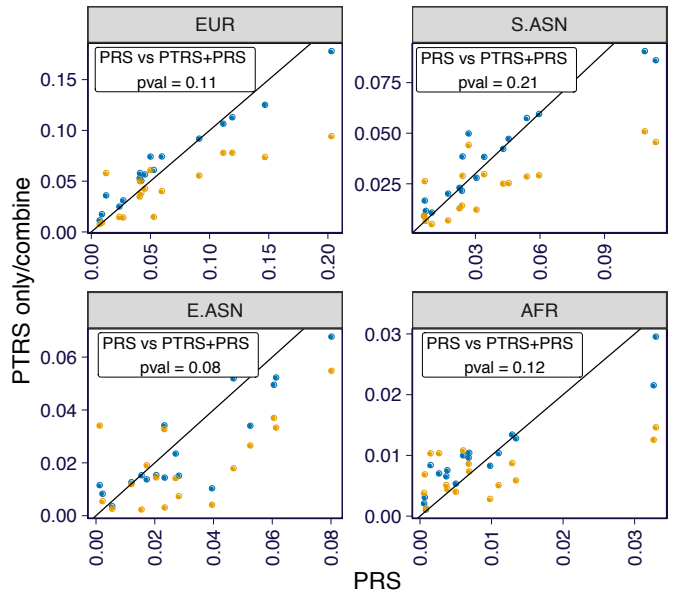



**(A)**




**(B)**

**Figure 3.10: Portability of PRS and MESA-based PTRSs.** The results of the 17 quantitative traits are summarized in the violin and box plots for each of the score types. **(A)** Results in all ancestry groups are shown. **(B)** A zoom-in plot focusing on results in African ancestry.

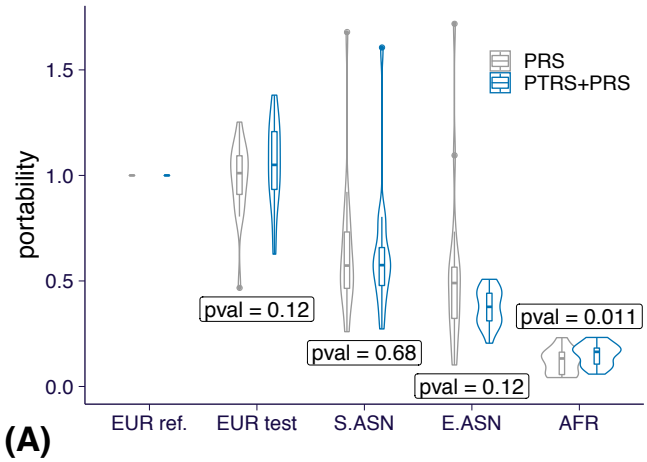



**(A)**

**(B)**

| sample | R2 (PRS – PTRS) | wilcox_pval |
|---|---|---|
| AFR | –0.000937 | 0.1200 |
| EUR test | –0.004660 | 0.1090 |
| EUR ref. | –0.000912 | 0.3060 |
| E.ASN | 0.004920 | 0.0797 |
| S.ASN | –0.001040 | 0.2070 |

**Figure 3.11: Prediction accuracy of the score combining PTRS and PRS.** Combing the clumping and thresholding-based PTRS and PRS, the results on the prediction accuracy are shown below. **(A)** The prediction accuracy of the PRS is shown on x-axis and it is compared against the prediction accuracy of the PTRS (yellow) or the combined score on y-axis. The results on all of the 17 quantitative traits are shown. Each panel corresponds to one ancestry group. The p-values are for comparing PRS accuracy versus the combined score accuracy via the paired Wilcoxon signed rank test. **(B)** A summary of the difference between the prediction accuracy of PRS and the combined score is shown for each of the ancestry group. The second column shows the mean difference and the third column shows the results of the paired Wilcoxon signed rank test comparing the accuracy of PRS versus the combined score.




| sample | portability (PRS – PTRS) | wilcox_pval |
|---|---|---|
| AFR | –0.03280 | 0.011 |
| EUR test | –0.07350 | 0.120 |
| E.ASN | 0.15900 | 0.120 |
| S.ASN | 0.00408 | 0.678 |

**(B)**

**Figure 3.12: Portability of the score combining PTRS and PRS.** Combing the clumping and thresholding-based PTRS and PRS, the results on the portability are shown below. **(A)** The prediction accuracy of the PRS is shown on x-axis and it is compared against the prediction accuracy of the PTRS (yellow) or the combined score on y-axis. The results on all of the 17 quantitative traits are shown. Each panel corresponds to one ancestry group. The p-values are for comparing PRS accuracy versus the combined score accuracy via the paired Wilcoxon signed rank test. **(B)** A summary of the difference between the prediction accuracy of PRS and the combined score is shown for each of the ancestry group. The second column shows the mean difference and the third column shows the results of the paired Wilcoxon signed rank test comparing the accuracy of PRS versus the combined score.

## 3.2 Obtaining PTRS from GWAS summary statistics

### *3.2.1 Abstract*

In the previous work (Section 3.1 and [83]), we proposed the elastic net-based PTRS (EN-PTRS) and examined its performance in terms of the predictive performance and the portability across populations. However, we trained EN-PTRS using individual-level data which largely limited the utility of PTRS in practice. This is because, most of the time, large

GWAS results only share summary statistics rather than individual-level information. To resolve this issue, we extend lassosum [93] in the context of PTRS which enables EN-PTRS to be built on the basis of GWAS summary statistics and a reference LD panel.

### *3.2.2 Introduction*

The performance of polygenic risk scores (PRSs) depends on the sample size, under the current availability of GWAS cohorts (tens to hundreds of thousands), the more samples being used in PRS training, the higher predictive power will the PRS has. However, it creates challenges on sharing and handling such a big amount of GWAS data which involves security issues, data storage and transferring burdens, computational costs, and etc. Luckily, many of the PRS methods [144, 93, 46, 88] are linear predictors so the PRS training only depends on two quantities (i.e. sufficient statistics): i) the marginal associations between each genetic variant and the phenotype (GWAS summary statistics) and ii) the correlation/covariance between genetic variations (LD information). For these PRS approaches, to develop summary statistics-based implementations is of great importance with twofold reasons. First, it avoids the requirement of individual-level data which simplifies the data acquiring process in terms of both permission and data transferring/downloading. Second, it bypasses some heavy computation which involves individual-level data. In this work, we propose a summary statistics-based implementation of elastic net-based PTRS which carries out EN-PTRS weights using GWAS summary statitstics and a reference LD panel. We call this summary statistics-based implementation of elastic net-based PTRS as S-EN-PTRS in short.

### *3.2.3 Results*

#### 3.2.3.1 Method overview

The S-EN-PTRS implementation is inspired by [93] which builds lasso/elastic net-based PRSs from GWAS summary statistics. To adapt [93] method for solving EN-PTRS, we first obtain gene-level summary statistics by applying S-PrediXcan [7] to GWAS summary statistics. We note that the S-PrediXcan results are equivalent to the marginal associations between each of the predicted gene expressions and a phenotype. Next, similar to lassosum, we carry out the training of a sequence of elastic net models (along the regularization path [40]) predicting the phenotype from predicted expressions by utilizing the gene-level marginal associations and gene/gene covariances (see details in Section 3.2.5.1). We provide an out-of-box implementation of the proposed method which takes S-PrediXcan results along with a reference LD panel and generates the corresponding EN-PTRS weights. The software is available at `https://github.com/liangyy/SPrediXcan2PTRS`. We also implemented a clumping-based PTRS (clump-PTRS) using the same interface (Section 3.2.5.2).

#### 3.2.3.2 S-EN-PTRS is sensitive to LD panel

| trait short name | ukb field | gwas id | gwas reference | gwas sample size |
|---|---|---|---|---|
| eosinophil | 30150 | Astle_et_al_2016_Eosinophil_counts | [5] | 173480 |
| lymphocyte | 30120 | Astle_et_al_2016_Lymphocyte_counts | [5] | 173480 |
| monocyte | 30130 | Astle_et_al_2016_Monocyte_count | [5] | 173480 |
| neutrophil | 30140 | Astle_et_al_2016_Neutrophil_count | [5] | 173480 |
| platelet | 30080 | Astle_et_al_2016_Platelet_count | [5] | 173480 |
| rbc | 30010 | Astle_et_al_2016_Red_blood_cell_count | [5] | 173480 |
| wbc | 30000 | Astle_et_al_2016_White_blood_cell_count | [5] | 173480 |
| bmi | 21001 | GIANT_2017_BMI_Active_EUR | | 109433 |
| height | 50 | GIANT_HEIGHT | [158] | 253288 |
| sbp | 4080 | ICBP_SystolicPressure | [35] | 203056 |
| dbp | 4079 | ICBP_DiastolicPressure | [35] | 203056 |

**Table 3.4: Information on the 11 quantitative traits being used for examining the performance of PTRS.** We used the GWAS harmonized in [6]. Column "gwas id" indicates the GWAS ID used in [6]. Column "gwas reference" shows the reference to the GWAS. Column "ukb field" shows the Field ID in the UK Biobank

To examine the performance of S-EN-PTRS, we applied our implementation to 11 quantitative traits (Table 3.4 and Section 3.2.5.3). We evaluated the performance of S-EN-PTRS using a random subset of 5,000 participants in UK Biobank who are of European ancestry (Section 3.2.5.4). To see the benefit of considering all the genes jointly rather than using marginal effects only, we calculated a p-value thresholding-based PTRS (naive-PTRS). Specifically, for a given p-value cutoff, naive-PTRS utilizes all the significant genes. We run S-EN-PTRS under a range of offset values (0.01, 0.1, 0.3, 0.5, 0.7, 0.9) and $\alpha = 1$. We also run clump-PTRS with squared correlation cutoff = 0.1 and a sequence of p-value cutoffs (applied after clumping), $10^{-7}$, $10^{-6}$, $10^{-5}$, $10^{-4}$: $10^{-3}$, 0.005, 0.01, 0.05, 0.1, 0.5, 1.

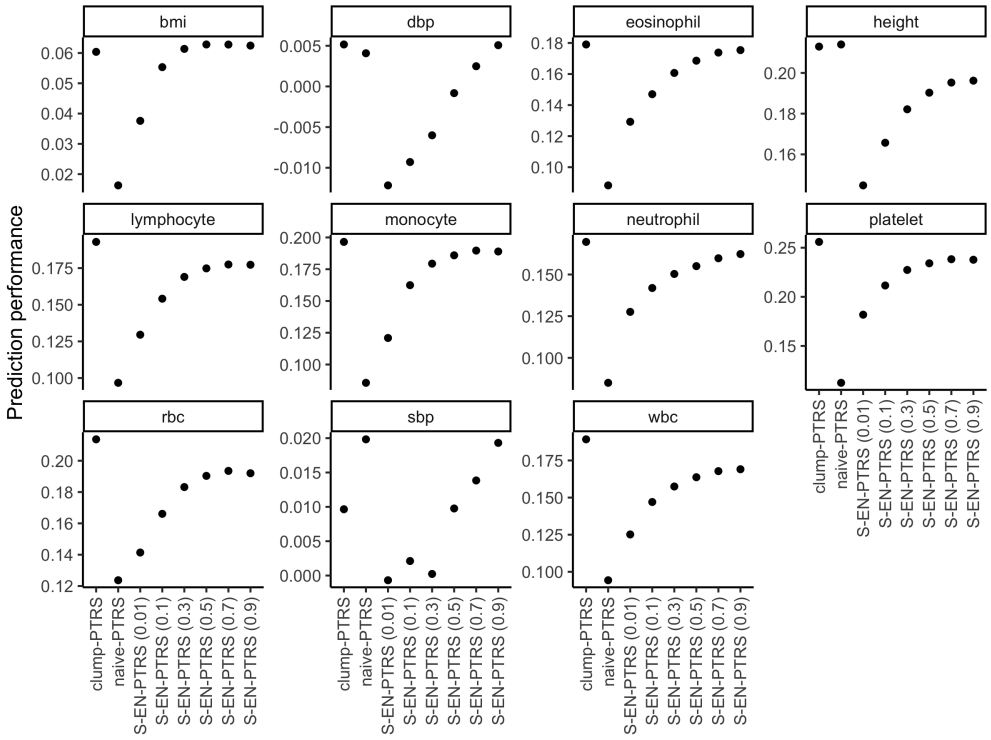



**Figure 3.13: Performance of S-EN-PTRS, clump-PTRS, and naive-PTRS.**

As shown in Figure 3.13, among most of these 11 quantitative traits, the performance ordering is clump-PTRS performs > S-EN-PTRS > naive-PTRS. This result is not consistent with the observation that EN-PTRS outperforms clump-PTRS when using individual level data for training [83]. Since here we used out-of-sample LD panel, such discrepancy in performance could be attributed to the LD mismatch. Moreover, we observed that S-EN-PTRS performance increases as the offset value becomes bigger. Since the offset value measures the tendency to treat all these genes as independent predictors as opposed to take LD information into account, this result also suggests that S-EN-PTRS is potentially sensitive to LD panel.

### *3.2.4 Discussion*

In this work, we proposed and implemented two summary statistics-based PTRS approaches: i) elastic net-based PTRS (S-EN-PTRS); ii) clumping-based PTRS (clump-PTRS). We examined the performance of these summary statistics-based PTRSs on 11 quantitative traits while using out-of-sample LD panel. We observed that, with out-of-sample LD panel, clump-PTRS outperforms S-EN-PTRS even though it's been reported that EN-PTRS outperforms clump-PTRS when using individual-level data. Such discrepancy in performance could be attributed to the LD mismatch and it indicates that, comparing to clump-PTRS, S-EN-PTRS is more sensitive to the accurary of LD panel. In practice, if high-quality LD panel (ideally in-sample LD) is available, it is recommended to use S-EN-PTRS; otherwise, clump-PTRS is preferred and one should use S-EN-PTRS with cautious. With mismatched LD, one could increase the offset parameter to reduce the effect of mismatch LD on prediction performance.

### *3.2.5 Methods*

#### 3.2.5.1 Details of S-EN-PTRS implementation

For $N$ samples and $K$ genes, let $y \in \mathbb{R}^{N\times 1}$ represent the phenotype and $G \in \mathbb{R}^{N\times K}$ be the predicted expression matrix. As described in Section 3.1.5.9, EN-PTRS is defined by the following problem:

$$\arg\min_{x} \|y - Gx\|_2^2 + \lambda_0 \cdot (\alpha\|x\|_1 + \frac{1-\alpha}{2}\|x\|_2^2) \tag{3.12}$$

The EN-PTRS weights correspond to the solution to Eq 3.12.

We note that the amount of the regularization on each gene depends on the scaling of the corresponding column in $G$. To resolve such ambiguity, we further assume that each column of $G$ is standardized and so is $y$. With this assumption, implicitly, the gene-level effect of gene $k$ ($x_k$) quantifies the amount of phenotypic increase (in the unit of standard

deviation of the phenotype) under one unit increase of the predicted expression level of gene $k$ (e.g. the predicted expression goes from the population mean to the population mean plus standard deviation).

Under this scaling, Eq 3.12 is equivalent to

$$\arg\min_x x'Rx - 2b'x + \lambda \cdot (\alpha\|x\|_1 + \frac{1-\alpha}{2}\|x\|_2^2) \tag{3.13}$$

$R$ is the sample correlation matrix of $G$ and $b = z_{\text{PrediXcan}}/\sqrt{N}$. Since, equivalently, we use GWAS cohort for EN-PTRS training, $N$ is essentially the GWAS sample size. And $\lambda = \lambda_0/N$.

In practice, we don't have access to $R$ (the in-sample sample correlation). Instead we calculate the sample correlation using a reference LD panel as $\widetilde{R}$. When the GWAS cohort and the reference LD panel are from the same ancestry group, $R \approx \widetilde{R}$ especially when both GWAS and reference LD panel have a lot of samples. In practice, similar to lassosum [93], to stabilize the model fitting, we add an offset term to $\widetilde{R}$ when approximating $R$. In other words, we let $R = (1-o)\cdot\widetilde{R} + o \cdot I$.

We treat $o \in (0,1)$, $\alpha \in [0,1]$, and $\lambda \in (0,\infty)$ as hyperparameters. For a given pair of $o$ and $\alpha$, we determine the maximum $\lambda$ by solving for the minimum possible $\lambda$ satisfying the KKT condition when $x = 0$ is the solution ([40]). In other words,

$$\lambda_{\max} = \arg\min_\lambda\{\lambda : |2R_{k\cdot}\cdot\vec{0} - 2b_k| \le \alpha\lambda,\ \forall k = 1,\cdots,K\} \tag{3.14}$$

$$= \frac{2\cdot\max_k|b_k|}{\alpha} \tag{3.15}$$

We solve for a sequence of equally spaced (in log-scale) $\lambda$ values from $\lambda_{\max}$ to $\lambda_{\min}$ and usually $\lambda_{\min} = \lambda_{\max}/100$.

Similar to [40], we solve Eq 3.13 by coordinate descent. For the sequence $\lambda_1 = \lambda_{\max}, \lambda_2, \cdots, \lambda_J = \lambda_{\min}$, we first set $x = \vec{0}$ for $\lambda = \lambda_{\max}$. And then, for each $\lambda = \lambda_j$, we initialize $x$

as the solution from $\lambda = \lambda_{j-1}$.

Additionally, to obtain $\widetilde{R}$, we need information more than the one required for S-PrediXcan since we need to know not only the variant/variant covariance for each variant pair within a gene but the variant pair from two different genes. We note that it is not suitable to pre-compute gene covariance since, usually, the GWAS may miss some genetic variants in the prediction models and we should exclude these variants for the PTRS training. Instead, similar to S-PrediXcan implementation, we calculate $\widetilde{R}$ on the fly using the pre-computed variant/variant covariance from the reference LD panel. In practice, we consider gene/gene covariance for each chromosome and assume the covariance between genes from different chromosomes equals to zero. With this assumption, we need to pre-compute and store variant/variant covariance for all the variants that appear in at least one of the prediction models and lie in the same chromosome. For chromosome 1 to 22, the number of variants to consider ranges from 4,000 to 30,000, for which we need to store all the pair-wise information. In fact, the number of samples in the reference LD panel $N_{\text{LD}}$ is usually smaller than the number of variants being considered $N_{\text{SNP}}$ and it is order of magnitude smaller when using GTEx data as the reference (sample size is about hundreds) The $N_{\text{SNP}}$-by-$N_{\text{SNP}}$ variant/variance covariance matrix can be represented by only $N_{\text{LD}}$ eigenvalues and eigenvectors. We make use of this property and save the eigenvalue decomposition of the variant/variant covariance matrix instead.

In principle, we need to loop over the genome-wide variants until convergence. But here, we also implement an alternative scheme in which we loop over variants within each chromosome (this is similar to lassosum [93] which solves elastic net models for each LD block while using the same $y$ for all blocks). We note that this chromosome-wise approach usually gives similar results as the genome-wide iteration and converges a little faster.

### 3.2.5.2 Details of clump-PTRS implementation

Given a list of genes with PrediXcan z-scores and $r^2$, a cutoff on squared correlation between genes, we want to perform clumping such that genes with duplicated signals (genes which are "highly" correlated with a more significant gene) are removed. The procedure is similar to Section 3.1.5.10 and it is outlined in Algorithm 3.

**Algorithm 3:** Clumping procedure

**Input :** Z-scores for a list of predictor candidates $z \in \mathbb{R}^{K\times 1}$, correlation between predictors $R \in \mathbb{R}^{K\times K}$, a squared correlation cutoff $r^2$.

**Output:** A list of predictors (in terms of index) among which $R^2_{ij} < r^2$ and the ones with high $|z|$ should always be preferred.

```
Function Clumping(z, R, r^2):
    sortedIndex = argsort(|z|, decreasing);
    Initialize a Hash table statusHash with keys 1, ···, K and value = 0;
    Initialize outputList as empty;
    for j ← 1 to K do
        i = sortedIndex[j];
        if statusHash[i] = -1 then
            continue;
        Append i to outputList;
        for k ← j to K do
            if statusHash[k] = -1 then
                continue;
            end
            if R^2_ik ≥ r^2 then
                statusHash[k] = -1;
            end
        end
    end
    return outputList
End
```

### 3.2.5.3 Applying S-EN-PTRS to 11 GWASs

We applied S-EN-PTRS to 11 GWASs listed in Table 3.4. These GWASs were harmonized previously and imputed in GTEx v8 WGS samples [6]. We used the CTIMP models, cross-

tissue gene expression imputation models [55], from Whole Blood which were trained on GTEx v8 European samples [8]. The variant/variant covariance was obtained from GTEx v8 European samples.

#### 3.2.5.4 Evaluating PTRS performance using UK Biobank data

We evaluated the prediction performance by calculating the Spearman correlation between the predicted and observed phenotype values on 5,000 randomly selected UK Biobank participants with European ancestry. To take covariates into account, we regressed out top 10 genetic PCs, age, and sex (and second order terms of age and sex) from both the observed and the predicted values before calculating the correlation. The corresponding UK Biobank phenotypes used for the evaluation are shown in Table 3.4.

To select the hyperparameters, we split the selected participants into two halves. We used the first half samples to determine the hyperparameters ($\lambda$ for S-EN-PTRS and the p-value cutoff for clump-PTRS) by selecting the best-performing models. And then we used the second half of the samples to calculate the prediction performance under the models with the selected hyperparameters.

# CHAPTER 4

# EXPLORING METHODS TO LEVERAGE INFORMATION IN PARENTAL PHENOTYPES TO FACILITATE GWAS

## 4.1 Abstract

To recruit cases for late-onset disease study is challenging since these diseases occur in elder people. Moreover, typically we have very limited number of late-onset disease cases in Biobank data. But, on the other hand, the parental disease status may be available by questionnaire. Because of this, methods have been developed to utilize parental disease status instead [86, 56]. In these approaches, the late-onset phenotype of the participant is imputed from parental statuses. And, at downstream, genome-wide association study (GWAS) is performed using the participant's genotype and imputed phenotype. In this paper, we take another view on utilizing parental phenotypes. We treat this problem as missing parental genotype rather than missing participant's phenotype. First, we propose an imputation scheme to infer the parental origin of the participant's genotype from a collection of extra parental phenotypes (non-focal phenotypes) and participant's genotype. Second, we propose a computationally efficient approach to incorporate the imputed parental origin information into the downstream GWAS. We explore the feasibility of the proposed two-step approach on simulated and real data. And we derive the power increase of GWAS as a function of imputation quality. These results indicates that the imputation scheme needs about 100 non-focal phenotypes to achieve enough accuracy to facilitate the GWAS in downstream.

## 4.2 Introduction

Late-onset diseases, by definition, occur in elder people, which increases the difficulty to recruit cases for GWAS. Nowadays, biobank data has become an important data source for

GWAS because of its ability to collect genome and phenome of a large population. And the GWASs of some complex diseases with relative high prevalence in a general population have achieved great successes in utilizing biobank-scale data set. However, the biobank usually has limited number of the late-onset disease cases due to the fact that the participants are typically at the middle-age and healthy. For instance, looking at late-onset Alzheimer's disease in UK Biobank. There are about 20% of UK Biobank participants older than 65 at the age of recruitment [132] and around hundreds of late-onset Alzheimer's disease cases. The reason of short of cases is two-fold. First, there are only about 100,000 elder people to start with without specifically targeting Alzheimer's disease research. Second, the recruitment could be biased towards healthier participants, which may give rise to a even lower prevalence than the Alzheimer's diease prevalence in a general population [59].

On the other hand, the parents of participants usually were or have been at the age of high prevalence of late-onset disease. It means that the parents of participants are more ideal objects to work with since the disease of interest is more prevalence. And when looking at the participant's parents, we don't need to limit the scope to elder participants. Moreover, in principle, the parental case/control status of the late-onset disease is possibly achievable from the participants via questionnaire. For example, UK Biobank has collected 12 parental phenotypes via web-based questionnaires, and, in particular, there are over 26,000 paternal Alzheimer's disease cases and over 49,000 maternal Alzheimer's disease cases [132]. In comparison, a recent GWAS which is dedicated to late-onset Alzheimer's disease [71] have around 35,000 cases.

Motivated by the potential richness of late-onset disease cases in parental phenotypes, methods have been developed to leverage parental phenotypes for the corresponding GWAS. In this context, the challenge is that we only have access to the genotype of the child (the participant) and the phenotype of the parents. The existing approaches treat it as a missing phenotype problem, in which they first impute the missing phenotype. By missing

phenotype, it means that the phenotype of the child is not yet observable since the participant has not reached the typical age of onset. And the subsequent GWAS is performed using child's genotype and imputed phenotype. To impute the child's phenotype from the parental phenotype, GWAX was proposed to use the parental phenotype as a proxy of the child's phenotype [86]. More recently, [56] imputed the genetic liability of the child from the family disease history by modeling case/control phenotype with the liability threshold model.

In this paper, we want to take another view of the problem. Here, we treat it as a missing genotype problem instead. Specifically, we consider parent's genotype as missing and we want to, first, impute parent's genotype and, second, perform GWAS with observed parental phenotype along with the imputed genotype of the parents. Since one of the child's alleles is from father and the other is from mother, essentially, the imputation is about inferring which allele is from father and which allele is from mother. Moreover, considering that we typically have access to phased genotype, the imputation problem could be further reduced to inferring the parental origin of the two haplotypes of the child. To infer the haplotype origin, we need to leverage the information buried in the parental phenotypes other than the one of interest (we call them non-focal phenotypes below). In particular, each haplotype carries the genetic risks of these non-focal phenotypes, which could provide information on which haplotype is more likely to come from which parent. Intuitively, if the genetic risks carried by one haplotype resemble the corresponding observed maternal phenotypes, then, this haplotype is more likely to come from mother as compared to father. A concrete example of the intuition is shown in Figure 4.1.

Subsequent to the imputation of haplotype origin, we propose an approach to integrate the imputed haplotype origin into the GWAS of the focal phenotype, which uses parental phenotypes and child's haplotypes. Our proposed approach includes GWAX approach as a special case at which the two child's haplotypes are both equally likely to come from each of the two parents. When the quality of the imputation is good, the proposed approach has

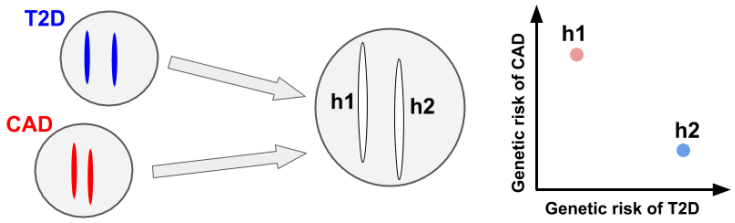



**Figure 4.1: An example on imputing haplotype origin from non-focal phenotypes of parents and genetic risks carried in child's haplotypes.** In this figure, a circle represents a cell with two chromosomes inside representing the sister chromosomes of a diploid (in practice, there are 23 pairs and here we only show one chromosome for the sake of simplicity). On the left, the two circles represent the cell from the father and the mother respectively with the chromosomes in father colored in blue and the ones in mother colored in red. And suppose the father is observed to have type II diabetes (T2D) and the mother is observed to have cardiovascular disease (CAD). The circle in the middle represents the cell of the child (the participant in the biobank) where the two chromosomes are labelled as "h1" and "h2". These chromosomes are uncolored since the haplotype origin is unknown and is the goal of the imputation. On the right, the plot shows the genetic risk of T2D and CAD carried in "h1" and "h2". Since "h1" carries higher CAD risk and lower T2D risk, which resembles the paternal phenotype, "h1" is colored with light red indicating that it is more likely that "h1" is from the mother. For similar reason, "h2" is colored with light red.

higher power to identify GWAS signals than GWAX.

In the following sections, we first describe the imputation scheme for the inference of haplotype origin (Section 4.5.1). Secondly, we describe the proposed GWAS approach that integrates imputation results and give the theoretical results on how the imputation quality affects the power of the downstream GWAS (Section 4.5.2). And then, we perform simulation to examine the imputation quality at different parameter settings, such as the number of non-focal phenotypes and the heritability (Section 4.3.1). Similarly, on the basis of simulated data, we test the proposed GWAS approach and verify the theoretical results (4.3.2). Following the analysis on simulated data, we apply the imputation scheme to trios in the Framingham transcriptome data and perform downsampling analysis to examine the imputation quality on real data (Section 4.3.3). Finally, we summarize the observations and the insights obtained from the theoretical results along with the applications to simulated and real data, and discuss the potential pitfall and future direction (Section 4.4)

## 4.3 Results

### *4.3.1 Examining imputation quality on simulated data*

As described in Section 4.5.1, we proposed two genetic models (Eq 4.5 and 4.6) to model parental phenotype $y^{\pi}$ given parental half-genotype $H^{\pi,1}$. To evaluate these approaches, we examined the imputation performance of these approaches on simulated data where $Z_i$ was set to 1 for all individuals (Section 4.5.3). We also included an "ideal" approach in the comparison which is, in principle, the "best" imputation we could get under the current model assumption. In the "ideal" approach, we pretend that both the genetic effect and the variance of the environmental effect are observed, which means that $\theta$ is known. So, the posterior of $Z$ can be evaluated under the true parameters, which bypass the uncertainty introduced by fitting model parameter via EM iteration.

We simulated data under different heritabilities and the number of phenotypes included in the simulation also varied. In Figure 4.2, we show the imputation accuracy (in terms of the probability that the posterior of $Z$ is correct) of the three approaches. As the number of phenotypes increases, the imputation accuracy increases for all the three approaches. Similarly, as phenotypes used for imputation are more heritable, *i.e.* the genetic effect captures more observed phenotypic variation, the imputation is more accurate. As expected, the "ideal" approach performs better than the other approaches in all settings. But, across all settings, there are some individuals with accuracy close to zero, especially when the number of phenotype is greater than 10 and heritability is between 0.01 and 0.25. These low-accuracy imputations also appear in the "ideal" approach which suggests that it is due to the intrinsic uncertainty buried in the data.

Comparing the "ideal" approach and the PRS-based approach, they have similar performance with PRS-based approach performs slightly worse which could be attributed to the noise introduced by PRS training. However, the on-the-fly approach performs differently,

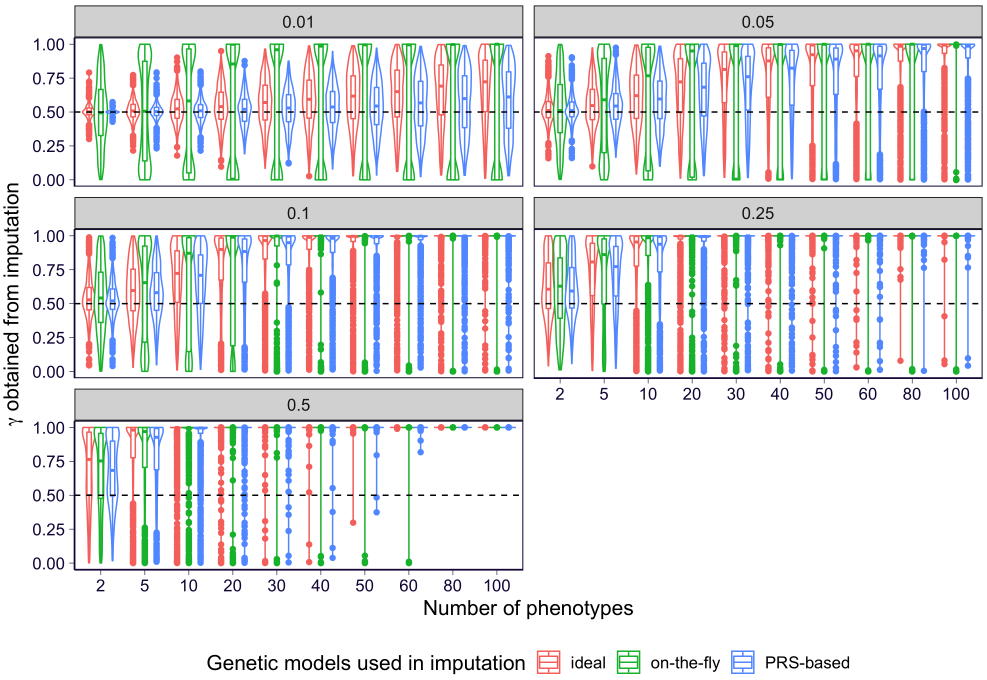



**Figure 4.2: The imputation performance on the basis of different genetic models.** The imputation performance under a different heritability is shown in each panel. Within each panel, the number of phenotypes included in the imputation is shown on x-axis and the imputation accuracy (the probability that the posterior $Z$ is correctly assigned) is shown on the y-axis. The violin/boxplot contains the results on all of the 1,000 individuals included in the imputation. The results of the "ideal", on-the-fly, and PRS-based (PRS trained with sample size 10,000) approaches are colored in red, blue, and green respectively.

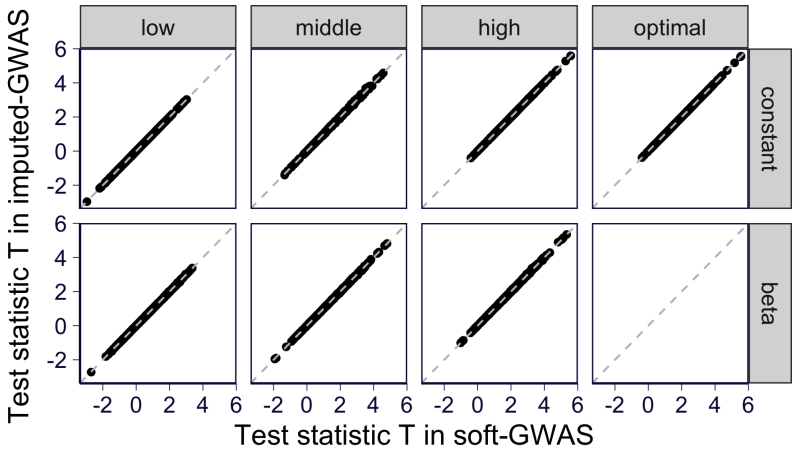
especially when the heritability is low (0.01 to 0.1). In particular, the on-the-fly approach has close to zero accuracy on many samples, which is most sever when the number of phenotypes is low or the heritability is low. Such bad performance could be attributed to the overfitting issue since the on-the-fly approach has the most flexible genetic model among the three. But, notice that, when the phenotype is highly heritable (stronger signal in the data) or there are many phenotypes being included (the genetic model is more constrained) in the imputation, the on-the-fly approach starts to perform similarly to the other two. And, moreover, the performance of the on-the-fly approach is even better than the PRS-approach when the phenotype is highly heritable. This is due to two reasons. First, the performance of the PRS-based approach is limited by the external PRS model and when there are more samples used in PRS training, the PRS-based approach performs better (Figure 4.6). And, secondly, the on-the-fly approach may learn a better genetic model via the EM iteration.

Then, we performed a another simulation to examine the utility of introducing non-negative coefficient in the PRS-based approach (Eq 4.6). Here we simulated data under a specific application context where the parental phenotypes are transcriptome profile (Section 4.5.3). The non-negative constraint on the coefficient $b$ can reduce the amount of error when the true $b$ is non-negative but not so when the $b$ has random or negative sign (Figure 4.7). In practice, if the PRS is at least able to predict the sign to some degree, the non-negative constraint is helpful in the sense of reducing the amount of overfitting.

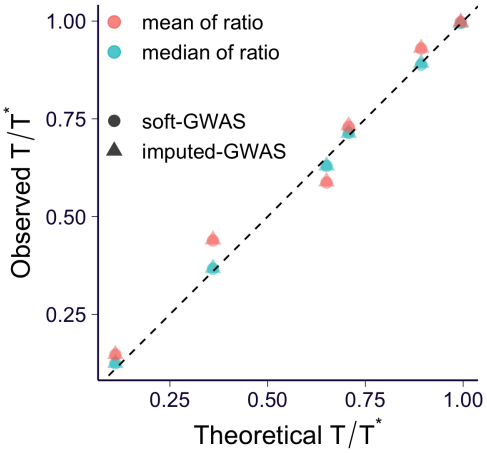
### *4.3.2 Verifying the proposed GWAS approach on simulated data*

As described in Section 4.5.2, we proposed two approaches to carry out association study between phenotype and genotype with the imputation results, posterior of $Z$ (also called $\gamma$), being integrated. One of the approach, soft-GWAS, estimates the effect size via EM iteration where the haplotype imputation result is treated as a prior which corresponds to the hidden variable in the EM. And the statistical significance of this approach is obtained

from likelihood ratio test. In the other approach, imputed-GWAS, we run association test between the posterior haplotype and the phenotype.




**(A)**




**(B)**

**Figure 4.3: The results of the proposed GWAS approaches, imputed-GWAS and soft-GWAS, on the simulated data. (A)** The test statistics are shown by panels where each panel presents the results under a specific $\gamma$ distribution (the distribution type is organized in rows and the accuracy of $\gamma$ is organized in columns). The test statistics of soft-GWAS are shown on x-axis and the ones of the imputed-GWAS are shown on y-axis. The gray dashed line is $y = x$. **(B)** The observed and theoretical relative power are shown. For each simulation setting, the theoretical result is calculated based on Eq 4.14. And the observed ratio is calculated for all replications separately and the median (in green) and mean (in red) across all replications are shown. Circle points indicate the results from soft-GWAS and the triangular ones indicate the results from imputed-GWAS. The black dashed line is $y = x$.

To examine the performance of the two approaches, we simulate some data and run these methods with various kinds of distribution on $\gamma$ which correspond to various imputation qualities (Section 4.5.4). We included two types of distributions for $\gamma$ where $\gamma$ is either a constant across individuals or drawn from a Beta distribution (labeled as "constant" and "beta"). Moreover, to consider different imputation qualities, we sampled $\gamma$ under three scenarios: 1) $\gamma$ is at low accuracy, *i.e.* posterior $Z$ has less than 50% chance to be correct (labeled as "low"); 2) $\gamma$ is at medium accuracy, *i.e.* around 50% chance to be correct (labeled as "medium"); 3) $\gamma$ is at high accuracy, *i.e.* more than 50% chance to be correct (labeled as "high"). We also included the "optimal" scenario where $\gamma$ is 100% accurate for all individuals.

Under the null data, *i.e.* no genetic effect, both soft-GWAS and imputed-GWAS have calibrated p-value (Figure 4.8). When including the genetic effect, the imputed-GWAS and soft-GWAS have similar test statistic $T$ (defined as $\hat{\beta}/\text{se}(\hat{\beta})$) across all settings (Figure 4.3A). And the similar trends are observed in the estimate effect size $\hat{\beta}$ and its standard error as well (Figure 4.9). These results suggest that imputed-GWAS and soft-GWAS performs quite similarly though imputed-GWAS is less interpretable from the first principle but computationally simpler. So, we can reliably treat imputed-GWAS as a good alternative approach to soft-GWAS, which approximates soft-GWAS in a much more computationally feasible manner.

Furthermore, we examined the bias and power introduced by soft-GWAS and imputed-GWAS. Here we measured the relative power by the ratio of test statistic over the test statistic in "optimal" case, $T^\star$. We found that the observed biases nicely agree with the theoretical result described in Section 4.5.2 (Figure 4.10A). Moreover, And similarly, the relative powers also agree with the theoretical result (Figure 4.3B and 4.10B). So, in practice, before running the GWAS, we can get a good insight about the potential bias and power gain upper bound via these theoretical results.

### *4.3.3 Applying the imputation scheme to trios Framingham transcriptomic study*

To investigate the utility of the proposed haplotype imputation scheme, we ran the scheme on the genotype and gene expression profile from Framingham Heart Study (Section 4.5.5). We selected Framingham data for four reasons: 1) it has the genotype data for the trio so we know the actual haplotype origin as the gold standard; 2) it has the whole transcriptome data so we observe hundreds of parental phenotypes for each chromosome; 3) the cis-regulation of gene expression captures a substantial amount of heritability so that we can rely on a local genetic model to avoid multiple-chromosome fitting; 4) it is straightforward to apply PRS-based imputation to gene expression since there are many high-quality and publicly available transcriptome predictors.

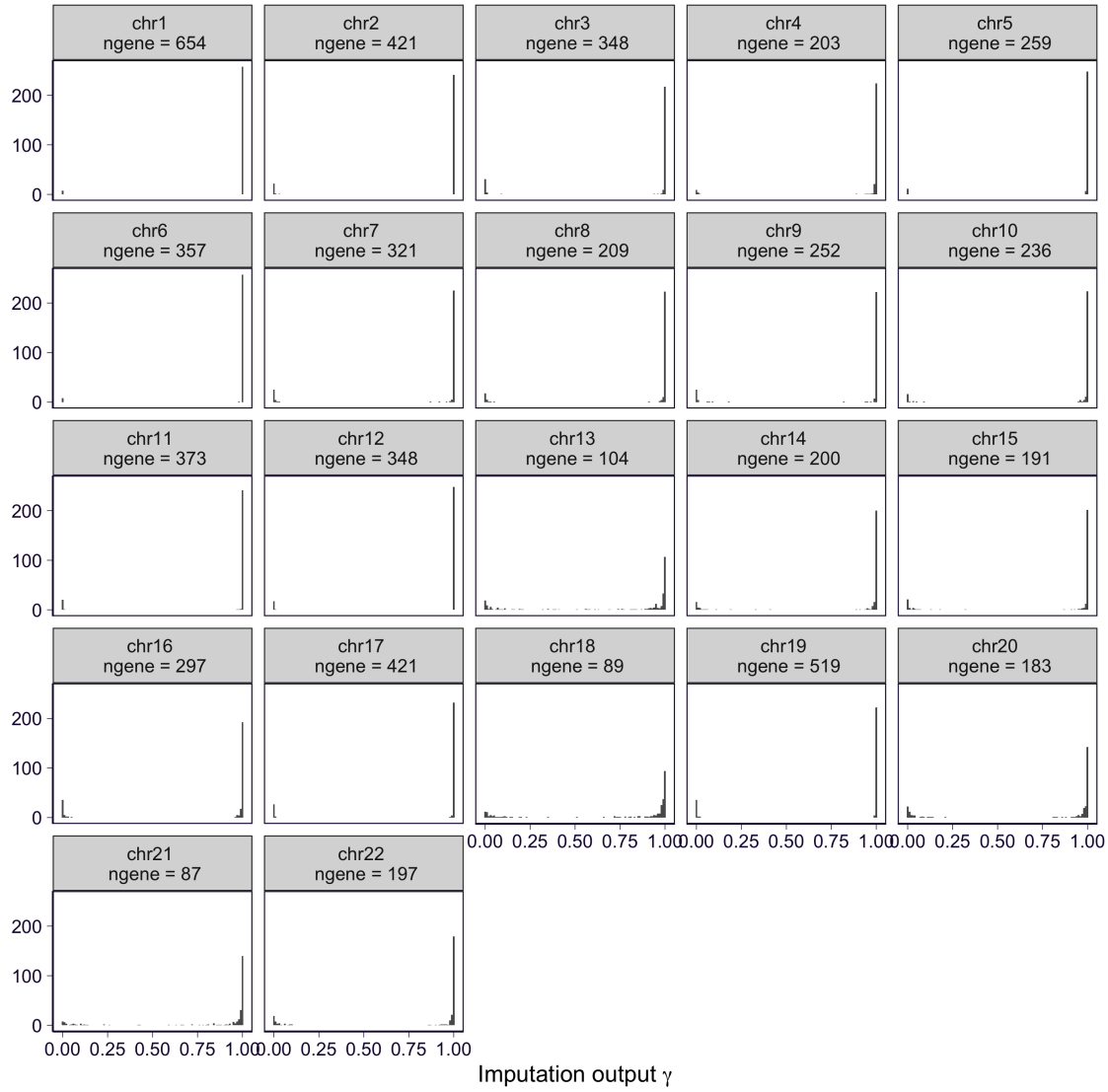



**Figure 4.4: PRS-based imputation results using EN DAPG models as the genetic predictor.** The histogram of the imputation output $\gamma$ (ground truth is 1) is shown for each chromosome in the panels. "ngene" represents the number of genes used in the imputation.

Overall, we extracted 266 trios and applied the imputation scheme to the child's haplotype and parental transcriptome. As we phased the genotype data using the trio information so the first haplotype of the child is, by design, from father. To verify this result, we calculated the genetic relatedness between the child's haplotypes and the parents' genotypes. The observed relatedness is consistent with the expectation (Section 4.5.5 and Figure 4.11). So, $\gamma = 1$ for all individuals is the ground truth of the imputation.

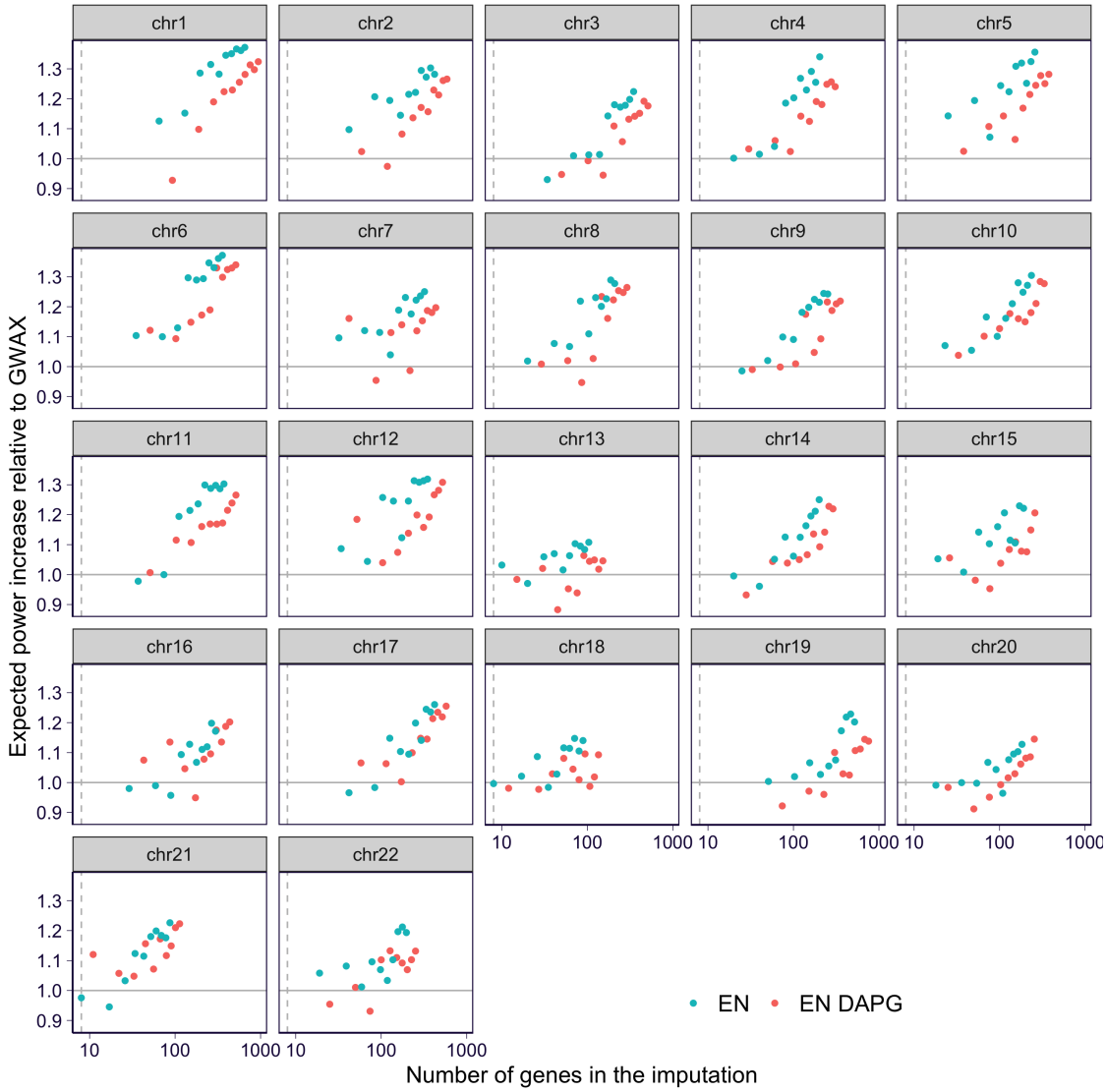



**Figure 4.5: The expected power increase of the imputed-GWAS using the PRS-based imputation relative to the approach without imputation (GWAX).** For each PRS-imputation run on either full data or downsampled data (downsampling genes), the number of genes used in the imputation is plotted on x-axis. And on y-axis, we show the corresponding relative power of the GWAS which uses the imputation result as input. In particular, the relative power is defined as the expected test statistic using the imputation over the expected test statistic using $\gamma = 0.5$ (*i.e.* the GWAX approach where no imputation is integrated). The results using EN models as genetic predictor are shown in green, and the ones using EN DAPG models are shown in red. The horizontal solid line is $y = 1$ corresponding to no power increase relative to GWAX. And the vertical dashed line is $x = 10$ which is approximately the number of parental phenotypes available in the UK Biobank.

As for the PRS-based approach, we used publicly available gene expression models as

described in Section 4.5.5. Specifically, we included models trained with elastic net [169] (EN) and models trained with fine-mapping (by DAPG [152]) informed elastic net (EN DAPG) which was developed recently [8]. For each chromosome, there are about 80-950 genes per chromosomes available for imputation. The PRS-based approach using EN models achieves good accuracy in all chromosomes and the quality is higher for those chromosomes that have more genes (Figure 4.4). For many individuals, the imputation output $\gamma$ is very close to 1 which is the ground truth with some strong errors which has $\gamma$ very close to 0. And the same trend is observed in the imputation runs using EN DAPG models as well (Figure 4.12) As we downsampled the genes for each chromosome, the imputation becomes less certain in the sense that the $\gamma$ values are pushed towards the middle (Figure 4.13 and 4.14). Since the imputation quality affects the downstream imputed-GWAS analysis, we calculated the expected power relative to no imputation scenario ($\gamma = 0.5$) which corresponds to the GWAX approach (Figure 4.5). As expected, the power increases when more genes are used. Moreover, the imputation using EN models performs better than the one using EN DAPG models. It could be attributed to the fact that EN DAPG is a sparse model so EN DAPG models are more sensitive to the imperfect variant overlapping between the HRC v1.1 and GTEx v8 (which is the variant panel used in EN and EN DAPG models). Roughly speaking, the imputed-GWAS is more powerful than the GWAX approach when the number of genes used in the imputation exceeds 100. In other word, we need more than 100 parental phenotypes to obtain a good enough imputation such that imputed-GWAS outperforms GWAX. As the current UK Biobank has only 12 parental phenotypes, we conclude that we have too few parental phenotypes to support the current imputation scheme.

## 4.4 Discussion

In this paper, we propose a two-approach to perform genotype-phenotype association study in the situation that we only have access to child's genotype and parental phenotype. Specif-

ically, in the first step, we propose a likelihood based imputation scheme to infer the parental origin of the child's haplotypes where the statistical model focuses on the non-focal parental phenotypes. And, in the downstream, we propose an efficient approach to integrate the imputation results to association test, which we call imputed-GWAS.

In the simulation study, we show that the imputation scheme can capture the parental origin when we observe tens to hundreds of heritable parental phenotypes. Furthermore, we show, theoretically and via simulation, that the power of the imputed-GWAS relies on the quality of the imputation. When the imputation is noisy, the imputed-GWAS doesn't typically benefit from such imputation.

Lastly, we perform the haplotype imputation on transcriptome and genotype data in trios from Framingham Heart Study. Here the imputation is performed within each chromosome and there are tens to hundreds of genes per chromosome. In this specific case, we show that the PRS-based imputation scheme is able to effectively impute the haplotype origin. And we observe that we need at least around 100 genes to have theoretical power gain in the downstream imputed-GWAS analysis (relative to the GWAS without no imputation, *e.g.* GWAX). From here, we conclude that this two-step approach is not applicable to the current biobank scale data set due to the lack of parental phenotypes.

Our paper has many limitations both in the proposed approach and in the data analysis. First, in the two-step procedure, we assume that all phenotypes are independent. This assumption simplifies the genetic model used in the imputation and the current framework is easily extendable to consider corrected non-focal phenotypes. However, the other complication comes from the correlation between the non-focal phenotypes and the focal phenotype (GWAS phenotype). It potentially introduces biases or false positives/negatives in the downstream GWAS, which we don't study in details in the paper. Secondly, some of our conclusions are based on simulation where all the limitations about simulation study apply. Especially, in our case, our conclusion about power and performance is always de-

pendent on parameter settings such as genetic architecture of the trait, heritability, and etc. Even though we simulate under carefully picked parameters so that it is not too far from the reality but someone should always interpret these results with cautious. Finally, due to the lack of data, we don't have a complete run of the two-step procedure on real data. With this, we want to emphasize the importance of parental phenotypes and we urge the development of a more systematic procedure for parental phenotype collection. In the context of studying late-onset disease, a rich set of parental phenotypes is of great importance. And, along this line, our paper provides an example that how these parental phenotypes can facilitate the research and enable us to dig deeper into the data.

## 4.5 Methods

### *4.5.1 Imputing haplotype origin*

#### 4.5.1.1 Modeling setups

We consider the following problem setup. All the variables below are at individual-level and, for simplicity, we omit the indexing on individual. The phased genotype of the child, $H^1$ and $H^2$ representing the two haplotypes, are observed. And we observe parental phenotypes, $y^\pi$ ($\pi$ = father or mother), which is a vector of length $P$ ($P$ phenotypes in total). Let $Z$ be an indicator variable representing the event that haplotype 1 is from father. The goal is to infer $Z$ from data.

Let "half-genotype" be half of the genotype where one of the two alleles is included for each variant, which is a weaker definition than the haplotype so that the alleles are not required to be on the same chromosome. To relate the observed phenotype with the child's haplotypes, we introduce $G_p^j$ representing the genetic effect of $j$th half-genotype on $p$th

phenotype for a given individual. And furthermore, we assume a genetic model as follow.

$$y_p = G_p^1 + G_p^2 + \epsilon_p \tag{4.1}$$

$$\epsilon_p \sim N(0, \sigma^2) \tag{4.2}$$

In the context of modeling parental phenotype, from Mendel's laws, we observe one of the two half-genotypes from a parent, which gives rise to one of the two child's haplotypes. So, Eq 4.1 is reduced to

$$y_p = G_p^1 + \epsilon_p^\star \tag{4.3}$$

$$\epsilon_p^\star \sim N(0, \sigma_\star^2) \quad , \tag{4.4}$$

where we assume $G_p^j \perp\!\!\!\perp \epsilon_p$ so that $G_p^2$ is absorbed into error term. Here the order of the two half-genotypes are arbitrary and we set the one being transmitted to the child as half-genotype 1.

In practice, we don't observe $G_p^j$ so we further assume that

$$G_p^j = \mu^p + H^j \beta^p \quad . \tag{4.5}$$

Under this model, we need to fit genetic effect $\beta^p$ along the way of inference so we refer to this approach as on-the-fly approach. And, moreover, if we have a genetic predictor (obtained from external data), *e.g.* a polygenic risk score model, we can model $G_p^j$ as

$$G_p^j = \mu^p + b^p \widetilde{G}_p^j \tag{4.6}$$

$$b^p \geq 0 \quad , \tag{4.7}$$

where $\widetilde{G}_p^j = H^j \widetilde{\beta}^p$ is the predicted value of $p$th phenotype and $\mu^p, b^p$ are scalars which

are introduced to account for the different scaling of the phenotype between the observed phenotype here and the one for predictor training. Besides, $b^p$ is set to non-negative to impose the constraint that predictor should at least predict the direction correctly. And we refer to this approach as PRS-based approach.

#### 4.5.1.2 Constructing the likelihood and the imputation

On the basis of Eq 4.3, Eq 4.5, and Eq 4.6, we propose the following likelihood function for the observed data $y^{\text{father}}, y^{\text{mother}}, H^1, H^2$.

$$\Pr(y^{\text{father}}, y^{\text{mother}} | H^1, H^2; \theta) = \prod_{i=1}^{N} \Pr(y_i^{\text{father}}, y_i^{\text{mother}} | H_i^1, H_i^2; \theta) \tag{4.8}$$

$$\Pr(y_i^{\text{father}}, y_i^{\text{mother}} | H_i^1, H_i^2; \theta) = \sum_{z \in \{0,1\}} \Pr(y_i^{\text{father}}, y_i^{\text{mother}} | Z_i = z, H_i^1, H_i^2; \theta) \Pr(Z_i = z) \quad , \tag{4.9}$$

where $\Pr(Z_i = 0) = \Pr(Z_i = 1) = 0.5$ for all individuals since the two child's haplotypes are equally likely to come from father and mother before observing the data. And $\theta$ represents the model parameter in the genetic models proposed in Eq 4.5 or Eq 4.6. Moreover, given $Z_i = 1$, $H_i^{\text{father},1} = H_i^1$ and $H_i^{\text{mother},1} = H_i^2$. And when $Z_i = 0$, similar equations follow but $H^1$ and $H^2$ flip the order. So, $\Pr(y_i^{\text{father}}, y_i^{\text{mother}} | Z_i = z)$ is straightforward to compute.

The goal is to update the distribution of $Z$ after observing the data, *i.e.* to obtain $\Pr(Z_i | y_i^{\text{father}}, y_i^{\text{mother}}, H_i^1, H_i^2; \theta)$. Since the true $\theta$ is unknown, we propose to plugin the maximum likelihood estimator (MLE) of $\theta$ under Eq 4.9. Algorithmically, we use EM algorithm to obtain the MLE where we iterate over updating posterior of $Z$ given current parameter $\theta^{(t)}$ and updating $\theta$ on the basis of the posterior of $Z$ (Section 4.8.1).

In practice, the imputation described above is performed one chromosome at a time. So, the genetic model only carries the heritability of the corresponding chromosome. As there are 22 autosomes with various length and gene density, the per-chromosome heritability is only a

fraction of the chip heritability, which also depends on genetic architecture. Because of this, the per-chromosome imputation is less effective than the imputation using all chromosomes jointly. But when fitting all chromosomes jointly, $Z$ becomes a 22-length vector and there are $2^{22}$ possible configurations of $Z$. To resolve this computation burden, we propose an alternative approach which updates one chromosome at a time (Section 4.8.2).

### *4.5.2 Integrating imputation results to GWAS*

#### 4.5.2.1 Problem overview

Here we consider performing GWAS using parental phenotype and genotype. As we only observe the genotype of the child and the parental genotype is missing, we, at most, observe one of the two alleles for each locus of the parent. So, similar to Eq 4.3, we consider using allelic test for GWAS [75]. Specifically, we consider modeling $y^{\pi}|H^{\pi,1}$, $\pi$ = father or mother, via: 1) linear model if quantitative phenotype, and 2) logistic model if case-control phenotype. And $H^{\pi,1}$ represents the parental half-genotype that has been transmitted to the child for a given locus.

#### 4.5.2.2 GWAS with softly assigned haplotype origin (soft-GWAS)

In practice, we don't observe $H^{\pi,1}$ directly but we know that $H^{\pi,1}$ is either $H^1$ or $H^2$. And from the imputation of haplotype origin (Section 4.5.1), we have some information about which haplotype is more likely to come from which parents. One way to incorporate this piece of information into the model is by modeling $\Pr(Y^{\pi}|H^1, H^2)$ instead, which, similar to Eq 4.9, gives rise to the following model

$$\Pr(y^{\pi}|H^1, H^2) = \sum_{z\in\{0,1\}} \Pr(y^{\pi}|H^1)\gamma + \Pr(y^{\pi}|H^2)(1-\gamma) \quad , \tag{4.10}$$

where $\gamma$ is the posterior $Z$ obtained from the imputation. To test whether the genetic effect is non-zero, we can use likelihood ratio test to obtain the statistical significance, which corresponds to running an EM algorithm similar to Section 4.8.1 (see Section 4.8.3). This approach may result in big computation burden when performing genome-wide analysis.

#### 4.5.2.3 GWAS with imputed half-genotype (imputed-GWAS)

Alternatively, we propose to plugin the posterior mean of the parental half-genotype instead. For instance, in the case of analyzing a quantitative trait using paternal phenotype, we perform linear regression with observed phenotype $y^{\text{father}}$ and imputed half-genotype $\widetilde{H}^{\text{father}}$ where $\widetilde{H}^{\text{father}} = \gamma H^1 + (1-\gamma)H^2$. By doing so, we avoid running EM algorithm and the computation complexity is the same as a convention GWAS. And GWAX could be considered as a special case of imputed-GWAS where $\gamma = 0.5$.

#### 4.5.2.4 The power and bias of imputed-GWAS

Here we focus on the power and bias of the imputed-GWAS approach. For the sake of simplicity, we analyze a simple scenario where phenotype is quantitative. Let $y$ be the observed phenotype and $X_i^1$ be the true half-genotype. The imputed half-genotype takes the form $r_i X_i^1 + (1-r_i)X_i^2$. Note that $r_i$ is similar to $\gamma_i$ but $r_i$ denotes the probability of assigning the haplotype correctly. So, $r_i = \gamma_i$ if haplotype 1 is from father and $r_i = 1 - \gamma_i$ otherwise. We assume the genetic model is

$$y_i = X_i^1 \beta + \epsilon_i \tag{4.11}$$

$$\epsilon_i \sim N(0, \sigma^2) \quad . \tag{4.12}$$

And we rely on linear regression to obtain statistical significance, *i.e.* $\widehat{\beta} = (\widetilde{X}'\widetilde{X})^{-1}\widetilde{X}'y$. Note that the optimal test is achieved when $r_i = 1$ for all individuals. And let $T^\star$ be the

test statistic ($\widehat{\beta}/\text{se}(\widehat{\beta})$) obtained at the optimal. We derive the following results on the bias and power of the imputed-GWAS approach (Section 4.8.4).

$$\mathrm{E}(\widehat{\beta}) = \frac{\overline{r}}{\overline{S}}\beta \tag{4.13}$$

$$\mathrm{E}(T) = \frac{\overline{r}}{\sqrt{\overline{S}}}\mathrm{E}(T^{\star}) \quad , \tag{4.14}$$

where $S_i = r_i^2 + (1 - r_i)^2$ and $\overline{x}$ means taking sample mean of variable $x$. As $r$ is a measure of imputation quality (the closer to 1, the higher quality), the bias and power both depend on the imputation quality. And the power monotonously increases as $r$ increases.

### *4.5.3 Simulation study of the imputation scheme*

#### 4.5.3.1 Simulation to comparing different genetic model and other parameter settings

The goal of this simulation study of the imputation scheme is to see how the imputation quality is affected by heritability, the number of non-focal phenotypes, the choice of the genetic model, and etc. We simulate parental phenotypes, genotypes, and child's genotype by the following procedure.

1. Simulate parental half-genotypes, $H^{\pi,1}$ and $H^{\pi,2}$, for $\pi$ = father and mother of each individual. Variants are independently sampled from Bernoulli distribution with minor allele frequency sampled from a uniform distribution $U[0.05, 0.45]$.

2. Simulate effect size with $\beta \sim (1 - \pi_0)\delta_0 + \pi_0 N(0, 1)$ where $\pi_0 = 0.5$.

3. Calculate parental genetic effect as $G^{\pi,j} = H^{\pi,j}\beta$.

4. Simulate environmental effect, $\epsilon^{\pi}$ for each parent where $\epsilon^{\pi} \sim N(0, \frac{1-h^2}{h^2}\widehat{\text{Var}}(G^{\pi,1} + G^{\pi,2}))$.

5. Calculate observed parental phenotype as $y^{\pi} = G^{\pi,1} + G^{\pi,2} + \epsilon^{\pi}$.

6. Transmit $H^{\text{father},1}$ and $H^{\text{mother},1}$ to child with $H^1 = H^{\text{father},1}$ and $H^2 = H^{\text{mother},1}$.

To simplify the message, we fix the number of variants to 50 and the sample size for imputation to 1,000. Step 2-5 is repeated for 100 parental phenotypes with heritability $h^2 = 0.01, 0.05, 0.1, 0.25, 0.5$.

To test the PRS-based approach, we also simulate a separate cohort, with 20,000 individuals, for PRS training. To see how PRS quality affects the imputation, we also train PRS using a subset of these individuals (sample size = 5,000 and 10,000).

#### 4.5.3.2 Simulation to examine the PRS-based approach

Here, we are specifically interested in the utility of PRS-based approach in the context that parental transcriptome is available. In this setting, we focus on building the predictor of cis-regulation. Typically, we have hundreds of samples in a transcriptome study and the cis-window contains thousands of variants. So, PRS-approach serves better in this scenario.

To examine the power of PRS-approach in this specific setting, we simulate data by the following procedure.

1. Simulate parental genetically determined gene expression $G^{\pi,1}$ and $G^{\pi,2}$.

2. Simulate the parameter $b$ in Eq 4.6 from: 1) $b = -0.1$, 2) $b = 0.1$, 3) $b \sim N(0,1)$, and 4) $b \sim \max(0, N(0,1))$.

3. Simulate some covariates $C_m \sim N(0,1)$ with effect size $a_m \sim N(0,1)$.

4. Calculate $y^{\pi} = (G^{\pi,1} + G^{\pi,2})b + \sum_m a_m C_m + \epsilon^{\pi}$ where the variance of $\epsilon^{\pi}$, $\sigma_e^2$, is set such that heritability $h^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2} = 10^{-4}, 10^{-3}, 0.01, 0.05$ ($\sigma_g^2$ is the genetic variation).

5. For the child, the genetic effect of haplotype 1, $G^1$, is set to $G^{\text{father},1}$ and, similarly, $G^2$ is set to $G^{\text{mother},1}$.

For simplicity, we fix sample size to 300, the number of phenotypes to 500, and the number of covariates to 4.

### *4.5.4 Simulation study of the proposed GWAS approaches*

Here we simulate data to test the performance of the proposed GWAS approaches, soft-GWAS and imputed-GWAS. The simulation procedure is as follow which generates one phenotype-genotype pair at a time (individual index is ignored).

1. Simulate $H^1$ and $H^2$ from Bernoulli distribution where minor allele frequency is samples from $U \sim U[0.05, 0.45]$.

2. Simulate phenotype $y = H^1\beta + \epsilon, \epsilon \sim N(0, \sigma_e^2)$ where $\beta$ is either 0 (the null) or 1 (the alternative). And when $\beta = 1$, $\sigma_e^2$ is set such that per-SNP heritability is 0.001; when $\beta = 0$, $\sigma_e^2 = 1$.

3. Simulate $\gamma \sim g(\cdot)$ where $g(\cdot)$ takes the following forms: $\delta_1$, $\delta_{0.9}$, $\delta_{0.1}$, $\delta_{0.5}$, Beta$(5, 2)$, Beta$(2, 5)$, and Beta$(2, 2)$.

We fix sample size to 5,000 and we repeat the procedure 500 times which results in 500 independent phenotype-genotype pairs for the association test. We refer to $\delta_\cdot$ as "constant" distribution and Beta$(\cdot, \cdot)$ as "beta" distribution. In this simulation, $H^1$ is the affecting haplotype, so the "optimal" test is obtained at $\gamma_i = 1$ which corresponds to $\delta_1$. Moreover, for each distribution type, we include three distributions which have $\gamma$ centered around high values (referred as "high"), around 0.5 (referred as "middle"), and around low values (referred as "low") So, $\delta_{0.9}$ and Beta$(5, 2)$ lie in category "high"; $\delta_{0.5}$ and Beta$(2, 2)$ lie in category "middle"; and $\delta_{0.1}$ and Beta$(2, 5)$ lie in category "low".

### *4.5.5 Analyzing Framingham Heart Study*

We obtained genotype and transcriptome data ([61, 165]) via dbGap accession number phs000007.v29.p1. The genotype data and gene-level expression quantification have been cleaned-up and processed previously in [155]. In brief, the genotype data was pre-phased locally by SHAPEIT [31] and imputed to HRC v1.1 [98] using Michigan Imputation Server [30]. Note that under this phasing procedure, the first haplotype of the child is from father and we also verified this result by checking the genetic relatedness between child's half-genotype and father/mother genotype. And the gene expression data was pre-processed by Affymetrix power tools suite.

We extracted the individuals in the study of whom both the genotype and the trancriptome data were available. Overall 4,838 individuals were selected. We constructed expression matrix with only these individuals being included and then we quantile normalized the expression within each gene. The expression matrix was used to run PEER factor analysis [131] to obtain the hidden confounding factors of the experiments where we set the number of factors to 40. We also ran PCA using the genotypes of these individuals via GCTA tools [161] where the first 20 PCs were kept. These 20 PCs and 40 PEER factors were used as the covariates in the haplotype imputation.

Based on the pedigree information, we extracted the trios where both parents have expression data available and the whole trio have genotype data available. If two trios share the same father and/or mother, we kept only one of them for the analysis. In the end, we collected 266 trios. We applied the PRS-based imputation scheme to these trios where the predicted expression was obtained from elastic net and DAPG weighted elastic net models trained on GTEx V8 whole blood European samples [6, 8]. In particular, we ran the imputation for each chromosome, And to examine the power of the imputation, within each chromosome, we downsampled the genes to a fraction of the original number and re-ran the imputation to see how the imputation quality depends on the downsampling fraction.

We defined the genetic relatedness between half-genotype and genotype or genotype and genotype in a unified way. First, we standardized haplotype. For an diploid, we defined $\widetilde{H} = \frac{H^1}{\sqrt{2}} + \frac{H^2}{\sqrt{2}}$ where $H^j$ is the $j$th haplotype. And for a haploid, we defined $\widetilde{H} = H$. The genetic relatedness is defined as $K_{p,q} = \frac{1}{K}\sum_{k=1}^{K} \widetilde{H}_k^p \widetilde{H}_k^q$ where $p$ and $q$ represents either diploids (from father's and mother's genotype) or haploids (from child's haplotype). If both $p$ and $q$ are diploid, this definition is consistent with conventional definition, i.e. $\mathrm{E}(K_{p,q}) = 1$ if $p = q$ and $\mathrm{E}(K_{p,q}) = 0$ if $p$ and $q$ are completely unrelated. For the genetic relatedness between haploid $\widetilde{H}^1$ and diploid $\widetilde{H}^p$,

$$\mathrm{E}(K_{1,p}) = \frac{1}{K}\sum_k \frac{1}{\sqrt{2}}[\ \mathrm{E}(H_k^1 H_k^{p,1}) + \mathrm{E}(H_k^1 H_k^{p,2})\ ] \tag{4.15}$$

$$= \begin{cases} \frac{1}{\sqrt{2}} & \text{, if } H^1 = H^{p,1} \text{ or } H^2 = H^{p,2} \\ 0 & \text{, otherwise} \end{cases} \tag{4.16}$$

## 4.6 Code Availability

The code used for this work is at `https://github.com/liangyy/haplotype-po`.

## 4.7 Supplementary Figures

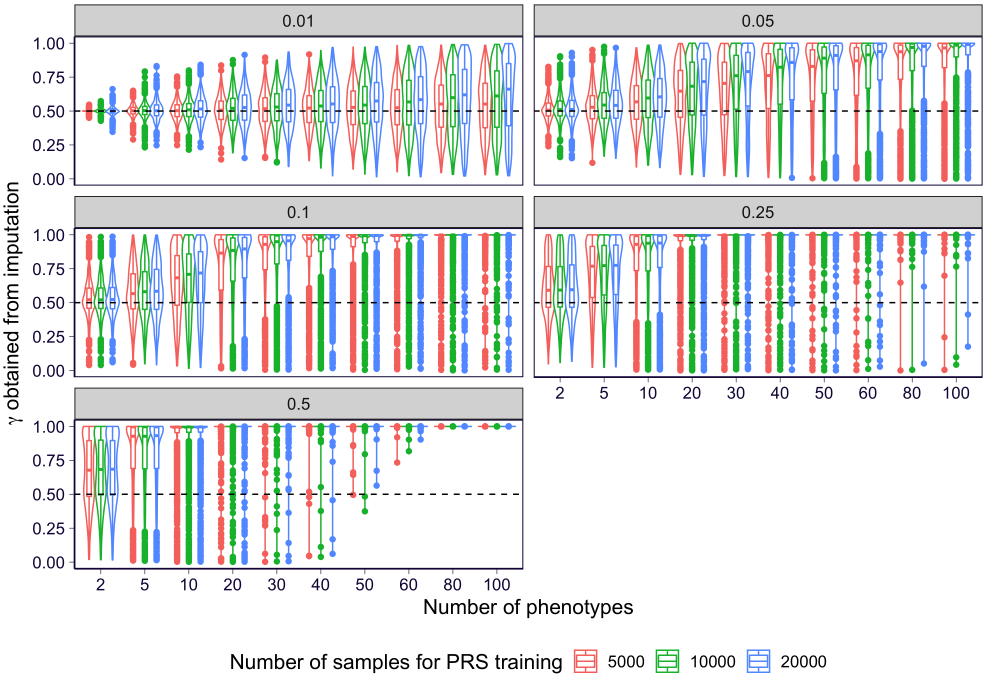



**Figure 4.6: The imputation performance of the PRS-based approach with PRS trained with different sample sizes.** The imputation performance under a different heritability is shown in each panel. Within each panel, the number of phenotypes included in the imputation is shown on x-axis and the imputation accuracy (the probability that the posterior $Z$ is correctly assigned) is shown on the y-axis. The violin/boxplot contains the results on all of the 1,000 individuals included in the imputation. The results of the PRS-based approaches that are based on PRSs trained in 5,000, 10,000, and 20,000 individuals are colored in red, blue, and green respectively.

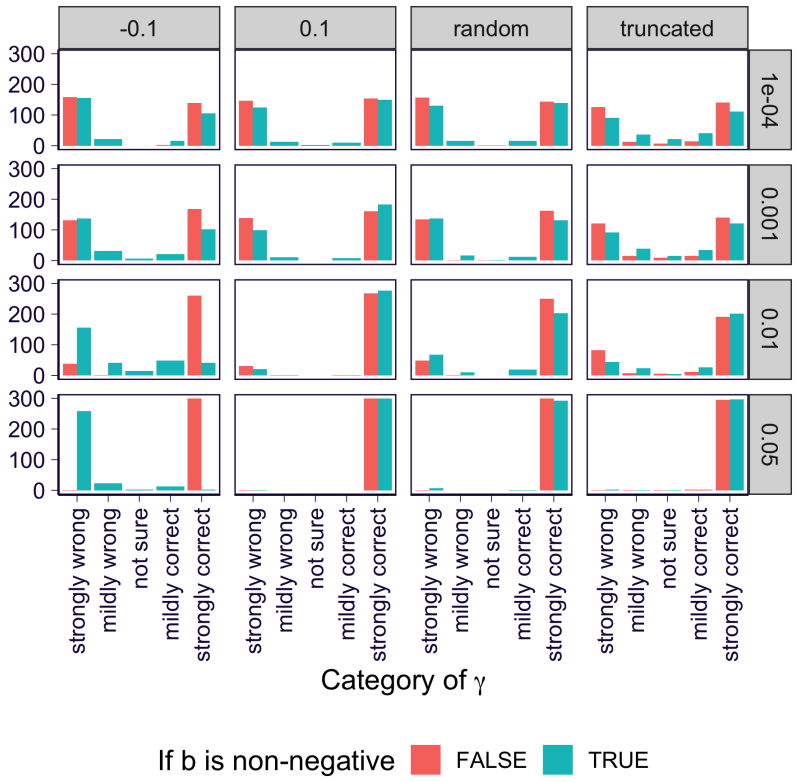



**Figure 4.7: Comparing the performance of PRS-based approach with/without non-negative constraint on the coefficient.** The simulated data used here is generated in the context of using parental transcriptome (second scheme in Section 4.5.3) with sample size = 300 and number of genes = 500. Each panel shows the imputation results under one simulation setting with per-gene heritability organized in rows and the distribution of the true $b$ in columns. Specifically, the panels labeled with "random" mean that $b \sim N(0, 1)$ so that it has random sign and, similarly, the ones with "truncated" mean that $b \sim \max(0, N(0, 1))$ so that it is non-negative. For illustration purpose, we binned the imputation results into 5 categories on x-axis, in which $r$ (the probability of being correct) are binned into $[0, 0.1)$, $[0.1, 0.4)$, $[0.4, 0.6]$, $(0.6, 0.9]$, and $(0.9, 1]$ representing "strongly wrong", "mildly wrong", "not sure", "mildly correct", and "strongly correct" respectively. And y-axis shows the count of the bin. The results of the PRS-based approaches with and without non-negative constraint on $b$ are shown in green and red.

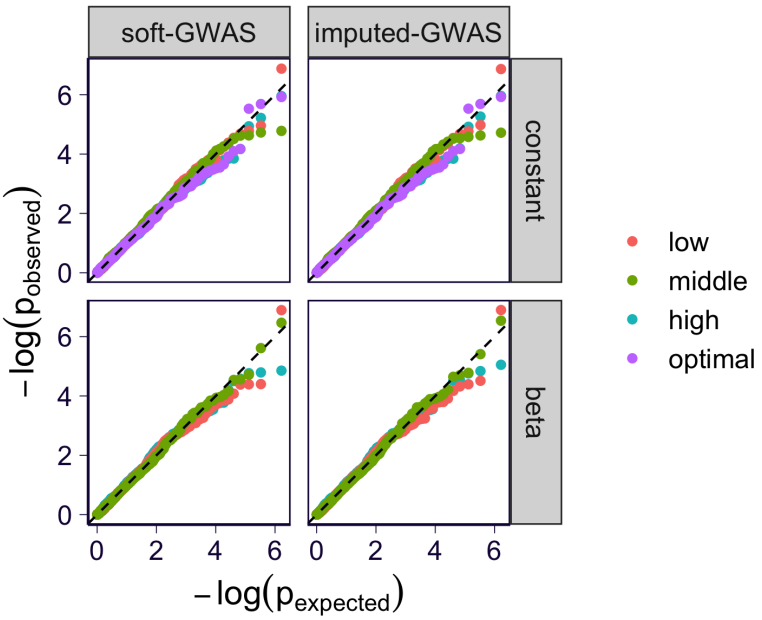
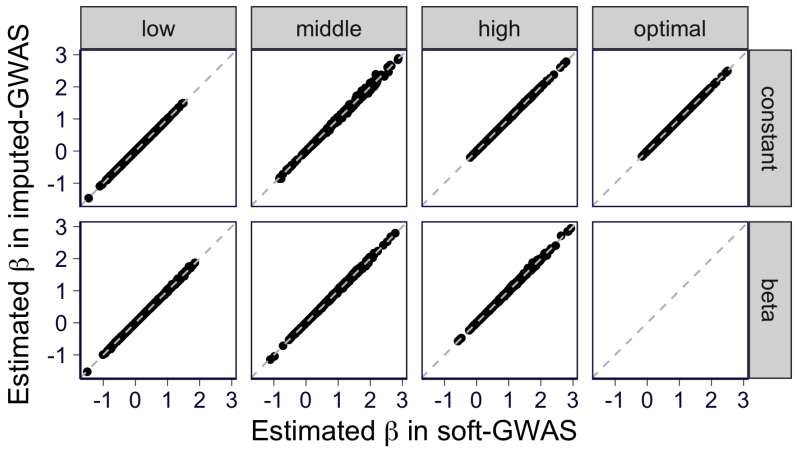
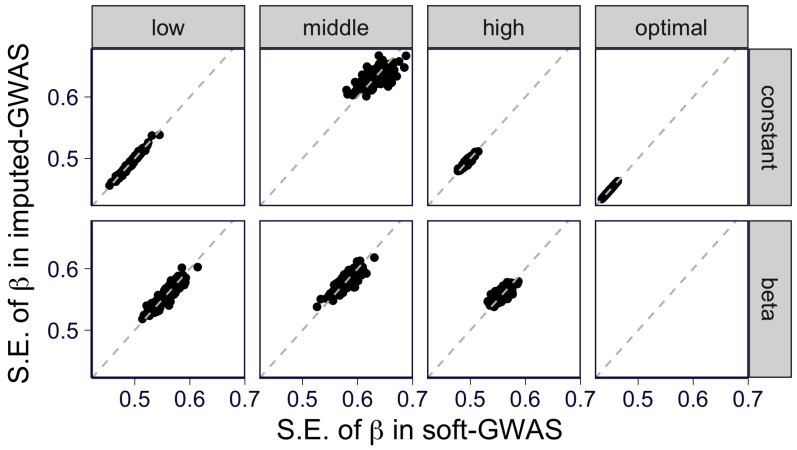



**Figure 4.8: QQ-plot of the proposed GWAS tests under the simulated null data.** QQ-plot of the association p-values (expected on axis versus observe on y-axis in $-\log$ scale) are shown. Each panel shows the results on one GWAS approach (soft-GWAS or imputed-GWAS in columns) and one distribution of $\gamma$ (constant or beta in rows). The QQ-plots corresponding to different $\gamma$ distributions are drawn and shown separately in different colors (see more details at Section 4.5.4).

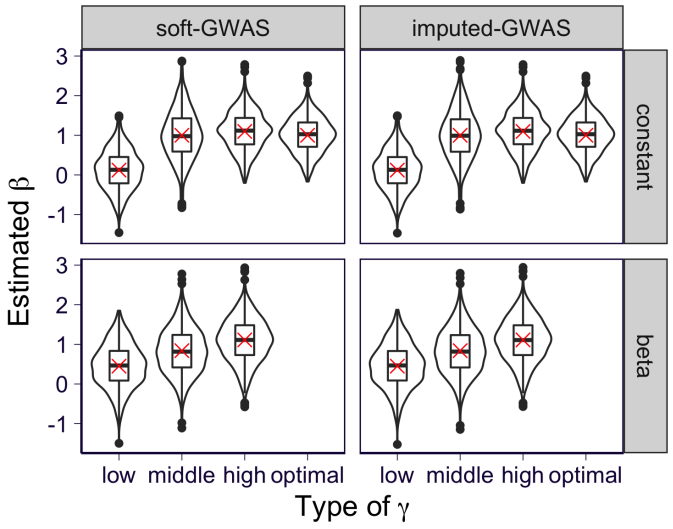



**(A)**

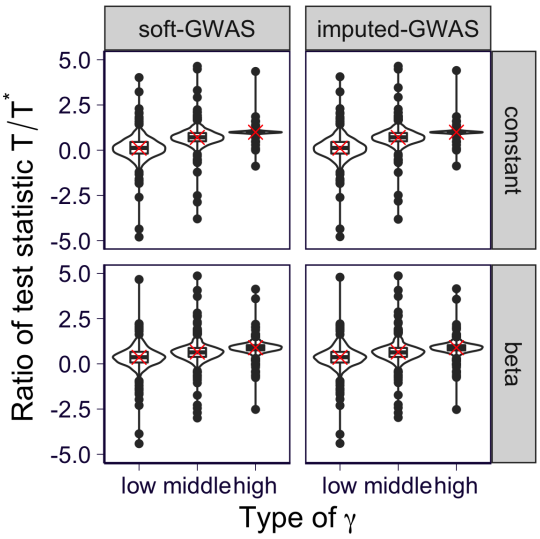



**(B)**

**Figure 4.9: Comparing the effect size estimates in soft-GWAS and imputed-GWAS on simulated data.** Each panel presents the results under a specific $\gamma$ distribution (the distribution type is organized in rows and the accuracy of $\gamma$ is organized in columns). The results of soft-GWAS are shown on x-axis and the ones of the imputed-GWAS are shown on y-axis. The gray dashed line is $y = x$. **(A)** The estimated effect sizes $\hat{\beta}$ are shown. **(B)** The standard error of $\hat{\beta}$ are shown.




**(A)**




**(B)**

**Figure 4.10: Comparing the theoretical and observed bias and relative power.** Each panel presents the results under a specific $\gamma$ distribution (the distribution type is organized in rows and the type of method is organized in columns). And the results are stratified by the accuracy of $\gamma$ on x-axis. **(A)** The estimated effect sizes $\hat{\beta}$ are shown in the violin/boxplot and the red cross indicates the expected effect size after accounting for the theoretical bias. **(B)** The observed across all replications are shown in the violin/boxplot where outliers with observed ratio outside $[-5, 5]$ are excluded. The red cross shows the theoretical relative power.

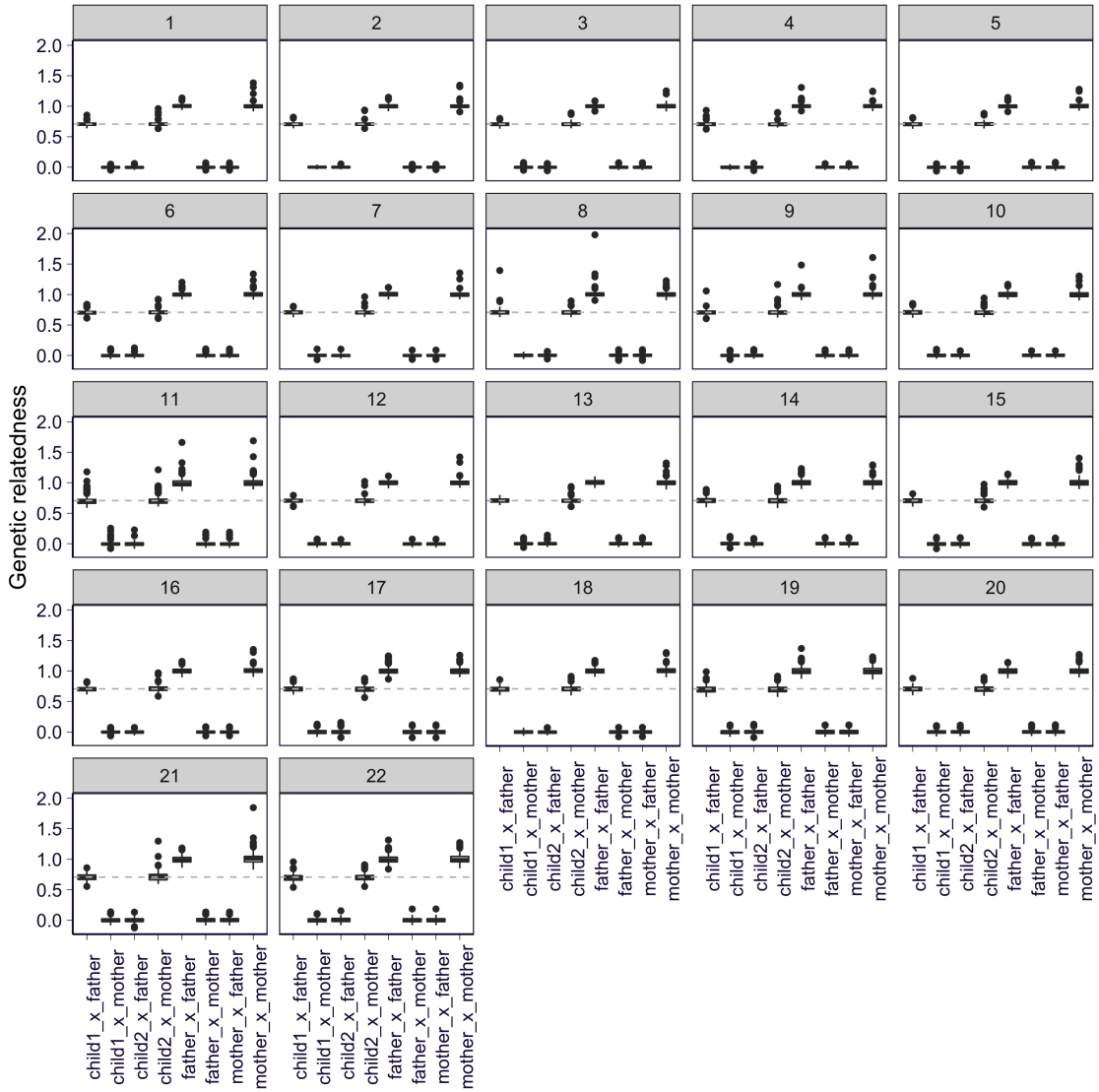



**Figure 4.11: Genetic relatedness between child's haplotypes and parents' genotypes in Framingham trios.** Each panel presents the results on one of the 22 autosomes. On x-axis, "sample1_x_sample2" means the genetic relatedness between sample1 and sample2 where sample1 and sample2 can be either haplotype or genotype. "child1" represents the first haplotype of the child and "child2" means the second one. "father" and "mother" represent the genotype of the father and mother. On y-axis, the genetic relatedness is shown (see detailed definition in Section 4.5.5). The horizontal dashed line is $y = \frac{1}{\sqrt{2}}$ which is the expected genetic relatedness between a haploid and a diploid if the haploid comes from the diploid.

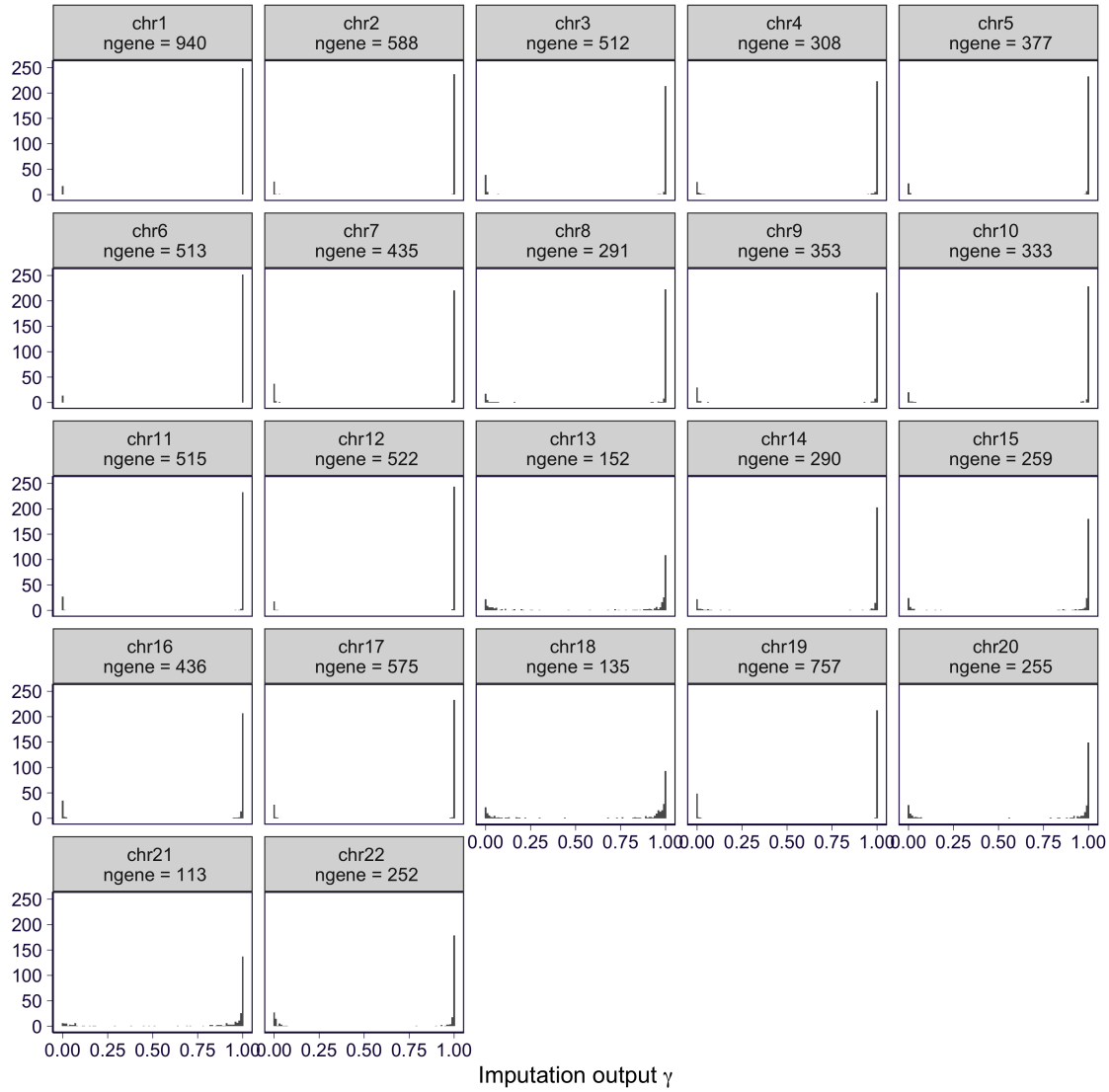



**Figure 4.12: PRS-based imputation results using EN DAPG models as the genetic predictor.** The histogram of the imputation output $\gamma$ (ground truth is 1) is shown for each chromosome in the panels. "ngene" represents the number of genes used in the imputation.

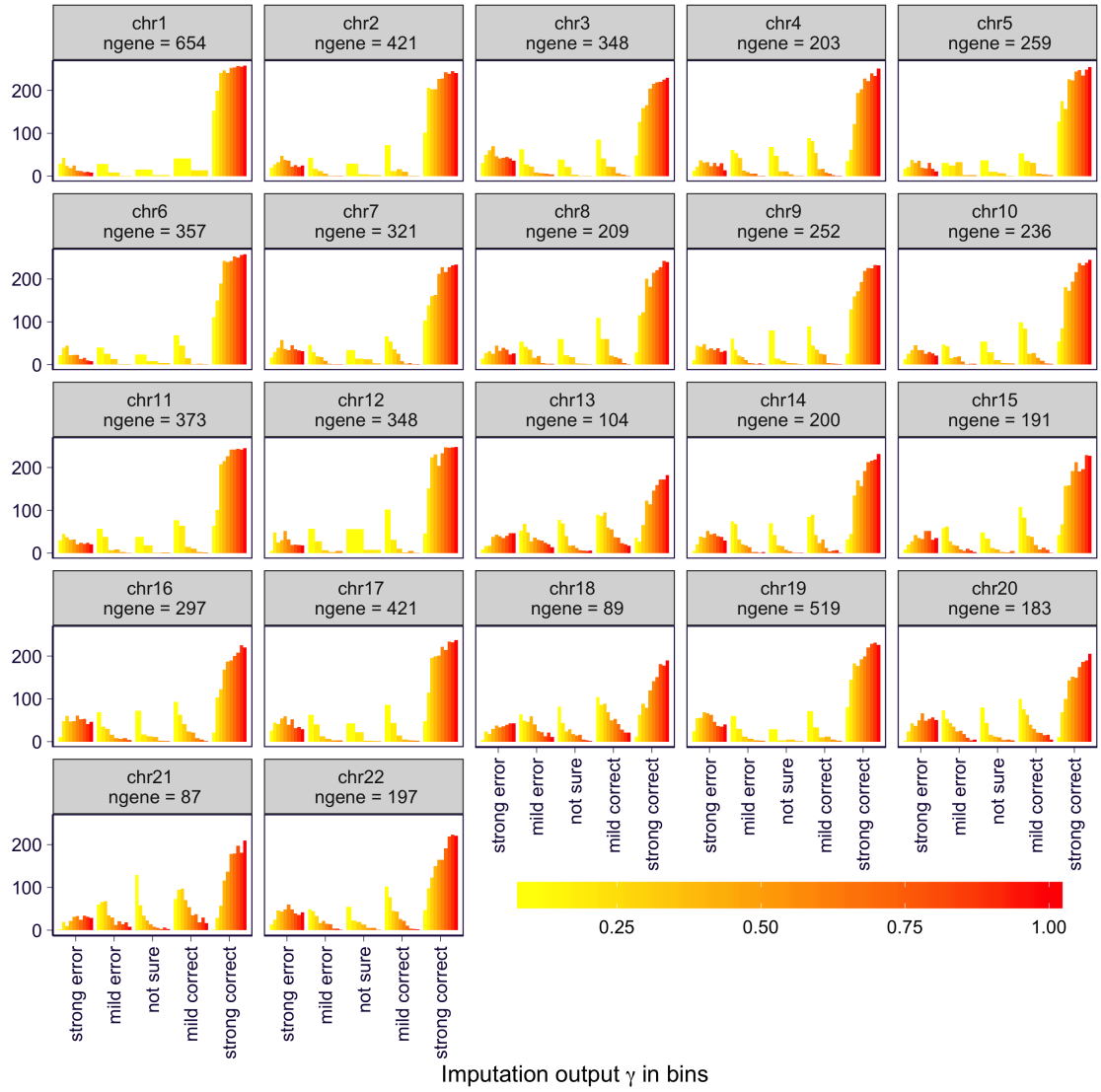



**Figure 4.13: PRS-based imputation results on the downsampled data using EN models as the genetic predictor.** The imputation results on the downsampled data is shown for each chromosome in the panels. The imputation output $\gamma$ is stratified into 5 bins $[0, 0.1)$, $[0.1, 0.4)$, $[0.4, 0.6]$, $(0.6, 0.9]$, and $(0.9, 1]$ representing "strongly wrong", "mildly wrong", "not sure", "mildly correct", and "strongly correct" respectively. And y-axis shows the count of the bin. The results are colored by the downsampling fraction relative to the full data. "ngene" represents the number of genes in the full data.

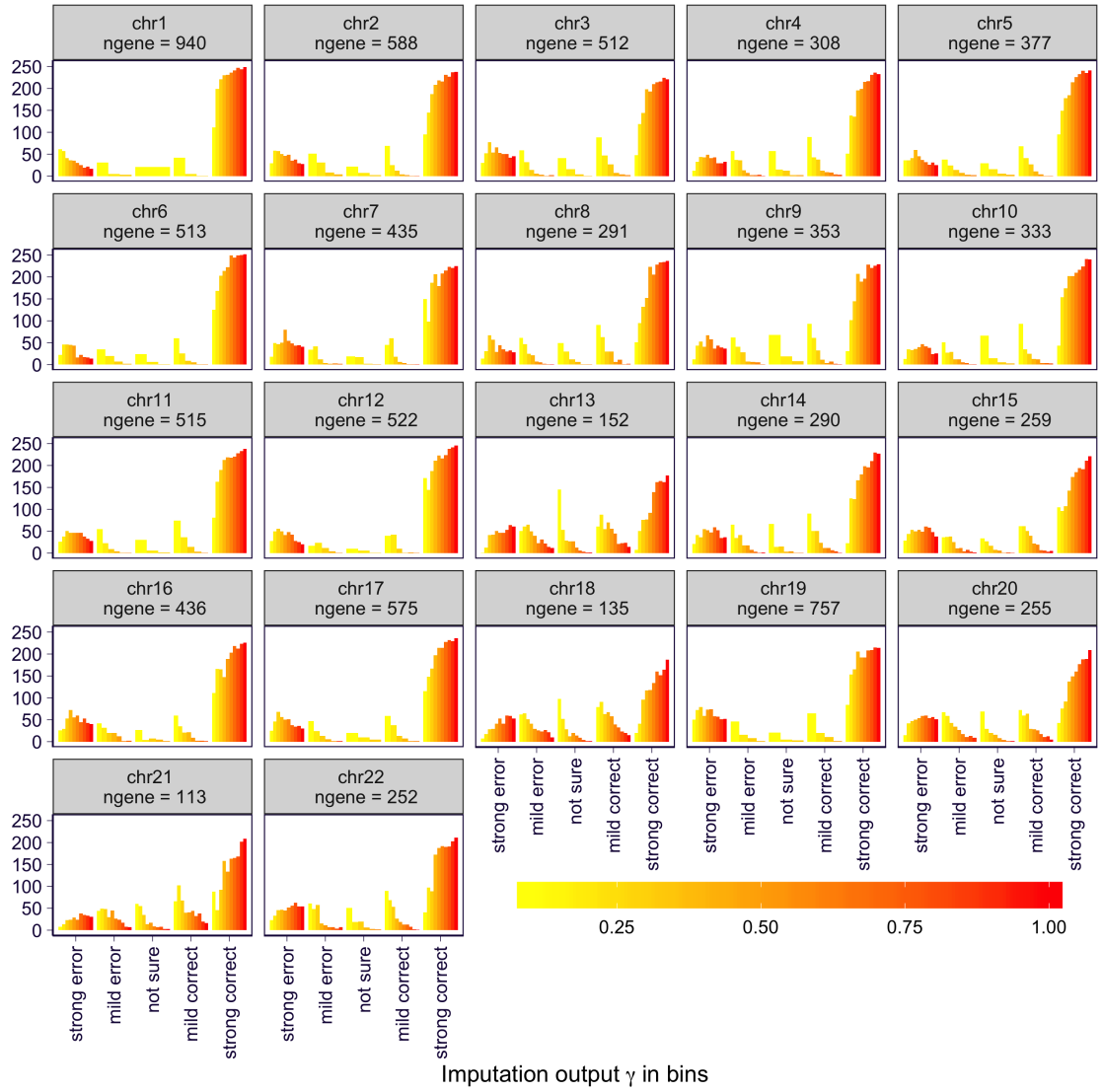



**Figure 4.14: PRS-based imputation results on the downsampled data using EN DAPG models as the genetic predictor.** The imputation results on the downsampled data is shown for each chromosome in the panels. The imputation output $\gamma$ is stratified into 5 bins $[0, 0.1)$, $[0.1, 0.4)$, $[0.4, 0.6]$, $(0.6, 0.9]$, and $(0.9, 1]$ representing "strongly wrong", "mildly wrong", "not sure", "mildly correct", and "strongly correct" respectively. And y-axis shows the count of the bin. The results are colored by the downsampling fraction relative to the full data. "ngene" represents the number of genes in the full data.

## 4.8 Supplementary Notes

### *4.8.1 The EM algorithm to impute haplotype origin*

Based on Eq 4.9, we have ($p$ indexes phenotypes and $i$ indexes individuals)

$$Q(\theta, \theta^{(t)}) := \mathrm{E}_{Z|y,H,\theta^{(t)}}[\log \Pr(y^{\text{father}}, y^{\text{mother}}, Z|H^1, H^2; \theta)] \tag{4.17}$$

$$= \sum_i \mathrm{E}_{Z_i|y_i,H_i,\theta^{(t)}}[\log \Pr(Z_i) + \sum_p \log \Pr(y_{i,p}^{\text{father}}, y_{i,p}^{\text{mother}}|Z_i, H_i^1, H_i^2; \theta)] \tag{4.18}$$

$$\text{(assume phenotypes are independent conditioning on haplotypes)} \tag{4.19}$$

$$= const. + \sum_i \mathrm{E}_{Z_i|y_i,H_i,\theta^{(t)}}[\sum_p \log \Pr(y_{i,p}^{\text{father}}, y_{i,p}^{\text{mother}}|Z_i, H_i^1, H_i^2; \theta)] \tag{4.20}$$

$$\text{(as prior knowledge, we have } \Pr(Z_i) = 0.5) \tag{4.21}$$

Since $Z$ indicates if haplotype 1 is from father, we have

$$\log \Pr(y_{i,p}^{\text{father}}, y_{i,p}^{\text{mother}}|Z_i, H_i^1, H_i^2; \theta) \tag{4.22}$$

$$= Z_i F_1(y_{i,p}^{\text{father}}, y_{i,p}^{\text{mother}}, H_i^1, H_i^2) + (1 - Z_i) F_2(y_{i,p}^{\text{father}}, y_{i,p}^{\text{mother}}, H_i^1, H_i^2) \tag{4.23}$$

where

$$F_1(y_{i,p}^{\text{father}}, y_{i,p}^{\text{mother}}, H_i^1, H_i^2) := \log \Pr(y_{i,p}^{\text{father}}, y_{i,p}^{\text{mother}}|Z_i = 1, H_i^1, H_i^2) \tag{4.24}$$

$$= l(y_{i,p}^{\text{father}}, H_i^1) + l(y_{i,p}^{\text{mother}}, H_i^2) \tag{4.25}$$

$$F_2(y_{i,p}^{\text{father}}, y_{i,p}^{\text{mother}}, H_i^1, H_i^2) := \log \Pr(y_{i,p}^{\text{father}}, y_{i,p}^{\text{mother}}|Z_i = 0, H_i^1, H_i^2) \tag{4.26}$$

$$= l(y_{i,p}^{\text{father}}, H_i^2) + l(y_{i,p}^{\text{mother}}, H_i^1) \tag{4.27}$$

and $l(y, H)$ represent the log-likelihood of observing phenotype $y$ and haplotype $H$. For simplicity, we use $F_1(i, p)$ and $F_2(i, p)$ as the short form of $F_1(y_{i,p}^{\text{father}}, y_{i,p}^{\text{mother}}, H_i^1, H_i^2)$ and

$F_2(y_{i,p}^{\text{father}}, y_{i,p}^{\text{mother}}, H_i^1, H_i^2)$.

Combining Eq 4.20 and 4.23, we have

$$Q(\theta, \theta^{(t)}) = const. + \sum_i \mathrm{E}_{Z_i|y_i,H_i,\theta^{(t)}}[\sum_p Z_i F_1(i,p) + (1-Z_i)F_2(i,p)] \tag{4.28}$$

$$= const. + \sum_i w_i \sum_p F_1(i,p) + \sum_i (1-w_i) \sum_p F_2(i,p) \tag{4.29}$$

where $w_i := \mathrm{E}_{Z_i|y_i,H_i,\theta^{(t)}}[Z_i] = \Pr(Z_i = 1|y_i, H_i, \theta^{(t)})$.

At E step, we update $w_i$ by applying Bayes rule

$$w_i = \frac{\Pr(y_i|Z_i=1, H_i, \theta^{(t)}) \Pr(Z_i = 1)}{\sum_{k\in\{0,1\}} \Pr(y_i|Z_i=k, H_i, \theta^{(t)}) \Pr(Z_i = k)} \tag{4.30}$$

$$= \frac{\exp(\sum_p F_1(i,p;\theta^{(t)}))}{\exp(\sum_p F_1(i,p;\theta^{(t)})) + \exp(\sum_p F_2(i,p;\theta^{(t)}))} \tag{4.31}$$

where, notice that, we need to use $\theta^{(t)}$ when evaluating $F_1$ and $F_2$.

At M step, we update $\theta$ by

$$\theta = \arg\max_\theta \sum_i w_i \sum_p F_1(i,p;\theta) + \sum_i (1-w_i) \sum_p F_2(i,p;\theta) \tag{4.32}$$

Specifically, for on-the-fly approach (Eq 4.5), it corresponds to solving weighted least squares. For instance, to obtain father-specific model parameters $\beta^{\text{father}}$ and $\sigma^2_{\text{father}}$, it is

equivalent to solve weighted least squares with the following settings

$$\text{response} = \begin{bmatrix} y^{\text{father}} \\ y^{\text{father}} \end{bmatrix} \tag{4.33}$$

$$\text{design matrix} = \begin{bmatrix} H^1 \\ H^2 \end{bmatrix} \tag{4.34}$$

$$\text{weight} = \begin{bmatrix} w \\ 1-w \end{bmatrix} \tag{4.35}$$

Similarly, to obtain mother-specific model parameters $\beta^{\text{mother}}$ and $\sigma^2_{\text{mother}}$, we solve with

$$\text{response} = \begin{bmatrix} y^{\text{mother}} \\ y^{\text{mother}} \end{bmatrix} \tag{4.36}$$

$$\text{design matrix} = \begin{bmatrix} H^1 \\ H^2 \end{bmatrix} \tag{4.37}$$

$$\text{weight} = \begin{bmatrix} 1-w \\ w \end{bmatrix} \tag{4.38}$$

For the PRS based approach (Eq 4.6), we can simply replace $H_1$ and $H_2$ with the haplotypic PRS $\widetilde{G}^1$ and $\widetilde{G}^2$ and the similar calculation can be applied to solve for father/mother-specific $b^p$ and other parameters. Since we assume $b^p$ to be non-negative, at the M step, we set $b^p = 0$ if the weight least squares gives negative $b^p$.

### *4.8.2 Fitting multiple chromosomes in iterative manner*

The haplotype phasing is done for each chromosome. So, the EM algorithm in Section 4.8.1 can only handle one chromosome since the likelihood requires phased haplotypes. Even though one could fit one chromosome at a time while treating the contribution from other

chromosomes as noise, to fit all chromosomes jointly is preferred. The reason is that the former captures at most per-chromosome heritability while the latter models the contribution from all chromosomes which, in other words, reduces the noise.

In principle, to handle multi-chromosome situation, instead of having $Z_i$ being a binary value, $Z_i$ should be a vector with each entry representing the parental origin of the corresponding chromosome. This creates complexities in the E step of the EM algorithm. At the E step, one needs to sum over all the possible configurations of $Z_i$. The number of terms in this summation is $2^2 2 = 4,194,304$, which is tractable but introduces heavy computation.

To avoid such computational burden, we apply an alternative approach which considers the contributions from other chromosomes while still handling one chromosome at a time. In this approach, we loop over chromosomes. At each chromosome, we apply the EM algorithm and then we update $y^{\text{father}}$ and $y^{\text{mother}}$ to the residual of the current model. In other words, the procedure is:

1. Initialization: $\beta = 0$ (or $b = 0$) and $Z^1 = \cdots = Z^{22} = 0.5$.
2. For chromosome 1 to 22, do the following until convergence
   (a) Run EM updates until convergence and update $Z^k$ correspondingly.
   (b) Let $y^{\text{father}} \leftarrow y^{\text{father}} - \hat{y}^{\text{father}}$ and $y^{\text{mother}} \leftarrow y^{\text{mother}} - \hat{y}^{\text{mother}}$ where $\hat{y}$ is the predicted values of the current EM fit.
3. Return $Z^1, \cdots, Z^{22}$

### *4.8.3 The algorithm for soft-GWAS*

In this paper, we are interested in testing the association between a focal phenotype and each of the SNPs. The likelihood of the GWAS problem is described in Eq 4.10 where $H^1$ and $H^2$ represent the haplotypes of a single SNP. And from the imputation results, we have

$\Pr(Z_i = 1) = \gamma_i$. Let's further assume that $y^\pi = H^\pi\beta + e,\ e \sim N(0, \sigma^2)$. The model parameters can be estimated by maximum likelihood estimation

$$\hat{\beta}, \hat{\sigma}^2 = \arg\max_{\beta,\sigma^2} \Pr(y^\pi | H^1, H^2; \beta, \sigma^2) \tag{4.39}$$

And we can construct a likelihood ratio test as follow

$$\lambda_{\text{LR}} = -2[\log \Pr(y^\pi | H^1, H^2; \beta = 0, \sigma^2 = \hat{\sigma}_0^2) - \log \Pr(y^\pi | H^1, H^2; \beta = \hat{\beta}, \sigma^2 = \hat{\sigma}^2)] \tag{4.40}$$

where $\hat{\sigma}_0^2 := \arg\max_{\sigma^2} \Pr(y^\pi | H^1, H^2; \beta = 0, \sigma^2)$.

Notice that solving Eq 4.39 is similar to the on-the-fly approach as described in Section 4.8.1. Here we need to replace the full-chromosome haplotypes in the on-the-fly approach with single-SNP haplotype and use $\Pr(Z_i) = \gamma_i$ instead of $\Pr(Z_i = 1) = 0.5$. Furthermore, the current EM scheme can be adapted to handling binary trait with minor modifications.

### *4.8.4 Derivation of the power and bias in imputed-GWAS*

Without loss of generality, we assume that the imputed-GWAS is on $y^{\text{father}}$ $\widetilde{X}$ where $\widetilde{X} = \gamma H_1 + (1 - \gamma)H_2$ (here we use $H_k$ in place of $H^k$ to avoid potential ambiguity). And the true model (allelic test model) is $y^{\text{father}} = H_1\beta + \epsilon$. So, the imputed-GWAS estimate is

$$\hat{\beta} := (\widetilde{X}'\widetilde{X})^{-1}(\widetilde{X}' y^{\text{father}}) \tag{4.41}$$

In the following, we derive $\text{E}(\hat{\beta})$ and $\text{Var}(\hat{\beta})$ in order to analyze the bias and power of the imputed-GWAS estimate.

$$\widetilde{X}'\widetilde{X} = (\Gamma H_1 + (I-\Gamma)H_2)'(\Gamma H_1 + (I-\Gamma)H_2) \tag{4.42}$$

$$\approx \sum_i \gamma_i^2 H_{1,i}^2 + \sum_i (1-\gamma_i)^2 H_{2,i}^2 \tag{4.43}$$

$$(\text{since } H_1 \perp\!\!\!\perp H_2) \tag{4.44}$$

$$\approx H^2[\ \sum_i \gamma_i^2 + (1-\gamma_i)^2\ ] \tag{4.45}$$

$$(H \perp\!\!\!\perp \gamma \text{ and } H_1, H_2 \sim iid) \tag{4.46}$$

$$= H^2 N \overline{S} \tag{4.47}$$

$$(\text{let } S = \gamma^2 + (1-\gamma)^2) \tag{4.48}$$

where $\Gamma = \text{diag}(\gamma)$ and $N$ is the sample size of the imputed GWAS.

$$\widetilde{X}' y^{\text{father}} = [\ \Gamma H_1 + (I-\Gamma)H_2\ ]'(\ H_1\beta + \epsilon\ ) \tag{4.49}$$

$$= (\ H_1'\Gamma H_1\beta + H_1'\Gamma\epsilon\ ) + (\ H_2'(I-\Gamma)H_1\beta + H_2'(I-\Gamma)\epsilon\ ) \tag{4.50}$$

$$\approx H_1'\Gamma H_1\beta + H_1'\Gamma\epsilon + H_2'(I-\Gamma)\epsilon \tag{4.51}$$

$$(\text{since } H_1 \perp\!\!\!\perp H_2) \tag{4.52}$$

$$\approx H^2 N\overline{\gamma}\beta + \widetilde{X}'\epsilon \tag{4.53}$$

Based on Eq 4.47 and 4.53, we have

$$\text{E}(\hat{\beta}) \approx \text{E}[\frac{H^2 N\overline{\gamma}\beta + \widetilde{X}'\epsilon}{H^2 N\overline{S}}] \tag{4.54}$$

$$= \frac{\overline{\gamma}}{\overline{S}}\beta \tag{4.55}$$

$$(\text{since } H_1 \perp\!\!\!\perp \epsilon \text{ and } H_2 \perp\!\!\!\perp \epsilon \text{ so that } \text{E}(\widetilde{X}'\epsilon) = 0) \tag{4.56}$$

$$\mathrm{Var}(X'\epsilon) = \mathrm{Var}(H_1'\Gamma\epsilon) + \mathrm{Var}(H_2'(I-\Gamma)\epsilon) \tag{4.57}$$

$$(\text{since } H_1 \perp\!\!\!\perp H_2) \tag{4.58}$$

$$= \sum_i \gamma_i H_{1,i}^2 \mathrm{Var}(\epsilon) + \sum_i (1-\gamma_i)^2 H_{2,i}^2 \mathrm{Var}(\epsilon) \tag{4.59}$$

$$\approx H^2 \mathrm{Var}(\epsilon)[\ \sum_i \gamma_i^2 + (1-\gamma_i)^2\ ] \tag{4.60}$$

$$(\text{since } H \perp\!\!\!\perp \gamma \text{ and } H_1, H_2 \sim iid) \tag{4.61}$$

$$= H^2 N\overline{S}\mathrm{Var}(\epsilon) \tag{4.62}$$

$$\mathrm{Var}(\hat{\beta}) = \mathrm{Var}(\ \frac{\widetilde{X}' y^{\text{father}}}{\widetilde{X}'\widetilde{X}}\ ) \tag{4.63}$$

$$\approx \frac{1}{[\ H^2 N\overline{S}\ ]^2} \mathrm{Var}(\widetilde{X}' y^{\text{father}}) \tag{4.64}$$

$$= \frac{1}{[\ H^2 N\overline{S}\ ]^2} \mathrm{Var}(H^2 N\overline{S} + \widetilde{X}'\epsilon) \tag{4.65}$$

$$= \frac{1}{[\ H^2 N\overline{S}\ ]^2} \mathrm{Var}(\widetilde{X}'\epsilon) \tag{4.66}$$

$$= \frac{1}{[\ H^2 N\overline{S}\ ]^2} H^2 N\overline{S}\mathrm{Var}(\epsilon) \tag{4.67}$$

$$= \frac{\mathrm{Var}(\epsilon)}{H^2 N\overline{S}} \tag{4.68}$$

Based on Eq 4.55 and 4.68, we have

$$\mathrm{E}(T) := E[\frac{\hat{\beta}}{\sqrt{\widehat{\mathrm{Var}}(\hat{\beta})}}] \tag{4.69}$$

$$\approx \frac{E(\hat{\beta})}{\sqrt{\mathrm{Var}(\hat{\beta})}} \tag{4.70}$$

$$= \frac{\overline{\gamma}}{\overline{S}} \frac{\sqrt{H^2 N \overline{S}}}{\sqrt{\mathrm{Var}(\epsilon)}} \beta \tag{4.71}$$

$$= \frac{\overline{\gamma}}{\sqrt{\overline{S}}} \underbrace{\sqrt{\frac{NH^2}{\mathrm{Var}(\epsilon)}} \beta}_{\mathrm{E}(T^\star) \text{ from } y^{father} \sim H_1} \tag{4.72}$$

$$= \frac{\overline{\gamma}}{\sqrt{\overline{S}}} E(T^\star) \tag{4.73}$$

Eq 4.73 follows since in the ideal case, $\widetilde{X} = H_1$, $\overline{S} = \overline{\gamma} = 1$ so that $\mathrm{E}(T^\star) = \sqrt{\frac{NH^2}{\mathrm{Var}(\epsilon)}}\beta$.

# CHAPTER 5

# BRAINXCAN IDENTIFIES BRAIN FEATURES ASSOCIATED WITH BEHAVIORAL AND PSYCHIATRIC TRAITS USING LARGE SCALE GENETIC AND IMAGING DATA

Material from: Liang, Yanyu, Owen Melia, Thomas Brettin, Andrew Brown, and Hae Kyung Im, "BrainXcan identifies brain features associated with behavioral and psychiatric traits using large scale genetic and imaging data.", medRxiv, preprint 2021, Cold Spring Harbor Laboratory Press [82]

## 5.1 Abstract

Advances in brain MRI have enabled many discoveries in neuroscience. Comparison of brain MRI features between cases and controls have highlighted potential causes of psychiatric and behavioral traits. However, due to the cost of collecting MRI data and the difficulty in recruiting particular patient groups, most studies have small sample sizes, limiting their reliability. Furthermore, interpretation is complicated by reverse causality, where many observed brain differences are the result of disease rather than the cause. Here we propose a method (BrainXcan) that leverages the power of large-scale genome-wide association studies (GWAS), reference brain MRI data, and methodological advances in causal inference using genetic instruments to discover new mechanisms of disease etiology and validate existing ones. BrainXcan tests complex traits for association with genetic predictors of brain MRI derived phenotypes to pinpoint relevant region-specific and cross-brain features. It also evaluates consistency and direction of the causal flow with Mendelian Randomization. As this approach requires only genetic data, BrainXcan allows us to test a host of hypotheses on mental illness, across many disorders and MRI modalities, using existing public data resources. Our method shows that reduced axonal density across the brain is associated with the risk

of schizophrenia, consistent with the disconnectivity hypothesis. We also find structural features hippocampus, amygdala, and anterior cingulate cortex among others associated with schizophrenia risk highlighting the potential of our approach to bring orthogonal lines of evidence to inform the biology of complex traits.

## 5.2 Introduction

Advances in brain MRI have enabled many discoveries in neuroscience. However, reproducibility of brain-wide associations studies (BWAS) is low due, in large part, to small sample sizes [95]. These small sample sizes are the result of the high cost of collecting MRI scans, as well as the difficulty in recruiting patients with particular mental illnesses. Also, unlike genome-wide association studies where disease status does not alter germline genetic variation, brain features can be altered by disease status and treatments, which can yield significant associations due to reverse causality.

The UK Biobank is in the process of measuring brain MRI in 100,000 individuals [84]. The unprecedented scale of the data, the automated uniform processing of the data, the availability of genetic and a myriad of phenotypic data will undoubtedly catalyze many discoveries in the coming years. The interim analysis of brain MRI image derived phenotypes (IDPs) found many genome-wide significant loci associated and established that most IDPs are heritable [129]. Zhao et al generated polygenic risk scores of 101 brain volumetric phenotypes using 19,629 UK Biobank participant data and showed they could explain more than 6% of the phenotypic variance in four independent studies [167].

The Psychiatric Genomics Consortium is a cooperative effort of investigators across the world that combines studies of many mental disorders and has enabled discoveries that would not have been possible within each of the studies. All their GWAS summary results are publicly available to allow other investigators to test their own hypotheses and extract new biological insight. The PGC studies 11 psychiatric disorders including ADHD, Alzheimer's

disease, autism, bipolar disorder, and schizophrenia.

Methods that leverage UK Biobank's large scale image data and the PGC's large scale GWAS data have the potential to unlock many insights into the biology of mental disorders. In this paper we propose one such method, BrainXcan, which leverages these two data resources to address some of the deficiencies in small scale MRI studies. Using UK Biobank data as a reference, we build models to predict brain IDPs from genetic data. These models can then be applied to from genome-wide association studies. For example, using the schizophrenia GWAS data collected by the PGC, our method tests for association between schizophrenia and a number of different functional, structural and diffusion MR modalities with size of $\sim 70,000$ cases and $\sim 240,000$ controls. Furthermore, by applying a Mendelian randomization approach we infer the direction of causality: whether the changes in IDP are the cause of disease or a consequence of it.

IDP-associated genetic markers have been used for causal inference with methods such as Mendelian Randomization to investigate the mediating role of brain features on behavioral phenotypes with both large sample sizes and protection from reverse causality. For instance, [60] studied the genomic loci and corresponding genes that are shared between brain volume IDPs and intelligence and they identified 92 shared genes which provided insight of the shared genetic etiology of brain volume and intelligence. [125] performed bi-directional MR analysis with depression and dMRI IDPs finding suggestive evidence that the change of the mean diffusivity in thalamic radiations could be a consequence of major depressive disorder. A related approach is one that correlates genetically predicted brain IDP/phenotype and the complex trait, an extension of transcriptome-based methods [43, 51] to IDPs. Based on this idea, [68] developed imaging-wide association study (IWAS) using 14 brain features from the Alzheimer's Disease Neuroimaging Initiative. They also used standard PRS approaches to generate prediction weights using the GWAS summary results from [36] (n=8,428).

In this paper, we perform an in-depth analysis of the genetic architecture of IDPs and

further process UK Biobank's IDPs to develop a framework that maximizes interpretability, robustness, computational efficiency, and user friendliness.

The high polygenicity of brain features imposes several challenges to existing methods limiting the power to detect their link to diseases; strong genetic instruments needed for Mendelian randomization based approaches are difficult to identify. We address these challenges by developing polygenic predictors of IDPs informed by their complex genetic architecture and correlation structure. To facilitate interpretation of the results, we develop region-specific and brain-wide predictors providing an in-depth analysis and quantification of potential biases. We make sure that the implementation is computationally efficient and scalable to genome-wide Biobank-scale data. We develop an extension of the association method that can infer the association using the increasingly available GWAS summary results, i.e. without the need to use individual level data. We add a Mendelian Randomization module to estimate the direction of the causal flow. We illustrate the power of the approach by applying it to 44 human traits. Finally, we provide the software, the recommended pipeline, and automated reports to improve usability and lower the barrier to adoption for users less familiar with genetic studies.

## 5.3 Results

### *5.3.1 Overview of the BrainXcan framework*

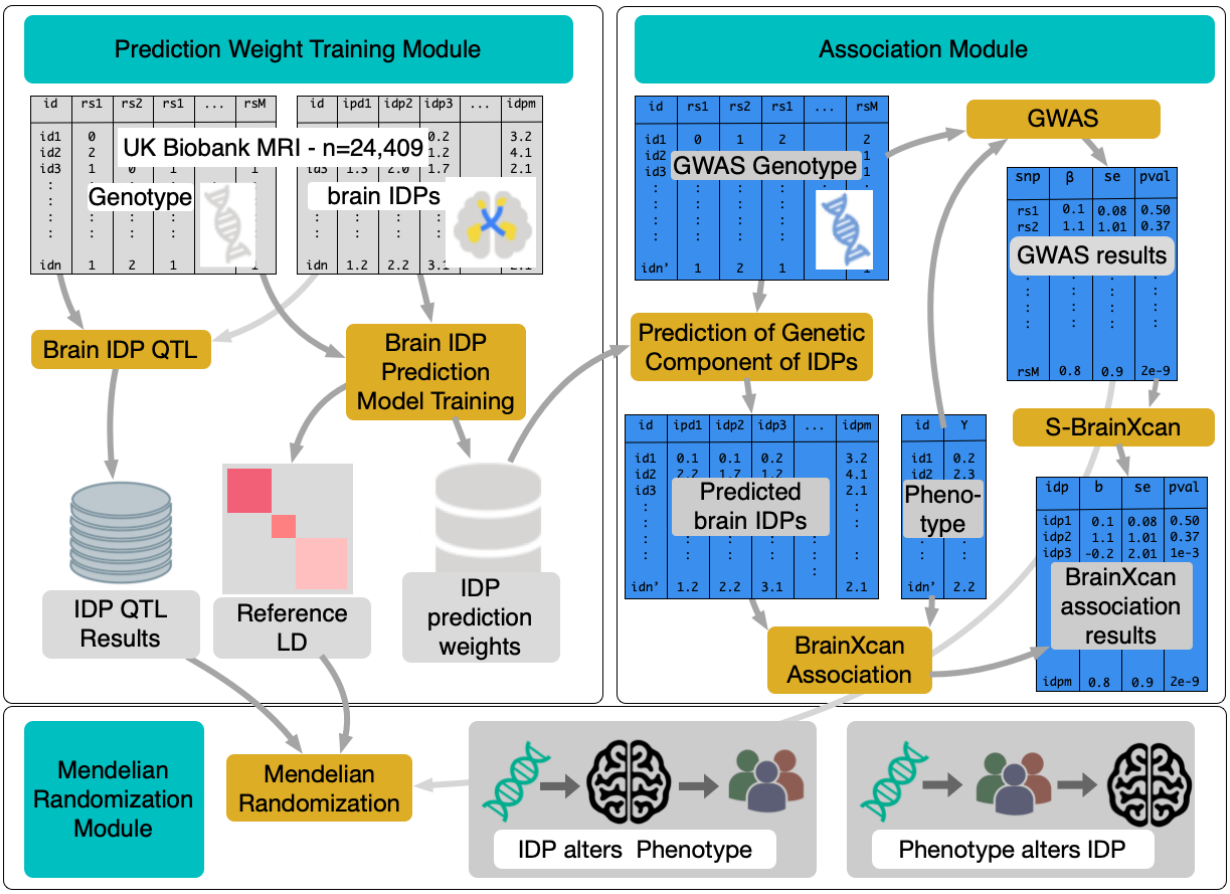



**Figure 5.1: The workflow for implementing BrainXcan framework.** The workflow is divided into three modules: "Prediction weight training", "Association between predicted IDP and phenotype", and "Mendelian randomization". The "Prediction weight training" module uses reference brain MRI data with genotype and IDPs and trains genetic predictors of brain IDPs. It also computes QTLs of brain IDPs and LD information, i.e. genotype covariance. This module only needs to be run once and the results are provided to users as part of the BrainXcan software package. The "Association" and "Mendelian randomization" modules are the main analysis components that will be performed by most BrainXcan users. In a typical use case, users first run "Association" module with either individual-level phenotype/genotype data or the GWAS summary statistics. The significant (with user-defined threshold) IDPs, will be processed with the "Mendelian randomization" module to examine the direction of the putative causal flow and the consistency across multiple loci.

The BrainXcan framework is organized in three modules as outlined in Figure 5.1. The **"Prediction weight training" module** trains linear genetic predictors of brain IDPs, performs brain IDP QTL analysis (association between brain IDPs and genotype) and calculates the sample covariance of the genotypes. These outcomes are saved for use in subsequent modules and shared publicly in `predictdb.org` with versioned and permanent record in zenodo.org.

The **"Association" module** operates on the GWAS of the phenotype of interest where genotype and phenotype data are available. First, it computes the genetic predictors of brain IDPs using the genotype data and the weights from the training stage. Next, it calculates the association statistics between the predicted brain IDPs and the phenotype of interest via linear regression. Generlized linear models to take into account binary or other types of outcomes can be easily accommodated. Significant associations pinpoint candidate causal relationship between brain features and the phenotype. As explained in the IDP processing section, we use further derived IDPs that represent region-specific features as well as brain-wide features.

For most large scale GWAS studies, the individual level data are not available but the BrainXcan framework can still be applied because the association statistics can be inferred using the summary results from GWAS, the IDP model weights, and the reference LD data generated by the first module (see Section 5.5.7). The feasibility of this approach was shown with the PrediXcan framework [7].

The **"Mendelian randomization" module** performs a number of multiple instrument-based Mendelian randomizations to determine the direction of the putative causal flow, i.e. whether alteration in brain features affects the complex trait or whether the phenotype status alters brain features. It provides bi-directional test of causal flow and effect size scatter plots to help assess the consistency of the results.

### *5.3.2 Preprocessing brain MRI derived phenotypes*

We downloaded uniformly preprocessed brain MRI derived phenotypes from the UK Biobank (IDPs). We focused here on 159 IDPs derived from structural images representing total and gray matter volumes of different regions of the brain and 300 diffusion MRI derived phenotypes representing neurite density, dispersion, and connectivity features. After excluding related individuals and those of non European ancestry, IDP data on 24,409 individuals remained. We adjusted for covariates including the first 10 genetic PCs, age at recruitment, sex, and four technical covariates indicating the location of the head in the scanner. See details in (Section 5.5.1) and the list of IDPs in Table 5.1.

We processed the structural and diffusion MRI modalities separately. We further categorized the structural IDPs into 5 subtypes: cortical gray matter volume, sub-cortical gray matter volume, sub-cortical total volume, cerebellum gray matter, and brain-stem. We also included total gray matter volume and total gray and white matter volume as additional IDPs. Diffusion MRI measures were also categorized into 4 subtypes: fractional anisotropy (FA, a measure of diffusion along the white matter tracts), intracellular volume fraction (ICVF, an estimate of neurite/axonal density), isotropic volume fraction (ISOVF, an index of the relative extra-cellular water diffusion), and orientation diffusion index (OD, a measure of neurite dispersion) [164]. Our main analysis focused on the Tract-based spatial statistics (TBSS) IDPs (192 in total) while the probability track based measures (108 in total) were left for internal validation.

For each subtype, we performed a principal components analysis. The first principal component was a weighted average of IDPs in the subtype suggesting that they could be used as proxies for the brain-wide feature (Figure 5.9). For diffusion MRI subtypes, the principal component explained 39% (FA), 53% (ICVF), 26% (ISOVF), and 20% (OD) of the total variance. For the structural MRIs, the first principal component explained 16% (cortical gray matter), 34% (sub-cortical gray matter), 37% (sub-cortical total volume), and

45% (cerebellum gray matter) of the total variability. The residual IDPs after adjusting for the first principal component within each subtype, had a much reduced correlation structure as shown in Figure 5.10 and 5.11.

### *5.3.3 Brain IDPs can be decomposed into common and region-specific features*

Since an important attribute of any method is the interpretability of the results, we sought to define brain features with the goal of facilitating interpretation. We prioritized the ability to tease out brain-wide effects from region-specific effects. i.e., determining whether a detected association with the phenotype was due to a feature that is common across the whole brain or specific to a region (e.g. thalamus, anterior limb of the internal capsule, etc.).

To distinguish between brain-wide effects and region-specific effects, we postulated the generative model shown in Figure 5.2a. The brain feature in each region ($F_k$) was modeled as the sum of two independent latent components: one region-specific ($R_k$) and one brain-wide ($L$). The observed value, IDP, was modeled as a noisy version of the region's feature ($\text{IDP}_k = F_k + \epsilon_k = L + R_k + \epsilon_k$). The parameter $s^2$ determined the scale of the region-specific component (modeled as a normal random variable with variance $s^2$) and $t^2$ was the variance of noise term $\epsilon_k$.

### *5.3.4 Attenuation and collider biases can be estimated*

Our effects of interest were represented as $\beta_k$ (region-specific) and $\alpha$ (brain-wide). Ideally, we would like to fit a regression model of the trait $Y$ jointly on the brain-wide ($L$) and the region-specific ($R_k$) components but since the latent variables were not available to us, we used the principal component (PC) and the residual IDPs (residual of $\text{IDP}_k$ after regressing out the subtype's PC) as proxies.

To examine the effect of using these proxies instead of the latent factors $R_k$ and $L$, we

derived analytical expressions for the bias when regressing the trait on one of the residual IDPs ($Y \sim \text{resIDP}_k$) and the subytpe's PC ($Y \sim \text{PC}$) (see details in Section 5.8.1).

The overall noise level within $\text{IDP}_k$ led to what is known as attenuation bias whereas adjusting for PCs (weighted sum of IDPs) led to what is known as collider bias [29].

$$\text{coef IDP}_k = \beta_k - \frac{t^2}{s^2+t^2} \cdot \beta_k - \frac{s^2}{s^2+t^2} \sum_{j \neq k} \frac{\beta_j}{m-1} + O_p\left(\frac{1}{n}\right), \tag{5.1}$$

where $m$ is the number of IDPs in the subtype (ranging from 14 to 96) and $n$ is the sample size of the GWAS study or equivalently of the BrainXcan association study (typically $\sim 100K - 1M$). The attenuation bias reduced the estimate of $\beta_k$ by a factor of $(1 - \frac{t^2}{s^2+t^2})$. The third term represents the collider bias, proportional to the average effect size across brain regions, which can be expected to be relatively small if the effect is specific to a region.

Regression on the subtype's PC instead of the latent variable $L$ yields a biased coefficient:

$$\text{coef PC} = \alpha + \sum_j \beta_j + O_p\left(\frac{1}{m}\right) + O_p\left(\frac{1}{n}\right). \tag{5.2}$$

This coefficient is the sum of the latent brain-wide effect ($\alpha$) and the effects of individual regions. The interpretation of this coefficient will depend on the significance of the individual region effects estimated in Eq.(5.1). If all region-specific effects are small and not significant, then we can assume that the brain-wide effect $\alpha$ is not biased. However, if the attenuation bias is very large, it is possible that the effect is missed in the first regression (Y on residaul IDP) but detected in Eq.(5.2). These considerations must be carefully taken into account for the interpretation.

Next, to better inform optimal prediction approaches, we proceeded to investigate the genetic architecture of these brain features by calculating their heritability and degree of polygenicity.

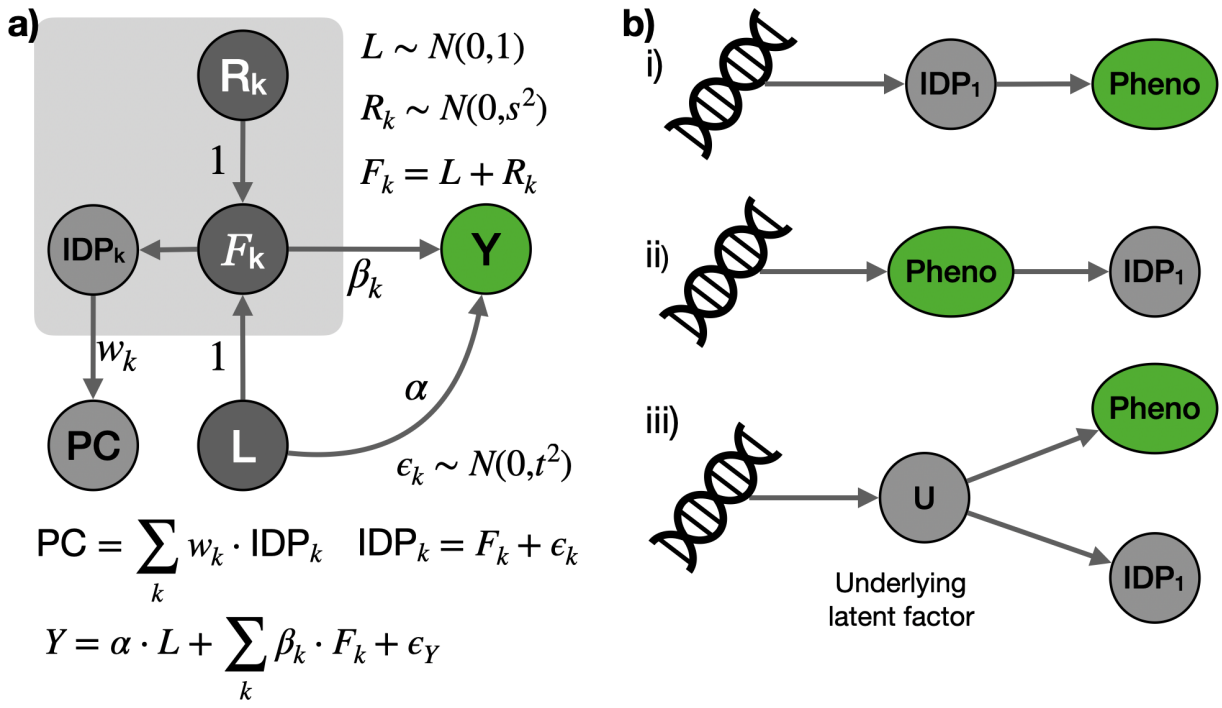



**Figure 5.2: Generative model of brain features and complex traits. a)** Brain features, $F_k$, were modeled as a sum of a region-specific effect $R_k$ and brain-wide component $L$. The IDPs were considered to be noisy measurements of brain features. Principal components of IDPs were weighted averages of IDPs. The complex trait $Y$ was modeled as the sum of brain-wide effect ($\alpha$) and region-specific effects ($\beta_k$) and an error term ($\epsilon_Y$). **b) Mendelizan Randomization causal flow interpretation.** Associations between brain IDPs and the phenotype can arise from multiple mediating scenarios. We considered i) brain IDP alters phenotype, ii) phenotype (disease status) alters brain IDP, iii) underlying latent factor alters both phenotype and brain IDP. Given the power differential with current GWAS and reference image datasets, significant scenario ii) may not rule out scenario i) or iii). See discussion in text.

### 5.3.5 Both global and region-specific brain features are heritable and highly polygenic

We calculated the heritability of brain IDPs using standard mixed effects modeling approach [160] (Section 5.5.3). Heritability estimates ranged from 5% to 43% with all the 95% confidence intervals above zero as shown in Figure 5.3a. Since principal components were heritable, the residual IDPs (PC-adjusted) were less heritable than the original IDPs. (Figure 5.12).

To quantify the degree of polygenicity of brain IDPs, we estimated the effective number of independently associated SNPs ($M_e$) using the stratified LD fourth moments regression [107] (Section 5.5.4 and Figure 5.13). Two hundred and sixty five out of 359 IDPs yielded a significant ($p < 0.05$) estimate of $M_e$ with values ranging from 1,035 to 24,675 with a median of 6,245 which was higher than the estimates for canonical examples of polygenic traits such as height or BMI. The estimates are shown in Figure 5.3b with common human traits added for reference.

### 5.3.6 Ridge regression predicts brain features better than elastic net

We trained genetic predictors of the original brain IDPs, the derived principal components, and the residual IDPs using penalized regression approaches, ridge regression and elastic net. We restricted our search to linear models so that BrainXcan could be applied using solely GWAS summary statistics even when the individual level genotype and phenotype data were not available. Given the high polygenicity of IDPs, we anticipated that many of the predictors would be based on ridge regression with all the SNPs having a nonzero weight. Therefore, to keep the computation manageable and to make the prediction models applicable to a broader set of GWAS studies where access to individual level data may not be possible, we restricted the training to HapMap3 SNPs, which tend to be imputed with high quality, with MAF $> 0.01$ in European samples. To avoid issues with strand mis-

specification, we excluded the ambiguous SNPs (e.g. AT, CG). These restrictions left us with a total of 1.07 million SNPs (Section 5.5.2) for the subsequent procedure.

We trained two sets of models, one with a ridge and the other one with a elastic net penalty (Section 5.5.5). Recall that ridge regression uses $l_2$ penalty and yields highly polygenic predictors (all non zero weights) whereas elastic net yields sparse models, setting the weights of most variants to 0. Given the high polygenicity of IDPs, we expected ridge regression to perform better than the sparsity-inducing elastic net penalty.

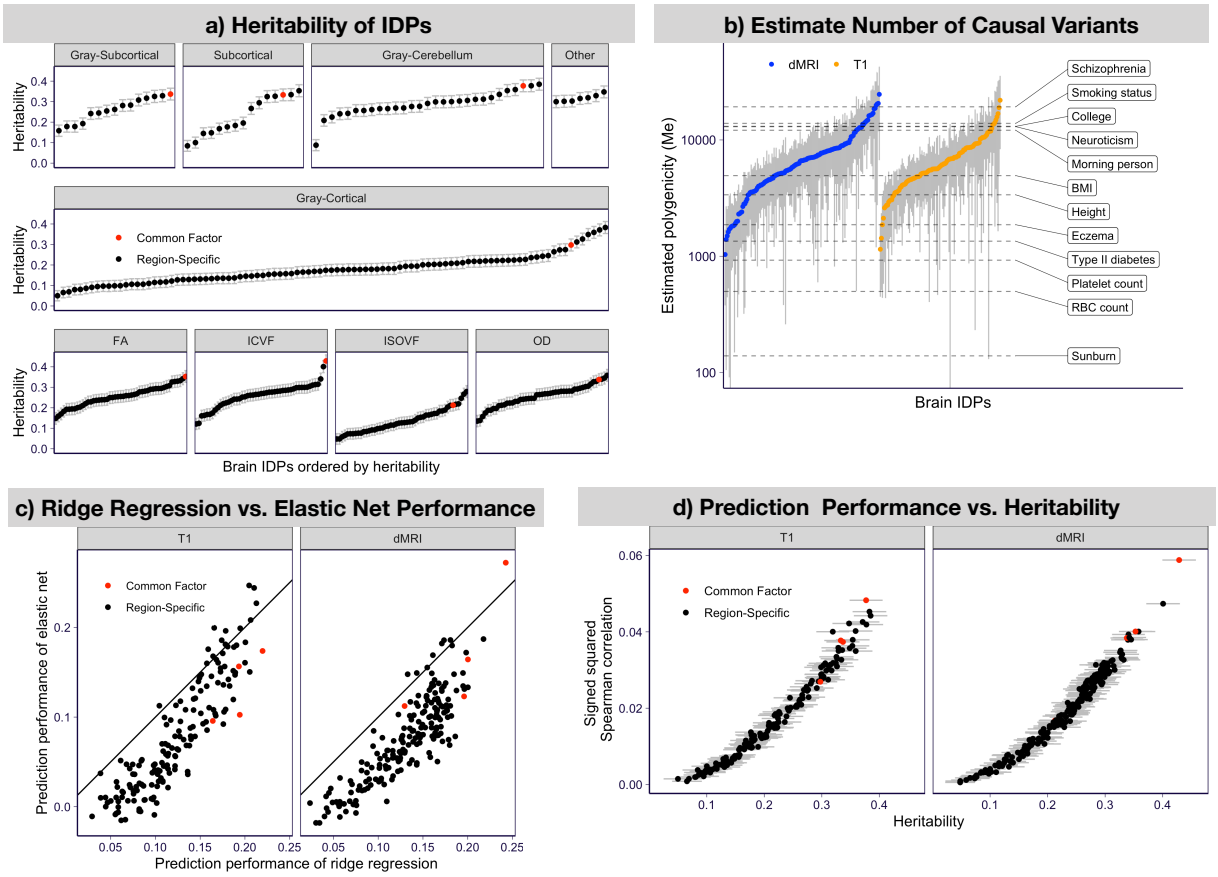



**Figure 5.3: Genetic architecture of IDPs and prediction. a) Heritability of IDPs.** Heritability and 95% confidence intervals are shown by IDP subtype: volumes of gray matter subcortical structures, total volume of subcortical structures, volume of cortical regions and other volumes among structural IDPs. The third row show heritability of diffusion IDPs. Red show common factors (principal components of each subtype) and black are the component IDPs in the subtype. FA: fractional anisotropy; ICVF: intracellular volume fraction; ISOVF: isotropic volume fraction; OD: orientation dispersion index. **b) Polygenicity of IDPs.** The effective number of independently association SNPs ($M_e$) estimated using the stratified LD fourth moments regression [107] for 522 brain IDPs are shown on the y-axis. Gray bars indicate 95% confidence intervals. Horizontal dashed lines indicate the estimated $M_e$ of 12 complex traits for reference. **c) Prediction performance of IDPs.** This panel compares the performance of ridge regression vs. elastic net approaches predictors measured by the Spearman correlation between the predicted and observed IDP values in a five-fold cross-validated scheme. For each brain IDP, the performance of ridge predictor is shown on the x-axis and the performance of elastic net predictor is shown on the y-axis. The black solid line is the identity line ($y = x$). The IDP PCs are in red and the remaining brain IDPs are in black. **d) Prediction performance vs. heritability.** The signed squared Spearman correlations between observed and predicted IDPs of the ridge regression are shown on y-axis and the estimated heritability are shown on x-axis. The signed squared correlation is defined as $\text{sign}(x) \cdot x^2$ for correlation $x$ to preserve the sign of the correlation while taking the square. The error bar indicates the 95% confidence interval of the estimated heritability. The IDP PCs are in red and the rest of the brain IDPs are in black.

To evaluate the model performance we calculated the Spearman correlation between the observed and predicted IDP values with a five-fold cross-validation scheme (Section 5.5.5.3), shown in Figure 5.3c. For ridge regression the prediction performance ranged from 0.024 to 0.24 with a median of 0.13. For elastic net the range was between -0.018 and 0.27 with a median of 0.075. For most IDPs, ridge regression yielded higher cross-validated performance than elastic net (See Figure 5.3c.)

We also found that the improved performance of ridge regression over elastic net predictions was correlated with the polygenicity ($M_e$) of the trait (Figure 5.14), corroborating our intuition that ridge regression performs better for polygenic traits whereas elastic net performs better for more sparse traits.

All the ridge predictors and 95% of the elastic net predictors showed positive cross-validated Spearman correlation (Figure 5.15 and Table 5.2) demonstrating the feasibility of genetically predicting IDPs. With these predictors in hand we moved to the next stage of the development of the framework where we predicted IDPs using genotype alone and correlated these predicted values with complex traits.

As expected, the prediction performance increased with the heritability of the IDP, i.e. more heritable traits were predicted better as shown in Figure 5.3d. However, we also noted that the median prediction $R^2$ was lower than 8% of the heritability (Figure 5.3c), an upper bound of how the performance of linear predictors. This low proportion of heritability captured by the genetic predictors highlights the need to increase the sample size of reference image data to reach the upper bound of the performance.

We filtered out unreliable predictors by keeping only the ones that showed prediction performance correlation greater than 0.1. Among structural IDP residuals, 105 ridge predictors and 54 elastic net predictors passed the threshold from a total of 159 trained. Among 192 diffusion IDP residuals, 148 ridge predictors and 62 elastic net predictors passed the threshold. All subtype-level PCs except the elastic net-based PC predictor of cortical region

volumes were well predicted and kept for the subsequent analysis.

### *5.3.7 Summary BrainXcan finds disease-associated brain features using GWAS summary statistics*

To expand the applicability of our tool, we extended the BrainXcan association module so that it could infer the association statistics using the GWAS summary results of the traits without the need to use individual level genotype and phenotype data. The ideas are similar to the S-PrediXcan method used for correlating genetically predicted transcriptome with complex traits. However, unlike gene expression prediction which only used variants in the vicinity of the gene, IDPs needed to handle a dense number of genetic predictors across the genome. To make this computation feasible, we developed a scalable method that could handle this added complexity as described in the Method section.

To enable the application of the method to situations where only summary statistics are available, we calculated the covariance between the HapMap3 SNPs to be used downstream and saved in a sparse format, setting correlations between SNPs that were separated by more than 200 SNPs to be 0. We also performed genetic associations of all the IDPs (GWAS of IDPs) and saved the results for the Mendelian randomization analysis downstream.

### *5.3.8 BrainXcan association: correlating genetically predicted IDPs with phenotypes*

We selected 9 phenotypes from the UK Biobank and performed BrainXcan association using data from 327,743 individuals of British ancestry. As described in the overview section above, we calculated the genetically predicted IDPs for all the individuals and correlated them with the phenotypes using linear regression. The phenotypes included alcohol consumption, smoking, coffee consumption, depression, parental depression, parental Alzheimer's disease,

handedness, BMI, and height. See detailed list on table 5.3. To avoid overfitting issues, we excluded individuals used for the training of the prediction models.

We also performed summary BrainXcan association analysis on 35 GWAS for which we did not have access to the individual level data. These phenotypes included behavioral, psychiatric and neurologic phenotypes, height, and body mass index (see Table 5.4).

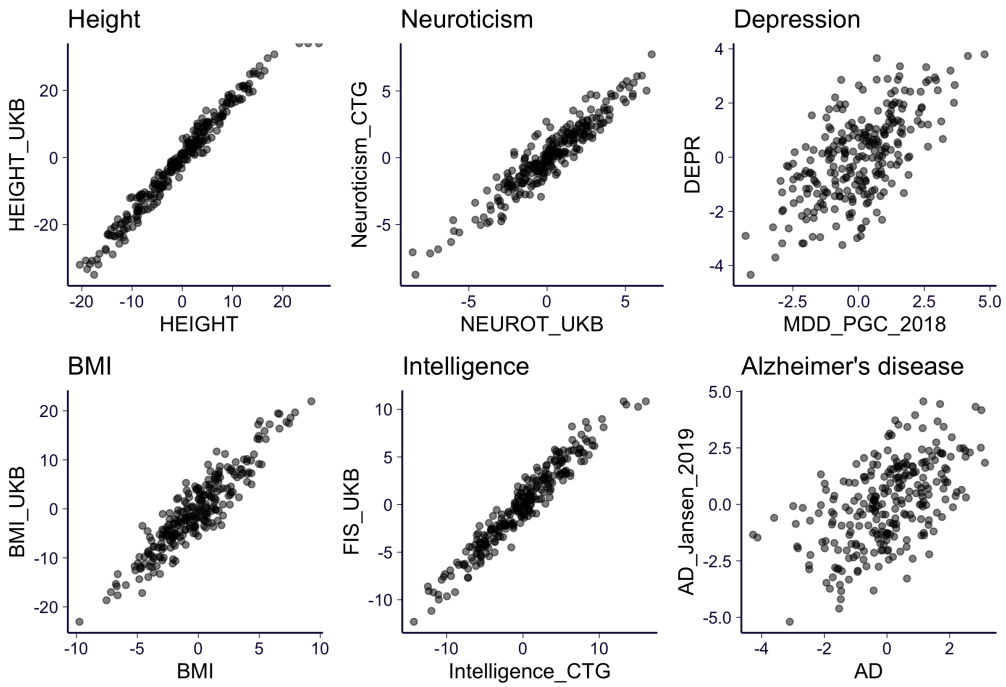
We confirmed the reliability of the summary version of BrainXcan (S-BrainXcan) by comparing the z-scores of the associations to the ones obtained from individual-level BrainXcan for standing height and body mass index in the UK Biobank. The highly concordant z-scores are shown in Figure 5.19. We also observed concordant BrainXcan z-scores (Figure 5.16) between ridge and elastic net predictors indicating robustness to the choice of prediction approach. Ridge predictors yielded more significant results, consistent with their increased prediction performance compared to elastic net.

Combining both summary level and UK Biobank traits, 98% of IDPs (257 out of the 261 IDPs in the main analysis) were significantly associated with at least one trait. As expected, better powered GWAS traits with larger number of significant associations also yielded more significant IDPs to trait associations.

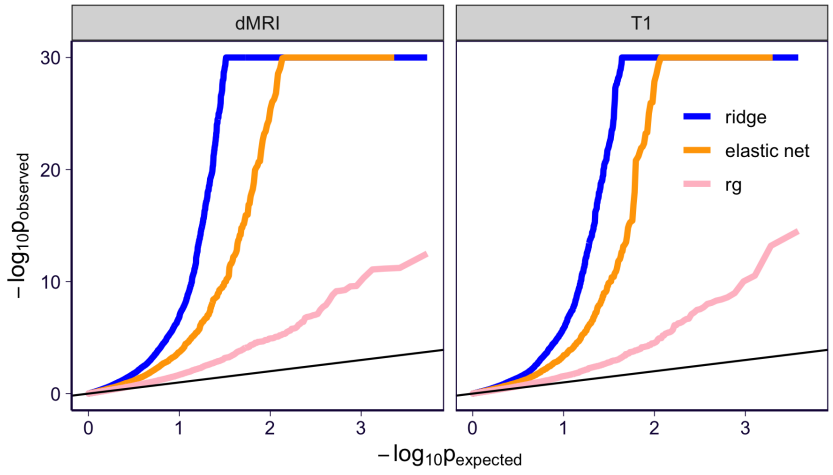
Better predicted IDPs were more significantly associated with phenotypes, likely due to reduced attenuation bias (Figure 5.17). Brain-wide features represented by PC's were more significantly associated with complex traits than region-specific features which could be explained by the higher predictability of the PC's but could also point to brain-wide effects being more prevalent among the selected GWAS traits (Figure 5.18).

### *5.3.9 Association results replicate in independent datasets*

For a number of traits, we had two independent GWAS of the same trait, which allowed us test the replication of our results. Reassuringly, as shown in Figure 5.4a, we found highly concordant association z-scores between the independent studies.




**(A)**




**(B)**

**Figure 5.4: S-BrainXcan association statistics for 35 GWAS. (a)** Six phenotypes which has multiple GWASs being analyzed by S-BrainXcan are shown. For each brain IDPs, the S-BrainXcan z-scores (from ridge models) from the two GWASs are shown on x-axis and y-axis respectively (see the GWAS label in Table 5.4). **(b)** The S-BrainXcan p-values are shown as the QQ-plot (against the expected p-values under the null). For visualization purpose, the observed p-values which is smaller than $1 \times 10^{-30}$ are set to $1 \times 10^{-30}$. Label 'rg' represents the genetic correlation result. The black lines are the identity line ($y = x$).

### *5.3.10 Genetic correlations yield similar but less significant associations*

Genetic correlation between brain features and complex traits can provide, in principle, similar information to BrainXcan association results. To compare the ability to identify associations, we computed the correlations for all pairs of IDP/complex traits and compared the results to BrainXcan associations. (Section 5.5.10 and Figure 5.4b). We found that the z-scores of the genetic correlation was highly correlated to the z-scores of BrainXcan associations (with correlations ranging from 0.51 to 0.97, with a median of 0.81 (Figure 5.21). However, BrainXcan yielded 8-fold larger number of significant IDP/trait associations compared to genetic correlation approach suggesting that optimal predictors of IDPs yield more power to identify putative causal links.

### *5.3.11 BrainXcan quantifies evidence for the direction of the causal flow*

Mendelian randomization evaluates the causal relationship between trait 1 and 2 by testing whether increased "exposure" to the first trait (represented as the trait 1 GWAS effect size) is associated with increased or decreased level of the second trait (indicated by the trait 2 effect size). In Mendelian randomization settings, individuals are thought to be randomized at meiosis to either inherit the risk increasing allele or not. Because of this parallel to randomized trials, the level of causal evidence derived from Mendelian randomization is considered to be higher than observational studies albeit lower than actual randomized trials (the gold standard for causal determination used in clinical trials).

It is possible to infer the direction of the causal flow by selecting variants that have strong effects on the first trait and testing for a significant association with the effect sizes of the second trait and vice versa, i.e. selecting the variants with strong effects on the second trait and testing whether they are associated with the effect sizes on the first trait. As described below, scatter plots of effect sizes showing the two selection strategies (by significance of trait 1 or 2), are added to the automated reports.

To determine the direction of the putative causal flow, i.e. whether changes in IDPs are affecting changes in the trait or vice versa, we applied a number of Mendelian randomization approaches with multiple instruments implemented in MR-BASE [53] including inverse variance weighted regression [19], weighted median method [15], and Egger regression [14]. For this purpose, we selected strong instruments, i.e. SNPs with very small p-values. For the complex phenotype, we selected SNPs with GWAS p-values $< 5 \times 10^{-8}$ and for brain IDPs, we selected SNPs with GWAS p-values $< 10^{-5}$. The more loose p-value for IDPs were necessary to have sufficient number of instruments, which could be tighten as the number of individuals in the reference image data increases. To streamline the interpretation of the multiple Mendelian randomization results, we combined the p-values of each Mendelian randomization output using an extension of the ACAT method [87] that take into account the concordance of the sign of the results.

### *5.3.12 Caveats on interpreting Mendelian randomization results*

There are two caveats that need to be considered when interpreting Mendelian randomization results. One is that we first select the IDP-trait pair based on their association. Therefore, the p-values of the Mendelian randomization will not be well-calibrated. Therefore, we propose using the p-values to discern between possible direction of the causal flow, not to quantify significance.

The second caveat relate to the power difference between reference image and GWAS studies. Currently, reference image data have much smaller sample sizes (n$\sim 30K$) compared to GWAS studies of complex traits (n$\sim 100K - 1M$). We consider three main mediating scenarios depicted in Figure 5.2b). In scenario i) the brain feature mediates the genetic association with the phenotype, i.e. genetic risk factors alter the brain feature which in turn alters the risk for the phenotype. In scenario ii) genetic factors affect the phenotype which in turn alter the brain feature. In scenario iii) genetic factors affect an underlying latent

factor which alter both the phenotype and the brain feature.

A significant i) and not significant ii) can be interpreted as evidence that the brain feature alteration is affecting the phenotype given the higher power of GWAS studies in general. However, a significant scenario ii) and non significant scenario i) could simply mean that the instruments (strongly associated SNPs and their effect sizes) for the brain feature are not reliable enough to yield significance. In this case, scenario ii) should be considered supported by the data but scenarios i) and iii) should not be ruled out.

### *5.3.13 BrainXcan use is simplified with an automated pipeline*

To facilitate the analysis to first-time users, we created an automated pipeline using snakemake [103]. The association and MR modules can be performed with default parameters or modified according to the study's specific needs. The pipeline tests all the IDPs for associations and Mendelian randomization test is performed for the top 10 (modifiable parameter of the pipeline). The pipeline will

- run BrainXcan association ($Y \sim$ resIDP and $Y \sim$ PC),
- run bidirectional Mendelian randomization for the top 10 significant IDPs, and
- generate automated report including figures, tables with top associations, Mendelian randomization figures, etc.

### *5.3.14 Application of BrainXcan to Schizophrenia*

To demonstrate the features of BrainXcan, we applied the full pipeline to the schizophrenia GWAS [120]. In this analysis, we tested 327 structural and diffusion MRI-derived phenotypes with cross validated Spearman correlation greater than 10%. For the main analysis we focused on 261 IDPs, which include 48 cortical gray matter volumes, 10 subcortical volumes, 13 subcortical gray matter volumes, fractional anisotropy (water diffusivity along nerve tracts)

in 46 regions, ICVF (intracellular volume fraction) in 44 regions, OD (orientation dispersion) in 45 regions, and ISOVF (isotropic volume fraction) in 13 regions. (Note that ISOVF was a less heritable index leading to fewer successful predictors). Brain-wide measures represented by principal components of each subtype were also included (gray volumes of cortical, cerebellum, and subcortical regions, subcortical total volumes, FA, ICVF, OD, and ISOVF).

Among the 261 IDPs, 46 were significantly associated with risk of schizophrenia after Bonferroni correction (raw p-value $< 0.05/261$). Figs. 5.7 and 5.8 provide a snapshot of the region-specific associations schizophrenia risk. We added an interactive annotation of different regions of the brain is added to the output of the automated pipeline. See example in `https://brainxcan.hakyimlab.org/post/2021/05/06/brainxcan-automated-reports/#interactive-annotation-of-regions`

Among diffusion MRI associations, the principal component of ICVF, a proxy for brain-wide axonal density, was the most significant association for schizophrenia risk, with lower axonal density associated with higher risk of schizophrenia (Figure 5.5). The principal component of fractional anisotropy, a brain-wide measure of water diffusion efficiency along nerve tracts, was also negatively associated with schizophrenia while the orientation dispersion index (dispersion of neurite orientation along tracts) did not show significant association. These results corroborate the hypothesis that schizophrenia is a disorder of disconnectivity. They are also consistent with reduced FA in schizophrenia cases compared to controls as reported by [65]. However, since our technique uses healthy individuals MRI-based genetic prediction of brain features, they are less likely to capture a consequence of the disease.

The total volume of the hippocampus (right side, relative to brain size) was positively associated with schizophrenia risk whereas the total volume of the thalamus (both sides, relative to brain size) were negatively associated. Gray matter volumes of the frontal orbital cortex, the anterior cingulate cortex, and posterior temporal fusiform cortex were positively associated whereas planum polare, amygdala's gray matter volumes were negatively asso-

ciated (Figure 5.6). Observed hippocampus, thalamus, amygdala, anterior cingular cortex volumes have all been reported to be associated with schizophrenia status [142, 126].

Among the top 10 IDPs associated with schizophrenia, we found that 2 IDP to schizophrenia risk causal flow were nominally significant ($p< 0.05$) and 5 schizophrenia to IDP causal flow were nominally significant. Notice that the IDPs used for the training of the prediction models did not have schizophrenia. Therefore, our design cannot lead to scenario ii in Figure 5.2b which needs the individuals in the IDP prediction training set to have been exposed to the schizophrenia. Given the caveats discussed in the Results section and the fact that ii) cannot be detected in our design, we conclude that our results provide support for scenario iii).

## 5.4 Discussion

We propose a robust and scalable framework we call BrainXcan, which leverages genetically predicted brain features trained in reference MRI datasets, genome-wide association studies of complex diseases, and computational and methodological advances to dissect the biology of behavioral, neurological, and psychiatric traits. Our approach addresses the sample size limitations of MRI studies by taking advantage of increasing cohorts of GWAS studies and large MRI data in predominantly healthy subjects. The use of genetic variation helps us circumvent the reverse causality problem.

Our association module identifies brain features likely to have an effect on behavioral and psychiatric traits but also features that can be modified by the disease. Our Mendelian randomization module quantifies the evidence for each of the direction of the putative causal flow (brain feature to disease or vice-versa). Naturally, both direction of the effects are informative. Understanding how human disease cause changes in brain features captured by MRI can help design better diagnostic tools. Brain features that modulate the risk to disease provides insights into the pathogenesis and can help identify preventive or therapeutic

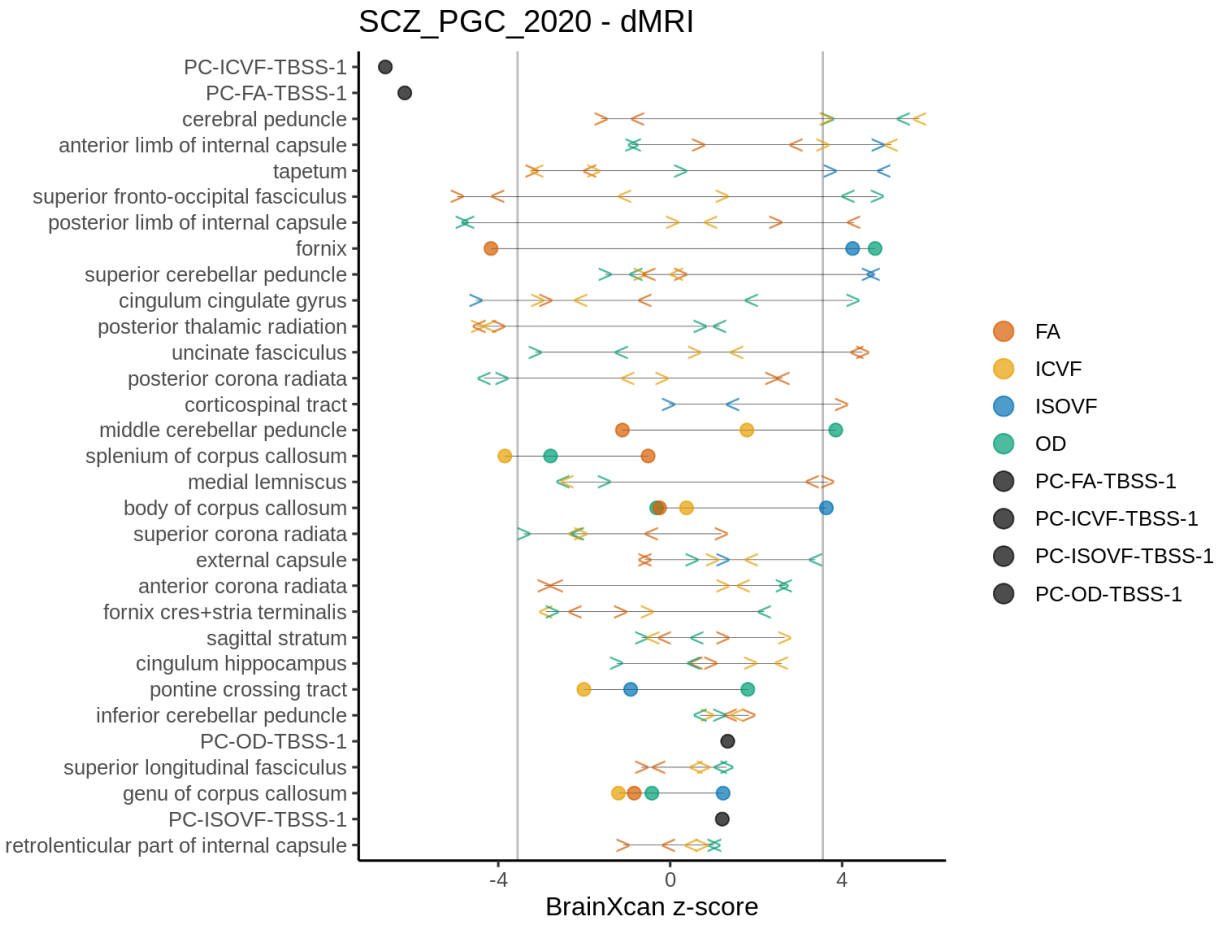



**Figure 5.5: Schizophrenia risk association with diffusion MRI.** Z-scores of region-specific associations of IDP with schizophrenia risk using GWAS effect sizes reported in [120]. The features starting with "PC" correspond to the principal components of IDP across regions and are proxies for the brain-wide feature for the subtype. FA: fractional anisotropy, ICVF: intracellular volume fraction, ISOVF: isotropic volume fraction, OD: orientation dispersion index. < indicates left, > indicates right, circles are used when sides are not defined

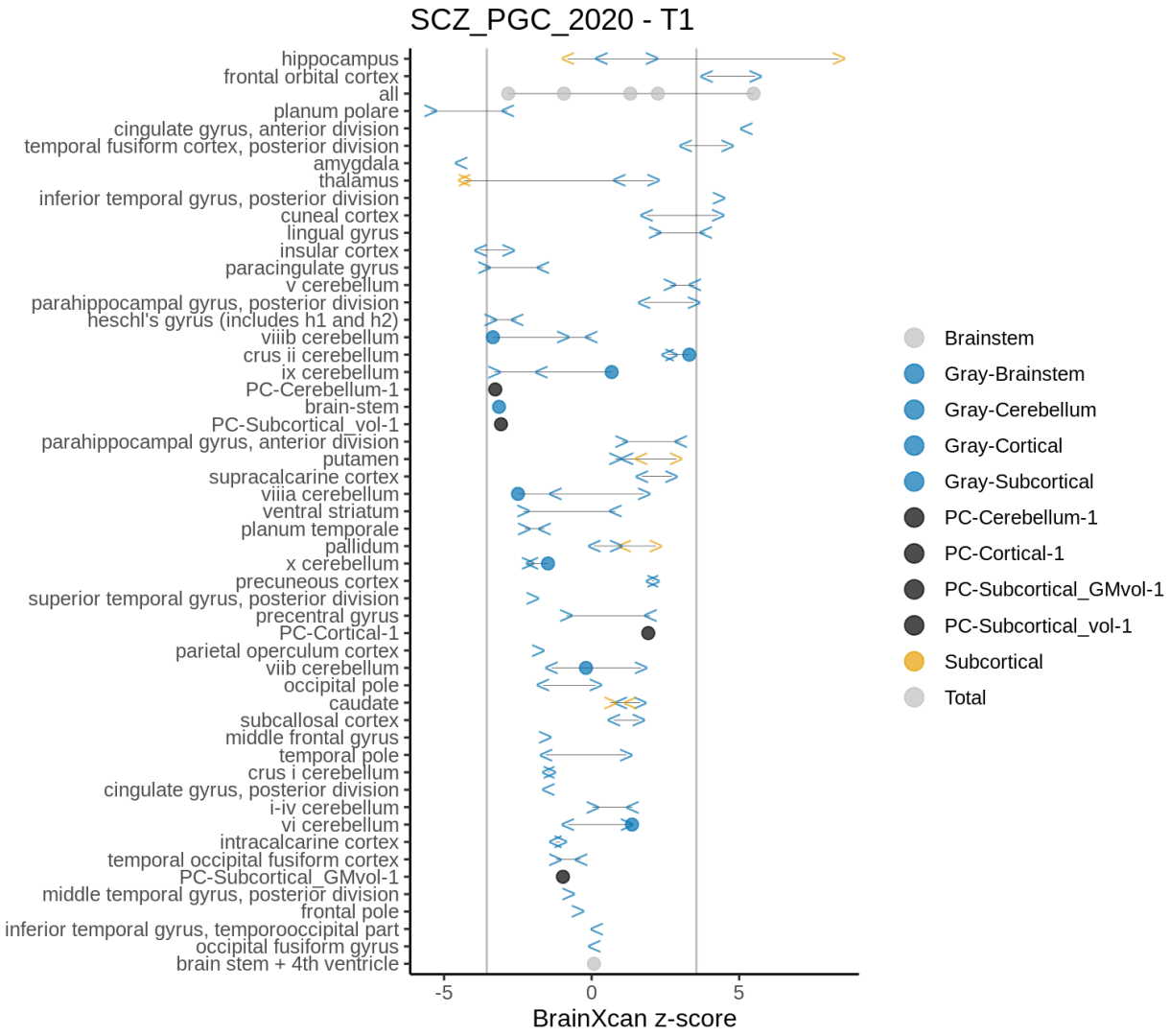



**Figure 5.6: Schizophrenia risk association with structural features.** Z-scores of region-specific association of IDP with schizophrenia risk using GWAS effect sizes reported in [120]. The features starting with "PC" correspond to the principal components of IDP across regions and are proxies for the brain-wide feature for the subtype. In blue are shown the gray matter volumes of the cortex, subcortex, brainstem, and cerebellum. In black are shown the principal components of each category. In yellow are shown the associations with subcortical total volumes quantified with FIRST. < indicates left, > indicates right, circles are used when sides are not defined.

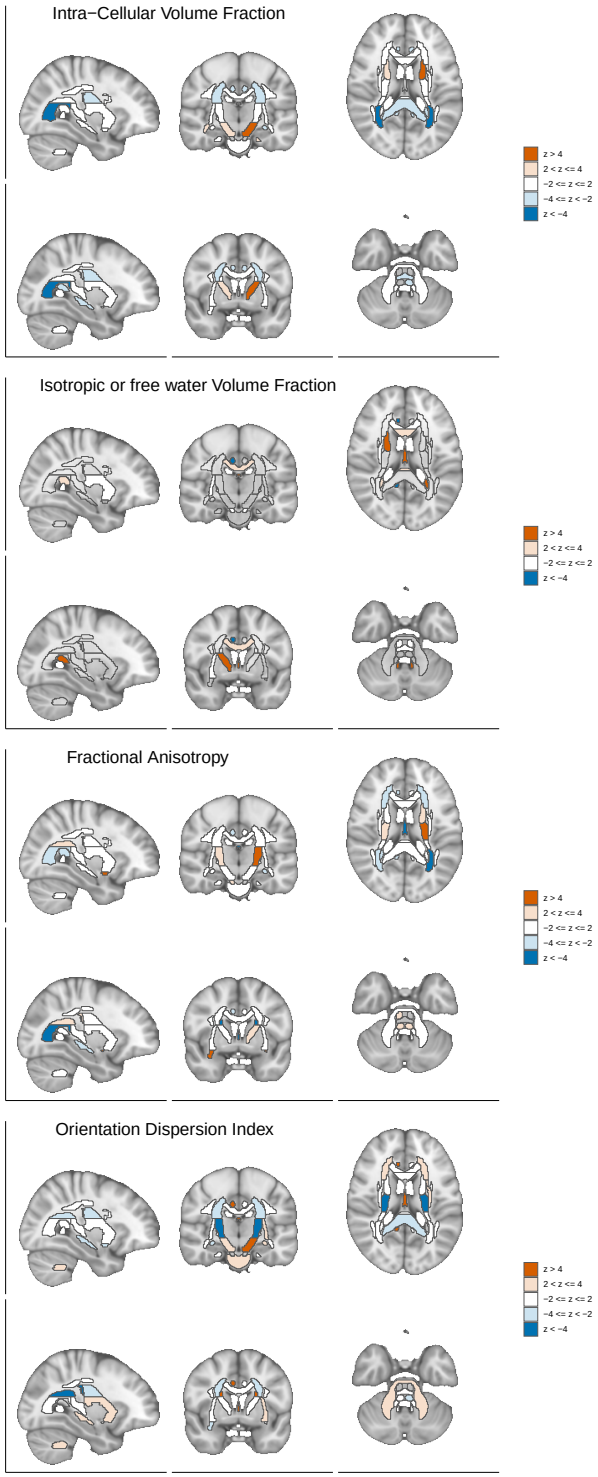



**Figure 5.7: Brain visualization of diffusion features associations with schizophrenia risk.** Z-scores of the associations between the brain region and schizophrenia risk are shown with different slices of the brain.

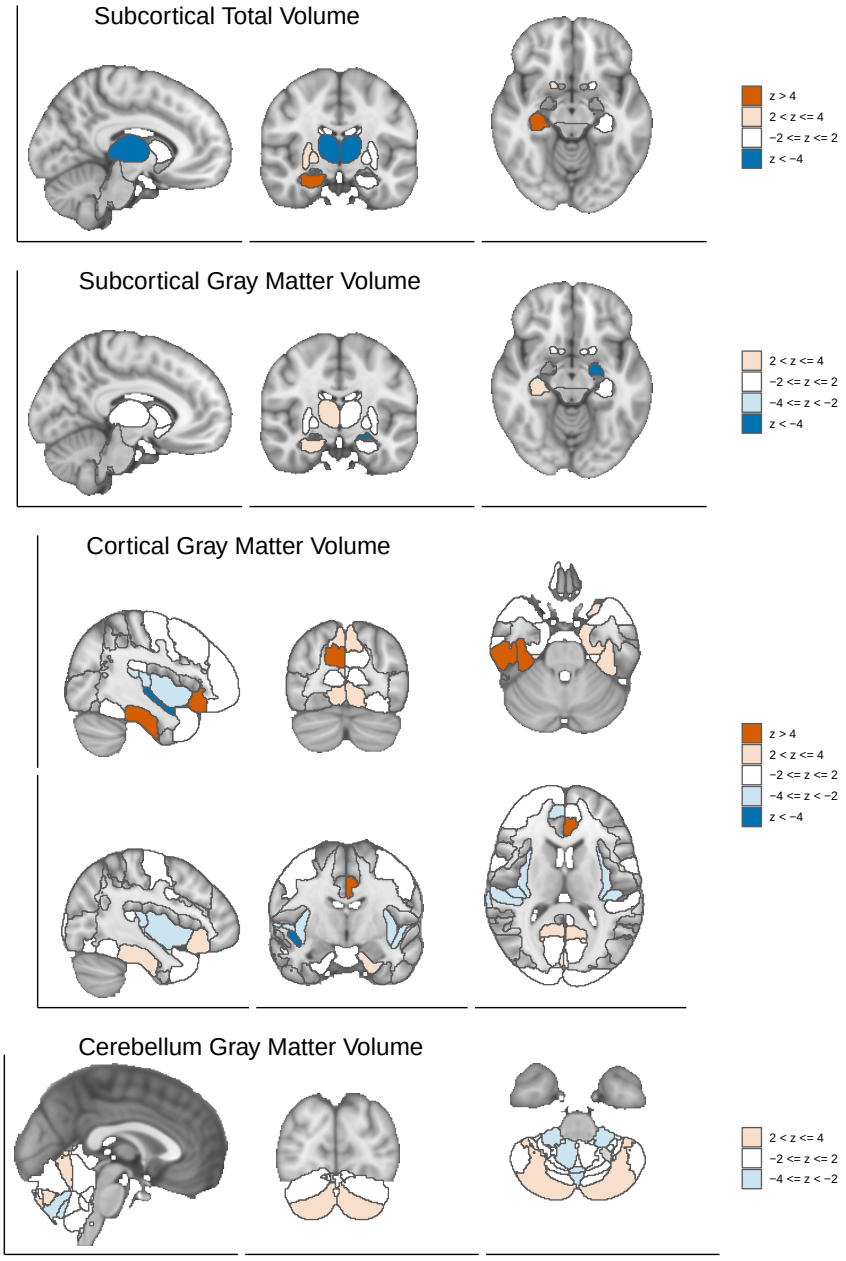



**Figure 5.8: Brain visualization of structural features associations with schizophrenia risk.** Z-scores of the associations between the brain region and schizophrenia risk are shown with different slices of the brain.

strategies.

To encourage broad adoption of our tools by users less familiar with genetic toosl, we provide a user friendly pipeline and an automated report. All the tools necessary to perform prediction, association, and causal flow assessment is provided (`https://github.com/hakyimlab/brainxcan`).

In the process of developing the tools, we learned that brain features are highly polygenic, in many cases even higher than human height (the canonical example of a polygenic trait), more similar to psychiatric and behavioral traits. The similarity in genetic architecture suggests that brain features can be useful endo-phenotypes to improve the classification of complex psychiatric diseases.

We present an application to schizophrenia to showcase the potential of our method. The significant association between low levels of brain-wide ICVF, proxy for axonal density, with high risk of schizophrenia corroborates the long standing disconnectivity hypothesis of schizophrenia. We also find the volumes of many regions of relevance associated with schizophrenia risk, including the amygdala, hippocampus, and the anterior cingular cortex. Our results add robust orthogonal lines of evidence to existing literature since our associations are based on the genetic components of the traits and are less likely to be confounded than studies with observed traits.

We provide a user-friendly software package to attract investigators less familiar with the genetic field. Our software is implemented in R and python with a streamlined pipeline coded in snakemake. An automated report in html format with pre-defined figures summarizing the results facilitates interpretation.

We note that there are several limitations in the current work. First, the prediction performance of the current genetic predictors are largely limited by the size of the training cohort (Figure 5.3d). As the UK Biobank is gradually collecting more brain imaging data [84], we expect the training cohort size will increase to at least 100,000 individuals. We will

be updating the the brain IDP predictors as the data becomes more available. Second, the S-BrainXcan calculation relies on the genotype covariance which is approximated as a banded matrix (Methods) here. This approximation may, particularly, affect the stability of the joint analysis since the jointly analysis relies heavily on the predicted IDP covariance which is derived from the genotype covariance. This was one of the reasons we decided not to pursue the joint model since they can lead to false positives. Third, the BrainXcan analysis cannot establish the causal relation between the brain IDPs and the complex phenotype. Although we run the Mendelian randomization in both the forward and the reverse directions for the IDP/phenotype candidate, the Mendelian randomization results should be interpreted with caution due to the following reasons: i) different Mendelian randomization tests may not give consistently significant results; ii) the Mendelian randomizations of the forward and the reverse directions typically have different power; iii) if the same GWAS is used for both BrainXcan association and Mendelian randomization, the Mendelian randomization p-value is not well-calibrated; iv) causality is valid only when the Mendelian randomization assumptions are hold. Given these limitations, we consider the Mendelian randomization results to be quantification of the evidence for the direction of the putative causal flow. Fourth, only linear prediction models are used in our implementation. More sophisticated models could be used for prediction but that will limit the application to cases where the full individual level data is available. Given the current availability of summary results and the need to use very large sample sizes to reliably estimate genetic effects makes the use of non-linear models less than optimal. Despite these limitations, we anticipate that BrainXcan, a user-friendly analysis tool, will be broadly adopted and help in identifying brain features important in the pathogenesis as well as diagnosis of complex phenotypes.

## 5.5 Methods

### *5.5.1 Preprocessing of UK Biobank IDP phenotypes*

We queried from UK Biobank database using ukbREST [115] to retrieve the list of 459 IDP phenotypes as shown in Table 5.1 [128]. Among these IDPs, 400 IDPs are diffusion MRI measurements including 192 TBSS-style measurements (TBSS) and 108 probabilistic-tractography-based measurements (ProbTrack). The remaining 159 IDPs are T1-weighted structural measurements, including 139 FAST-based grey matter segmentation based measurements and 14 FIRST-based subcortical structures measurements and 6 T1 structural brain MRIs measuring the total volumes of the peripheral cortical grey matter, ventricular cerebrospinal fluid, brain grey matter, brain white matter, brain grey + white matter, and brain stem. In total, we collected 24,409 European-descent individuals in UK Biobank with non-missing IDPs and non-missing values for other covariates such as genetic PCs, sex, and age at recruitment.

We scaled the structural IDP's using the volumetric scaling factor from T1 head image (UK Biobank Data-Field 25000) so that the measurement of the brain region volume was relative to the total brain volume. We regressed out the following scanner position covariates of IDPs: UK Biobank data rields 25756, 25757, 25758, and 25759. We also regressed out the first 10 genetic PCs, age, sex, squared age, age $\times$ sex, and squared age $\times$ sex.

To adjust for the correlation between IDPs and to extract the common factor of IDPs, we performed principal component analysis on the IDP matrices (individual-by-IDP matrices). Here, the PCA was done for each IDP subtype. For T1 IDPs, the subtypes were gray matter volume of the cortical regions, gray matter volume of the subcortical regions, total volume of subcortical regions, and gray matter of the cerebellum regions. For dMRI IDPs, the subtypes were defined by the four measure types (FA, ICVF, ISOVF, and OD) of TBSS and ProbTrack IDPs respectively. We obtained the first PC as the measure of the common factor

for each IDP subtype. We regressed out the first PC from the IDPs and obtained the IDP residuals. Finally, we inverse-normalized the PC-adjusted IDP residuals and also the IDP PCs for the subsequent analysis. We refer the IDP residuals and IDP PCs as brain IDP's.

### *5.5.2 Selecting variants from UK Biobank imputed genotypes*

We extracted the list of common variants (minor allele frequency $> 0.01$) among CEU individuals in HapMap 3 data [58]. And then, we excluded ambiguous variants which have reversely complementary bases as the reference and the alternative alleles (AT and CG pairs). In total 1,078,323 variants passed the criteria and among these, 1,071,650 variants appeared in the UK Biobank imputed genotype data (by SNP rs ID). In the subsequent analysis, we limited the computations to this set of variants.

### *5.5.3 Estimating the heritability*

We estimated the heritability of IDPs assuming random effects for SNPs [160]. We used the EMMA algorithm proposed in [62] to avoid the repeated calculations when dealing with multiple phenotypes on the same cohort. In the analysis, the genetic relatedness matrix (GRM) was built using the pre-selected HapMap 3 SNPs with minor allele frequency $> 0.05$. Since we accounted for the covariates in the preprocessing steps, we included no covariates other than the intercept in the heritability calculation.

### *5.5.4 Estimation of polygenicity*

We estimated the effective number of independently associated SNPs using the stratified LD fourth moments regression method [107]. We downloaded the pre-computed LD scores and LD fourth moments from `https://www.dropbox.com/sh/iiyftw01gdpt6un/AACU7AmWK45RxTmDJvRkdKhIa` and used the scores stored in `baselineLD.1kg.l2l4.mat`. These scores were based on baselineLD annotations [44]. The stratified LD fourth mo-

ments regression was performed in MATLAB by calling `SLD4M` function shared in `https://github.com/lukejoconnor/SLD4M` [107]. To obtain the effective number of independently associated common SNPs, we aggregated the estimated $M_e$ values across 10 MAF bins which correspond to the common variants (MAF $> 0.05$). The aggregation was done by setting `report_annot_mat` variable accordingly inside the `SLD4M` function.

### *5.5.5 Building polygenic predictors for IDPs*

We built polygenic predictors using both ridge regression models and elastic net models for the brain IDPs. The models were fitted using the pre-selected 1,078,323 genome-wide variants described above. Since covariates had already been adjusted for, we included only the intercept as covariate in the model training.

#### 5.5.5.1 Ridge regression models

In the ridge regression, we fit the following optimization problem:

$$\arg\min_{w} \|B - Xw\|_2^2 + \lambda\|w\|_2^2, \tag{5.3}$$

where $B$ is the mean-centered brain IDP (fitting one IDP at a time), $X$ is the standardized genotype matrix of the variants, and $w$ is the prediction model weights. To reduce the computation complexity and take advantage of the fact that the number of variants is much larger than the number of samples, we used the formula similar to the one in OmicKriging [154]. Specifically, the ridge regression estimator $\widehat{w}^{\text{ridge}} = (X'X + \lambda I)^{-1}X'B$ (where $\lambda$ is the hyperparameter) involves solving a linear system with $P \times P$ matrix where $P$ is the number of variants. Alternatively, it can be re-arranged as $\widehat{w}^{\text{ridge}} = X'(XX' + \lambda I)^{-1}B$ instead and, correspondingly, the predicted value of $B$ is $\widehat{B} = (XX')(XX' + \lambda)^{-1}B$. Let $\Sigma$ represent the GRM matrix and we have $\Sigma = XX'/P$. The expression for the $\widehat{B}$ can be re-parameterized as

$\widehat{B} = \theta\Sigma^{-1}(\theta\Sigma + (1-\theta)I)^{-1}B$ with $\theta = \frac{P}{P+\lambda} \in (0, 1]$. We performed 5-fold cross-validation on a grid of $\theta$ values 0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.95 to choose the hyperparameter $\theta$.

#### 5.5.5.2 Elastic net models

The elastic net models were fitted using the R package `snpnet` which implements the BSAIL algorithm proposed in [119]. Specifically, we fit the following optimization problem:

$$\arg\min_{w} \frac{1}{2N}\|B - Xw\|_2^2 + \lambda(\frac{1-\alpha}{2}\|w\|_2^2 + \alpha\|w\|_1) \tag{5.4}$$

We fixed the hyperparameter $\alpha = 0.1$ and determined the value of $\lambda$ by the $R^2$ on the validation set. First, we randomly split the data into two parts: the training set (containing 80% of the samples) and the validation set (containing 20% of the samples). Then, using the training set, we trained a series of elastic net models under a grid of $\lambda$ values. All these models were evaluated on the validation set by calculating the $R^2$ between the observed and predicted $B$. We selected the $\lambda$ with the largest validation $R^2$ and trained the final model using the full data (combining the training and validation sets).

When using `snpnet`, we set the maximum number of iterations (`niter`) to 100 and the number of SNPs to consider in each batch (`num.snps.batch`) to 200. The phenotypes were fitted one at a time.

#### 5.5.5.3 Calculating the prediction performance

For both ridge and elastic net models, we evaluated the prediction performance by 5-fold cross validation. At each split, we used the 4 folds as if the full data and trained the models using the procedures described above. And then, we made the prediction on the held-out one fold of the data. This procedure was repeated for all five splits and the predicted values

were aggregated across all folds. The prediction accuracy was evaluated in terms of $R^2$, Pearson correlation, and Spearman correlation.

### *5.5.6 BrainXcan with individual-level data*

We performed individual-level BrainXcan as described above on a set of 327,743 unrelated, European-descent UK Biobank participants who were not included in the brain IDP model training. We included sex, age, squared age, age × sex, squared age × sex, and the first 10 genetic PCs as covariates.

### *5.5.7 BrainXcan with summary statistics*

When the individual-level information is not available, we calculated the BrainXcan statistic approximately using the GWAS summary statistic and the genotype covariance from a reference panel. The formulas are similar to the ones proposed in [7]. Let $\widehat{b}_j$ and $\text{se}(\widehat{b}_j)$ be the estimated effect size and the corresponding standard error for variant $j$ from the GWAS. And $z_{\text{GWAS},j}$ is the GWAS z-score of SNP $j$. Let $\widehat{R}$ represent the genotype sample covariance matrix where $\widehat{R}_{jj'} = \widehat{\text{Cov}}(X_j, X_{j'})$, namely the sample covariance between variant $j$ and $j'$. We can calculate the marginal test statistics using the following results (see derivations in Section 5.8.3):

$$\hat{\beta}_k = \frac{\sum_{j=1}^{P} \widehat{w}_{kj} \widehat{R}_{jj} \widehat{b}_j}{\widehat{\sigma}_k^2} \tag{5.5}$$

$$z_{\text{BrainXcan},k} \approx \frac{\sum_{j=1}^{P} \widehat{w}_{kj} \sqrt{\widehat{R}_{jj}} z_{\text{GWAS},j}}{\widehat{\sigma}_k} \tag{5.6}$$

$$\widehat{\sigma}_k^2 = \sum_{j=1}^{P} \sum_{j'=1}^{P} \widehat{w}_{kj} \widehat{w}_{kj'} \widehat{R}_{jj'}, \tag{5.7}$$

where $z_{\text{BrainXcan},k}$ represents the BrainXcan marginal test z-score for the $k$th brain IDP. We refer this summary statistic-based BrainXcan as S-BrainXcan.

In principle, we need to consider the genotype covariance for all the genome-wide variant pairs. In this paper, to reduce the computation burden, we first assumed that the between-chromosome covariance is zero. Moreover, we considered the per-chromosome genotype covariance matrix as a banded matrix with bandwidth equal to 200. In other words, any variant pairs with more than 200 variants in-between are considered having zero covariance. We used the set of 24,409 UK Biobank individuals used for the brain IDP model training as the reference panel for genotype covariance calculation.

### *5.5.8 Performing GWAS for brain IDPs*

We performed genome-wide association studies for all the brain IDPs using Python package `tensorqtl` [136]. Since we adjusted covariates in the preprocessing of the brain IDPs (see previous sections), we did not include any covariates other than the intercept in the GWAS runs.

### *5.5.9 Mendelian randomization analysis of IDP/phenotype pairs*

To investigate the direction of the effect, we performed Mendelian randomization (MR) analysis for the significant IDP/phenotype pairs identified in the BrainXcan association stage. The MR analysis was performed for both directions: i) brain IDP $\rightarrow$ phenotype, treating the brain IDP as the mediating trait and the phenotype of interest as the outcome trait; ii) phenotype $\rightarrow$ brain IDP, the phenotype of interest as the mediating trait and the brain IDP as the outcome trait. As the inputs of the analysis, we used the brain IDP GWAS results as described in the above section. For the phenotype of interest, we used the GWAS results which were also used for the S-BrainXcan analysis.

For the mediating trait, the instrument variants were selected using LD clumping function (`ld_clump`) in the R package `ieugwasr` [37]. We used the EUR super-population in 1000 Genomes data [1] as the LD reference panel and the data was downloaded from `http:`

//fileserve.mrcieu.ac.uk/ld/1kg.v3.tgz. The LD clumping parameters were clump_kb = 10000 and clump_r2 = 0.001. The p-value parameter (clump_p) in the LD clumping was $10^{-5}$ for IDP GWAS and $5 \times 10^{-8}$ for phenotype GWAS, which gave approximately independent and significant variant instruments.

The MR analysis was performed using the R package TwoSampleMR [53]. We reported the MR results using three MR methods: i) inverse variance weighted MR [19]; ii) median-based estimator: weighted median [15]; iii) MR Egger analysis [14], which corresponds to mr_ivw , mr_weighted_median, and mr_egger_regression in TwoSampleMR. We further reported a meta-analyzed p-value summarizing the results of the three MR tests being performed. The meta-analysis is based on an extension of ACAT method [87] that takes into account the direction of the effects. See Section 5.8.4 and Figure 5.22 for additional detail.

### *5.5.10 Calculating the genetic correlation for IDP/phenotype pairs*

The genetic correlation between a brain IDP and the phenotype of interest was calculated using the cross-trait LD Score regression [16]. The actual calculation was performed using the Python package ldsc (https://github.com/bulik/ldsc). The pre-computed LD-scores were downloaded from https://storage.googleapis.com/broad-alkesgroup-public/LDSCORE/eur_w_ld_chr.tar.bz2 which are based on the 1000 Genomes European data. We used the brain IDP GWAS results as described in the above section and the GWAS of the phenotype was the same as the one used for the S-BrainXcan analysis.

## 5.6 Supplementary Figures

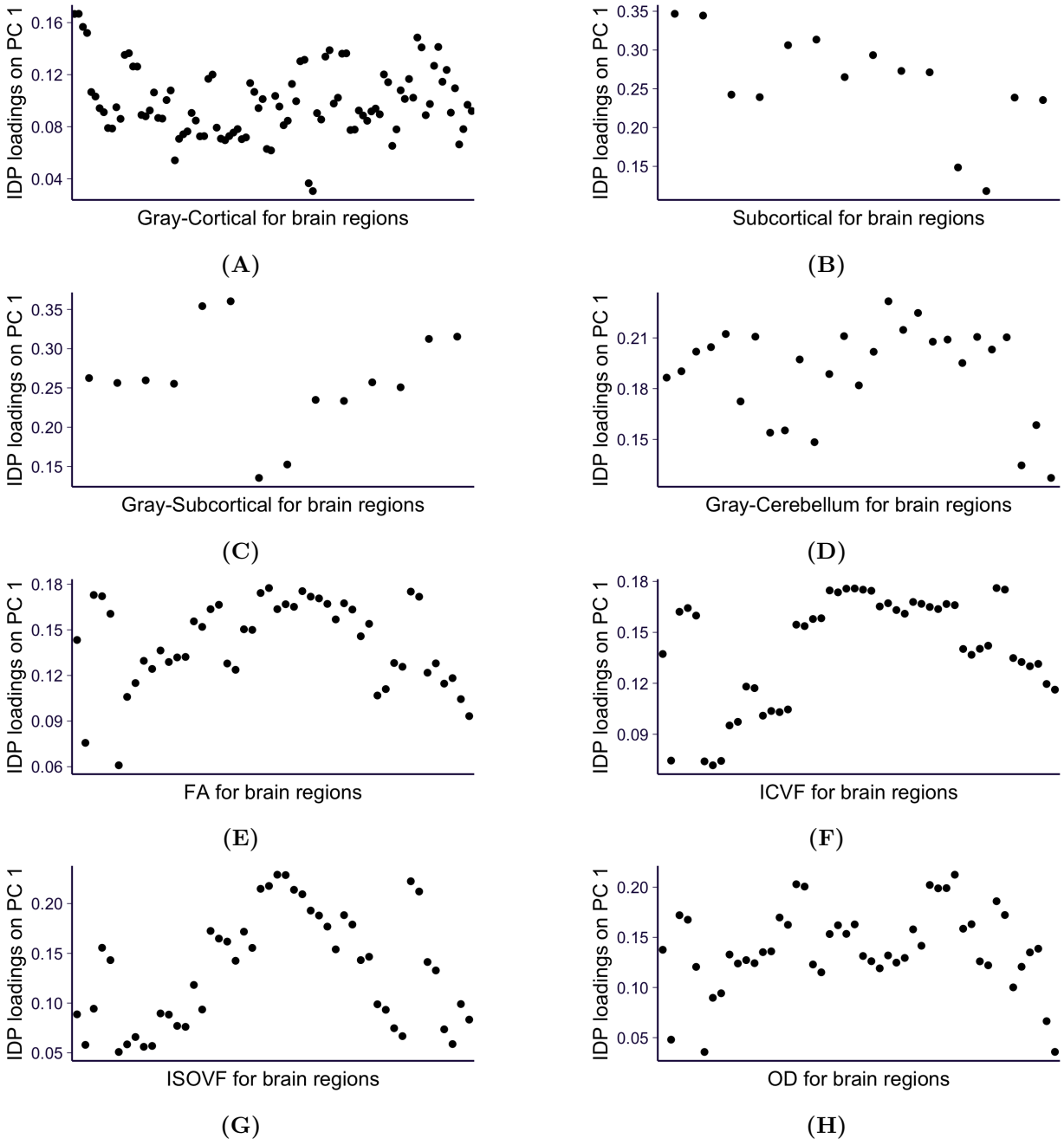



**Figure 5.9: The first PC of each IDP modality.** For each IDP modality, the contribution of PC1 to each of the brain IDPs within the modality group is shown. Panel **a)** to **d)** show results for T1 modalities: gray matter volume of cortical regions, total volume of subcortical regions, gray matter volume of subcortical regions, and gray matter volume of cerebellum regions. Panel **e)** to **h)** show results for TBSS-based dMRI modalities: FA, ICVF, ISOVF, and OD.

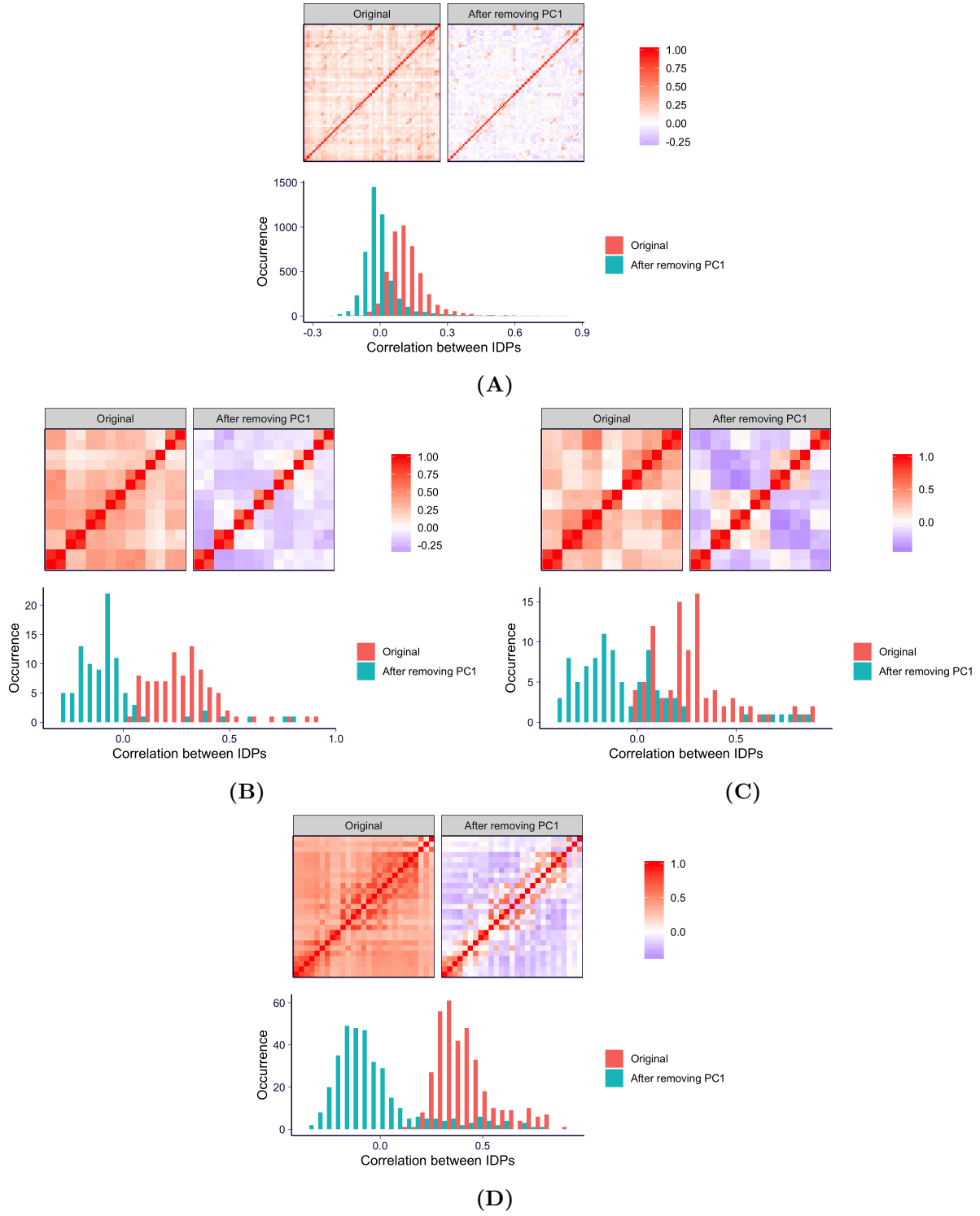



**Figure 5.10: The correlation between IDPs for T1 modalities.** For each T1 modality, the correlation between brain IDPs are shown before and after removing PC1 by the heatmaps and the histrogram. Panel **a)** to **d)** show results for T1 modalities: gray matter volume of cortical regions, total volume of subcortical regions, gray matter volume of subcortical regions, and gray matter volume of cerebellum regions.

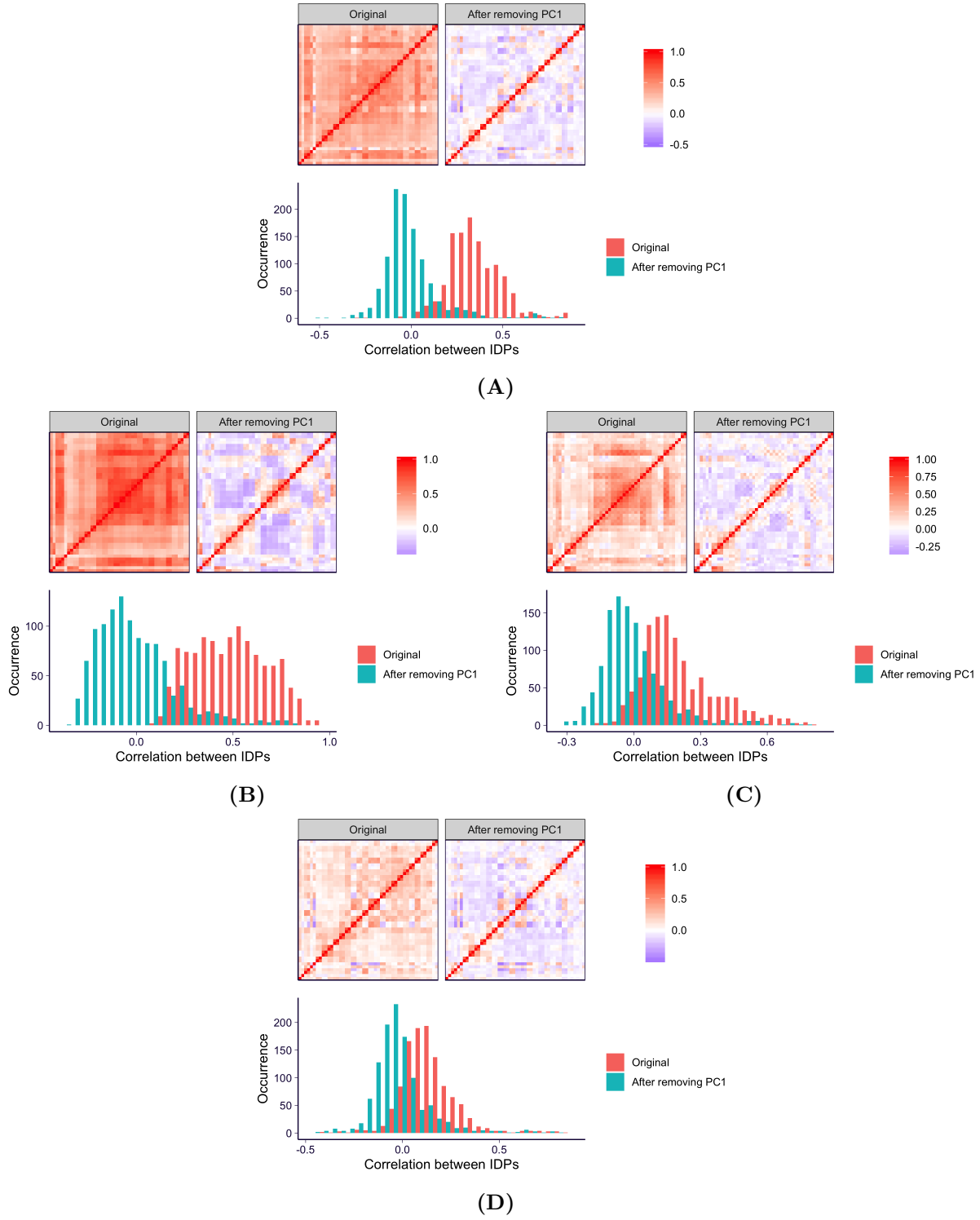



**Figure 5.11: The correlation between IDPs for dMRI modalities.** For each dMRI modality, the correlation between brain IDPs are shown before and after removing PC1 by the heatmaps and the histrogram. Panel **a)** to **d)** show results for TBSS-based dMRI modalities: FA, ICVF, ISOVF, and OD.

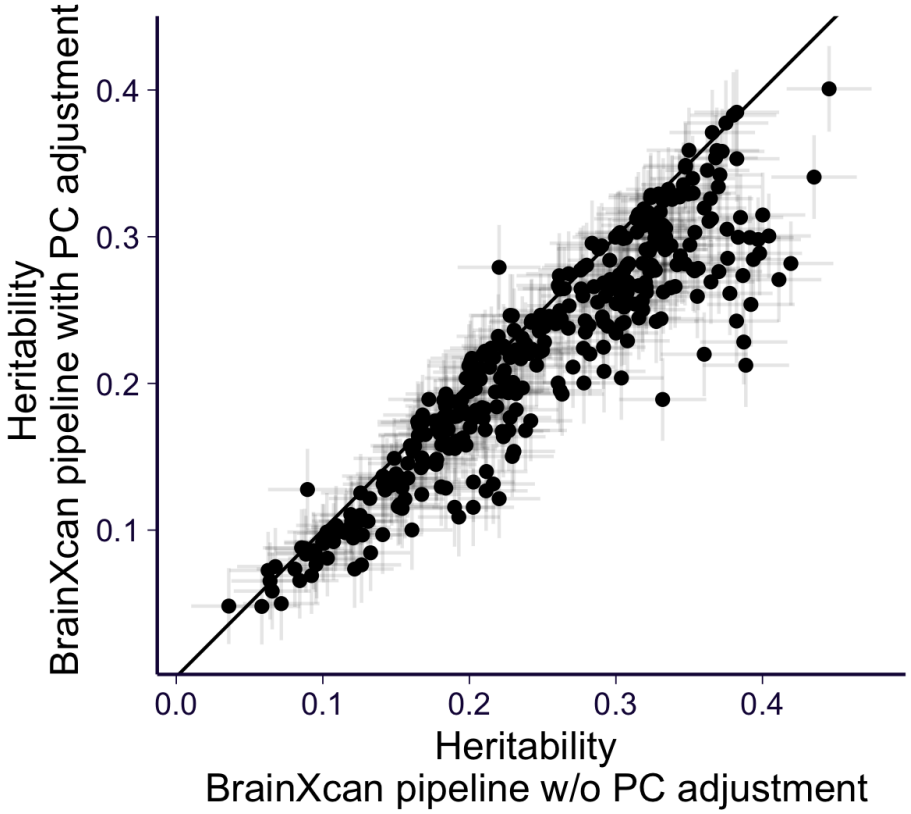



**Figure 5.12: Heritability of PC-adjusted vs. non adjusted brain IDPs.** The heritability of brain IDPs after the PC adjustment (y-axis) is compared to the heritability of brain IDPs before the PC adjustment (x-axis). The error bars indicate 95% confidence interval.

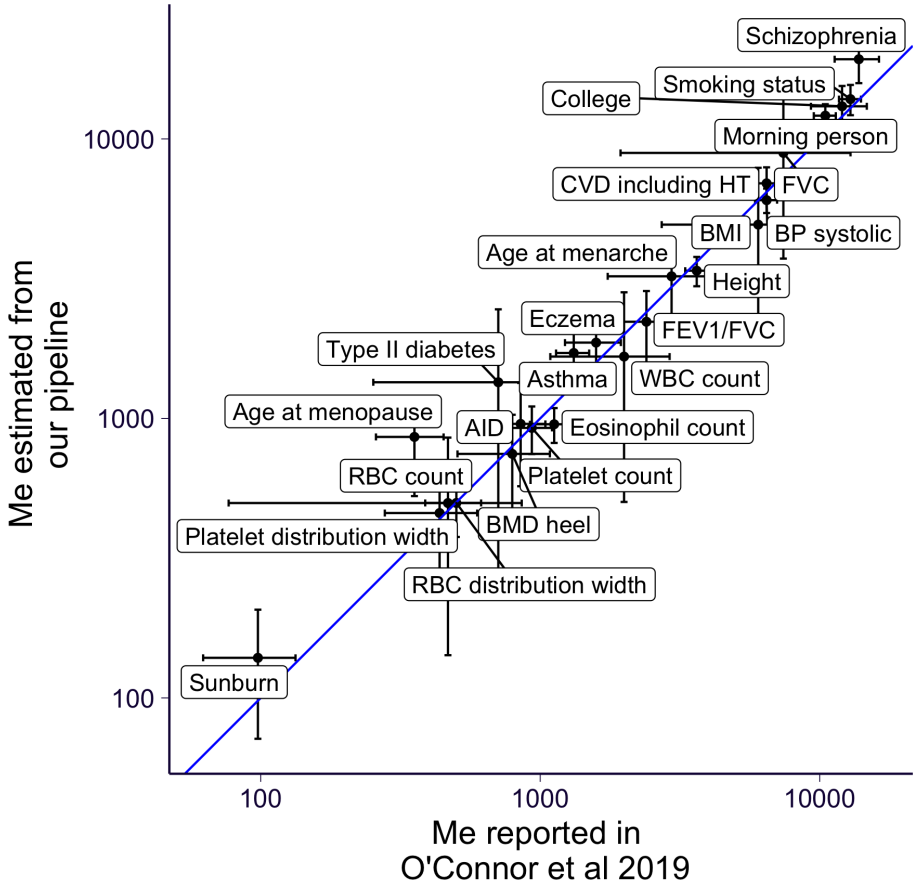



**Figure 5.13: Comparing estimated $M_e$ from [107] and our pipeline.** Our $M_e$ estimation pipeline is slightly different from the one being used in [107] (see more details in Section 5.5.4). To check the robustness of our pipeline, we compared the estimated $M_e$ from [107] (x-axis) and our pipeline (y-axis) for 24 traits. See definition of trait abbreviations from Table 1 and Table S4 of [107]. The error bar indicates the 95% confidence interval. The blue line is the identity line ($y = x$).

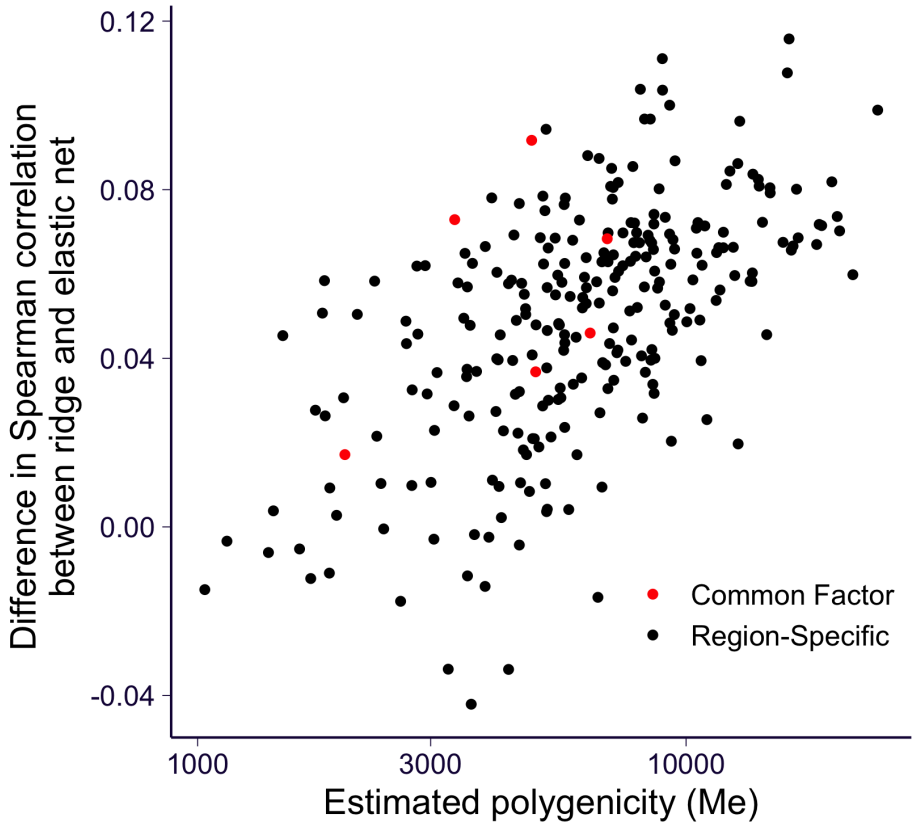
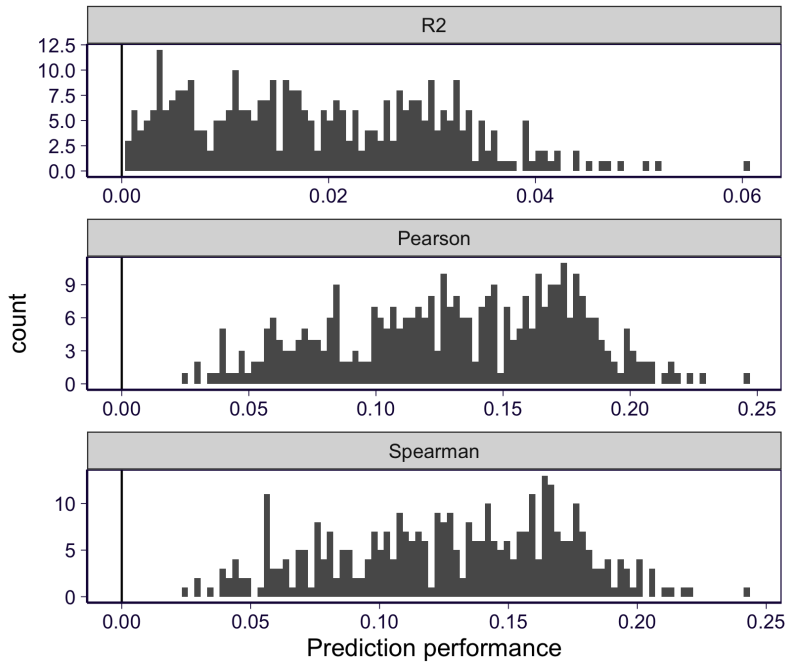
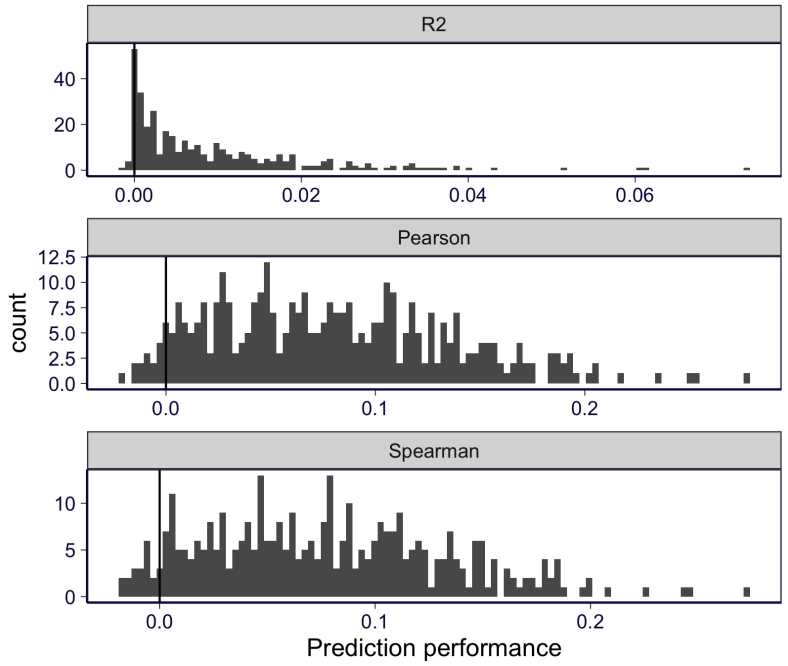



**Figure 5.14: Ridge predictor gain in performance vs. estimated polygenicity, $M_e$.** The x axis shows the estimated polygenicity, $M_e$, for the 522 IDPs with values significantly greater than 0 ($p < 0.05$). $M_e$ is the "the effective number of independently associated SNPs", a proxy for number of causal SNPs. The y axis shows the difference in performance between ridge predictors and elastic net predictors (in terms of the difference in Spearman correlation). The IDP PCs are in red and the rest of the brain IDPs are in black.




**(A)**




**(B)**

**Figure 5.15: The histogram of the prediction performance across all brain IDPs.** The prediction performance of the regression models, ridge regression (in **(a)**) and elastic net regression (in **(b)**), are shown in terms of the $R^2$, Pearson correlation, and Spearman correlation.

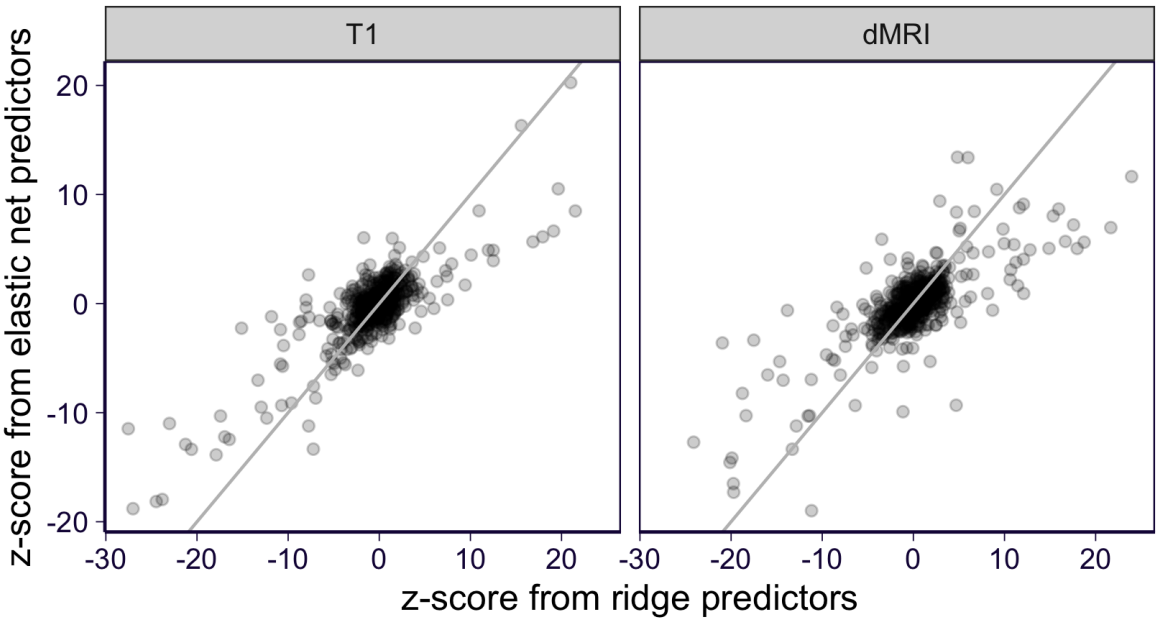



**Figure 5.16: Comparing the ridge and elastic net based individual-level BrainXcan results.** We compare the BrainXcan z-scores among the brain IDPs which have both ridge predictor and elastic net predictor with high quality (Spearman correlation $> 0.1$). The gray lines are the identity line ($y = x$).

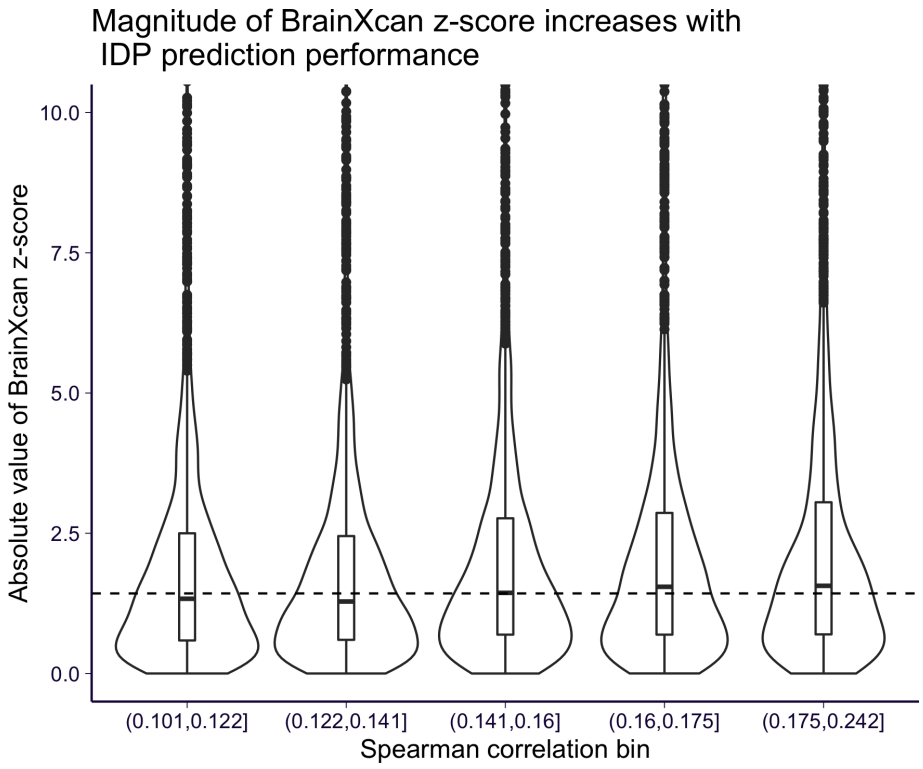



**Figure 5.17: Comparing the BrainXcan significance versus the performance of IDP predictors.** BrainXcan significance is defined as the absolute value of BrainXcan z-score and the performance of IDP predictors is defined as the Spearman correlation between predicted and observed values by cross-validation. We compare the BrainXcan significance (y-axis) versus the performance of IDP predictors (x-axis) among the brain IDPs which have ridge predictor in high quality (Spearman correlation $> 0.1$). These IDPs are grouped into 5 equal-size bins according to their prediction performance. The dashed line shows the median of the BrainXcan significance.

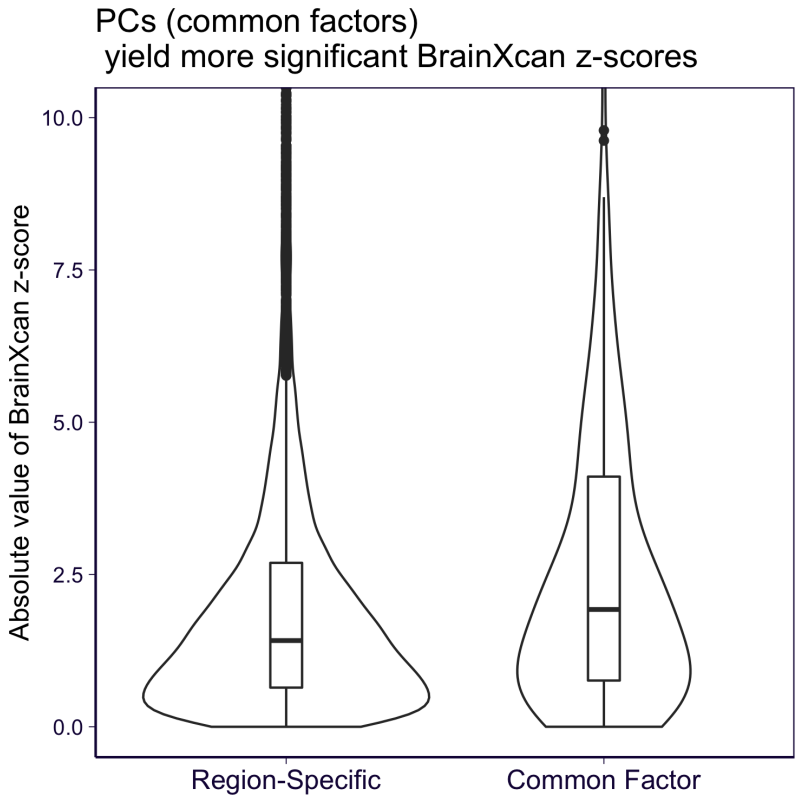



**Figure 5.18: Comparing the BrainXcan significance between region-specific IDPs and common factors.** BrainXcan significance is defined as the absolute value of BrainXcan z-score. We compare the BrainXcan significance (y-axis) of region-specific IDPs and common factors (PC1 of each IDP subtype) among the brain IDPs which have ridge predictor in high quality (Spearman correlation $> 0.1$).

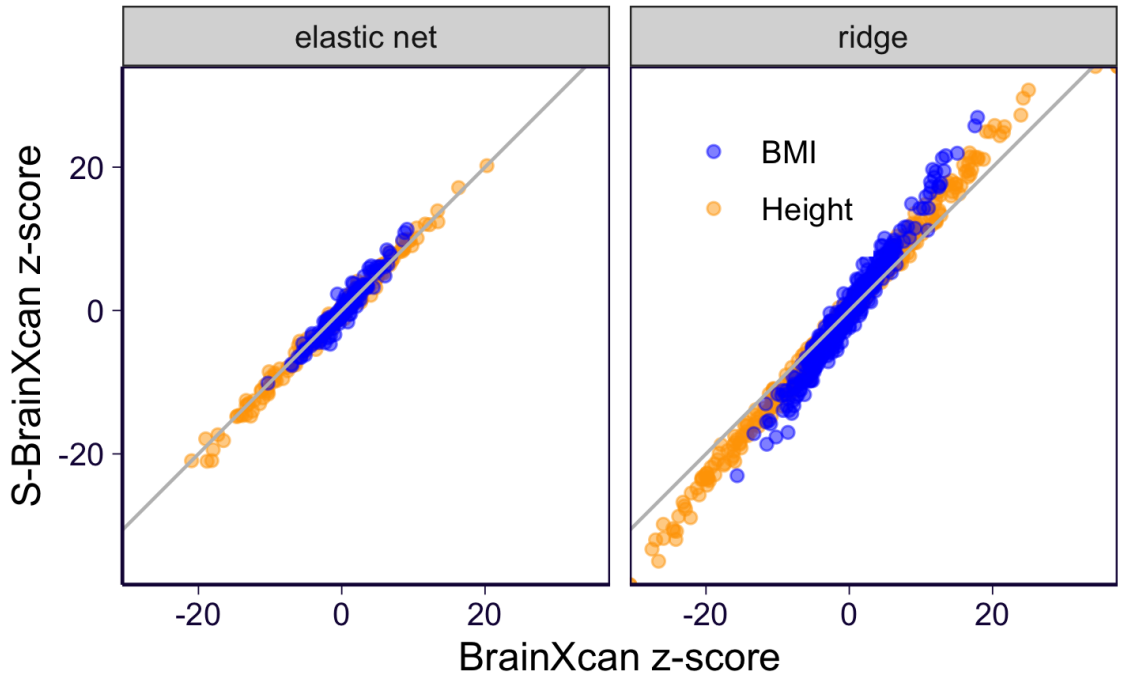



**Figure 5.19: Comparing individual-level BrainXcan and S-BrainXcan results on UK Biobank standing height and BMI.** We compare the BrainXcan z-scores of UK Biobank being calculated from the individual-level BrainXcan (on x-axis) and S-BrainXcan (on y-axis). IDP models with prediction performance greater than 0.1 (Spearman correlation) are shown.

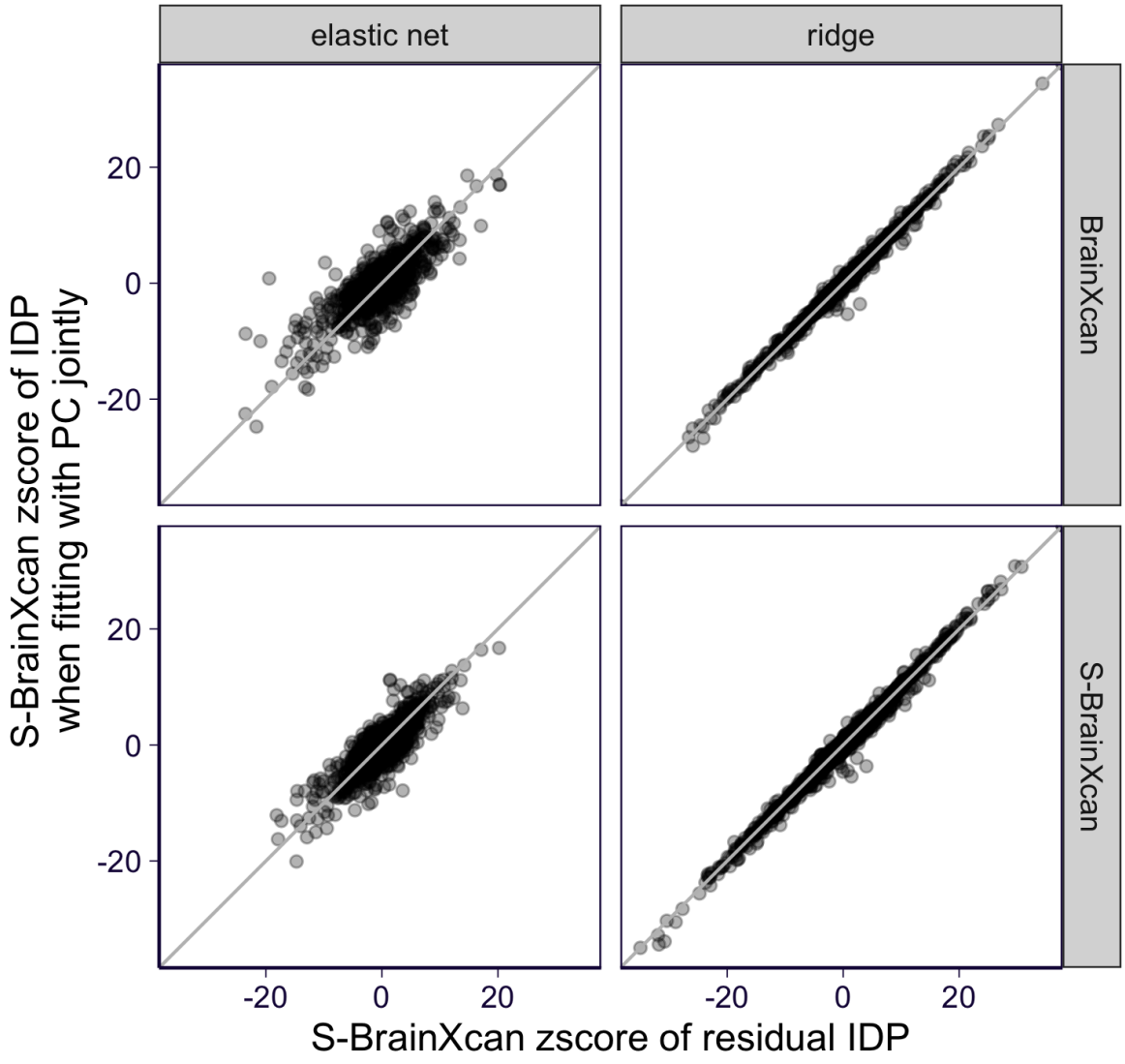



**Figure 5.20: Comparing BrainXcan results from residual IDP and IDP adjusted by PC.** We compare the BrainXcan z-scores obtained from $Y \sim$ resIDP (x-axis) and $Y \sim$ IDP+PC (y-axis). The top row shows results from individual-level BrainXcan and the bottom row shows results from summary BrainXcan. IDP models with prediction performance greater than 0.1 (Spearman correlation) are shown. The gray lines are the identity line ($y = x$). All GWASs are shown.

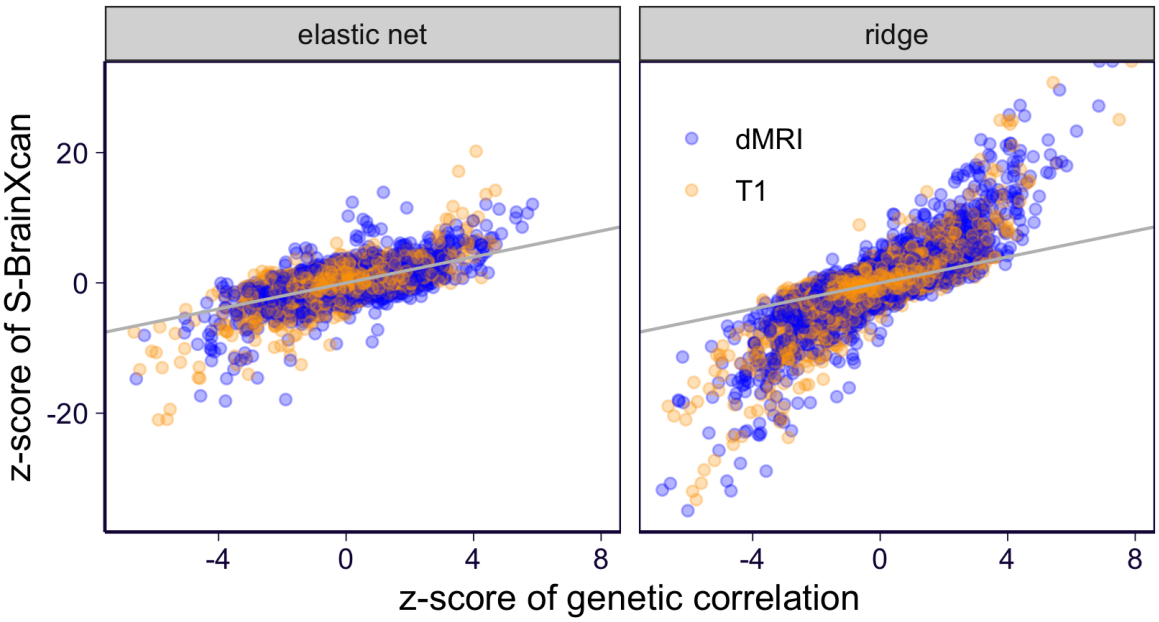



**Figure 5.21: Comparing z-scores of the genetic correlation and S-BrainXcan.** The z-scores of the genetic correlation (on x-axis) and the S-BrainXcan (on y-axis) are shown.

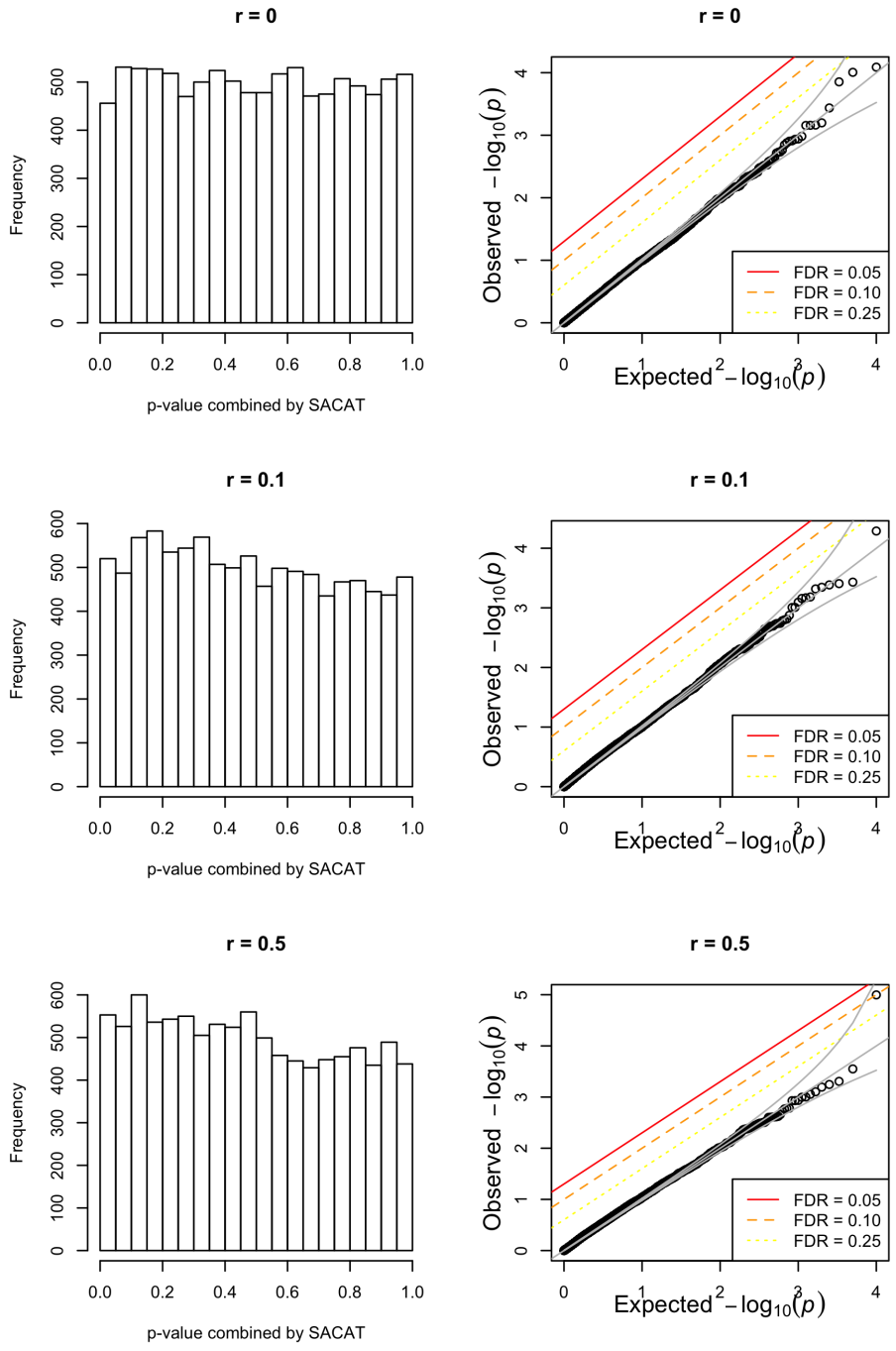



**Figure 5.22: SACAT based p-value distribution under the global null.** The simulation results of signed-ACAT (SACAT) are shown. The distribution of the SACAT-based meta-analyzed p-values are shown with histogram and QQ-plot (against uniform distribution). Each row shows the results from simulating the global null with dependency structure controlled by $r$. As $r$ increases, the p-values being combined are more dependent and when $r = 0$, the p-values being combined are independent. These figures show that the SACAT is relatively well calibrated for different values of correlation between entries. See more simulation details in Section 5.8.4.

## 5.7 Supplementary Tables

**Table 5.1: UK Biobank brain IDPs being analyzed.** The table is available at `https://docs.google.com/spreadsheets/d/1xzLvwtIshyX10orGnGVyDNuyDHU-zgoYhGYemf-6I10`. It contains the meta information about the brain IDPs analyzed in this paper. The columns are listed below. **ukb_field**: Field ID in UK Biobank; **modality**: indicates whether the IDP is from T1 MRI or diffusion MRI; **region**: anatomical region of the MRI measurement; **side**: indicates whether the measurement is from the left or right part of the brain; **measurement_type**: type of the measurement which also depends on the image processing procedure; **dmri_measure**: for diffusion MRI, indicate the type of statistic extracted from the measurement; **t1_anatomy_group**: for T1 MRI, indicate the anatomical grouping; **notes**: description of the IDP extracted from the UK Biobank database; **ukb_link**: link to the Field in UK Biobank website; **subtype**: IDP subtype.

**Table 5.2: The prediction performance of the ridge and elastic net predictors.** The table is available at `https://docs.google.com/spreadsheets/d/1fzFJA0hhNc0VBmclVCpKojqo9Sl4Vnm3cU03p1qgdN8`. It contains prediction performance (via cross-validation scheme) of the ridge regression and elastic net based genetic predictors of brain IDPs. **IDP_type**: Indicates whether the brain IDP is T1 IDP or dMRI IDP; **IDP**: The IDP code (in the format of "IDP-XXXXX" with XXXXX being the Field ID in UK Biobank); **model_name**: indicates whether the genetic predictor is based on ridge regression or elastic net; **Spearman**: Spearman correlation between the predicted and observed IDP values; **Pearson**: Pearson correlation between the predicted and observed IDP values; **R2**: $R^2$ measured from the predicted and observed IDP values; **is_kept**: indicates whether the genetic predictor passes the quality control: Spearman correlation $> 0.1$.

**Table 5.3: The list of 9 UK Biobank phenotypes analyzed by individual-level BrainXcan.** The table is available at `https://docs.google.com/spreadsheets/d/150mFxAV1p_IK1l36GKjazBt0pIb5eSC0LO7JeFaTS1I`. It contains the definition of these 9 UK Biobank phenotypes. For some phenotypes, they were constructed from multiple UK Biobank fields and the aggregation method across these fields was either taking the sum ('sum') or taking logical OR ('or'). For binary phenotype, we defined its value as 1 if the field takes the desired value ('target_value'). Missing values were either treated as 0 or removed. **phenotype**: name of the phenotype; **ukb_field**: UK Biobank fields from which the phenotype was constructed; **target_value**: for binary phenotype, we defined its value as 1 if the field takes the desired value; for quantitative traits, use the value as it is (labelled as 'asis'); **aggregate_method**: for some phenotypes, they were constructed from multiple UK Biobank fields and the aggregation method across these fields was either taking the sum ('sum') or taking the logical OR ('or'); **missing_values**: missing values in a field were either treated as 0 or removed.

**Table 5.4: The list of 35 GWAS analyzed by S-BrainXcan.** The table is available at `https://docs.google.com/spreadsheets/d/1jjDZZBsRNTvgidXXfJqXNOvFQ1U9ReeOG-WFcXtWGAg`. It contains the information about the 35 GWASs used in S-BrainXcan analysis. **phenotype**: The name of the phenotype. **phenotype_id**: The phenotype identifiers. **short_name**: The short names of the phenotypes. **sample_size**: The sample size of the GWAS. If there are different sample sizes for different SNPs, the mean sample size is shown. **portal**: The website from which the GWAS was downloaded. **filename**: The name of the raw GWAS file being downloaded.

## 5.8 Supplementary Notes

### *5.8.1 Deriving bias of BrainXcan estimates*

#### 5.8.1.1 A generative model of IDP/phenotype association

Recall that we consider the following generative model describing the effect of brain IDPs within a subtype on a complex trait $Y$:

$$Y = \alpha \cdot L + \sum_k \beta_k \cdot F_k + \epsilon_Y,\ \epsilon_Y \sim N(0, \sigma^2) \tag{5.8}$$

where $L$ represents brain-wide factor which universally affects all IDPs with the subtype. And $F_k$ represents the IDP of a specific brain region within the subtype. Eq 5.8 assumes that both the brain-wide factor and region-level IDPs may affect the complex trait and the effect sizes are $\alpha$ and $\beta_k$'s respectively. In the BrainXcan analysis, ideally, we are interested in both brain-wide effects ($\alpha$) and region-specific effects ($\beta_k$).

As described above, $F_k$ is partially determined by the brain-wide factor and we introduce $R_k$ to represent the rest of the variation in $F_k$ which is region-specific. Furthermore, we assume that the region-specific variations $R_k$ are independent to all other region's. More

specifically, we assume

$$F_k = L + R_k,\ R_k \sim N(0, \sigma_k^2) \tag{5.9}$$

$$R_k \perp\!\!\!\perp R_j,\ \forall k \neq j \tag{5.10}$$

And in this setup, the relative contribution of $L$ to $F_k$ is determined by $\sigma_k^2$.

In practice, we don't observe $F_k$. Instead, MRI imaging pipeline measures a noisy version of $F_k$. And we build genetic predictors using the noisy $F_k$ so that we manage to (partially) capture the genetically determined variation in $F_k$. In other words, let $\text{IDP}_k$ represent the predicted value of $F_k$ and we assume that

$$\text{IDP}_k = F_k + \epsilon_k,\ \epsilon_k \sim N(0, \tau_k^2) \tag{5.11}$$

where $\tau_k^2$ is the amount of noise when using $\text{IDP}_k$ as the proxy for $F_k$.

To further simplify the derivation, we let $\tau_j^2 = t^2,\ \forall j$ and $\sigma_j^2 = s^2,\ \forall j$. These assumptions imply equal contribution of the brain-wide factor to all regions. And they also assume that the quality of $F_k$ proxies (*i.e.* $\text{IDP}_k$) is the same across all regions. These assumptions simplify the notation and the qualitative conclusion still holds when this assumption is relaxed.

Notice that $L$ is also an unobserved latent factor. We try to capture $L$ by averaging over all $\text{IDP}_k$. In practice, we use the first principal component (PC1), which essentially is a weighted average of $\text{IDP}_k$, to approximate $L$. Using PC1 could account for the fact that the brain-wide factor does not contribute equally to all regions and predictor quality is not the same across all regions. But since here we assume equal contribution for $L$ and equal quality for $\text{IDP}_k$, we simply use unweighted average of $\text{IDP}_k$ as the proxy for $L$, which is

shown below:

$$\mathrm{PC} = \frac{1}{m}\sum_k \mathrm{IDP}_k \tag{5.12}$$

$$= L + \frac{1}{m}\sum_k (R_k + \epsilon_k) \tag{5.13}$$

where $m$ is the number of regions within the subtype being considered.

#### 5.8.1.2 Variances and covariances among variables

Here we list the variance and covariance among model variables.

$$\mathrm{Cov}(F_i, F_j) = \begin{cases} 1 & , i \neq j \\ 1 + s^2 & , i = j \end{cases}$$

$$\mathrm{Cov}(\mathrm{IDP}_i, \mathrm{IDP}_j) = \begin{cases} 1 & , i \neq j \\ 1 + s^2 + t^2 & , i = j \end{cases} \tag{5.14}$$

$$\mathrm{Var}(\mathrm{PC}) = 1 + \frac{1}{m^2}\sum_k (s^2 + t^2) \tag{5.15}$$

$$\mathrm{Cov}(\mathrm{PC}, \mathrm{IDP}_j) = 1 + \frac{1}{m}(s^2 + t^2) \tag{5.16}$$

$$\mathrm{Cov}(L, \mathrm{IDP}_j) = \mathrm{Cov}(L, F_j) = 1$$

$$\mathrm{Cov}(Y, \mathrm{IDP}_j) = \alpha + \sum_k \beta_k + s^2\beta_j \tag{5.17}$$

$$\mathrm{Cov}(Y, \mathrm{PC}) = \alpha + \sum_k \beta_k + \frac{1}{m}\sum s^2\beta_k \tag{5.18}$$

#### 5.8.1.3 Biases of the BrainXcan associations

As described in Section 5.8.1.1, $\mathrm{IDP}_k$ is a proxy of $F_k$ and PC is a proxy of $L$. In BrainXcan analysis, we fit linear regression model $Y \sim \mathrm{IDP}_k + \mathrm{PC}$, in which we seek to test whether there exists the region-specific effect ($\beta_k \neq 0$) so we focus on testing if the coefficient of

$\text{IDP}_k$ is zero. Similarly, we also fit $Y \sim \text{PC}$ to test for a brain-wide effect of the subtype. In this section, we derive the expected value of these coefficients (coefficient of $\text{IDP}_k$ in $Y \sim \text{IDP}_k + \text{PC}$ and coefficient of PC in $Y \sim \text{PC}$) to determine how they related to the parameters of interest ($\beta_k$ and $\alpha$).

## Coefficient of $\text{IDP}_k$

Consider fitting the linear model $Y \sim \text{IDP}_k + \text{PC}$. The coefficient of $\text{IDP}_k$ is

$$\begin{bmatrix} \text{coef IDP}_k \\ \text{coef PC} \end{bmatrix} = \begin{bmatrix} \widehat{\text{Var}}(\text{IDP}_k) & \widehat{\text{Cov}}(\text{IDP}_k, \text{PC}) \\ \widehat{\text{Cov}}(\text{IDP}_k, \text{PC}) & \widehat{\text{Var}}(\text{PC}) \end{bmatrix}^{-1} \begin{bmatrix} \widehat{\text{Cov}}(\text{IDP}_k, Y) \\ \widehat{\text{Cov}}(\text{PC}, Y) \end{bmatrix} \tag{5.19}$$

Taking the expected value of the coefficient, we have

$$\text{E}\left(\begin{bmatrix} \text{coef IDP}_k \\ \text{coef PC} \end{bmatrix}\right) = \begin{bmatrix} \text{Var}(\text{IDP}_k) & \text{Cov}(\text{IDP}_k, \text{PC}) \\ \text{Cov}(\text{IDP}_k, \text{PC}) & \text{Var}(\text{PC}) \end{bmatrix}^{-1} \begin{bmatrix} \text{Cov}(\text{IDP}_k, Y) \\ \text{Cov}(\text{PC}, Y) \end{bmatrix} + O_p\left(\frac{1}{n}\right) \tag{5.20}$$

, where $n$ is the sample size of the linear regression. And the $O_p(\cdot)$ is introduced when plugging-in variances and covariances in the places of sample variances and covariances (see more detailed discussion in [4] Appendix A).

Substituting Eq 5.14-5.18 for quantities in Eq 5.20, we have

$$\text{E}(\text{coef IDP}_k) = \underbrace{\frac{s^2}{s^2 + t^2}}_{\text{Attenuation bias}} \cdot \Bigg[\beta_k - \underbrace{\frac{1}{m-1} \sum_{j \neq k} \beta_j}_{\text{Collider effect}}\Bigg] + O\left(\frac{1}{n}\right) \tag{5.21}$$

$$= \beta_k - \frac{t^2}{s^2 + t^2} \cdot \beta_k - \frac{s^2}{s^2 + t^2} \sum_{j \neq k} \frac{\beta_j}{m-1} + O\left(\frac{1}{n}\right) \tag{5.22}$$

We note that there are two sources of bias. First, $\frac{s^2}{s^2+t^2}$ term is introduced by the fact that

$\text{IDP}_k$ is a noisy version of the actual affecting variable $F_k$, which is the so called attenuation bias. Second, $\frac{1}{m-1}\sum_{j\neq k}\beta_j$ term is introduced by the fact that PC is used instead of $L$ as a covariate. As shown in Eq 5.13, PC captures not only $L$ but $R_k$'s which makes PC a collider variable in testing association between $Y$ and $\text{IDP}_k$.

In summary, the coefficient of $\text{IDP}_k$ in $Y \sim \text{IDP}_k + \text{PC}$ mainly captures the effect of region $k$ on $Y$ ($\beta_k$) but it also captures the average effect from all other regions. In a situation where only a few regions have non-zero effects, the second term is usually small.

### Coefficient of PC

Consider fitting the linear model $Y \sim \text{PC}$. The coefficient of PC is

$$\text{coef PC} = \frac{\widehat{\text{Cov}}(\text{PC}, Y)}{\widehat{\text{Var}}(\text{PC})} \tag{5.23}$$

Similarly to Eq 5.20, to work out the expected value of PC coefficient, we plug-in variances and covariances in the places of sample variances and covariances.

$$\begin{aligned}
\text{E(coef PC)} &= \frac{\text{Cov}(\text{PC}, Y)}{\text{Var}(\text{PC})} + O(\frac{1}{n}) && (5.24)\\
&= \frac{m\alpha + (s^2 + m)\sum_j \beta_j}{s^2 + t^2 + m} && (5.25)\\
&= \alpha + \sum_j \beta_j + O(\frac{1}{m}) + O(\frac{1}{n}) && (5.26)
\end{aligned}$$

In summary, the coefficient of PC in $Y \sim \text{PC}$ captures the overall effect of the subtype ($\alpha + \sum_j \beta_j$).

#### *5.8.2 Using IDP residual instead of fitting IDP and PC jointly*

Fitting $\text{IDP}_k$ and PC jointly requires estimating the sample covariance between the predicted $\text{IDP}_k$ and PC. In summary-based BrainXcan, this estimation relies on an external LD panel,

which may contain some noise or even error especially when the LD panel is not representative of the GWAS cohort. In this case, the joint model fitting is sensitive to the quality of the sample covariance estimation. To ensure the robustness of the BrainXcan test, we take an alternative approach which avoid estimating the sample covariance.

In the alternative approach, we fit $Y \sim \text{resIDP}_k$ instead where $\text{resIDP}_k$ is the predicted value of $\text{IDP}_k$ residual (after regressing out PC). In this section, we show that the coefficient of $\text{IDP}_k$ in the joint model $Y \sim \text{IDP}_k + \text{PC}$ is approximately equivalent to the result of a two-step approach:

1. Regress out PC from $\text{IDP}_k$ and keep the residual $r_k$.
2. Obtain coefficient of $r_k$ in $Y \sim r_k$.

First of all, we can calculate $r_k$ from model $\text{IDP}_k \sim \text{PC}$.

$$r_k = \text{IDP}_k - a_k \cdot \text{PC} \tag{5.27}$$

$$a_k = \frac{\widehat{\text{Cov}}(\text{IDP}_k, \text{PC})}{\widehat{\text{Var}}(\text{PC})} \tag{5.28}$$

So, we have

$$\text{Cov}(Y, r_k) = \text{Cov}(Y, \text{IDP}_k) - a_k \cdot \text{Cov}(Y, \text{PC}) \tag{5.29}$$

$$\text{Var}(r_k) = \text{Var}(\text{IDP}_k) - 2 \cdot a_k \cdot \text{Cov}(\text{IDP}_k, \text{PC}) + a_k^2 \cdot \text{Var}(\text{PC}) \tag{5.30}$$

And in step 2, we can obtain the coefficient of $r_k$ from $Y \sim r_k$.

$$\text{coef } r_k = \frac{\widehat{\text{Cov}}(Y, r_k)}{\widehat{\text{Var}}(r_k)} \tag{5.31}$$

$$\text{E}(\text{coef } r_k) \approx \frac{\text{Cov}(Y, r_k)}{\text{Var}(r_k)} \tag{5.32}$$

$$= \frac{s^2}{s^2 + t^2} \cdot \left[ \beta_k - \frac{1}{m-1} \sum_{j \neq k} \beta_j \right] \tag{5.33}$$

Comparing Eq 5.21 and 5.33, we can conclude that E(coef IDP$_k$) $\approx$ E(coef $r_k$). And this result indicates that regressing the outcome $Y$ on IDP$_k$ and PC jointly is approximately equivalent to regressing $Y$ on the residual of IDP$_k$.

In practice, instead of calculating the residual IDP$_k$ by regressing out predicted PC from predicted IDP$_k$ (which still relies on covariance between PC and IDP$_k$), we build the genetic predictor of IDP$_k$ residual and predict the residual IDP$_k$ (which is called resIDP$_k$) directly. We show empirically that the coefficient from $Y \sim$ resIDP$_k$ is similar to coef IDP$_k$ (Figure 5.20).

### *5.8.3 Deriving summary-based BrainXcan*

Recall that the individual-level BrainXcan fits the linear model

$$Y = \widehat{B}_k \beta_k^\star + \sum_{l=1}^{L} a_l C_l + \epsilon,\ \epsilon \sim N(0, \tau^2) \tag{5.34}$$

where $\widehat{B}_k$ is the predicted value of an IDP residual or IDP PC and $C_l$ represents the covariates. And the main interest is to test if $\beta_k^\star$ is equal to zero.

To account for the covariates, we regress out covariates from both $Y$ and $\widehat{B}_k$ and the model is reduced to

$$\widetilde{Y} = \widetilde{B}_k \beta_k^\star + \widetilde{\epsilon},\ \widetilde{\epsilon} \sim N(0, \widetilde{\tau}^2) \tag{5.35}$$

$\widetilde{Y}$ and $\widetilde{B}_k$ are the residuals after regressing out the covariates. From Eq 5.35, we construct

the test statistics for the marginal test as follow:

$$\widehat{\beta}_k^\star = \frac{\widetilde{B}_k' \widetilde{Y}}{\widetilde{B}_k' \widetilde{B}_k} \tag{5.36}$$

$$\text{se}(\widehat{\beta}_k^\star) = \sqrt{\frac{\widehat{\tau}^2}{\widetilde{B}_k' \widetilde{B}_k}} \tag{5.37}$$

$$\widehat{\tau}^2 = \frac{(Y - \widetilde{B}_k \widehat{\beta}_k^\star)'(Y - \widetilde{B}_k \widehat{\beta}_k^\star)}{N - L} \tag{5.38}$$

When the individual-level data is not available, the test statistics can also be calculated approximately from the GWAS of the phenotype $Y$ as described in [7]. Here we consider the situation where there is no covariate in Eq 5.34.

Since $\widehat{B}_k = \sum_j \widehat{w}_{kj} X_j = X \widehat{\mathbf{w}}_k$ where $\widehat{\mathbf{w}}_k = (w_{k1}, \cdots, w_{kP})'$ and $X$ is the individual-by-variant genotype matrix, we can re-write Eq 5.36 as:

$$\widehat{\beta}_k^\star = \frac{\widehat{\mathbf{w}}_k' X' Y}{\widehat{\mathbf{w}}_k' X' X \widehat{\mathbf{w}}_k} \tag{5.39}$$

$$\text{se}(\widehat{\beta}_k^\star) \approx \sqrt{\frac{\widehat{\sigma}_Y^2 (1 - h^2_{\text{BrainXcan}})}{\widehat{\mathbf{w}}_k' X' X \widehat{\mathbf{w}}_k}} \tag{5.40}$$

where the estimated residual variation $\widehat{\tau}^2$ is approximated by $\widehat{\sigma}_Y^2 (1 - h^2_{\text{BrainXcan}})$, *i.e.* the total variation multiplies by the proportion of the variation that cannot be explained by a predicted brain IDP.

Consider the GWAS of phenotype $Y$, the resulting summary statistics are:

$$\widehat{b}_j = \frac{X_j' Y}{X_j' X_j} \tag{5.41}$$

$$\text{se}(\widehat{b}_j) \approx \sqrt{\frac{\widehat{\sigma}_Y^2 (1 - h^2_{\text{GWAS},j})}{X_j' X_j}} \tag{5.42}$$

Let the sample covariance of the genotype be $\widehat{R}$ where $\widehat{R}_{jj'} = \widehat{\text{Cov}}(X_j, X_{j'}) = X_j' X_{j'}/(N-1)$

and $\widehat{R}_{jj} = \widehat{\mathrm{Var}}(X_j) = X_j' X_j/(N-1)$. We can re-arrange Eq 5.41 and 5.42:

$$X'Y = (N-1) S_{\widehat{R}} \widehat{\mathbf{b}} \tag{5.43}$$

$$\sqrt{\widehat{\sigma}_Y^2} = \mathrm{se}(\widehat{b}_j) \sqrt{\frac{(N-1)\widehat{R}_{jj}}{1 - h^2_{\mathrm{GWAS},j}}} \tag{5.44}$$

, where $S_{\widehat{R}}$ has the same diagonal values as $\widehat{R}$ but its all off-diagonal values are zero and $\widehat{\mathbf{b}} = (\widehat{b}_1, \cdots, \widehat{b}_P)'$. Substituting Eq 5.43 and 5.44 to Eq 5.39 and 5.40, we have

$$\widehat{\beta}_k^{\star} = \frac{\widehat{\mathbf{w}}_k' \; S_{\widehat{R}} \; \widehat{\mathbf{b}}}{\widehat{\mathbf{w}}_k' \; \widehat{R} \; \widehat{\mathbf{w}}_k} \tag{5.45}$$

$$\mathrm{se}(\widehat{\beta}_k^{\star}) \approx \mathrm{se}(\widehat{b}_j) \sqrt{\frac{\widehat{R}_{jj}}{1 - h^2_{\mathrm{GWAS},j}}} \sqrt{\frac{1 - h^2_{\mathrm{BrainXcan}}}{\widehat{\mathbf{w}}_k' \; \widehat{R} \; \widehat{\mathbf{w}}_k}} \tag{5.46}$$

Note that in principle we could use any variant $j$ to calculate $\mathrm{se}(\widehat{\beta}_k^{\star})$. Here, specifically, we proceed by evaluating $\widehat{b}_j/\mathrm{se}(\widehat{b}_j)$ first.

$$\frac{\widehat{b}_j}{\mathrm{se}(\widehat{\beta}_k^{\star})} \approx \frac{\widehat{b}_j}{\mathrm{se}(\widehat{b}_j)} \sqrt{\frac{1 - h^2_{\mathrm{GWAS},j}}{\widehat{R}_{jj}}} \sqrt{\frac{\widehat{\mathbf{w}}_k' \; \widehat{R} \; \widehat{\mathbf{w}}_k}{1 - h^2_{\mathrm{BrainXcan}}}} \tag{5.47}$$

$$\approx z_{\mathrm{GWAS},j} \sqrt{\frac{\widehat{\mathbf{w}}_k' \; \widehat{R} \; \widehat{\mathbf{w}}_k}{\widehat{R}_{jj}}}, \text{ since } \frac{1 - h^2_{\mathrm{GWAS},j}}{1 - h^2_{\mathrm{BrainXcan}}} \approx 1 \tag{5.48}$$

Using Eq 5.48, we can write the z-score of the BrainXcan test as follow:

$$z_{\text{BrainXcan},k} = \frac{\widehat{\beta}_k^\star}{\text{se}(\widehat{\beta}_k^\star)} \tag{5.49}$$

$$= \frac{\widehat{\mathbf{w}}_k' \, S_{\widehat{R}}}{\widehat{\mathbf{w}}_k' \, \widehat{R} \, \widehat{\mathbf{w}}_k} \frac{\widehat{\mathbf{b}}}{\text{se}(\widehat{\beta}_k^\star)} \tag{5.50}$$

$$\approx \frac{\widehat{\mathbf{w}}_k' \, S_{\widehat{R}}}{\widehat{\mathbf{w}}_k' \, \widehat{R} \, \widehat{\mathbf{w}}_k} \sqrt{\widehat{\mathbf{w}}_k' \, \widehat{R} \, \widehat{\mathbf{w}}_k} \sqrt{S_{\widehat{R}}}^{-1} \mathbf{z}_{\text{GWAS}} \tag{5.51}$$

$$= \widehat{\mathbf{w}}_k' \sqrt{\frac{S_{\widehat{R}}}{\widehat{\mathbf{w}}_k' \, \widehat{R} \, \widehat{\mathbf{w}}_k}} \mathbf{z}_{\text{GWAS}} \tag{5.52}$$

where $\mathbf{z}_{\text{GWAS}} = (z_{\text{GWAS},1}, \cdots, z_{\text{GWAS},P})'$.

In summary, we can calculate the BrainXcan test statistics using GWAS summary statistics using the following results:

$$\widehat{\beta}_k^\star = \frac{\widehat{\mathbf{w}}_k' \, S_{\widehat{R}} \, \widehat{\mathbf{b}}}{\widehat{\mathbf{w}}_k' \, \widehat{R} \, \widehat{\mathbf{w}}_k} \tag{5.53}$$

$$z_{\text{BrainXcan},k} \approx \widehat{\mathbf{w}}_k' \sqrt{\frac{S_{\widehat{R}}}{\widehat{\mathbf{w}}_k' \, \widehat{R} \, \widehat{\mathbf{w}}_k}} \mathbf{z}_{\text{GWAS}} \tag{5.54}$$

### *5.8.4 Meta-analyzing Mendelian Randomization tests by extending ACAT method [87]*

The ACAT method [87] is a meta-analysis approach which can combine potentially dependent p-values with an analytical formula as described in below:

$$T = \sum_i \tan[(\frac{1}{2} - p_i) \cdot \pi] \tag{5.55}$$

$$p_{\text{ACAT}} = \frac{1}{2} - \frac{\arctan(T/N)}{\pi} \tag{5.56}$$

where $N$ is the total number of p-values being combined and $T$ is the test statistic which follows Cauchy distribution under the null. So, we want to apply ACAT to combine the

results of multiple Mendelian Randomization tests.

In the original use case of ACAT, the p-values are combined without considering the direction of the effect. But in the use case of combining Mendelian Randomization tests, since all the tests are analyzing the mediation effect of the same IDP/phenotype pair, we expect that when there is signal being present in the data, these MR tests should show consistent sign. So, we want to construct a variation of ACAT such that the direction of the effect is considered. Specifically, to combine p-values $p_1, \cdots, p_N$ with signs $s_1, \cdots, s_N$, if we assume the "+1" direction is the direction of true signal, we take the signs into consideration by modifying Eq 5.55:

$$T_{+1} = \sum_i \{s_i \cdot \tan[(\frac{1}{2} - p_i) \cdot \pi]\} \tag{5.57}$$

Or more generally, considering $s \in \{-1, +1\}$ as the true direction,

$$T_s = s \cdot \sum_i \{s_i \cdot \tan[(\frac{1}{2} - p_i) \cdot \pi]\} \tag{5.58}$$

In practice, both "+1" and "-1" directions are possible so we should test both and combine the two p-values at the end. We propose the following meta-analysis approach, signed ACAT (SACAT), which takes p-values $p_1, \cdots, p_N$ with signs $s_1, \cdots, s_N$ and return the combined p-value and sign:

$$p_s = \frac{1}{2} - \frac{\arctan(T_s/N)}{\pi} \tag{5.59}$$

$$p_{\text{SACAT}} = 2 \cdot \min_s p_s \tag{5.60}$$

$$s_{\text{SACAT}} = \arg\min_s p_s \tag{5.61}$$

We note that $p_{+1}$ and $p_{\text{-}1}$ can be combined into one p-value as shown in Eq 5.60 since $p_{+1} + p_{\text{-}1} = 1$ which is the result of $T_{+1} = -T_{\text{-}1}$.

To examine the calibration of SACAT, we perform a simulation study in which we simulate from the global null using the procedure in below:

1. For $i = 1, \cdots, 20$, we simulate $Z_i \sim N(0,1)$ where $\text{Cov}(Z_i, Z_j) = r$ except $\text{Cov}(Z_1, Z_j) = -r$.

2. Calculate $p_i = 2 \times \Pr(|Z| > Z_i; Z \sim N(0,1))$.

3. Repeat the above two steps 10000 times for $r = 0, 0.1, 0.5$ respectively.

We note that this simulation study considers the dependent p-values. In particular, the dependency exists when $r \neq 0$. We show, in Figure 5.22, that the proposed SACAT is well calibrated under the global null.

# CHAPTER 6

# CONCLUSION

Going back to the big picture of human genetics research, we are mainly dealing with three data categories on a daily basis: i) genotypes, ii) complex phenotypes including disease statuses and quantitative measures of human which are all at macro-scale, and iii) intermediate phenotypes including a wide range of human features such as molecular and organ characteristics which potentially mediate complex phenotypes.

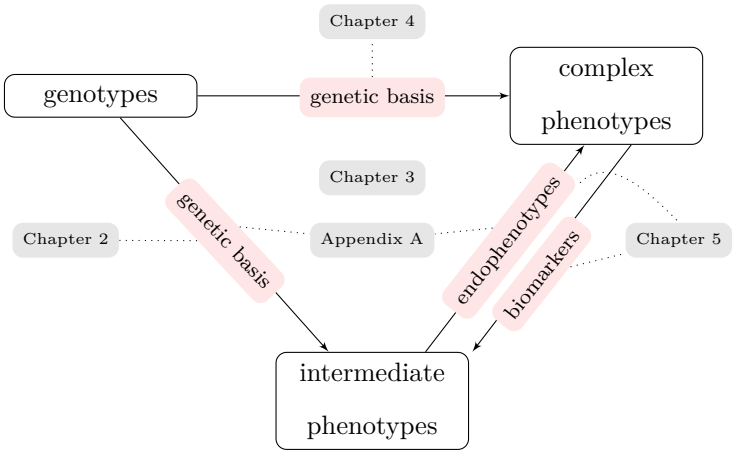



**Figure 6.1: A high-level recapitulation of the thesis.**

For human geneticists, we focus on the relation of these data types (Figure 6.1). In this dissertation, I developed computational and statistical methods to approach various methodological challenges and scientific questions which can all be fit into this big picture. Both Chapter 2 and Chapter 4 focus on the genetic basis of phenotypes. However, methodologically, the underlying statistical and computational challenges were quite different and the proposed methods were highly tailored to the specific problems. Chapter 3 is about applying what we have learned in this diagram and translating it into something with practical use. Whereas, the theme of Chapter 5 is endophenotype identification. Besides method developments, Appendix A centers on applications which integrate state-of-the-art datasets to prioritize causal gene candidates mediating genetic effects of variants. In addition to the

problems being touched in this dissertation, there are many other important challenges being folded in the big picture. Within each data type, there are various challenges awaiting for solutions. In terms of the type of the problems, they include data acquisition/quality (e.g. genotype imputation), pattern recognition (e.g. disease subtyping), network inference/analysis (e.g. regulatory network construction and phenotype comorbidity), and many others. As we fill in the blanks in the picture, on the application side, we also need to develop computational methods to leverage what we have learned so that they can be translated into practical use.

Returning to my dissertation, in the following, I briefly summarize each of my dissertation works and discuss the limitations and future directions.

Chapter 2 focused on the computational burden in cis-quantitaive trait locus (cis-QTL) study. On the basis of RNA-seq data, we can in principle leverage signals from both total reads and allele-specific reads to study the genetic basis of cis-regulation. However, the existing statistical frameworks suffer from prohibitively large computational burden when the sample size is large. To resolve this issue, I proposed a linear approximation to the original statistical model which is easy to solve and unifies the modeling of single-variant and multi-variant effects. The proposed framework is suitable to solve three common problems in cis-QTL study: i) marginal QTL calling; ii) QTL fine-mapping; iii) predicting cis-regulation from genotypes. I proposed and implemented computational procedures to solve these problems in a computationally efficient manner even for a large-scale QTL study. Comparing to the standard approaches which only leverage the signals from total reads, the proposed method requires similar amount of computational power yet achieves significantly higher power since it also considers signals from allele-specific effects. This work enabled the joint analysis of both total and allele-specific reads in large-scale cis-QTL analyses helping scientists to capture more genetic signals of cis-regulation which could have been missed previously. However, regarding the limitations of this work, to approximate count distribution

with lognormal requires the count to be large enough (far from zero count). For bulk RNA-seq data, this is usually not a problem. Recently, single-cell RNA-seq data has also been applied in cis-QTL study by forming pseudo-bulk data from each cell cluster [156]. In this scenario, the proposed approaches may suffer from data sparsity especially when the cell type of interest is rare. Nonetheless, with the simplicity of the proposed model, one can easily extend this framework to study context-specific and dynamic cis-QTLs which have been implicated to be more disease-relevant than cis-QTL being detected at static state. With more and more data being collected nowadays, one can envision that the proposed framework will be widely applicable to answer various questions sitting around cis-QTL and the proposed model can serve as a foundation for more sophisticated modelings of cis-regulation.

Chapter 3 focused on the portability issue of polygenic risk scores (PRSs). The existing PRS approaches are usually not transferable across ancestry groups. This discrepancy may result in healthcare disparity when applying PRSs in a clinical setting. The lack of portability could be caused by the fact that the phenotype is affected by different genetic factors in different ancestry groups (e.g. via the gene-by-environment interactions). However, even though the genetic basis is the same across populations, the portability may still be bad since PRS can overfit the linkage disequilibrium structure and/or minor allele frequencies of training ancestry. Chapter 3 focused on tackling the overfitting issue by proposing a biologically interpretable PRS approach. Motivated by the fact that a proportion of genetic effect acts through transcriptomic regulation, I implemented the predicted transcriptome risk scores (PTRSs) which use the genetically determined gene expression as features to predict complex phenotypes. Unlike existing PRS approaches which predict genotypes to phenotypes directly via a magic black-box, PTRS leveraged the molecular mechanisms underlying and modelled the "genotypes $\rightarrow$ genes $\rightarrow$ phenotypes" pathway explicitly. I characterized the predictive power of PTRS across different ancestry groups and showed that PTRS is more portable than PRS when the score is trained with European-descent individuals and tested

in African-descent individuals. Moreover, we showed that the portability remains when PTRS and PRS are combined and the combined score achieves comparable performance as a PRS. This work showed the potential of developing biologically interpretable PRSs, especially predicted omics-based PRS approaches, for improving the PRS portability. Guided by this work, in the future, one could include additional omics data such as RNA splicing, metabolome, proteome in additional to transcriptome. As we gradually build up the omics-based PRS family, another emerging challenge is how to combine these scores in an "organic" way. Machine learning techniques and knowledge from systems biology may be required in order to solve this problem.

Chapter 4 focused on a very specific GWAS setup where the phenotype is collected from participant's parents instead of the participants themselves. Here the key question is to resolve the missing genotype/phenotype problem. Specifically, Chapter 4 attempted to improve the GWAS of late-onset diseases. Usually, it is a lot harder to recruit patients with late-onset diseases than recruiting patients with high-prevalence diseases. This makes the GWAS of late-onset diseases challenging. Whereas, in some biobank data, the disease status of the parents of the participants are available via questionnaires. With these information, statistical approaches have been proposed to predict the participant's phenotype (i.e. the probability of getting the disease at certain age or beyond) and use the predicted phenotype along with the participant's genotypes to perform GWAS. In this work, I explored an alternative approach, in which I treated the parent's genotype as missing instead. I proposed an EM algorithm leveraging the parental phenomic data (from questionnaires) and the phased genotypes data from the offspring (the participant) to impute the parental origin of the genotypes which is also the haplotype of the parent. Besides, I proposed a downstream GWAS approach taking the observed phenotype and imputed haplotypes (with uncertainty) as inputs. I performed theoretical and simulation studies analyzing the statistical power of the imputation problem and the downstream GWAS problem. The results suggest that

hundreds of parental phenotypes are required in order to obtain enough imputation quality for downstream GWAS, which is not feasible with the present biobank data. Even though this work does not result in a better method improving the GWAS with parental phenotypes, it does provide us insight on the statistical nature of this problem which may guide the future studies with similar essence. Besides, for the problem of GWAS with parental phenotypes specifically, there are still unanswered questions. It has been shown that to treat it as a missing phenotype problem and impute the phenotype from family data achieves good performance in practice [56]. So, it will be interesting to think about the relation between imputing genotypes and imputing phenotypes. And furthermore, we'd like to figure out why the latter yields better results than the former and, besides, whether there is an alternative approach for imputing genotypes which can give improved performance comparing to the current one.

Chapter 5 focused on the identification of image-based brain features as endophenotypes for complex phenotypes. Previously, computational approaches have been proposed to identify associated genes for a complex phenotype which associate the genetically determined gene expressions to the complex phenotype. In this work, I proposed an extension of previous works to study the potentially mediating brain features of a complex phenotype. I leveraged biobank-scale image-based brain phenome. First, I showed that these brain features are highly polygenic. Second, I developed genetic predictors of brain features utilizing the genome-wide genetic variants. Lastly, I implemented a computational pipeline to perform: i) association tests between genetically determined brain features and the complex phenotype; ii) Mendelian randomization for establishing the causal flow between a brain feature and the complex phenotype. This work provides an alternative approach to the observation-based studies to look for the relation between brain features and phenotypes. Comparing to the small-scale observational studies, this work leverages the genetic evidences which takes the advantages of: i) large-scale GWAS studies; ii) genetics-based causal flow

identification. Even though we identified several interesting hits when applying the current models to some existing psychiatric phenotype GWASs, there are still many limitations in the current implementation. First of all, the sample size for training the genetic predictors is still quite small considering the fact that we are predicting highly polygenic phenotypes. Besides, the preprocessing of brain features requires more validations and careful interpretation. For instance, the first principal component (PC1) was regressed out from each modality. PC1 was interpreted as the common factor and the residuals were interpreted as region-specific features. More biological evidence should be collected to further validate this interpretation. And similarly, it will be interesting to examine how other principal components should be interpreted (like [166]), i.e. whether they bury "super-region" effects which affect a few spatially related regions though not all regions. Also, standardizing volumes with brain size could be problematic. Even though it is likely that PC1 will capture brain size if using brain features without standardization, it still requires further investigation to see how this choice affects the results. Another line of approach associates the polygenic score of complex phenotypes with the observed brain features. In the future, it will be interesting to perform empirical and theoretical comparison among these two strategies along with the observational study. Additionally, it will be nice to bring molecular features into the picture via some joint analysis such as partial least squares. On one hand, it provides extra information to support the endophenotype identification. On the other hand, it extends the organ-level mechanism to further down to molecular level, which may be of great importance in therapeutics discovery.

Throughout this dissertation, the main theme is to develop computational methodologies for analyzing genomic, transcriptomic, and phenomic data. Each chapter of my dissertation solve different problems in statistical genetics and human genetics, paving the road for geneticists to understand the biology of complex phenotypes underlying the massive amount of data.

# APPENDIX A

# EXPLOITING THE GTEX RESOURCES TO DECIPHER THE MECHANISMS AT GWAS LOCI

Material from: Barbeira, Alvaro N., Rodrigo Bonazzola, Eric R. Gamazon, Yanyu Liang, YoSon Park, Sarah Kim-Hellmuth, Gao Wang et al., "Exploiting the GTEx resources to decipher the mechanisms at GWAS loci", Genome Biology, published 2021, Springer Nature [6]

## A.1 Abstract

The resources generated by the GTEx consortium offer unprecedented opportunities to advance our understanding of the biology of human diseases. Here, we present an in-depth examination of the phenotypic consequences of transcriptome regulation and a blueprint for the functional interpretation of genome-wide association study-discovered loci. Across a broad set of complex traits and diseases, we demonstrate widespread dose-dependent effects of RNA expression and splicing. We develop a data-driven framework to benchmark methods that prioritize causal genes and find no single approach outperforms the combination of multiple approaches. Using colocalization and association approaches that take into account the observed allelic heterogeneity of gene expression, we propose potential target genes for 47% (2,519 out of 5,385) of the GWAS loci examined. Our results demonstrate the translational relevance of the GTEx resources and highlight the need to increase their resolution and breadth to further our understanding of the genotype-phenotype link.

## A.2 Introduction

In the last decade, the number of reproducible genetic associations with complex human traits that have emerged from genome-wide association studies (GWAS) has substantially

grown. Many of the identified associations lie in non-coding regions of the genome, suggesting that they influence disease pathophysiology and complex traits via gene regulatory changes. Integrative studies of molecular quantitative trait loci (QTL) [106] have established gene expression as a key intermediate molecular phenotype, and improved functional interpretation of GWAS findings, spanning immunological diseases [50], various cancers [159, 49], lipid traits [113, 21], and a broad array of other complex traits.

Large-scale international efforts such as the Genotype-Tissue Expression (GTEx) Consortium have provided an atlas of the regulatory landscape of gene expression and splicing variation in a broad collection of primary human tissues [22, 138, 139]. Nearly all protein-coding genes in the genome now have at least one local variant associated with expression changes and the majority also have common variants affecting alternative splicing (FDR ¡ 5%) [139]. In parallel, there has been an explosive growth in the number of genetic discoveries across a large number of traits, prompting the development of integrative approaches to characterize the function of GWAS findings [9, 42, 168, 51, 151]. Nevertheless, our understanding of underlying biological mechanisms for most complex traits substantially lags behind the improved efficiency of the discovery of genetic associations, made possible by large-scale biobanks and GWAS meta-analyses.

One of the primary tools for the functional interpretation of GWAS associations has been the integrative analysis of molecular QTLs. Colocalization approaches that seek to establish shared causal variants (e.g., eCaviar [54], *enloc* [153], and *coloc* [47]), enrichment analysis (S-LDSC [17] and QTLEnrich [42]) or mediation and association methods (SMR [168], TWAS [51] and PrediXcan [43]) have provided important insights, but they are often used in isolation, and there have been limited prior assessments of power and error rates associated with each [146]. Their applications often fail to provide a comprehensive, biologically interpretable view across multiple methods, traits, and tissues or offer guidelines that are generalizable to other contexts. Thus, a comprehensive assessment of expression and splicing

QTLs for their contributions to disease susceptibility and other complex traits requires the development of novel methodologies with improved resolution and interpretability.

Here, we present methods and resources that help elucidate how genetic variants associated with gene expression (cis-eQTLs) or splicing (cis-sQTLs) contribute to, or mediate, the functional mechanisms underlying a wide array of complex diseases and quantitative traits. Since splicing QTLs have largely been understudied, we perform a comprehensive integrative study of this class of QTLs, in a broad collection of tissues, and disease associations. We provide predictions of functional mechanisms for 74 distinct complex traits from 87 GWA study results and demonstrate independent validation and evaluation of findings using likely causal gene-disease relationships in the Online Mendelian Inheritance of Man (OMIM) database. Notably, we find widespread dose-dependent effects of cis-QTLs on traits through multiple lines of evidence. We examine the importance of considering, or correcting for, false functional links attributed to GWAS loci due to neighboring but distinct causal variants. We call this confounding LD contamination for the remainder of the paper. To identify predicted causal effects among the complex trait associated QTLs, we conduct systematic evaluation across different methods. Furthermore, we provide guidelines for employing complementary methods to map the regulatory mechanisms underlying genetic associations with complex traits.

## A.3 Results

### *A.3.1 Mapping the regulatory landscape of complex traits*

The final GTEx data release (v8) included 54 primary human tissues, 49 of which included at least 70 samples with both whole genome sequencing (WGS) and tissue-specific RNA-seq data. A total of 15,253 samples from 838 individuals were used for cis-QTL mapping (Figure A.1) [139]. In addition to the expression quantitative trait loci (eQTL) mapping, we also

evaluated genetic variation associated with alternative splicing (sQTL) and their impact on complex traits.

We downloaded and processed 114 publicly available GWAS datasets with genome-wide variant association summary statistics (here onwards, summary statistics). After data harmonization, format standardization, missing data imputation and other quality assurance steps (Figure A.8, A.9, and A.10), we retained 87 datasets representing 74 distinct complex traits including cardiometabolic, hematologic, neuropsychiatric and anthropometric traits (Figure A.7). We provide the full list of datasets used in our study and all processing scripts as a resource to the community (Table A.1).

Using these resources, we sought to identify likely causal associations among these gene- and alternatively spliced transcript-associated variants (eVariants and sVariants, respectively). For this purpose, we applied colocalization, enrichment, and association analyses, and provide a resource to enable investigations into gene prioritization approaches for disease associations.

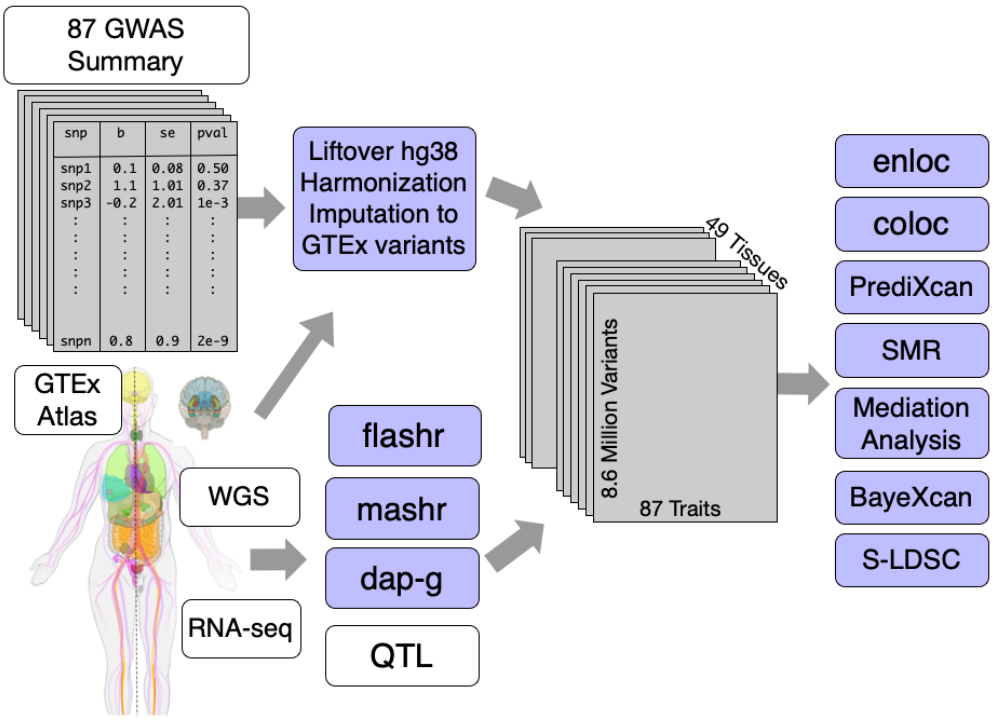



**Figure A.1: Overview of workflow for mapping complex trait associated QTLs.** Full variant association summary statistics results from 114 GWAS were downloaded, standardized, and imputed to the GTEx v8 WGS variant calls (maf¿0.01) for analyses. A total of 8.87 million imputed and genotyped variants were investigated to identify trait-associated QTLs. A total of 49 tissues, 87 studies (74 distinct traits), and 23,268 protein-coding genes and lncRNAs remained after stringent quality assurance protocols and selection criteria. A wide array of complex trait classes, including cardiometabolic, anthropometric, and psychiatric traits, were included.

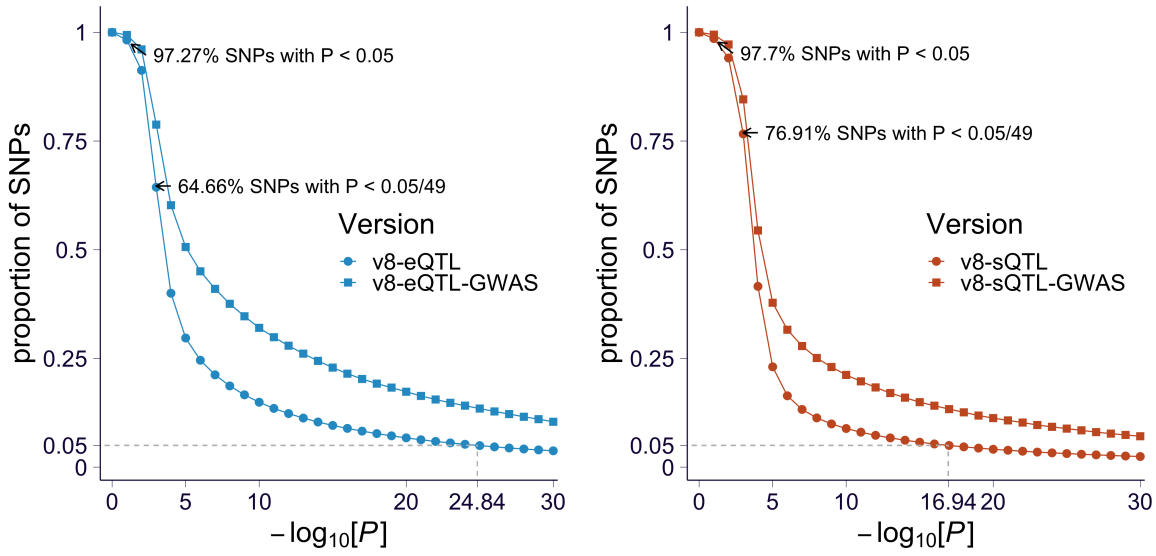



**Figure A.2: Expression and splicing QTL enrichment among GWAS variants.** The proportion of genetic variants associated with gene expression **(A)** and splicing **(B)** of at least one gene in at least one tissue for each p-value cutoff (on x-axis in $-\log_{10}(p)$ scale) is shown. The proportions for all tested variants are shown as circles and the proportions for the GWAS catalog variants are shown as squares.

Gene expression and alternative splicing dysregulations have been proposed as the underlying mechanism of the association signals in many diseases [113, 134, 123, 80, 42, 7]. Similar to previous reports [138], we observed robust and widespread enrichment of eQTLs and sQTLs among disease-associated variants (Figure A.2). This observation suggests a causal role for expression and splicing regulation in complex traits. Figure A.2 also illustrates the dangers of using a naive approach to assigning causal genes to GWAS variants that are associated with expression or splicing, especially when using loose p-value thresholds. For example, with a p-value threshold of 0.05, over 97% of common variants will be assigned some gene in some tissue associated at that level.

### *A.3.2 Dose-dependent regulatory effects of expression and alternative splicing on complex traits*

Nevertheless, enrichment studies can be confounded by many unknown factors. Therefore, we sought to gather stronger evidence for a causal link by testing whether there is a dose-dependent effect of expression and splicing QTLs on complex traits. Figure A.3A illustrates schematically our approach. We examined whether expression or splicing associated variants (referred to as e/sVariants for the remainder of the paper) with higher impact on gene expression or splicing lead to higher impact on a complex trait, i.e. a larger GWAS effect (Figure A.3A). The impact of the regulation of a gene on a trait is quantified by the slope $\beta_{\text{gene}}$. That is, a null hypothesis of no dose-dependent effect is equivalent to $\beta_{\text{gene}} = 0$.

To reduce unnecessary noise in the analysis, we included only the most likely causal e/sVariant within each credible set as determined by the e/sQTL fine-mapping (denoted "fine-mapped variants" throughout the remainder of the paper. See Methods on QTL fine-mapping).

First, we quantified dose-dependent effect of expression and splicing regulation on the trait as the average mediating effect size, $\bar{\beta}$. We calculated this average effect using the Pearson correlation between the absolute values of the molecular and complex trait effect sizes ($\text{cor}(|\gamma|, |\delta|)$) across all fine-mapped variants (for any gene) for each trait-tissue pair. As hypothesized, we found, consistently across all tissue-trait pairs, a positive correlation between the GWAS and QTL effects, which was significantly larger than the permuted null with matched local LD. The average correlations were 0.18 (s.e. $= 0.004$, $p < 1 \times 10^{-30}$) and 0.25 (s.e. $= 0.006$, $p < 1 \times 10^{-30}$) for expression and splicing, respectively with the distribution of the median correlation across tissues for each trait shown in Figure A.3B. Averages and standard errors were calculated taking into account correlation between tissues, and p-values were calculated against permuted null with matched local LD (Section A.5.8.2). The non-negative permuted correlation values indicate that local LD contributed to inflate

the estimated mediation effect. These results provide the first line of evidence of the dose-response effect.

To test and account for mediation effect heterogeneity (different slope/dosage sensitivity for different genes), we modeled the gene-specific mediation effect, $\beta_g$, as a random variable following a normal distribution $\beta_g \sim \mathcal{N}(0, \sigma^2_{\text{gene}})$. Under this random effects model, the null hypothesis can be stated as $\sigma^2_{\text{gene}} = 0$ (Section A.5.8.3 and A.5.8.4; Figure A.3C). As shown in Figure A.3C, these effects were significantly larger than expected from the permuted null (expression $p = 1.8 \times 10^{-9}$; splicing $p = 2.5 \times 10^{-7}$). These results indicate that strong genetic effects on expression or splicing are more likely to have a strong association to complex traits, adding strong support to a dose-dependent relationship between gene regulation and downstream traits.

Importantly, by averaging across all genes, the estimates, from both the average and the random-effects approach, of the mediating effect are robust to confounding due to LD, as discussed in Section A.5.8.5.

Another way to account for mediation effect heterogeneity is to make use of the allelic series of independent eQTLs identified for over half of the eGenes [139]. We examined whether the mediating effect ($\beta = \delta/\gamma$) inferred from the primary eQTL ($\beta_{\text{prim}}$) was consistent with the one inferred from the secondary eQTL ($\beta_{\text{sec}}$). Among the independent eQTLs for a given gene, we called primary the one with the larger effect size. We considered only fine-mapped eQTLs given the low power to detect multiple independent sQTLs. We confirmed this concordance, as reported by the GTEx consortium [139], demonstrating that the correlation between the primary and secondary mediating effects is larger than expected given the LD between them. To better visualize this concordance, we plotted the estimated mediating effects of primary against the secondary eQTLs (whole blood shown here but other tissues look similar) in Figure A.3D and showed that they cluster in the first and third quadrants. All gene-trait pairs with relatively high regional colocalization probability (rcp¿0.10, see colo-

calization details below) are shown here to facilitate visualization, but the clustering around the diagonal line was observed even without the filtering. This provides a third confirmatory evidence for the widespread dose-dependent effects of eQTLs on complex traits.

Note that genes with discordant effects within the allelic series would be harder to detect and suggest more complex causal relationship or context specificity.

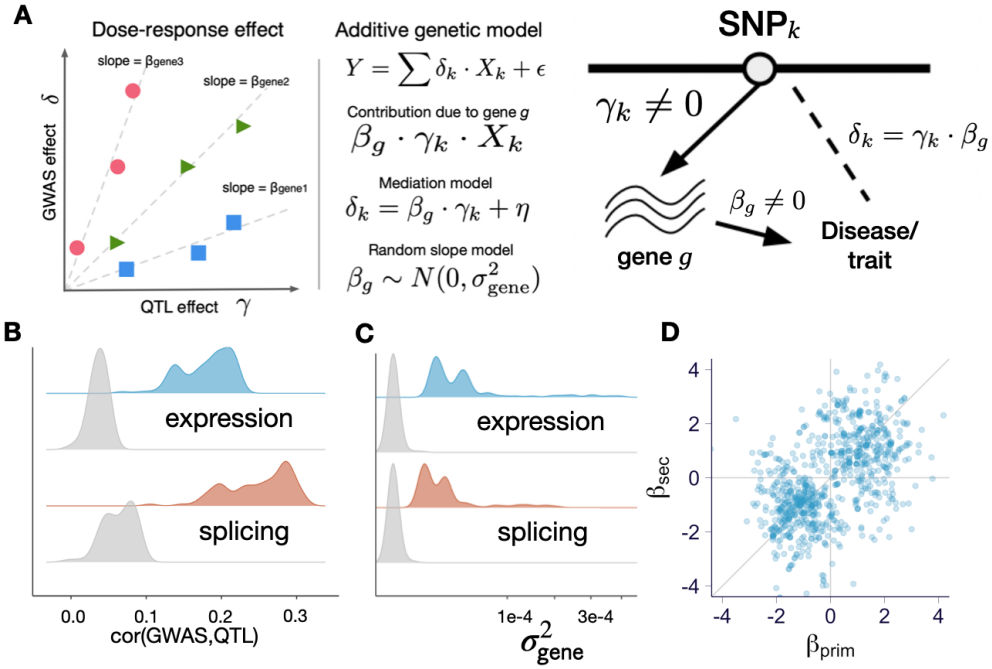



**Figure A.3: Dose-dependent effects of QTLs on complex traits.** Here all analyses were performed with fine-mapped variants (QTL with highest posterior inclusion probability). **(A)** Schematic representation of dose-response model. **(B)** Correlation between QTL and GWAS effects, Cor($|\hat{\delta}|, |\hat{\gamma}|$). Gray distribution represents permuted null with matched local LD. Each data point corresponds to the median correlation for the trait across 49 tissues. **(C)** Average mediated effects from mediation model ($\sigma^2_{\text{gene}}$, median across tissues). Gray distribution represents permuted null with matched local LD. **(E)** Mediated effects of secondary vs primary eQTLs of genes with colocalization probability (rcp) $> 0.10$. in whole blood, genes for all 87 traits are shown.

### *A.3.3 Causal gene prediction and prioritization*

In addition to genome-wide analyses that shed light on the molecular architecture of complex traits, QTL analysis of GWAS data can identify potential causal genes and molecular changes in individual GWAS loci. Towards this end, we performed association analysis with

genetically predicted regulation and colocalization (Figure A.4A). After evaluating the performance of *coloc* and *enloc* [153, 47], we chose *enloc* as our primary approach, due to its use of hierarchical models to estimate colocalization priors [153] and its ability to account for multiple causal variants. The *coloc* assumption of a single causal variant drastically reduces performance especially in large QTL datasets such as GTEx with widespread allelic heterogeneity (Figure A.26). For a more extensive discussion on the benefits of Bayesian colocalization methods and comparison of *enloc* to other colocalization approaches including SMR-HEIDI see [57]. We estimated the posterior regional colocalization probability (rcp), using *enloc*, for 12,072,964 tissue-gene-GWAS locus-trait tuples and 67,943,800 tissue-splicing event-GWAS locus-trait tuples. For the tally of colocalized genes, we used rcp¿0.5 as a stringent cutoff as demonstrated below with the low colocalization probabilities of height loci using two different datasets.

In total, we identified 3,477 (15% of 23,963) unique genes colocalizing with GWAS hits (rcp ¿ 0.5) across all traits and tissues analyzed (Figure A.14A). Similarly, 3,157 splicing events (1% out of 310,042) colocalized with GWAS hits, corresponding to 1,226 genes with at least one colocalized splicing event (5% of 23,963, Figure A.14B).

Colocalization of e/sQTLs with GWAS variants provides important causal support for molecular traits. However, we found their estimates to be overly conservative. To illustrate this point, we tested the colocalization of height with itself, using two large-scale studies of individuals of European-ancestry individuals: GIANT [158] and UK Biobank. We started by performing fine-mapping of both GWAS results using *susieR* [149]. Notably, only 416 (39%) of GIANT's fine-mapped credible sets overlapped with the corresponding UK Biobank credible sets. We estimated the colocalization probability as the sum of the product of posterior inclusion probabilities of variants for each of the 1069 independent credible sets in GIANT, which is similar to the approach used by eCAVIAR [54]. Two thirds of the GIANT credible sets (66.2%) had a colocalization probability below 0.01 and about half (48.9%)

had a colocalization probability below 0.001. In other words, two thirds of the loci found by GIANT would be considered not to be colocalized with UK Biobank's loci when using a seemingly very loose colocalization probability cutoff of 0.01. Given the larger sample size of the UK Biobank GWAS (n=337,119 UKB GWAS vs. n=253,288 for GIANT), the low colocalization cannot be attributed to lack of power. This result is likely due in part to the sensitivity to small LD differences between different EUR populations that make up large GWAS meta-analysis cohorts such as GIANT. Our analysis illustrates the fact that colocalization probability estimates are highly conservative and may miss many causal genes, and low colocalization probability should not be interpreted as evidence of lack of a causal link between the molecular phenotype and the GWAS trait. Notice that this limitation is not inherent to the colocalization method itself but the limitation of currently available large-scale GWAS meta analysis results.

A complementary approach to colocalization is to estimate the GWAS trait association with genetically predicted gene expression or splicing [43]. The GTEx v8 data provides an important expansion of these analyses, allowing generation of prediction models in 49 tissues with whole genome sequencing data to impute gene expression and splicing variation. We trained prediction models using a variety of approaches and selected the top performing one based on precision, recall, and other metrics [8]. Briefly, the optimal model uses fine-mapping probabilities for feature selection and exploits global patterns of tissue sharing of regulation (Section A.5.12.1; Figure A.29) to improve prediction. In-depth comparison of these fine-mapped models with Elastic Net-based and CTIMP [55] models is described in [8]. The analysis presented here uses these improved models (fine-mapped-*mashr*) instead of Elastic Net as reported in the main GTEx publication [139]. Multi-SNP prediction models were generated for a total of 686,241 gene-tissue and 1,816,703 splicing event-tissue pairs. The larger sample size and improved models led to an increase in the number of expression models to a median across tissues of 14,062, from a median of 4,776 GTEx v7 Elastic Net

models (median increase at 191%, Figure A.13). Splicing models are available only for the v8 release.

Next, we computed the association between an imputed molecular phenotype (expression or splicing) and a trait to estimate the genic effect on the trait, using the summary statistics based PrediXcan [7]. Given the widespread tissue-sharing of regulatory variation [138], we also computed MultiXcan scores to integrate patterns of associations from multiple tissues and increase statistical power [9]. Out of the 22,518 genes tested with PrediXcan, 6,407 (28%) showed a significant association with at least one of the 87 traits at Bonferroni-corrected p-value threshold ($p < 0.05/686,241$, where the denominator is the number of gene-tissue pairs tested; Figure A.14). For splicing, about 15% (20,364 of 138,890) of tested splicing events showed a significant association ($p < 0.05/1,816,703$, where the denominator is the number of intron-tissue pairs tested). Nearly all traits (94%; 82 out of 87) showed at least one significant gene-level PrediXcan association in at least one tissue (Figure A.18 and A.19); the median number of associated genes across traits was 974. This resource of PrediXcan associations can be used to prioritize a list of putatively causal genes for follow-up studies.

To replicate the PrediXcan expression associations in an independent dataset, BioVU, which is a large-scale biobank tied to Electronic Health Records [122, 33], we selected seven traits with predicted high statistical power. Out of 947 gene-tissue-trait discoveries tested, 458 unique gene-tissue-trait triplets (48%) showed replication in this independent biobank (PrediXcan association p ¡ 0.05; see Section A.5.12.5). Further confirming this statistical replication in BioVU, we used the PheWAS [33] catalog as the silver standard and found an AUC curve of 0.62. [116].

Altogether, these results provide abundant links between gene regulation and GWAS loci. To further quantify this, we split the genome into approximately LD-independent blocks [12] and identified blocks with a significant GWAS variant for each trait (at Bonferroni threshold adjusted for number of variants $0.05/8.8 \times 10^6 \sim 5.7 \times 10^{-9}$); we refer to any such region-

trait pair by "GWAS locus". We calculated the proportion of GWAS loci that contain a significantly associated gene via PrediXcan or a colocalized gene via *enloc* (rcp ¿ 0.5). Briefly, the LD blocks are defined by analyzing empirical patterns of LD observed in 1000 Genomes [1] and variants in different regions are unlikely to be correlated, thus providing us with a data-driven criterion to distinguish independent genomic signals.

Across the traits, 72% (3,899/5,385) of GWAS loci had a PrediXcan expression association in the same LD block, of which 55% (2,125/3,899) had evidence of colocalization with an eQTL (Table A.3). For splicing, 62% (3,345/5,385) had a PrediXcan association of which 34% (1,135/3,345) colocalized with an sQTL (Figure A.17). From the combined list of eGenes and sGenes, 47% of loci have a gene with both *enloc* and PrediXcan support. The distribution of the proportion of associated and colocalized GWAS loci across 87 traits is summarized in Figure A.4-C; for a typical complex trait, about 20% of GWAS loci contained a colocalized, significantly associated gene while 11% contained a colocalized, significantly associated splicing event. These results propose function for a large number of GWAS loci, but most loci remain without candidate genes, highlighting the need to expand the resolution of transcriptome studies.

A recent report estimates that the proportion of trait variance explained by the assayed transcriptome is on average 11% [162]. Even though this number is not directly comparable with the proportion of loci with support from PrediXcan and *enloc*, some discussion is warranted. Differences may arise with our analysis from the fact that 1) GTEx v8 doubles the number of samples with both genotype and RNAseq relative to v7, 2) we include links based on splicing in addition to expression, 3) a variant may act through both regulation of expression levels and other undetected mechanisms (pleiotropy), and 4) attenuation bias may reduce the estimates given the error in eQTL effect sizes.

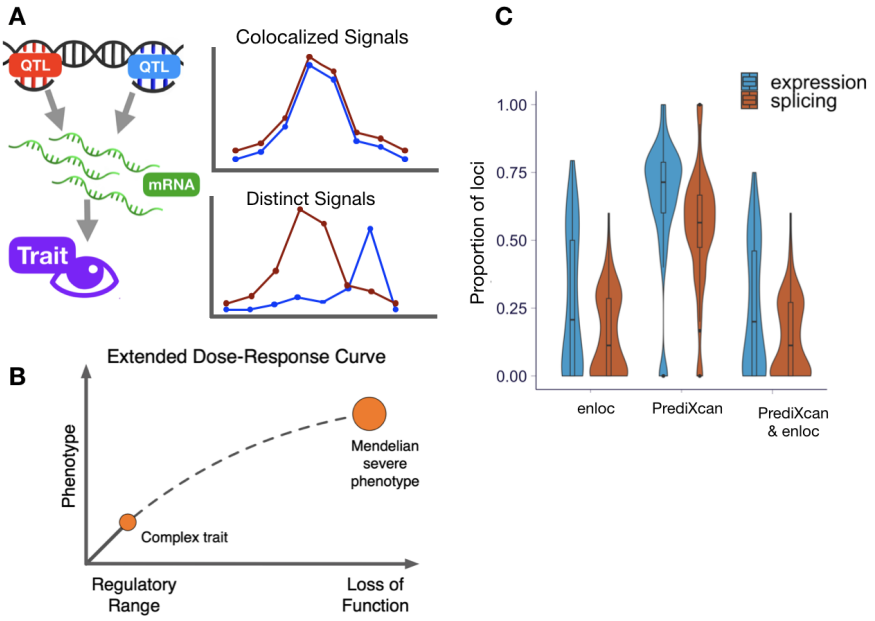



**Figure A.4: Identifying and validating predicted causal genes. (A)** Schematic representation of association and colocalization approaches. **(B)** Schematic representation of extrapolating the dose-response curve to the Mendelian end of phenotypic variation spectrum [117]. **(C)** Proportion of GWAS-associated loci per trait that contain colocalized and PrediXcan-associated signals for expression and splicing.

Of note, two members of the sterolin family, *ABCG5* and *ABCG8*, showed highly significant predicted causal associations using both PrediXcan and *enloc* for LDL-C levels and self-reported high cholesterol levels. *ABCG8* showed more significant associations in both datasets (chr2: 43838964 - 43878466; UKB self-reported high cholesterol: -log10($p_{\text{PrediXcan}}$) = 38.43, rcp = 0.985; GLGC LDL-C: -log10($p_{\text{PrediXcan}}$) = 71.40, rcp = 0.789), compared to *ABCG5* (chr2: 43812472 - 43838865; -log10($p_{\text{PrediXcan}}$) = 36.85, rcp = 0.941; -log10($p_{\text{PrediXcan}}$) = 80.80, rcp = 0.705). Mutations in either of the two ATP-binding cassette (ABC) half-transporters, *ABCG5* and *ABCG8*, lead to reduced secretion of sterols into bile, and ultimately, obstruct cholesterol and other sterols exiting the body [67]. In mice with disrupted *Abcg5* and *Abcg8* (G5G8-/-), a 2- to 3-fold increase in the fractional absoprtion of dietary plan steols and extrememly low biliary cholesterol levels was observed, indicating that disrupting these genes contribute greatly to plasma cholesterol levels [163].

The overexpression of human *ABCG5* and *ABCG8* in transgenic Ldlr-/- mice resulted in 30% reduction in hepatic cholesterol levels and 70% reduced atherosclerotic legion in the aortic root and arch [157] after 6-months on a Western diet.

Several other lipid-associated loci were also consistently predicted as causal across OMIM, the rare variant derived set, PrediXcan and *enloc*. Rare protein-truncating variants in *APOB* have been previously associated with reduced LDL-C and triglyceride levels and reduced coronary heart disease risk [114]. Interestingly, *APOB* has been predicted as a causal gene in four related traits, coronary artery disease, LDL-C levels, triglyceride levels, and self-reported high cholesterol levels. Among the four traits, PrediXcan showed the highest association to LDL-C levels (-log10($p_{\text{PrediXcan}}$) = 130.89; rcp = 0.485) while self-reported high cholesterol showed the strongest evidence using *enloc* at nearly maximum posterior probability (-log10($p_{\text{PrediXcan}}$) = 93.66; rcp = 0.969). Although *APOB* has been suggested as a better molecular indicator of predicted cardiac events in place of LDL-C levels [147, 25], its translation has been surprisingly slow in clinical practice [77]. Here, we provide an additional support for the crucial role *APOB* may play in predicting lipid traits.

### *A.3.4 Performance for identifying "ground truth" genes*

To compare the ability of different approaches to identify the causal gene that mediates the association between GWAS loci and the traits, we sought to curate sets of "ground truth" genes using information that is independent of GWAS results. We call these sets "silver standards" as a reminder of their imperfect nature. The first silver standard was based on the OMIM (Online Mendelian Inheritance in Man) database [52] and the second one was based on publicly available rare variant tests from exome-wide association studies [96, 85, 89] (Figure A.20, Table A.5), resulting in 1,592 OMIM gene-trait pairs and 101 rare variant based gene-trait pairs (Table A.11, Table A.12, Figure A.21.)

The rationale behind the choice of the OMIM database is the comorbidity among

Mendelian and complex diseases suggesting that genes whose loss of function cause Mendelian diseases also manifest in milder phenotypic variation when modified to a lesser degree by regulatory variation [92, 13]. In other words, that the dose-response curve at the regulatory range may be extrapolated to the rare, loss-of-function end (Figure A.4B). The rationale behind the use of the rare variant association study results is the excess of deleterious rare variants associated with complex traits in genes that are in the vicinity of common variants associated with the same trait [96, 41, 64]. Note that rare variant associations are nearly independent of common variants due to the allele frequency difference between them.

For the analysis, we partitioned the genome into approximately independent LD blocks [12] and considered all the blocks where a silver standard gene was available for the trait. Since only genes in the vicinity of an index gene can be discovered with cis-regulatory information, we only considered the LD blocks with a GWAS significant variant. This selection resulted in 228 OMIM gene-trait pairs (28 distinct traits) and 80 rare variant-associated genes-trait pairs (5 distinct traits) that are located within the same LD block as the GWAS locus for a matched trait (Figure A.22).

Both PrediXcan and *enloc* based on expression and splicing showed good sensitivity and specificity for identifying the silver standard genes as demonstrated by the ROC curves in Figs. A.5A-B. These are well above the gray random guess lines indicating the predictive ability of these methods to find causal genes (see comparison with permuted null in Figure A.27).

For applications such as target selection for drug development or follow-up experiments, another relevant metric is the precision or, equivalently, positive predictive value (PPV) – the probability that the gene-trait link is causal given that it is called significant or colocalized. Precision recall curves for expression and splicing based predictions are shown in Figure A.5C-D. With more stringent threshold (towards the left in the recall axis), higher precision is obtained.

For example, 8.7% of genes with PrediXcan significant genes ($p < 0.05/49 \times$ number of gene/trait pairs) were OMIM genes and 14.8% of genes with high colocalization probability (rcp¿0.5) were also OMIM genes for matched traits.

Multiple factors contribute to the rather low precision. One of them is the widespread molecular pleiotropy [139], i.e. multiple genes affected by the same trait-associated variants. Another factor reducing the overall causal gene detection performance is the inherent bias of the OMIM gene list. Our current understanding of gene function is biased towards protein-coding variants with very large effects, as reflected in the list of OMIM genes. Genes associated to rare severe disease tend to be depleted of regulatory variation [63, 102], which will decrease the performance of a QTL-based method [102].

Among the 206 loci with at least one OMIM gene (a few loci contained multiple OMIM genes), an OMIM gene was the closest to the top GWAS SNP in 31.6% of the loci, it was the most colocalized in 24.8% of the loci, and it was the most significant in 20.4% of the loci (Figure A.5E-F).

To further investigate whether this predictive power could be improved by combining multiple criteria, we performed a joint logistic regression of OMIM gene status on 1) the proximity of the top GWAS variant to the nearest gene (distance to the gene body), 2) posterior probability of colocalization, and 3) PrediXcan association significance between QTL and GWAS variants. To make the scale of the three features more comparable, we used their respective ranking. When genes did not have an enloc or PrediXcan score, they were assigned to the last position in the ranking. All three features were significant predictors of OMIM gene status, with better ranked genes more likely to be OMIM genes (proximity $p = 2.0 \times 10^{-2}$, *enloc* $p = 6.1 \times 10^{-3}$, PrediXcan $p = 2.5 \times 10^{-4}$), indicating that each method provides an additional source of causal evidence even after conditioning on the others. Similar results were obtained using splicing colocalization and association scores and the rare variant based silver standard, as shown in Table A.7. These results provide further

empirical evidence that a combination of colocalization and association methods will perform better than individual ones. The significance of the proximity score even after accounting for significance and colocalization indicates missing regulatory events, i.e. mechanisms that may be uncovered by assaying other tissue or cell type contexts, larger samples, and other molecular traits, underscoring the need to expand the size and breadth of QTL studies. Proximity criterion also helps resolve cases when QTL data indicates multiple genes with similar significance.

Predicted OMIM genes included well-known findings such as *PCSK9* for LDLR, with *PCSK9* significant and colocalized for relevant GWAS traits (LDL-C levels, coronary artery disease, and self-reported high cholesterol), and *Interleukins* and *HLA* subunits for asthma, both significant and colocalized for related immunological traits. Significantly associated and colocalized genes that predicted OMIM genes also included *FLG* (eczema), *TPO* (hypothyroidism), and *NOD2* (inflammatory bowel disease) (see Table A.11 for complete list). Analysis with rare variant-based silver standard yielded similar conclusions (Section A.5.13.2; Figure A.25).

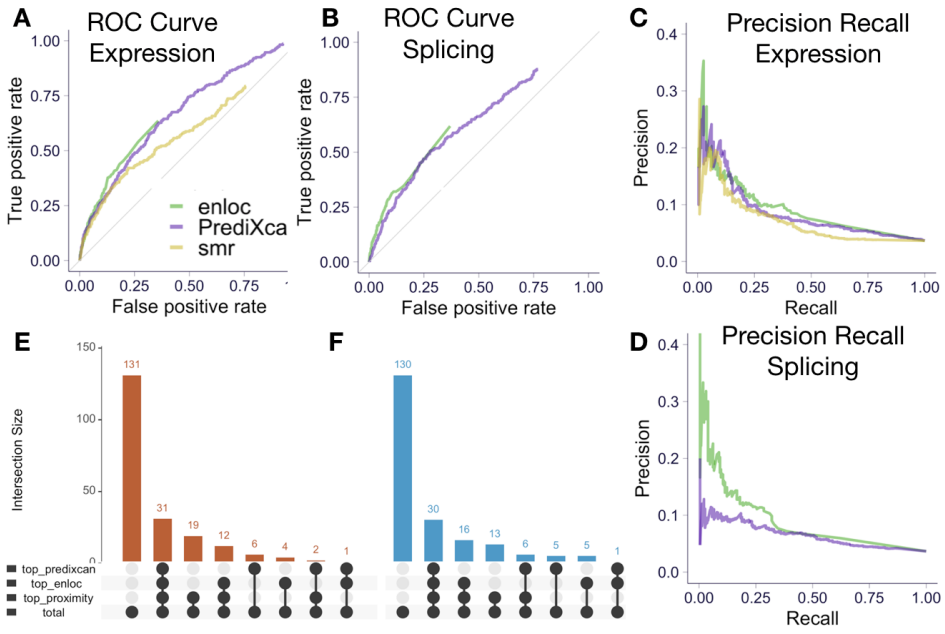



**Figure A.5: Causal gene identification performance.** ROC curves of *enloc* and PrediXcan statistics to identify the 'causal' genes (OMIM silver standard) using expression **(A)** and splicing **(B)** are shown. Precision recall curves of *enloc* and PrediXcan to identify silver standard genes using expression **(C)** and splicing **(D)** (We show the precision in the range 0 to 0.4 to improve visualization). The number of GWAS loci (LD block-trait pairs) where the OMIM gene was ranked at the top by proximity, *enloc*, and PrediXcan using expression **(E)** and splicing **(F)**. In 131 loci out of 206 the OMIM gene was not ranked at the top by either proximity, significance, or colocalization. In thirty one of the loci, the OMIM gene was ranked first by all three criteria. In nineteen loci, the OMIM gene was closest gene (to the top GWAS variant) but not the top gene by PrediXcan significance nor *enloc*'s colocalization proability.

### *A.3.5 Tissue enrichment of GWAS signals*

The broad sharing of regulatory variation across tissues and the reduced significance of tissue-specific eQTLs, makes causal tissue identification challenging. To address this problem, we devised a novel approach to identify tissues of relevance for the etiology of complex traits. We investigated the patterns of tissue specificity and tissue sharing of PrediXcan association results across 49 tissues. For each trait-gene pair, the PrediXcan z-score can be represented as a $49 \times 1$ vector with each entry being the gene-level z-score in the corresponding tissue (if the prediction model of the gene is not available in that tissue, we filled in zero). To explore the tissue-specificity of the PrediXcan z-score vector, we proceeded by assigning

the z-score vector to a tissue-pattern category and tested whether certain tissue-pattern categories were over-represented among colocalized PrediXcan genes as compared to non-colocalized genes. We used the FLASH factors identified from matrix factorization applied to the cis-eQTL effect size matrix, as PrediXcan and cis-eQTL shared similar tissue-sharing pattern (see Section A.5.9). To obtain a set of detailed and biologically interpretable tissue-pattern categories from the 31 FLASH factors, we manually merged them into 18 categories as shown in Figure A.29. For each trait, we projected the z-score vector of each gene to one of the 31 FLASH factors (as described in Section A.5.9) so that the gene was assigned to the corresponding tissue-pattern category. We defined a 'positive' set of genes as the ones with PrediXcan p-value that meets Bonferroni significance at $\alpha = 0.05$ in at least one tissue and *enloc* rcp ¿ 0.01 in at least one tissue, which could be thought as a set of candidate genes affecting the trait through expression level. We chose a rather low threshold used for the rcp due to the stringent conservative nature of colocalization probabilities. We also constructed a 'negative' set of genes with *enloc* rcp = 0, which could be thought as a set of genes whose expressions were unlikely to affect the trait. We proceeded to test whether certain tissue-pattern categories were enriched in 'positive' set as compared to 'negative' set. Since the main focus of this analysis was tissue-specific patterns, we excluded *Factor1* (the cross-tissue factor) and *Factor25* (likely to be a tissue-shared factor capturing tissues with large sample size). Additionally, we excluded *Factor7* (testis), as it was unlikely to be the mediating tissue but might introduce false positives. We tested the enrichment of each tissue-pattern category by Fisher's exact test ('positive'/'negative' sets and in/not in tissue-patter category). Among 87 traits, 82 traits had *enloc* signal and the enrichment of these was calculated accordingly.

Using the pattern of tissue classes of non-colocalized genes (rcp = 0) as the expected null, we assessed whether significantly associated and colocalized genes (PrediXcan significant and rcp ¿ 0.01) were over-represented in certain tissue classes (Figure A.6). Consistent

with previous reports [42, 109], we identified several instances in which the most significant tissue is supported by current biological knowledge. For example, blood cell count traits were enriched in whole blood, neuroticism and fluid intelligence in brain/pituitary, hypothyrodism in thyroid, coronary artery disease in artery, and cholesterol-related traits in liver. Taken together, these results show the potential of leveraging regulatory variation to help identify tissues of relevance for complex traits.

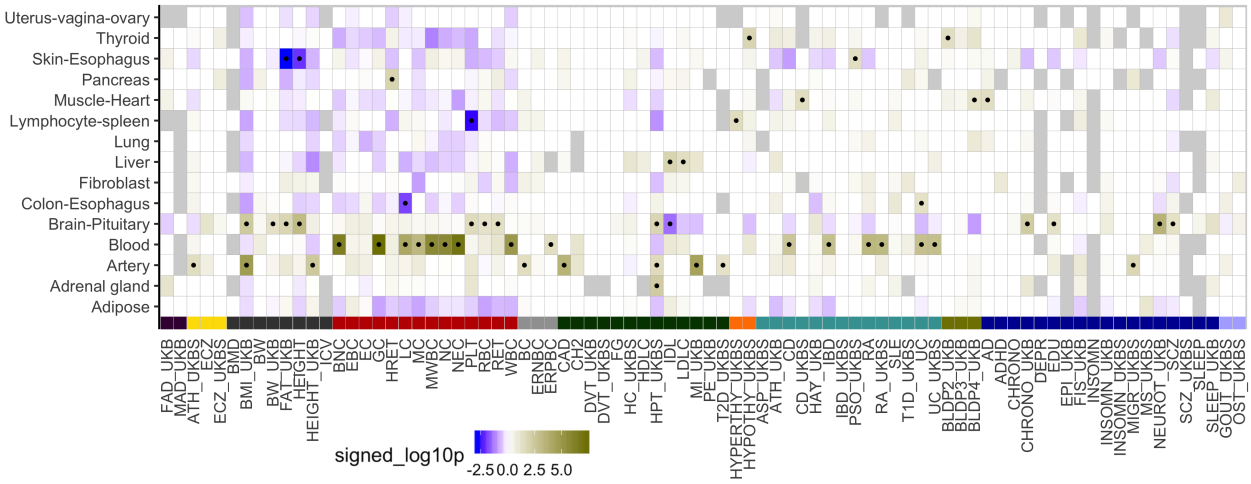



**Figure A.6: Identifying trait-relevant tissues using tissue-specific enrichment.** Enrichment of tisssue-specific association and colocalization compared to the pattern of tissue-specificity of non-colocalized genes. Over-representation of the tissue class for PrediXcan-significant and colocalized genes is indicated by dark yellow while depletion is indicated by blue. Black dots label the tissue class-trait pairs passing the nominal p-value significance threshold of 0.05. Abbreviation: A.1. Trait category colors: A.7.

## A.4 Discussion

We performed in-depth examination of the phenotypic consequences of the genetic regulation of the transcriptome and provide data-driven analytical approaches to benchmark methods that assign function to GWAS loci and best-practice guidelines for using the GTEx resources to interpret GWAS results. We provide a systematic empirical demonstration of the widespread dose-dependent effect of expression and splicing on complex traits, i.e., variants with larger impact at the molecular level have larger impact at the trait level. Furthermore,

we found that target genes in GWAS loci identified by *enloc* and PrediXcan were predictive of OMIM genes for matched traits, implying that for a proportion of the genes, the dose-response curve can be extrapolated to the rare and more severe end of the genotype-trait spectrum. The observation that common regulatory variants target genes also implicated by rare coding variants underscores the extent to which these different types of genetic variants converge to mediate a spectrum of similar pathophysiological effects and may provide a powerful approach to drug target discovery.

We implemented association and colocalization methods that leverage the observed allelic heterogeneity of expression traits. After extensive comparison using two independent sets of silver standard gene-trait pairs, we conclude that combining *enloc*, PrediXcan, and proximity ranking outperforms the individual approaches. The significance of the proximity ranking is a sign of the "missing regulability" emphasizing the need to expand the resolution, sample size, and range of contexts of transcriptome studies as well as to examine other molecular mechanisms.

We caution that the increased power offered by this release of the GTEx resources also brings higher risk of false links due to LD contamination and that naive use of eQTL or sQTL association p-values to assign function to a GWAS locus can be misleading. Colocalization approaches can be used to weed out LD contamination but given the lack of LD references from source studies, they can also be overtly conservative. General purpose reference LD from publicly available sources are not sufficient for fine-mapping and colocalization approaches, which can be highly sensitive to LD misspecification when only summary results are used [11]. The GWAS community has made great progress in recognizing the need to share summary results, but to take full advantage of these data, improved sharing of LD information from the source study as well as from large sequencing reference datasets, is also required.

Finally, we generated several resources that can open the door for addressing key questions in complex trait genomics. We present a catalog of gene-level associations, including

potential target genes for nearly half of the GWAS loci investigated here that provides a rich basis for studies on the functional mechanisms of complex diseases and traits. We provide a database of optimal gene expression imputation models that were built on the fine-mapping probabilities for feature selection and that leverage the global patterns of tissue sharing of regulation to improve the weights. These imputation models of expression and splicing, which to date has been challenging to study, provide a foundation for transcriptome-wide association studies of the human phenome – the collection of all human diseases and traits – to further accelerate discovery of trait-associated genes. Collectively, these data thus represent a valuable resource, enabling novel biological insights and facilitating follow-up studies of causal mechanisms.

## A.5 Supplementary Materials

### *A.5.1 Terminology*

For clarity and to reduce ambiguities, we provide the definition of some of the key terms used in the manuscript.

**Trait:** Here, trait (or complex trait) is used for observable, quantitative trait of individuals, such as presence of a disease or an anthropometric measurement. When speaking about traits, we do not include molecular phenotypes like gene expression or intron splicing quantification.

**LD block/LD region:** Region of the genome containing variants in LD among themselves, as determined from empirical LD patterns observed in 1000 Genomes [12]. Variants in different LD blocks are unlikely to be correlated.

**GWAS locus:** This term is, in general, used somewhat loosely to refer to a region with a significantly associated variant which may span from tens to hundreds of kilobases depending on the LD of the region. However, here for quantification, we define it as one of the approximately independent LD blocks from [12] that harbor a GWAS significant association. If multiple traits exist for a GWAS significant association in the block, we count them as distinct.

**eQTL, eVariant:** Here an eVariant is a genetic variant that is associated (FDR$<$ 0.05) with the expression of a gene. eQTL refers to the variant-gene pair, in which the variant is an eVariant for the gene.

**sQTL, sVariant:** An sVariant is a genetic variant that is associated (FDR$<$ 0.05) with the splicing (quantified as intron excision ratio) of a gene. sQTL refers to the variant-gene pair, in which the variant is an sVariant for the gene.

**Fine-mapped variant:** We call fine-mapped variant to the proxy for causal variant which we selected using *dap-g*'s posterior inclusion probabilities. These variants that are within credible sets with total posterior inclusion probability of at least 0.25 and have variant-level pip$>$ 0.01. Within each credible set, one such variant is selected for our analysis.

**LD contamination:** This phenomenon occurs when the variant that alters the expression or splicing is distinct from the one that alters the complex trait, but they are in LD. In these circumstances, the QTL will be associated with the GWAS trait and the GWAS variant will be associated with the molecular trait, but there is no causal relationship between the gene and the complex trait.

**Posterior inclusion probability (pip):** This is the probability that a variant has a causal effect on a trait. These probabilities are calculated by Bayesian fine-mapping approaches

such as *dapg* and *susier*.

**PrediXcan:** This term refers to the family of methods that seeks to identify causal genes by correlating the genetic component of gene expression (mRNA level and splicing) with the trait. This family includes S-PrediXcan (which uses GWAS summary statistics rather than individual level data) and MultiXcan (which aggregates evidence of associations across all tissues leveraging the fact that e/sQTLs are shared across tissues). We use PrediXcan as a generic term to refer to this family of methods.

**Silver standard genes:** To test the ability of colocalization and association methods to identify true causal genes, we curated a set of 'causal' genes. To emphasize the imperfect nature, we use the term silver standard genes. In this context, the term OMIM gene is used as the causal gene for the trait.

### *A.5.2 Genotype-Tissue Expression (GTEx) Project*

All processed Genotype-Tissue Expression (GTEx) Project v8 data have been made available on dbGAP (accession ID: phs000424.v8). Primary and extended results generated by consortium members are available on the Google Cloud Platform storage accessible via the GTEx Portal.The GTEx Project v8 data, based on 17,382 RNA-sequencing samples from 54 tissues of 948 post-mortem subjects, has established the most comprehensive map of regulatory variation to date. In addition to the larger sample size and greater tissue coverage compared to v6, v8 data also included whole-genome sequencing data, facilitating high resolution QTL map of 838 subjects for 49 tissues with at least 70 samples. The GTEx consortium mapped complex trait associations for 23,268 cis-eGenes and 14,424 cis-sGenes [139]. We did not include trans QTLs in our analyses due to limited power after correcting for confounders and potential pleiotropic effect in complex trait associations. Below, we briefly describe the whole-genome sequencing, RNA-sequencing and QTL data processing protocols. Detailed

description of subject ascertainment, sample procurement, and sequencing data processing are available elsewhere [139].

### A.5.2.1 Whole-genome sequence data processing and quality control

Out of 899 WGS samples sequenced at an average coverage of 30x on HiSeq200 (68 samples) and HiSeqX (all other samples), variant call files (VCF) for 866 GTEx donors were included in downstream analyses after excluding one each from 30 duplicate samples and three donors. Of these, 838 subjects with RNA-seq data were included for QTL mapping and subsequent complex trait association analyses in our study. All whole-genome sequencing data were mapped to GRCh38/hg38 reference.

### A.5.2.2 RNA-Seq data processing and quality control

Whole transcriptome RNA-Seq data were aligned using STAR (v2.5.3.a; [34]). For STAR index, GENCODE v26 (GRCh38) was used with the sjdbOverhang 75 for 76-bp paired-end sequencing protocol. Default parameters were used for RSEM [78] index generation. GTEx utilized Picard to mark and remove potential PCR duplicates and RNA-SeQC [32] to process post-alignment quality control. RSEM was then used for per-sample transcript quantification. Subsequently, read counts were normalized between samples using TMM [121]. For eQTL analyses, latent factor covariates were calculated using PEER as follows: 15 factors for N $<$ 150 per tissue; 30 factors for 150 $\leq N <$ 250; 45 factors for 250 $\leq N <$ 350; and 60 factors for N $\geq$ 350. Finally, fastQTL [110] was used for cis-eQTL mapping in each tissue. Only protein-coding, lincRNA, and antisense biotypes as defined by Gencode v26 were considered for further analyses. To study alternative splicing, GTEx applied LeafCutter (version 0.2.8; [79]) using default parameters to quantify splicing QTLs in cis with intron excision ratios [139].

### *A.5.3 Genome-wide association studies (GWAS) data*

#### A.5.3.1 Harmonization of GWAS summary statistics

The process followed for the harmonization and imputation are depicted in Figure A.8. For each standardized GWAS summary statistics, we mapped all variants to hg38 (GRCh38) references using *pyliftover*. For missing chromosome or genomic position information in the original GWAS summary statistics file, we queried dbSNP build 125 (hg17), dbSNP build 130 (hg18/GRCh36), and dbSNP build 150 (hg19/GRCh37) using the provided variant rsID information and the original reference build of the GWAS summary statistics file. Variants with missing chromosome, genomic position, and rsID information were excluded from further analyses. Only autosomal variants were included in our analyses. Missing allele frequency information was filled using the allele frequencies estimated in the GTEx (v8) individuals of genotype-based European genetic ancestry (here onwards, GTEx-EUR) whenever possible. We excluded variants with discordant reference and alternate allele information between GTEx and the GWAS study. We included only the alleles with the highest MAF among multiple alternate alleles if the variant was reported as multiallelic in GTEx. When more than one GWAS variant mapped to a given GTEx variant (i.e., the same chromosomal location in hg38), only the one with the highest significance was retained. For binary traits, if the sample size was present but the number of cases was missing, we filled the missing count with the sample size and number of cases reported in the paper. For continuous traits, if the file contained the sample size for each variant, the reported number was used. If not, we filled this value using the number reported in the corresponding publication. If only some variants were missing sample size information, we filled the missing value with the median of all reported values.

### A.5.3.2 Imputation of GWAS summary statistics

To standardize the number of variants across tissue-trait pairs, all processed GWAS results were imputed. We implemented the Best Linear Unbiased Prediction (BLUP) approach [74, 112] in-house (`https://github.com/hakyimlab/summary-gwas-imputation`) to impute z-scores for those variants reported in GTEx without matching data in the GWAS summary statistics. This algorithm does not impute raw effect sizes ($\beta$ coefficients). The imputation was performed in specific regions assumed to have sufficiently low correlations between them, defined by approximately independent linkage disequilibrium (LD) blocks [12] lifted over to hg38/GRCh38.

Only GTEx variants with MAF $> 0.01$ in GTEx-EUR subjects were used in downstream analyses. Covariance matrices (reference LD information) were estimated on these GTEx-EUR subjects. The corresponding (pseudo-)inverse matrices for covariances $C$ were calculated via Singular Value Decomposition (SVD) using ridge-like regularization $C + 0.1I$. To avoid ambiguous strand issues homogeneously, palindromic variants (i.e. CG) were excluded from the imputation input. Thus, an imputed z-score was generated for palindromic variants available in the original GWAS; for them, we report the absolute value of the original entry with the sign from the imputed z-score. The sample size that we report for the imputed variants is the same as the sample size for the observed ones if it is reported as constant across variants, or their median if it changes across the observed variants, which occurs in the case of meta-analyses.

We initially considered publicly available GWAS summary statistics for 114 complex traits provided by large-scale consortia and the UK Biobank [20] (table A.8). Of these, 27 studies with a relatively small intersection of variants with the GTEx panel (number of variants$< 2 \times 10^6$, compared to almost $9 \times 10^6$ variants available in GTEx) exhibited significant deflation of their association p-values (Figure A.10). Thus, all analyses focused on 87 traits where missing variants could be properly imputed unless otherwise stated explicitly

(table A.1). We observed noteworthy association prediction performance across the selected 87 traits (e.g., with a median $r^2 = 0.90$ (IQR = 0.0268) between the original and imputed zscores on chromosome 1). The median slope was 0.94 (IQR = 0.0164), as the imputed zscore values tend to be more conservative than the original ones. Imputation quality was consistent across traits, depending strongly on the number of input available variants (Figure A.9). The main reason to drop the 27 traits were to keep the consortium's multiple papers consistent. None of the conclusions in this paper changed when including all 114 traits.

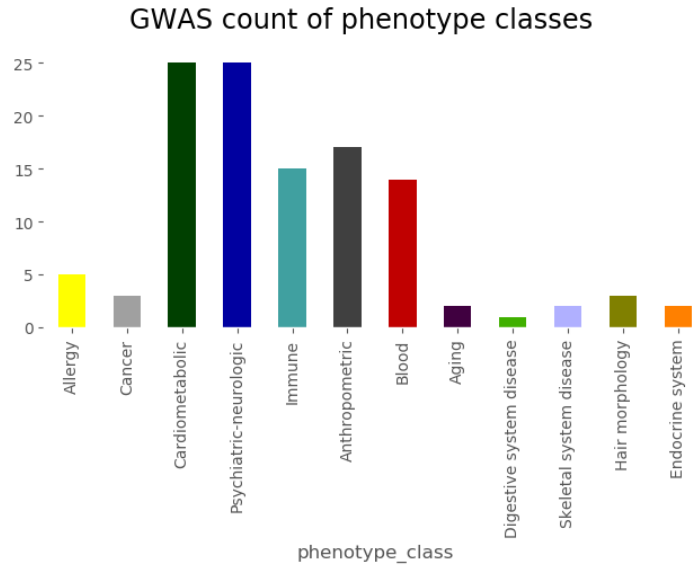



**Figure A.7: GWAS trait categories.** Categories of the traits with full GWAS summary statistics used in the analysis. See list of traits in A.1.

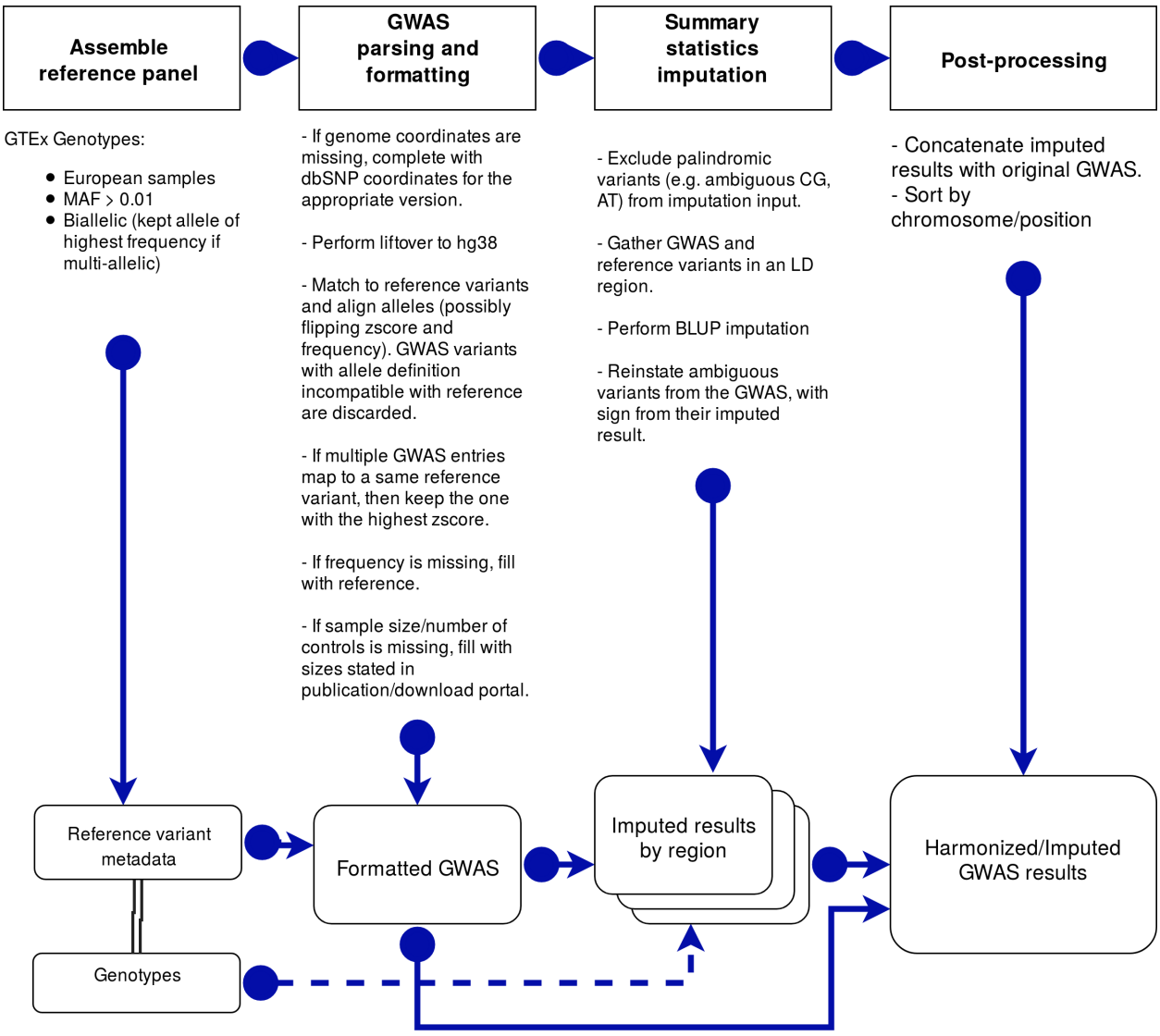



**Figure A.8: Workflow of GWAS results processing.**

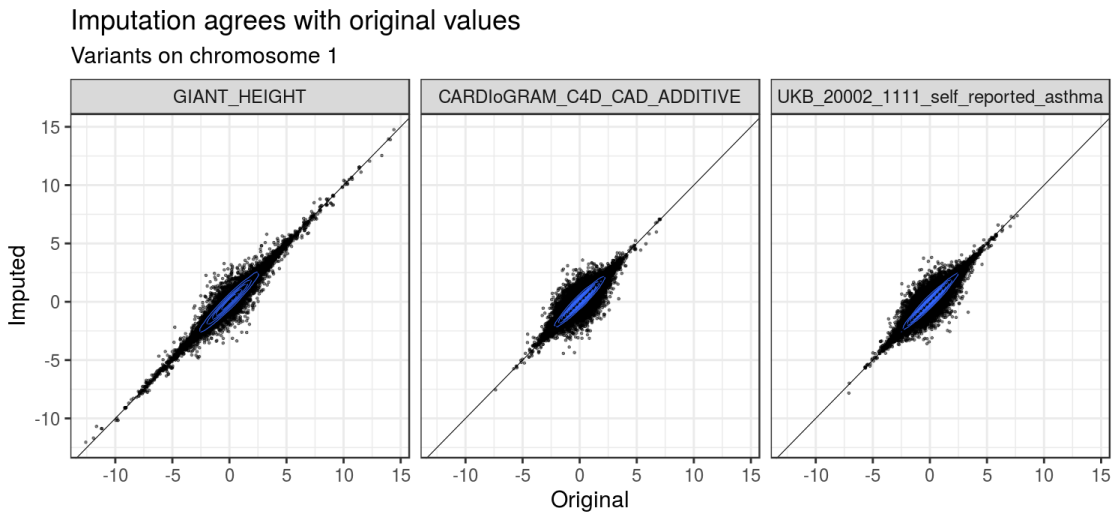



**Figure A.9: GWAS imputation quality** Original versus imputed zscores for palindromic variants in chromosome 1 for 3 traits.

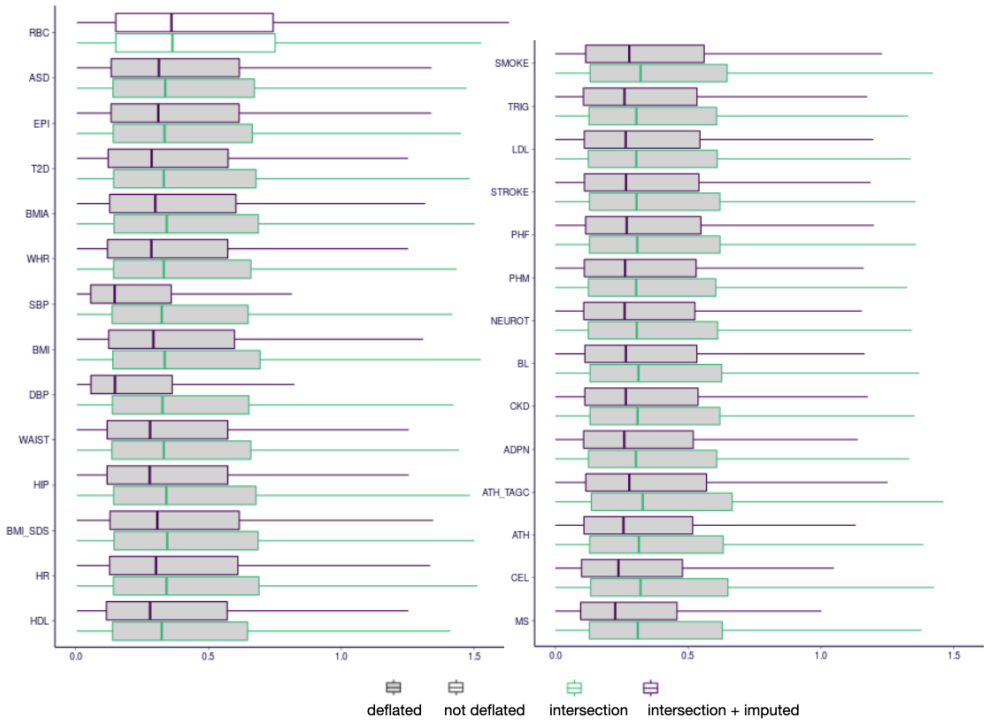



**Figure A.10: GWAS imputation deflation** This figure compares the distribution of p-values for 28 GWAS traits before and after imputation. Vertical scale shows -log10(p-value) of variant association. The 27 traits that exhibited deflation are filled in gray. An undeflated trait (e.g., Red Blood Cell count) is included for comparison. See trait abbreviation list in table A.8.

| Category | Trait | Abbreviation | Sample_Size |
|---|---|---|---|
| Psychiatric-neurologic | Alzheimers Disease | AD | 54162 |
| Psychiatric-neurologic | Attention Deficit Hyperactivity Disorder | ADHD | 53293 |
| Psychiatric-neurologic | Chronotype | CHRONO | 128266 |
| Psychiatric-neurologic | Chronotype UKB | CHRONO_UKB | 337119 |
| Psychiatric-neurologic | Depressive Symptoms | DEPR | 180866 |
| Psychiatric-neurologic | Education Years | EDU | 293723 |
| Psychiatric-neurologic | Epilepsy UKB | EPI_UKB | 337119 |
| Psychiatric-neurologic | Fluid Intelligence Score UKB | FIS_UKB | 337119 |
| Psychiatric-neurologic | Insomnia In Both Sexes | INSOMN | 113006 |
| Psychiatric-neurologic | Insomnia UKB | INSOMN_UKB | 337119 |
| Psychiatric-neurologic | Insomnia UKBS | INSOMN_UKBS | 337119 |
| Psychiatric-neurologic | Migraine UKB | MIGR_UKB | 337119 |
| Psychiatric-neurologic | Migraine UKBS | MIGR_UKBS | 337119 |
| Psychiatric-neurologic | Multiple Sclerosis UKBS | MS_UKBS | 337119 |
| Psychiatric-neurologic | Neuroticism UKB | NEUROT_UKB | 337119 |
| Psychiatric-neurologic | Parkinsons Disease UKBS | PD_UKBS | 337119 |
| Psychiatric-neurologic | Psychological Problem UKBS | PSY_UKBS | 337119 |
| Psychiatric-neurologic | Schizophrenia | SCZ | 150064 |
| Psychiatric-neurologic | Schizophrenia UKBS | SCZ_UKBS | 337119 |
| Psychiatric-neurologic | Sleep Duration | SLEEP | 128266 |
| Psychiatric-neurologic | Sleep Duration UKB | SLEEP_UKB | 337119 |
| Anthropometric | BMI UKB | BMI_UKB | 337119 |
| Anthropometric | Birth Weight | BW | 143677 |
| Anthropometric | Birth Weight UKB | BW_UKB | 337119 |
| Anthropometric | Body Fat Percentage UKB | FAT_UKB | 337119 |
| Anthropometric | Bone Mineral Density | BMD | 49988 |
| Anthropometric | Height | HEIGHT | 253288 |
| Anthropometric | Intracraneal Volume | ICV | 30717 |
| Anthropometric | Standing Height UKB | HEIGHT_UKB | 337119 |
| Cardiometabolic | CH2DB NMR | CH2 | 24154 |
| Cardiometabolic | Coronary Artery Disease | CAD | 184305 |
| Cardiometabolic | Deep Venous Thrombosis UKB | DVT_UKB | 337119 |
| Cardiometabolic | Deep Venous Thrombosis UKBS | DVT_UKBS | 337119 |
| Cardiometabolic | Fasting Glucose | FG | 46186 |
| Cardiometabolic | Fasting Insulin | INSUL | 38238 |
| Cardiometabolic | HDL Cholesterol NMR | HDLC | 19270 |
| Cardiometabolic | Heart Attack UKB | MI_UKB | 337119 |
| Cardiometabolic | High Cholesterol UKBS | HC_UKBS | 337119 |
| Cardiometabolic | Hypertension UKBS | HPT_UKBS | 337119 |
| Cardiometabolic | LDL Cholesterol NMR | LDLC | 13527 |
| Cardiometabolic | Pulmonary Embolism UKB | PE_UKB | 337119 |
| Cardiometabolic | Triglycerides NMR | IDL | 21559 |
| Cardiometabolic | Type 2 Diabetes UKBS | T2D_UKBS | 337119 |
| Blood | Eosinophil Count | EC | 173480 |
| Blood | Granulocyte Count | GC | 173480 |
| Blood | High Light Scatter Reticulocyte Count | HRET | 173480 |
| Blood | Lymphocyte Count | LC | 173480 |
| Blood | Monocyte Count | MC | 173480 |
| Blood | Myeloid White Cell Count | MWBC | 173480 |
| Blood | Neutrophil Count | NC | 173480 |
| Blood | Platelet Count | PLT | 173480 |
| Blood | Red Blood Cell Count | RBC | 173480 |
| Blood | Reticulocyte Count | RET | 173480 |
| Blood | Sum Basophil Neutrophil Count | BNC | 173480 |
| Blood | Sum Eosinophil Basophil Count | EBC | 173480 |
| Blood | Sum Neutrophil Eosinophil Count | NEC | 173480 |
| Blood | White Blood Cell Count | WBC | 173480 |
| Cancer | Breast Cancer | BC | 120000 |
| Cancer | ER-negative Breast Cancer | ERNBC | 120000 |
| Cancer | ER-positive Breast Cancer | ERPBC | 120000 |
| Allergy | Asthma UKBS | ATH_UKBS | 337119 |
| Allergy | Eczema | ECZ | 116863 |
| Allergy | Eczema UKBS | ECZ_UKBS | 337119 |
| Immune | Ankylosing Spondylitis UKBS | ASP_UKBS | 337119 |
| Immune | Asthma UKB | ATH_UKB | 337119 |
| Immune | Crohns Disease | CD | 20833 |
| Immune | Crohns Disease UKBS | CD_UKBS | 337119 |
| Immune | Hayfever UKB | HAY_UKB | 337119 |
| Immune | Inflammatory Bowel Disease | IBD | 34652 |
| Immune | Inflammatory Bowel Disease UKBS | IBD_UKBS | 337119 |
| Immune | Psoriasis UKBS | PSO_UKBS | 337119 |
| Immune | Rheumatoid Arthritis | RA | 80799 |
| Immune | Rheumatoid Arthritis UKBS | RA_UKBS | 337119 |
| Immune | Systemic Lupus Erythematosus | SLE | 23210 |
| Immune | Type 1 Diabetes UKBS | T1D_UKBS | 337119 |
| Immune | Ulcerative Colitis | UC | 27432 |
| Immune | Ulcerative Colitis UKBS | UC_UKBS | 337119 |
| Aging | Fathers Age At Death UKB | FAD_UKB | 337119 |
| Aging | Mothers Age At Death UKB | MAD_UKB | 337119 |
| Digestive system disease | Irritable Bowel Syndrome UKBS | IBS_UKBS | 337119 |
| Endocrine system disease | Hyperthyroidism UKBS | HYPERTHY_UKBS | 337119 |
| Endocrine system disease | Hypothyroidism UKBS | HYPOTHY_UKBS | 337119 |
| Skeletal system disease | Gout UKBS | GOUT_UKBS | 337119 |
| Skeletal system disease | Osteoporosis UKBS | OST_UKBS | 337119 |
| Morphology | Balding Pattern 2 UKB | BLDP2_UKB | 337119 |
| Morphology | Balding Pattern 3 UKB | BLDP3_UKB | 337119 |
| Morphology | Balding Pattern 4 UKB | BLDP4_UKB | 337119 |

**Table A.1: List of 87 GWAS datasets**

### A.5.3.3 IGAP GWAS

We used summary results from an Alzheimer's Disease study from International Genomics of Alzheimer's Project (IGAP).

IGAP is a large two-stage study based upon genome-wide association studies (GWAS)

on individuals of European ancestry. In stage 1, IGAP used genotyped and imputed data for 7,055,881 single nucleotide polymorphisms (SNPs) to meta-analyze four previously-published GWAS datasets consisting of 17,008 Alzheimer's disease cases and 37,154 controls (The European Alzheimer's disease Initiative - EADI the Alzheimer Disease Genetics Consortium - ADGC The Cohorts for Heart and Aging Research in Genomic Epidemiology consortium - CHARGE The Genetic and Environmental Risk in AD consortium - GERAD). In stage 2, 11,632 SNPs were genotyped and tested for association in an independent set of 8,572 Alzheimer's disease cases and 11,312 controls. Finally, a meta-analysis was performed combining results from stages 1 & 2.

#### A.5.3.4 NHGRI-EBI GWAS catalog

In addition to the GWAS summary statistics described above, we obtained the list of trait-associated SNPs from the GWAS catalog [18] (downloaded on 9/7/2018), which, at download, contained 80,727 entries. To measure the enrichment of e/sQTL in the GWAS Catalog, we computed the proportion of e/sQTL in the GWAS catalog relative to the proportion of e/sQTL among all GTEx V8 variants. We then obtained a measure of the uncertainty in the proportion and enrichment-fold using block jackknife. See [139] for details.

### *A.5.4 Correlated t-test to summarize across traits and tissues*

Most statistics shown in these analyses are at the tissue-trait level. There are 4,263 statistics, generated from 49 tissues and 87 traits. Typical statistical tests assume the data from which the statistic is computed is sampled independently and identically distributed (IID). Among different tissues, there are wide ranges of standard errors and different patterns of correlation. Because of this, the IID assumption can not be applied to the tissue-trait statistics. Therefore, we describe our derivation of standard errors when statistics are summarized across traits for a given tissue, and when statistics are summarized across tissue and trait pairs. In the following paragraphs, we use $S_{tp}$ to indicate a statistic estimated in tissue $t$

and trait $p$. This statistic has standard error $\text{se}(S_{tp})$.

**Summarizing across traits for a given tissue.** When we have one statistic per tissue-trait pair and summarize across traits in a given tissue, we assume the traits are independent, but we take into account the differences in standard errors. For each tissue $t$, we summarized $S_{t1}, \cdots, S_{tP}$ by fitting the following linear model:

$$S_{tp} = \mu_S^t + \epsilon_{tp} \tag{A.1}$$

$$\epsilon_{tp} \sim N(0, \text{se}(S_{tp})^2 \times \sigma_t^2) \tag{A.2}$$

So $\hat{\mu}_S^t$ is an estimate for the statistic $S$ summarized across all traits in tissue $t$, and this estimate has standard error $\text{se}(\hat{\mu}_S^t)$. This is essentially a weighted average across traits.

**Summarizing across trait and tissue pairs.** When we summarize across all tissue-trait pairs, $S_{11}, \cdots, S_{tp}, \cdots, S_{TP}$, we fit a similar linear model, which allows for correlation between tissues and correlation between traits, and corrects for differences in the standard errors.

$$S_{tp} = \mu_S + \mu_S^t + \mu_S^p + \epsilon_{tp} \tag{A.3}$$

$$\mu_S^t \sim N(0, \sigma_T^2) \tag{A.4}$$

$$\mu_S^p \sim N(0, \sigma_P^2) \tag{A.5}$$

$$\epsilon_{tp} \sim N(0, \text{se}(S_{tp})^2 \times \sigma^2), \tag{A.6}$$

Here, $\mu_S^t$ is the tissue-specific random intercept, and $\mu_S^p$ is the trait-specific random intercept. These components account for features common across traits that are specific to tissue $t$ and features common across tissues that are specific to trait $p$ respectively. The estimate $\hat{\mu}_S$ is the weighted average of $S_{tp}$ across all tissue-trait pairs, and its standard error

is $\text{se}(\hat{\mu}_S)$.

**Testing whether two statistics have different mean.** We would often like to test whether two statistics are different, *e.g.* enrichment signal measured for sQTL as $\mu_{S_1}$ versus enrichment signal measured for eQTL as $\mu_{S_2}$. For this, we need to construct a test aggregating pairwise differences across all tissue-trait pairs. For this purpose, we constructed the following paired test. Our test statistic is $T^{tp} := S_{1,tp} - S_{2,tp}$ with $\text{se}(T^{tp}) = \sqrt{\text{se}(S_{1,tp})^2 + \text{se}(S_{2,tp})^2}$. We calculate $\hat{\mu}_T$ by summarizing across all tissue-trait pairs as described in the previous paragraph. Under the null $\mathcal{H}_0 : \mu_{S_1} = \mu_{S_2}$ and $\hat{\mu}_T \sim N(0, \text{se}(\hat{\mu}_T))$.

### *A.5.5 Enrichment of QTLs among trait-associated variants*

To estimate the proportion of SNPs considered as associated with expression (for at least one gene) at various p-value thresholds, we used the most significant p-value (tested using all GTEx individuals) for each SNP from all associations in all tissues (including all genes and variants tested). We observed that the proportion of variants associated with expression and splicing at different significance threshold was much larger for trait-associated variants from the GWAS catalog than for the full set of tested common variants (Figure A.2). At a nominal threshold, the proportion of common variants associated with the expression of a gene in some tissue increased from 92.7% in the V6 release [138] to 97.3% in V8. For splicing, the proportion was 97.7%. These results should serve as a cautionary note that assigning function to a GWAS locus based on QTL association p-value alone, even with a more stringent threshold, could be misleading.

### *A.5.6 Cis-region and covariates used in fine-mapping and prediction of expression and splicing traits*

For each gene, we considered all variants within the cis-window (1Mbps) with MAF>0.01, and used the same covariates as in the GTEx v8 main eQTL analysis: sex, WGS plaform,

WGS library preparation protocol, top 5 genetic principal components, and PEER factors. The number of PEER factors was determined from the sample size: 15 for $n < 150$, 30 for $150 \leq n < 250$, 45 for $250 \leq n < 350$, 60 for $350 \leq n$.

### *A.5.7 Fine-mapping expression and splicing QTLs*

We applied *dap-g* [151] to the 49 tissues to estimate the degree to which a variant might exert a causal effect on expression or splicing levels, using default parameter values. First, we selected genes annotated as protein-coding, lincRNA or pseudogenes. We used the covariates listed in the Section A.5.6. This yielded a list of clusters (variants related by LD), and posterior inclusion probabilities (*pip*) that provide an estimate of the probability of a variant being causal. We repeated this process for splicing ratios from Leafcutter, using a cis-window ranging from 1Mbps upstream of the splicing event start location to 1Mbps downstream of the end location. We used individual-level data for GTEx-EUR subjects both for expression and splicing. We note that the main report of the GTEx v8 included individuals of non-European descent and reported only expression QTL fine-mapping. Sample sizes ranged from 65 in kidney cortex to 602 in skeletal muscle tissues. All results are made publicly available (`https://github.com/hakyimlab/gtex-gwas-analysis`).

### *A.5.8 Mediation analysis to quantify the dose-dependent effects of expression and splicing on traits*

Enrichment of expression and splicing QTLs suggest a causal role of molecular trait regulation on complex traits. However, confounders such as LD contamination could be inflating these results limiting their interpretation. Here, we sought to gather stronger evidence for a causal link. We tested whether there is a dose-dependent effect of expression and splicing QTLs on complex traits and also whether independent QTLs provided similar measures of the mediated effects.

#### A.5.8.1 Selection of fine-mapped variants as instrumental variables and their effect sizes

To investigate the relationship between GWAS and QTL effect sizes in the transcriptome, we generated a set of fine-mapped QTL signals derived from *dap-g* fine-mapping performed in the GTEx-EUR individuals to serve as proxy for causal QTLs. For splicing, we utilized sQTLs at the splicing event/variant level rather than the gene/variant level. We considered only variants within credible sets with at least 25% total probability. Within each credible set, the variant with highest posterior inclusion probability was selected as the fine-mapped variant. Only variants with variant-level *pip* of at least 0.01 were considered.

For each of the selected QTLs, we used the QTL effect size estimated from the marginal test (using the GTEx-EUR individuals) and the GWAS effect size reported by the study or if missing, calculated from the imputed z-score from the GWAS imputation by $\hat{\beta} \approx z/\sqrt{f(1-f)N}$, where $f$ is the allele frequency and $N$ is the GWAS sample size.

#### A.5.8.2 Correlation between GWAS and QTL effect sizes

To get a first-order approximation to the mediated effect sizes without imposing any modeling assumptions, we calculated the Pearson correlation of the magnitude of observed GWAS effect size and of cis-eQTL effect size, $\widehat{\text{Cor}}(|\hat{\delta}_k|, |\hat{\gamma}_k|)$, for the list of selected fine-mapped QTLs. This was done for each tissue-trait pair separately. The observed Pearson correlation captures the mediated effect (see details in Section A.5.8.5). To obtain a null distribution for the correlation that accounts for the potential confounding effect of different local LD score values, we computed the Pearson correlation under the shuffled data within each LD-score bin defined by quantiles (100 bins were used). The significance of the difference between observed and null distribution was calculated using the correlated t-test method described in Section A.5.4.

### A.5.8.3 Modeling effect mediated by regulatory process

We compared the magnitude of GWAS and cis-QTL effect sizes, which is the basis of multi-SNP Mendelian randomization approaches [14].

To formalize the relationship between the GWAS effect size ($\delta$) and the QTL effect size ($\gamma$), we assumed an additive genetic model for the GWAS trait. Specifically, for variant $k$,

$$Y = \sum_k \delta_k \cdot X_k + \epsilon, \tag{A.7}$$

where $X_k$ is the allele count of variant $k$, $Y$ is the trait, and $\epsilon$ is the un-explained variation. We decomposed GWAS effect size into its mediated and un-mediated components,

$$\delta_k = \sum_{g \in \mathcal{G}_k} \beta_g \gamma_{k,g} + \nu_k, \tag{A.8}$$

where $\mathcal{G}_k$ represents the set of genes regulated by variant $k$ with corresponding QTL effect size as $\gamma_{k,g}$, and $\nu_k$ is the un-mediated effect of variant $k$ on trait. And $\beta_g$ is the downstream effect of gene $g$ on the trait.

### A.5.8.4 Transcriptome-wide estimation of mediated effects

To estimate the transcriptome-wide contribution of the mediated effects on complex traits, we proposed a mixed-effects model on the basis of Eq. A.8,

$$|\delta_k| = \beta_g \cdot (\text{sign}(\delta_k) \cdot \gamma_{k,g}) + b_0 + b_1 \cdot \sqrt{\text{LD-score}_k} + \epsilon \tag{A.9}$$

$$\beta_g \sim N(0, \sigma^2_{\text{gene}}) \tag{A.10}$$

$$\epsilon \sim N(0, \sigma^2), \tag{A.11}$$

where $b_0, b_1$ are the fixed effect capturing the un-mediated effect and $\beta_g$ is the mediated effect of the gene or splicing event $g$. In short, we assumed a random effects model to account for the heterogeneity of the $\beta$'s and aimed at estimating $\sigma^2_{\text{gene}}$ as the transcriptome-wide average of the mediated effect. For each tissue-trait pair, we fitted the model using selected fine-mapped QTLs, as described in Section A.5.8.1, along with the corresponding $\hat{\delta}_k$ (GWAS

effect for variant $k$), $\hat{\gamma}_{k,g}$ (QTL effect for variant $k$, gene $g$). To obtain the distribution of $\sigma^2_{\text{gene}}$ under the null, we performed the same calculation using shuffled GWAS effect sizes. The effect allele choice is arbitrary and we chose them so that all the GWAS effects are positive. This choice made the modeling of the effect of local LD more straightforward since we expect that variants in high LD regions may tag more causal variants and end up with a larger estimated GWAS effect, which would result in a positive $b_1$. The square root of LD-score represents better the potential effect of LD score on the effect size. Using absolute value of the GWAS effects ($|\delta_k|$) and $\text{sign}(\delta_k) \cdot \gamma_{k,g}$ in equation (A.9) is a convenient way to implement the recoding of the effect allele.

#### A.5.8.5 Robustness of the estimation of the mediating effect to LD contamination

We illustrate the intuition behind the LD-contamination correction when the average mediated effects are estimated using the approximate method (correlation of absolute values) or the mixed-effects approach.

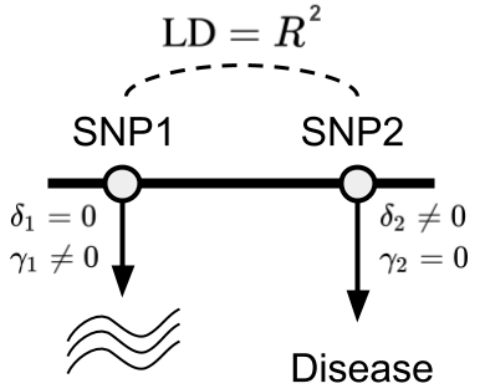



**Figure A.11: Schematic representation of LD contamination.** SNP1 has a causal effect on the expression level of a gene but not on the trait (Disease here), $\delta_1 = 0$ and $\gamma_1 \neq 0$. SNP2 has a causal effect on the trait but not on the expression of the gene, $\delta_2 \neq 0$ and $\gamma_2 = 0$.

Consider the LD-contamination scenario where SNP 1 and SNP 2 are in LD with correlation $R^2$ (suppose LD is fixed) and have a non-zero effect on gene expression and trait,

respectively (as shown in Figure A.11). The marginal effect estimates of SNP 1, *i.e.* $\hat{\delta}_1$ and $\hat{\gamma}_1$, are given by

$$\hat{\delta}_1 = R\delta_2 + \epsilon_{\text{GWAS}} \tag{A.12}$$

$$\hat{\gamma}_1 = \gamma_1 + \epsilon_{\text{QTL}}, \tag{A.13}$$

where Eq. A.12 holds because the marginal effect size depends on LD. To determine the covariance of the magnitude of the GWAS and QTL estimates for SNP 1, we consider $\text{E}(|\hat{\delta}_1||\hat{\gamma}_1|)$.

$$\begin{aligned}
\text{E}(\hat{\delta}_1\hat{\gamma}_1 \,|R) &= \text{E}((R\delta_2 + \epsilon_{\text{GWAS}}) \cdot (\gamma_1 + \epsilon_{\text{QTL}}) \,|R) && \text{(A.14)}\\
&= \text{E}(R\delta_2\gamma_1 \,|R) + \text{E}(\epsilon_{\text{GWAS}}\gamma_1) + \text{E}(R\delta_2\epsilon_{\text{QTL}} \,|R) + \text{E}(\epsilon_{\text{GWAS}}\epsilon_{\text{QTL}}) && \text{(A.15)}\\
&= R \cdot \text{E}(\delta_2\gamma_1), && \text{(A.16)}
\end{aligned}$$

where Eq. A.16 holds since the last three terms in the previous line are zeros, due to the independence among $\epsilon_{\text{GWAS}}$, $\epsilon_{\text{QTL}}$, and true effect sizes, $\delta$ and $\gamma$.

Hence, the covariance of the GWAS and QTL effect sizes under the LD contamination scenario is

$$\begin{aligned}
\text{Cov}(\hat{\delta}_1, \hat{\gamma}_1 \,|R) &= \text{E}(\hat{\delta}_1\hat{\gamma}_1 \,|R) - \text{E}(\hat{\delta}_1 \,|R) \cdot \text{E}(\hat{\gamma}_1 \,|R) && \text{(A.17)}\\
&= R \cdot \text{E}(\delta_2\gamma_1) - \text{E}(\hat{\delta}_1 \,|R) \cdot \text{E}(\hat{\gamma}_1 \,|R) && \text{(A.18)}\\
&= R \cdot \text{E}(\delta_2\gamma_1) - \text{E}(R\delta_2 + \epsilon_{\text{GWAS}}) \cdot \text{E}(\gamma_1 + \epsilon_{\text{QTL}}) && \text{(A.19)}\\
&= R \cdot \text{E}(\delta_2\gamma_1) - R \cdot \text{E}(\delta_2) \cdot \text{E}(\gamma_1) && \text{(A.20)}\\
&= R \cdot \text{Cov}(\delta_2, \gamma_1), && \text{(A.21)}
\end{aligned}$$

which implies that conditioning on LD, the observed correlation between $\hat{\delta}$ and $\hat{\gamma}$ should be very small.

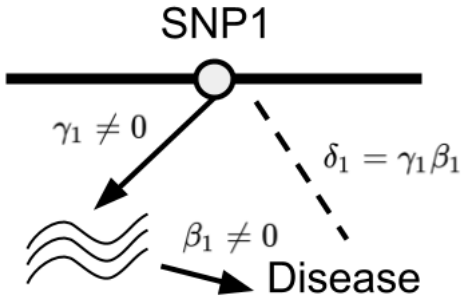



**Figure A.12: Diagram representation of mediation model.**

Similarly, we can derive the correlation between GWAS and QTL effect size estimates under the simple mediation model shown in Figure A.12, where we have

$$\hat{\delta}_1 = \beta_1\gamma_1 + \epsilon_{\text{GWAS}} \tag{A.22}$$

$$\hat{\gamma}_1 = \gamma_1 + \epsilon_{\text{QTL}}, \tag{A.23}$$

where Eq. A.22 follows by definition of the mediation model considering no direct effect. So,

$$\text{Cov}(\hat{\delta}_1, \hat{\gamma}_1 \,|\beta_1) = \beta_1\text{E}(\gamma_1^2) - \beta_1\text{E}(\gamma_1)^2 \tag{A.24}$$

$$= \beta_1\text{Var}(\gamma_1) \tag{A.25}$$

So, if we consider a gene locus, which naturally conditions on local LD and gene-level effect $\beta$, we can conclude that

$$\text{Cov}(\hat{\delta}_1, \hat{\gamma}_1 \,|\text{gene locus}) = \text{Cov}(\hat{\delta}_1, \hat{\gamma}_1 \,|\beta_1, R) \tag{A.26}$$

$$= \begin{cases} 0 & \text{LD contamination} \\ \text{Var}(\gamma_1) & \text{Mediation model} \end{cases} \tag{A.27}$$

### A.5.8.6 Concordance of mediated effects for allelic series of independent eQTLs

Under the mediation model in Eq. A.8, we expect that for a given gene with multiple QTL signals, these signals should share the same downstream effect, $\beta_g$. Since the number

of splicing events with multiple QTL signals was limited, we restricted this analysis to eQTLs only. We tested for concordance of downstream effect size obtained from the primary and secondary eQTL of a gene (ranked by QTL significance or QTL effect size estimate). Specifically, for a given trait and gene $g$, we defined the observed downstream effect for the $k$th variant as $\hat{\beta}_{k,g} = \hat{\delta}_k/\hat{\gamma}_{k,g}$. Thus, for each gene, we obtained $\hat{\beta}_{prim}$ and $\hat{\beta}_{sec}$ as the observed downstream effect for the primary and secondary eQTLs if more than one eQTL signal was detected by *dap-g*. Ideally, for a mediating gene in a causal tissue (or a good proxy tissue), we would expect that $\hat{\beta}_{prim}$ and $\hat{\beta}_{sec}$ should be similar. We measured the concordance in two ways: 1) correlation between $\hat{\beta}_{prim}$ and $\hat{\beta}_{sec}$; 2) percent concordant, defined as the fraction of eQTL pairs having the same sign in $\hat{\beta}_{prim}$ and $\hat{\beta}_{sec}$. The results of 1) were reported in [139].

To visualize the concordance of $\hat{\beta}_{prim}$ and $\hat{\beta}_{sec}$, we first scaled $\hat{\delta}$ and $\hat{\gamma}$ by their standard deviation among all eQTLs selected in Section A.5.8.1. Then, we extracted the set of genes with at least two *dap-g* eQTLs (defined in A.5.8.1) and labelled the top two eQTLs (rank by QTL effect size magnitude) as primary and secondary based on QTL significance or QTL effect size. We computed $\hat{\beta}_{prim}$ and $\hat{\beta}_{sec}$ and removed the genes with $\hat{\beta}_{prim}$ or $\hat{\beta}_{sec}$ in the top and bottom 5%. As a control, we also simulated random $\delta$ to compute simulated $\beta_{sim}$ for downstream analysis. We further filtered the genes by selecting only those with *enloc* rcp $> 0.1$.

### *A.5.9 Identifying patterns of regulation of expression across tissues*

We used FLASH Sparse Factor Analysis [150] to identify latent factors specific to different tissue clusters. We ran `flashr` on a set of top eQTLs (obtained from all GTEx individuals) per gene which had been tested in all 49 tissues (around 16,000 eQTLs in total were selected) and shown strong evidence of being active in at least one tissue. Then, for each selected variant-gene pair, the marginal effect size estimates were extracted for all 49 tissues regardless of whether it was significant in that tissue or not. The resulting estimated effect-size matrix

(of dimension $\sim 16,000 \times 49$) was the input to `flashr` (with normal prior on loading and uniform with positive support as prior on factor) to obtain the sparse factors. The `flashr` run yielded 31 FLASH factors (Figure A.29), which were used to assign the tissue-specificity of an eQTL.

We defined the eQTL cross-tissue patterns by projecting the estimated effect-size vector across 49 tissues onto the FLASH factors and computed the quality of the projection, PVE, as $\text{PVE}_k = \frac{\|\vec{\beta}_k\|_2^2}{\|\vec{\beta}\|_2^2}$. $\text{PVE}_k$ represented the quality score for using FLASH factor $k$ to explain the cross-tissue pattern of eQTL. The eQTL was assigned to a FLASH factor $k$ if $\text{PVE}_k$ was maximal among all FLASH factors and $\text{PVE}_k > 0.2$ and for those with $\text{PVE}_k \leq 0.2$ in all FLASH factors, `NA` (short for not assigned) was assigned instead. These "not assigned" eQTLs had more complex tissue-sharing pattern than the factors captured in the FLASH analysis. To obtain an interpretable tissue-specificity category, we labeled *Factor1* as the shared factor, *Factor2*, *Factor13*, *Factor14*, *Factor29*, and *Factor30* as brain-specific factors, and the rest of the factor assignment as *other factors*.

We applied the multivariate adaptive shrinkage implemented in *mashr* [140] to smooth cis-eQTL effect size estimates (obtained from all GTEx individuals) by taking advantage of correlation between tissues. To fit the *mashr* model, we used the set of $\sim 16,000$ cis-eQTLs as stated in Section A.5.9 to learn the *mashr* prior, and then fit the *mashr* model using $\sim 40,000$ randomly selected variant-gene pairs for the same set of eGenes. We learned data-driven *mashr* priors in three ways: 1) FLASH factors as described above; 2) PCA with number of PC = 3; 3) empirical covariance of observed z-scores. The data-driven covariances were further denoised by calling `cov_ed` in `mashr`. Furthermore, we included the set of canonical covariances as described in [140] as an additional *mashr* prior. We fit the *mashr* model using the set of randomly selected variant-gene pairs with the error correlation estimated by applying `estimate_null_correlation` function in `mashr` and the priors obtained above. The resulting *mashr* model was used to compute the posterior mean,

standard deviation, and local false sign rate (LFSR) for any variant-trait pair.

### *A.5.10 Causal gene prioritization*

Two classes of methods can be used to identify the target genes of GWAS loci. One class is based on the colocalization of GWAS and QTL loci, which seeks to determine whether the causal variant for the trait is the same as the causal variant for the molecular phenotype. The other class is based on the association between the genetically regulated component of gene expression (or splicing) with the trait. We applied representative examples of each class of methods.

#### A.5.10.1 Colocalization

For a given variant associated with multiple traits such as gene expression (eQTL) and complex disease (trait-associated variant), extensive LD makes it challenging to identify the underlying true causal mechanisms. Colocalization approaches attempt to address this problem. Here, we conducted colocalization analysis using two independent approaches: *coloc* [47] and *enloc* [153]), to estimate whether a gene's expression or a splicing event shares a causal variant with a trait.

#### A.5.10.2 enloc

We computed Bayesian regional colocalization probability (rcp) using *enloc*, to estimate the probability of a GWAS region and a gene's cis window sharing causal variants. We used the *dap-g* results described in A.5.7, which was based on EUR individuals only. We split the GWAS summary statistics into approximately LD-independent regions [12], each region defining a GWAS locus. For each tissue-trait combination, we computed the rcp of every overlapping GWAS locus to a gene's or splicing event's cis window with *enloc*'s default execution mode.

For each trait, we counted the number of GWAS loci that contain a GWAS significant hit, and among these, the number of loci that additionally contain a gene with *enloc* colocalization

$rcp > 0.5$. As shown in Figure A.16C, across traits, a median 29% of loci with a GWAS signal contain an *enloc* colocalized signal. Given *enloc*'s conservative nature, we caution that $rcp < 0.5$ does not mean that there is no causal relationship between the molecular phenotype and the complex trait; rather, it should be interpreted as lack of sufficient evidence with current data. We summarize the findings in Figure A.17. We observed a smaller proportion of GWAS loci containing a colocalized splicing event (median 11% across traits).

### A.5.10.3 coloc

We computed *coloc* on all cis-windows with at least one eVariant (cis-eQTL per-tissue q-value$< 0.05$) or sVariant. For each gene's cis-window, we used summary statistics from the GWAS traits and the main GTEx eQTL/sQTL analysis. For binary traits, case proportion and 'cc' trait type parameters were used. For continuous traits, sample size and 'quant' trait type parameters were used. In both cases, imputed or calculated z-scores were used as effect coefficients in Bayes factor calculations.

*Coloc* is very sensitive to the choice of priors. We used *enloc*'s enrichment estimates to define data-based priors in a consistent manner. First, we defined likely LD-independent blocks of variants using definitions provided previously [12]. The probability of eQTL signal, $\Pr(d_i = 1)$, was estimated using *dap-g* [151]. Subsequently, we calculated priors $p_1$, $p_2$, and $p_{12}$ for colocalization analyses as follows:

$$p_1 := \Pr(\gamma_i = 1, d_i = 0) = \frac{\exp(\alpha_0)}{1 + \exp(\alpha_0)} \times (1 - \Pr(d_i = 1)),$$
$$p_2 := \Pr(\gamma_i = 0, d_i = 1) = \frac{1}{1 + \exp(\alpha_0 + \alpha_1)} \times \Pr(d_i = 1), \text{ and}$$
$$p_{12} := \Pr(\gamma_i = 1, d_i = 1) = \frac{\exp(\alpha_0 + \alpha_1)}{1 + \exp(\alpha_0 + \alpha_1)} \times \Pr(d_i = 1),$$

where $\alpha_0$ and $\alpha_1$ indicate intercept effect estimate and log odds ratio estimate for the enrichment using *enloc*, respectively.

We ran *coloc* using variants in the cis-window for each gene and the intersection with each GWAS trait, obtaining five probabilities for each gene-tissue-trait tuple: **P0** for the

probability of neither expression nor GWAS having a causal variant; **P1** for the probability of only expression having a causal variant; **P2** for only the GWAS having a causal variant; **P3** for the GWAS and expression traits to have distinct causal variants; **P4** for the GWAS and expression traits to have a shared causal variant. We repeated this process using sQTL results.

### *A.5.11 Fine-mapping of height GWAS using summary statistics*

To investigate the robustness of fine-mapping, we fine-mapped "height" from the GIANT GWAS meta-analysis and "standing height" from the UK Biobank using `susieR` [149]. We performed fine-mapping using `susie_bhat` within each LD block [12]. We used GWAS effect sizes $\tilde{\beta}$ imputed from z-scores by $\tilde{\beta} = z/\sqrt{Nf(1-f)}$ and $\text{se}(\tilde{\beta}) = \tilde{\beta}/z$, where $f$ is allele frequency and $N$ is GWAS sample size. The GTEx-EUR individuals were used to calculate the reference LD panel. We recorded 95% credible set which has posterior probability 95% to capture a causal signal. To compare the fine-mapping results of two GWASs, we defined their 95% credible sets as "overlapped" if they shared at least one variant. To see how 95% in GIANT GWAS is colocalized with UK Biobank GWAS, we calculated colocalization probability as $\sum_{i:\text{variant}_i \in 95\%\text{CS of GIANT}} \text{PIP}_{i,\text{GIANT}} \times \text{PIP}_{i,\text{UKB}}$.

### *A.5.12 Association to predicted expression or splicing*

#### A.5.12.1 Predicting the genetically regulated components of expression and splicing

To predict expression, we constructed linear prediction models [8], using only individuals of European ancestry, and variants with MAF $> 0.01$, for genes annotated as protein-coding, pseudo-gene, or lncRNA. For each gene-tissue pair, we selected the variants with highest *pip* in their cluster, and kept those achieving pip $> 0.01$ in *dap-g* [151]. We used *mashr* [140] effect sizes (as computed in A.5.9) for each selected variant. For each model, we computed the covariance matrix between variants using only individuals of European ancestries, with sample sizes ranging from 65 (kidney - cortex) to 602 (skeletal muscle). This allowed us to

build LD panels for every tissue. For every gene, we also computed the covariance of all the variants present across the different tissue models, compiling a cross-tissue LD panel to compute the correlation between predicted expression levels across tissues. We refer to these models as fine-mapped-*mashr* models. We compared the number of *mashr* models to the number of Elastic Net models from GTEx version 7 (Figure A.13). We generated analogous prediction models for splicing ratios, as computed by Leafcutter [79], applying the same model-building methodology to the data from the sQTL analysis.

Expression phenotypes were adjusted for unwanted variation using the following covariates: sex, sequencing plaform, the top 3 principal components from genotype data, and PEER factors. The number of PEER factors was determined from the sample size: 15 for $n < 150$, 30 for $150 \leq n < 250$, 45 for $250 \leq n < 350$, 60 for $350 \leq n$. We obtained 686,241 models for different (gene, tissue) pairs.

We also generated analogous prediction models for splicing ratios, with the same model-building methodology applied to the data from the sQTL analysis, obtaining 1,816,703 (splicing event, tissue) pairs.

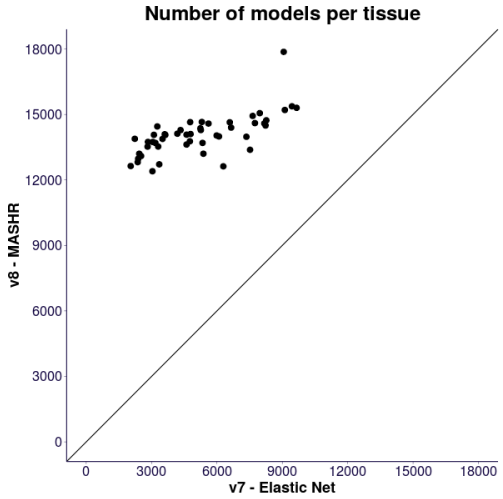



**Figure A.13: Number of models available in v8 fine-mapped-*mashr* family of models, compared to v7 Elastic Net family.** The point with 17,867 models is Testis, consistently with the high levels of expression observed in the eQTL analysis [139].

### A.5.12.2 PrediXcan

We performed PrediXcan analysis [7] on the 87 complex traits, using the GWAS summary statistics described in A.5.3.2, to identify trait-associated genes (typically $p < 2.5 \times 10^{-7}$). We used the 49 models and LD panels described in A.5.12.1, separately on each trait, to obtain 59,485,548 gene-tissue-trait tuples. Repeating this process to generate splicing event ratio models, we obtained 154,891,730 splicing event-tissue-trait tuples; for each trait, the Bonferroni-significance threshold was $p < 9.5 \times 10^{-8}$.

### A.5.12.3 Colocalized and significantly associated genes

We assessed how many genes present evidence of trait association and colocalization, using both expression and splicing event. First, we counted the proportion of genes that showed a colocalized expression signal with any trait in any tissue, and observed 15% such genes at rcp$>$ 0.5. Then, for each gene, we considered the splicing event with highest colocalization value in any trait or tissue, and found evidence for 5% at rcp$>$ 0.5.

Then we repeated this process for PrediXcan associations at different signifcance thresholds. About 30% of genes showed a significant PrediXcan association to any trait, and only 8% when filtered for associations with rcp$>$ 0.5. When using the highest splicing association and colocalization value for a gene, these proportions were 20% and 3%, respectively.

These proportions gauge our power to predict causal genes affecting complex traits on the GTEx resource, with expression yielding more findings than splicing.

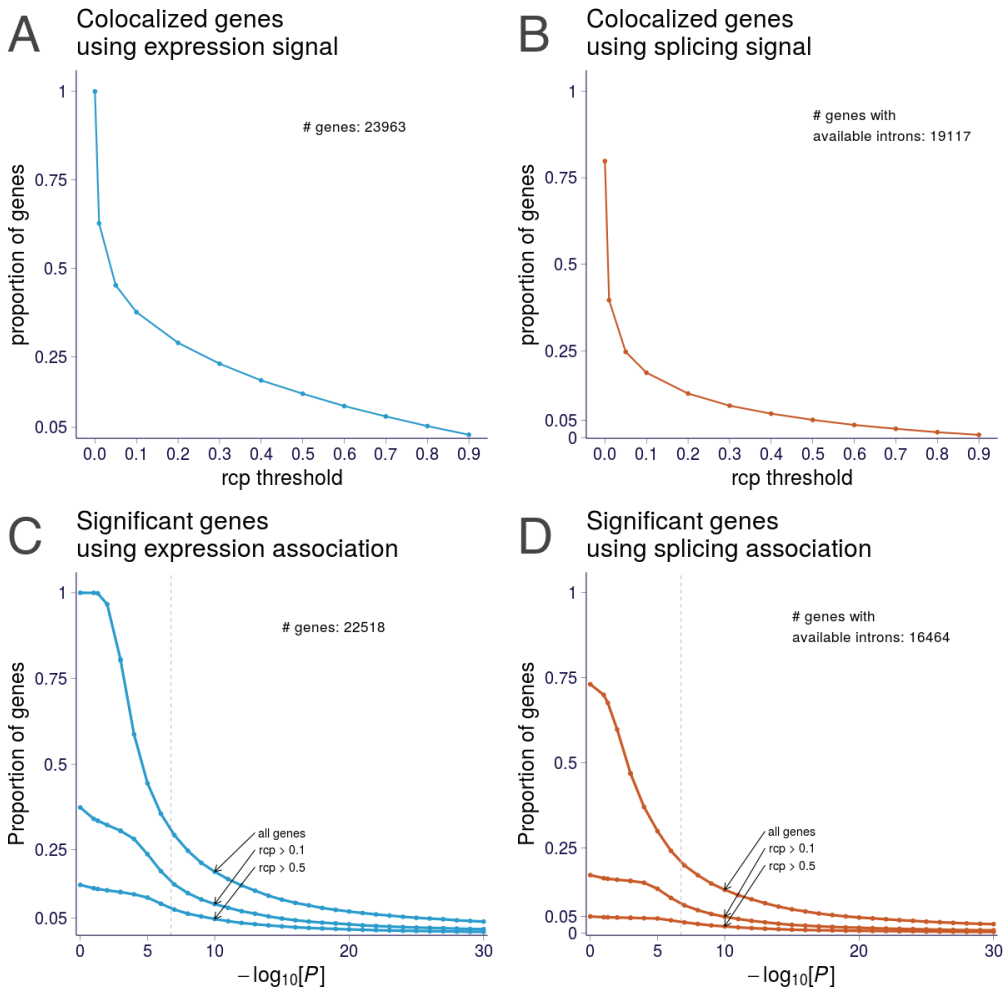



**Figure A.14: Proportion of genes with a colocalized or associated signal using expression or splicing event.**
**A** shows the proportion of genes with colocalization evidence in expression data, for different rcp thresholds. 3,477 genes show evidence at rcp$>0.5$ (15% out of 23,963 genes with *enloc* results). **B** shows the proportion of genes with colocalization evidence in splicing data; 1,277 genes (5% of all 23,963) show evidence at rcp$>0.5$.
**C** shows the proportion of genes with association evidence in expression data, additionally filtered by colocalization on different thresholds. About 30% of genes show associations at the bonferroni threshold ($p < 0.05/686,241$), while 8% also show colocalization evidence.
**D** shows the proportion with association and colocalization evidence in splicing data; about 20% show association evidence ($p < 0.05/1,816,703$) and 3% are also colocalized.

### A.5.12.4 S-MultiXcan

Given the substantial sharing of eQTLs across tissues [138], we aggregated PrediXcan results across tissues using S-MultiXcan [9]. MultiXcan has been shown to exploit the tissue sharing

of regulatory variation, to improve our ability to identify trait-associated genes. The method extends the single-tissue PrediXcan approach, leveraging GWAS summary statistics and taking into account the correlation between tissues. We obtained association statistics for 1,958,220 gene-trait pairs and 11,986,329 splicing event-trait pairs.

#### A.5.12.5 PrediXcan replication in BioVU

We replicated the significant gene-level associations for a prioritized list of traits (table A.15) using BioVU [33], Vanderbilt University's DNA Biobank tied to a large-scale Electronic Health Records (EHR) database. We sought BioVU replication in the exact discovery tissues for the significant gene-trait associations. We restricted our analysis to subjects of European ancestries, using principal component analysis as implemented in EIGENSOFT (version 7.1.2; [118]). First, we estimated the genetically determined component of gene expression in the BioVU individuals using the PrediXcan imputation models. We then conducted association analysis for the prioritized traits using logistic regression, with sex and age as covariates.

Among replicated loci are *SORT1* (liver, coronary artery disease rcp = 0.952; dicovery p = $2.041 \times 10^{-19}$ BioVU p = $3.475 \times 10^{-4}$), which has a well-established associations to lipid metabolism and cardiovascular traits [105]. Chromosome 6p24 region, which contains *PHACTR1*, has been previously associated with a constellation of vascular diseases, including coronary artery disease [137] and migraine headache [3]. Notably, *PHACTR1* was significant in three different arteries (aorta artery, coronary artery and tibial artery) in two traits (coronary artery disease and migraine) in the replication analysis. In all six tissue-trait pairs, *PHACTR1* showed very high posterior probabilities in discovery analyses (rcp = 0.992 to 1.00). In our replication analysis, *PHACTR1* remained significant only for coronary artery disease associations (table A.15, aorta artery, discovery p = $2.246 \times 10^{-39}$, BioVU p = $7.484 \times 10^{-8}$; coronary artery, discovery p = $1.952 \times 10^{-37}$, BioVU p = $2.047 \times 10^{-7}$; tibial artery, discovery p = $1.559 \times 10^{-33}$, BioVU p = $9.880 \times 10^{-7}$).

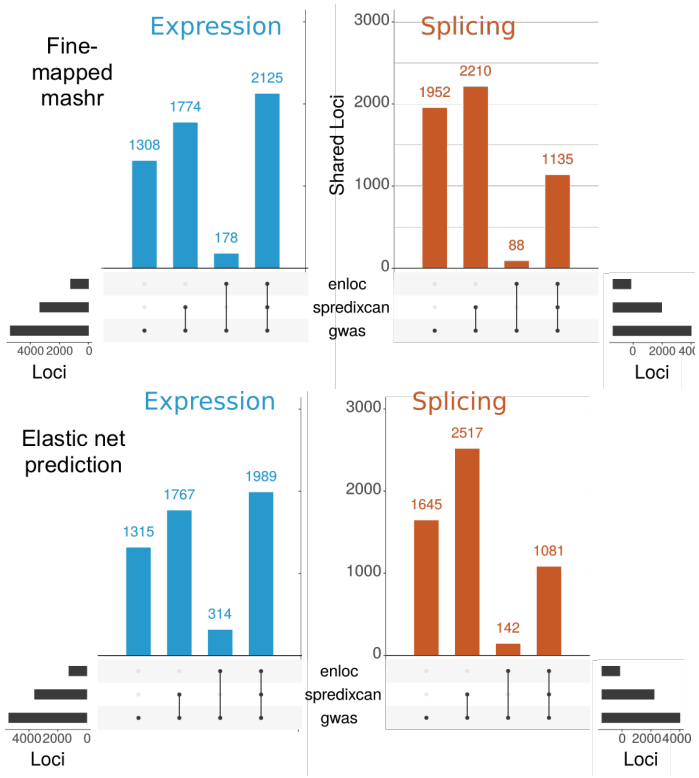



**Figure A.15: Causal gene prioritization using PrediXcan and *enloc*.**
**(A)** Fine-mapped-*mashr* model predictions: summary of GWAS loci that also contain an associated PrediXcan or colocalized signal, for expression (left) and splicing (right).
Significance was defined at Bonferroni-adjusted threshold for number of tests in each trait: $p < 0.05/(\text{gene-tissue pairs}) = 7.28\times10^{-8}$ for expression, $p < 0.05/(\text{intro-tissue pairs}) = 2.75\times10^{-8}$ for splicing. Colocalization status was defined as *enloc* rcp$> 0.5$.
**(B)** Elastic net predictions: summary of GWAS loci that also contain an associated PrediXcan or *enloc* signal, for expression (left) and splicing (right), using Elastic Net models.
Significance was defined at Bonferroni-adjusted threshold for number of tests in each trait: $p < 0.05/(\text{gene-tissue pairs}) = 1.77\times10^{-7}$ for expression, $p < 0.05/(\text{intro-tissue pairs}) = 9.51\times10^{-8}$ for splicing. Colocalization status was defined as *enloc* rcp$> 0.5$.

The number of loci with potential target genes according to both S-PrediXcan and enloc went up from 1989 with Elastic Net models to 2125 with the improved fine-mapped-*mashr* models. The number of S-PrediXcan backed loci increased only 7, so that the added loci were mostly due to an increased overlap with enloc.

| name | european samples | abbreviation | expression models | splicing models |
|---|---|---|---|---|
| Adipose - Subcutaneous | 491 | ADPSBQ | 14732 | 42912 |
| Adipose - Visceral (Omentum) | 401 | ADPVSC | 14640 | 41720 |
| Adrenal Gland | 200 | ADRNLG | 13622 | 36754 |
| Artery - Aorta | 338 | ARTAORT | 14396 | 40474 |
| Artery - Coronary | 180 | ARTCRN | 13878 | 40579 |
| Artery - Tibial | 489 | ARTTBL | 14493 | 40690 |
| Brain - Amygdala | 119 | BRNAMY | 12814 | 24236 |
| Brain - Anterior cingulate cortex (BA24) | 135 | BRNACC | 13528 | 28806 |
| Brain - Caudate (basal ganglia) | 172 | BRNCDT | 14118 | 32127 |
| Brain - Cerebellar Hemisphere | 157 | BRNCHB | 13771 | 39862 |
| Brain - Cerebellum | 188 | BRNCHA | 13992 | 40747 |
| Brain - Cortex | 184 | BRNCTXA | 14284 | 35086 |
| Brain - Frontal Cortex (BA9) | 158 | BRNCTXB | 14091 | 32031 |
| Brain - Hippocampus | 150 | BRNHPP | 13526 | 27437 |
| Brain - Hypothalamus | 157 | BRNHPT | 13741 | 30326 |
| Brain - Nucleus accumbens (basal ganglia) | 181 | BRNNCC | 14062 | 32670 |
| Brain - Putamen (basal ganglia) | 153 | BRNPTM | 13694 | 28461 |
| Brain - Spinal cord (cervical c-1) | 115 | BRNSPC | 13096 | 28883 |
| Brain - Substantia nigra | 101 | BRNSNG | 12637 | 23677 |
| Breast - Mammary Tissue | 337 | BREAST | 14654 | 44613 |
| Cells - Cultured fibroblasts | 417 | FIBRBLS | 13976 | 36809 |
| Cells - EBV-transformed lymphocytes | 116 | LCL | 12398 | 37627 |
| Colon - Sigmoid | 274 | CLNSGM | 14363 | 41581 |
| Colon - Transverse | 306 | CLNTRN | 14582 | 41215 |
| Esophagus - Gastroesophageal Junction | 281 | ESPGEJ | 14285 | 41004 |
| Esophagus - Mucosa | 423 | ESPMCS | 14589 | 37186 |
| Esophagus - Muscularis | 399 | ESPMSL | 14603 | 40376 |
| Heart - Atrial Appendage | 322 | HRTAA | 14035 | 36322 |
| Heart - Left Ventricle | 334 | HRTLV | 13200 | 29470 |
| Kidney - Cortex | 65 | KDNCTX | 11164 | 24571 |
| Liver | 183 | LIVER | 12714 | 27011 |
| Lung | 444 | LUNG | 15058 | 44346 |
| Minor Salivary Gland | 119 | SLVRYG | 13884 | 38380 |
| Muscle - Skeletal | 602 | MSCLSK | 13381 | 31855 |
| Nerve - Tibial | 449 | NERVET | 15373 | 45478 |
| Ovary | 140 | OVARY | 13738 | 40857 |
| Pancreas | 253 | PNCREAS | 13695 | 31203 |
| Pituitary | 219 | PTTARY | 14647 | 42343 |
| Prostate | 186 | PRSTTE | 14450 | 41991 |
| Skin - Not Sun Exposed (Suprapubic) | 440 | SKINNS | 14932 | 42005 |
| Skin - Sun Exposed (Lower leg) | 517 | SKINS | 15204 | 42219 |
| Small Intestine - Terminal Ileum | 144 | SNTTRM | 14065 | 39864 |
| Spleen | 186 | SPLEEN | 14073 | 40290 |
| Stomach | 269 | STMACH | 14102 | 36624 |
| Testis | 277 | TESTIS | 17867 | 67784 |
| Thyroid | 494 | THYROID | 15303 | 45217 |
| Uterus | 108 | UTERUS | 13199 | 39485 |
| Vagina | 122 | VAGINA | 12969 | 36931 |
| Whole Blood | 573 | WHLBLD | 12623 | 24568 |
| total | | | 686241 | 1816703 |

**Table A.2: Expression and splicing prediction models using fine-mapped-*mashr* models.** Training sample size and number of genes predicted for expression and splicing traits.

| | | |
|---|---|---|
| GWAS-significant (loci, trait) associations | | 5385 |
| GWAS-significant unique loci | | 1167 |
| enloc (loci, trait) colocalizations | expression | 2303 |
| enloc (loci, trait) colocalizations | splicing | 1223 |
| PrediXcan (loci, trait) associations | expression | 3899 |
| PrediXcan (loci, trait) associations | splicing | 3345 |
| PrediXcan & enloc (loci, trait) detections | expression | 2125 |
| PrediXcan & enloc (loci, trait) detections | splicing | 1135 |

**Table A.3: GWAS loci with colocalized or significant genes assigned.** Numbers of loci-trait associations with associated/colocalized genes/splicing event detected by each method. A locus is said to have a GWAS association to a trait if it contains at least one variant with $p < 0.05/(\text{variants tested}) = 5.7 \times 10^{-9}$. We list here how many such loci-trait associations have an S-PrediXcan association or *enloc* signal. Significant S-PrediXcan associations were defined at Bonferroni-adjusted threshold for number of tests in each trait: $p < 0.05/(\text{gene-tissue pairs}) = 7.28 \times 10^{-8}$ for expression, $p < 0.05/(\text{intro-tissue pairs}) = 2.75 \times 10^{-8}$ for splicing. Colocalization status was defined as *enloc* rcp$> 0.5$.

#### A.5.12.6 Summary-data-based Mendelian Randomization (SMR) and HEIDI

For comparison, we also performed top-eQTL based Summary-data-based Mendelian Randomization (SMR) [168] analysis of the 4,263 tissue-trait pairs. SMR, which integrates summary statistics from GWAS and eQTL data, has been used to prioritize genes underlying GWAS associations.

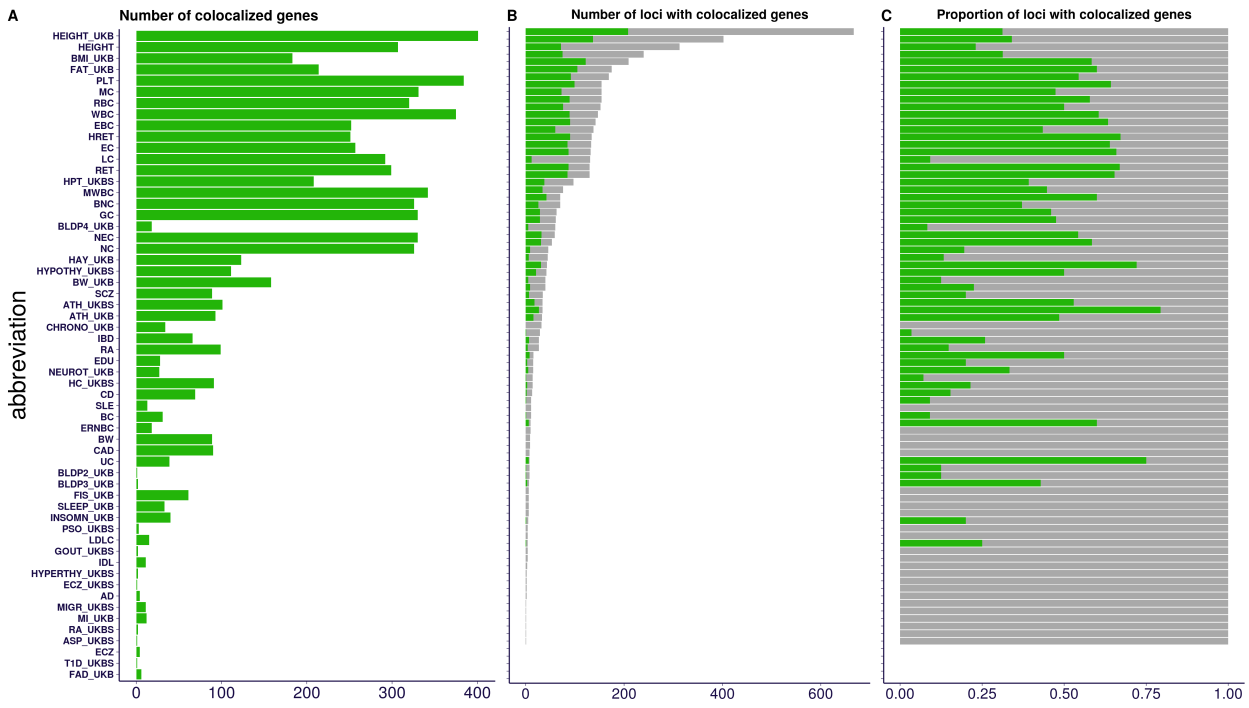



**Figure A.16: Colocalization of expression QTLs Colocalization for each of the 87 GWAS traits aggregated across the 49 tissues.** GWAS loci are shown in gray, colocalized results are shown in dark green. The traits are ordered by number of GWAS-significant variants.
**Panel A** shows the number of colocalized genes, achieving *enloc rcp* $> 0.5$ in at least one tissue, for each GWAS trait. The number of colocalized results tends to increase with the number of GWAS-significant variants.
**Panel B** shows the number of loci (approximately independent LD regions from [12]) with at least one GWAS-significant variant (dark gray), and among them those with at least one gene reaching $rcp > 0.5$ (dark green).
**Panel C** shows the proportion of loci with at least one GWAS-significant hit that contain at least one colocalized gene. Across traits, a median of 21% of the GWAS loci contain colocalized results. See trait abbreviation list in table A.1. These results were also presented in [139] and are shown here for completeness.


**Figure A.17: Colocalization of splicing QTLs for each of the 87 GWAS traits aggregated across the 49 tissues.** The traits are ordered by number of GWAS-significant variants. GWAS loci are shown in gray, colocalized results are shown in dark green.

**Panel A** shows the number of colocalized splicing event, achieving *enloc* $rcp > 0.5$ in at least one tissue, for each GWAS trait. As with gene expression results, the number of colocalized results tends to increase with the number of GWAS-significant variants.

**Panel B** shows the number of loci (approximately independent LD regions from [12]) with at least one GWAS-significant variant (dark gray), and among them those with one splicing event achieving $rcp > 0.5$ (dark green).

**Panel C** shows the proportion of loci with at least one GWAS-significant hit loci with at least one colocalized splicing event. Across traits, a median of 11% of the GWAS loci contain a colocalized result, lower than the gene expression counterpart (29%), indicating a decreased power in the sQTL study. See trait abbreviation list in table A.1.

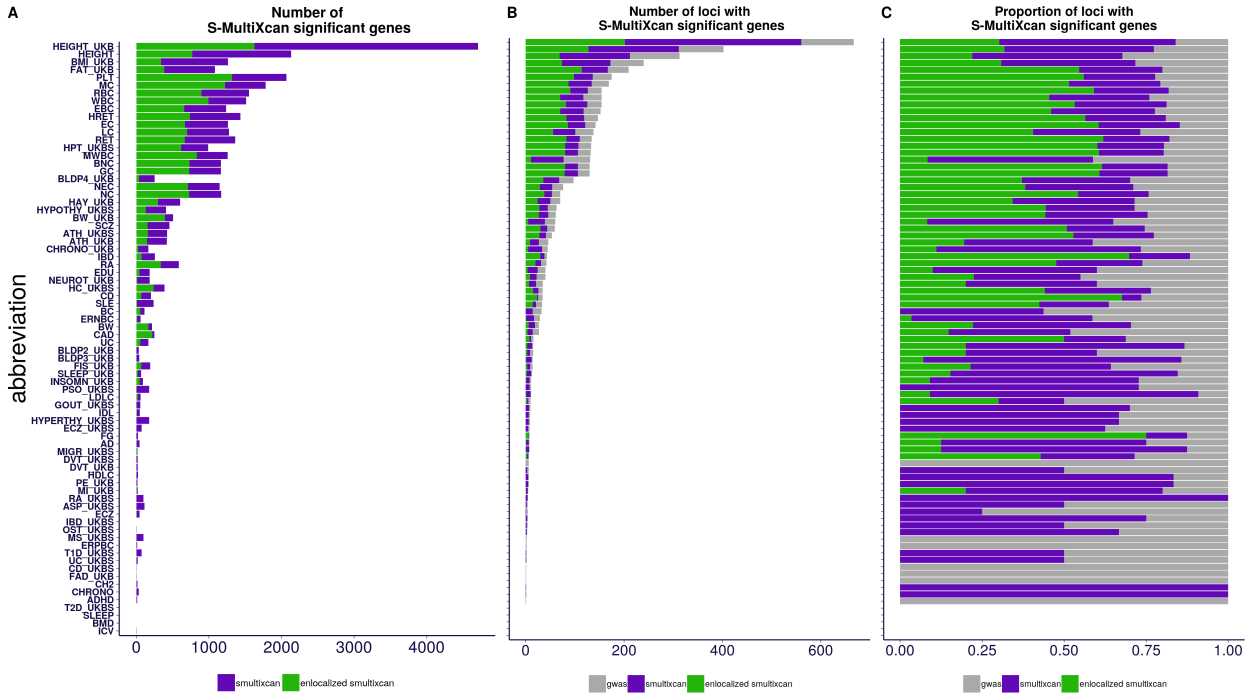



**Figure A.18: PrediXcan expression associations aggregated across tissues.** This figure summarizes S-MultiXcan associations for each of the 87 traits using the gene expression models. The traits are ordered by number of GWAS-significant variants.

**Panel A)** shows in purple the number of S-MultiXcan significant genes, and in dark green the subset also achieving *enloc rcp* $> 0.5$ in any tissue. S-MultiXcan has a high power for detecting associations, but 12% (median across traits) of these genes show evidence of colocalization.

**Panel B)** shows the number of loci (approximately independent LD regions [12]) with a significant GWAS association (gray), a significant S-MultiXcan association (purple), and a significant S-MultiXcan association that is colocalized (dark green). Anthropometric and Blood traits tend to present the largest number of associated loci, with Height from two independent studies leading the number of associations.

**Panel C)** shows the proportion of loci with significant GWAS associations (gray) that contain S-Multixcan (purple) and colocalized S-MultiXcan associations (dark green). Across traits, a median of 70% of GWAS-associated loci show a S-MultiXcan detection, while 19% show a colocalized S-MultiXcan detection.

See trait abbreviation list in table A.1.

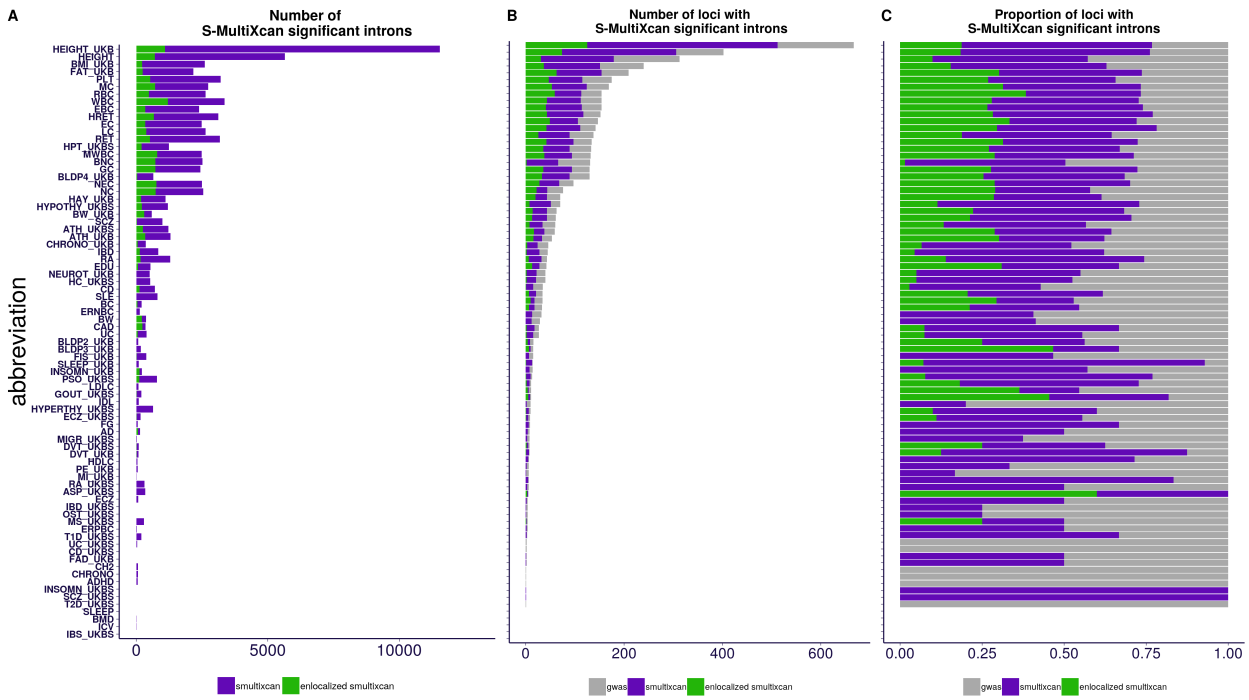



**Figure A.19: PrediXcan splicing associations aggregated across tissues.** This figure summarizes S-MultiXcan associations for each of the 87 traits using splicing models. The traits are ordered by number of GWAS-significant variants.
Panel A) shows in purple the number of S-MultiXcan significant splicing events, and in dark green the subset also achieving *enloc rcp* $> 0.5$ in any tissue. The proportion of colocalized, significantly associated splicing events is typically 2%, much lower than the proportion from gene expression (12%).
Panel B) shows the number of loci (approximately independent LD regions [12]) with a significant GWAS association (gray), a significant S-MultiXcan association (purple), and a significant S-MultiXcan association that is colocalized (dark green). As in the case of expression models, Anthropometric and Blood traits tend to present the largest number of associated loci.
Panel C) shows the proportion of loci with significant GWAS associations (gray) that contain S-Multixcan (purple) and colocalized S-MultiXcan associations (dark green). Across traits, a median of 63% of GWAS-associated loci show an S-MultiXcan association, while 11% show a colocalized S-MultiXcan association. These proportions are lower than the corresponding ones for expression (70% and 19% respectively).
See trait abbreviation list in table A.1.

### *A.5.13 Assessing the performance of association and colocalization methods to identify causal genes*

To assess the performance of colocalization and association methods to identify causal genes, we curated two sets of 'causal' gene-trait pairs. One set is based on the OMIM database and the other one is based on rare variant association results from exome-wide association

studies. To quantify the performance, we framed the causal gene identification problem as one of classification and used the standard tools such as ROC and precision recall curves, which have the advantage of not needing ad-hoc thresholds and show the full trade-off between true positives and false positives as well as precision vs. power. Throughout this section, we limited our scope to only the protein-coding genes.

### A.5.13.1 OMIM-based curation of causal genes

To obtain a curated set of trait-gene pairs from the OMIM database [52], we mapped our GWAS traits to the OMIM traits and linked them to the corresponding genes in the OMIM database. The mapping process is illustrated in Figure A.20 for a specific example GWAS trait, fasting glucose by the MAGIC consortium. First, the GWAS trait was mapped to the GWAS catalog trait names by searching for relevant keywords (defined manually A.4) in the description field of the GWAS catalog. Second, the GWAS catalog trait names were linked to phecodes using the mapping in the phewas catalog [33]. Third, we mapped phecodes to OMIM traits ids (MIM) as described in [10]. Finally, in step 4, we mapped OMIM traits to OMIM genes using the OMIM gene to phenotype map (`genemap2.txt`) in the OMIM database.

The keywords used for the each of the initial selected set of 114 GWAS traits is listed in tab A.4). For a subset of datasets with GWAS results from more than one source (public GWAS vs UKB) in our collection, we kept the dataset with higher number of GWAS loci to avoid double counting. The number of GWAS loci was determined based on counting the lead variants, using the PLINK V1.9 command `--clump-r2 0.2 --clump-p1 5e-8` at genome-wide significance $5 \times 10^{-8}$) for each trait. Furthermore, for this analysis, we excluded GWAS traits with fewer than 50 GWAS loci. The full list of OMIM based trait-gene pairs is listed in A.9.

With this procedure, we curated a list of 1,592 gene-trait pairs with evidence of causal associations in the OMIM database (hereafter, **OMIM genes**), which was downloaded from

omim.org/downloads (accessed on Aug 12th 2019). After matching traits, we retained 29 unique traits and 631 unique genes that were within the same LD block [12] as the GWAS hit (table A.9).

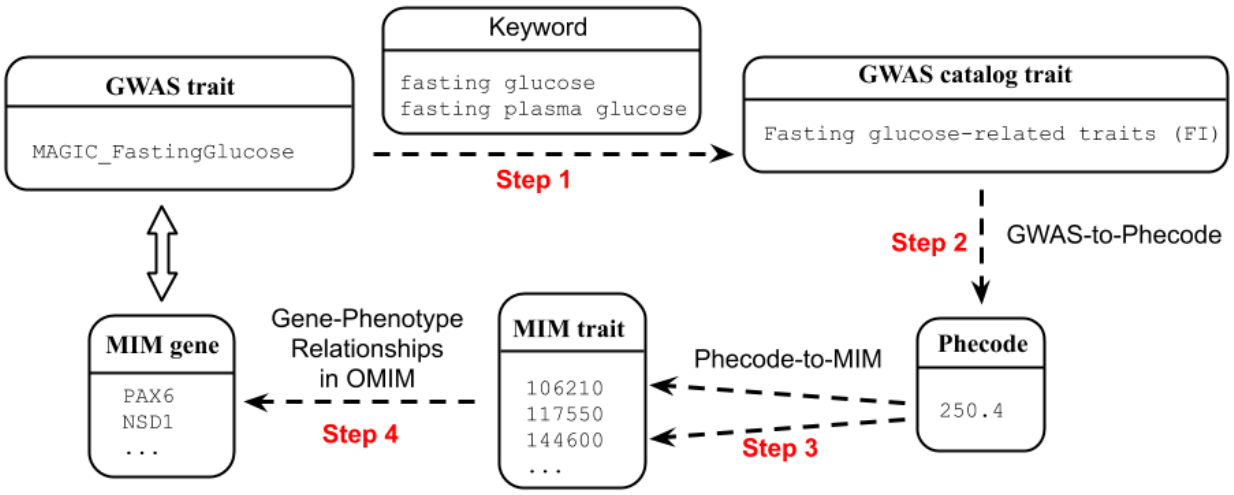



**Figure A.20: Workflow of OMIM-based curation of causal genes.** The workflow of OMIM-based causal gene curation is shown where each box represents the trait description/identifier in different databases. The steps to obtain OMIM genes for MAGIC_FastingGlucose, one of our GWAS traits, is shown as a concrete application of the workflow.

| Abbreviation | Keyword | Abbreviation | Keyword |
|---|---|---|---|
| Sleep_Duration_UKB | sleep duration | Sum_Eosinophil_Basophil_Ct | |
| Chronotype_UKB | chronotype | Sum_Neutrophil_Eosinophil_Ct | |
| Insomnia_UKB | insomnia | White_Blood_Cell_Count | white blood cell count |
| Fathers_Age_At_Death_UKB | aging | Coronary_Artery_Disease | coronary heart disease |
| Deep_Venous_Thrombosis_UKB | venous thromboembolism | Chronic_Kidney_Disease | chronic kidney |
| Asthma_UKB | asthma | Insomnia_In_Both_Sexes | insomnia |
| Irritable_Bowel_Syndrome_UKB | irritable bowel | Type_2_Diabetes | type 2 diabetes |
| Type_1_Diabetes_UKB | type 1 diabetes | Eczema | atopic dermatitis |
| Type_2_Diabetes_UKB | type 2 diabetes | Birth_Length | |
| Hyperthyroidism_UKB | hyperthyroidism | BMI_Childhood | bmi;body mass index |
| Hypothyroidism_UKB | hypothyroidism | Birth_Weight | |
| Psychological_Problem_UKB | psychiatric;psychological | Pubertal_Height_Female | |
| Multiple_Sclerosis_UKB | multiple sclerosis | Pubertal_Height_Male | |
| Parkinsons_Disease_UKB | Parkinson's | Intracraneal_Volume | intracranial volumn |
| Migraine_UKB | migraine | Asthma | asthma |
| Schizophrenia_UKB | Schizophrenia | Bone_Mineral_Density | bone mineral density |
| Osteoporosis_UKB | osteoporosis | BMI_Active_Inds | bmi;body mass index |
| Ankylosing_Spondylitis_UKB | ankylosing spondylitis | BMI_EUR | bmi;body mass index |
| Eczema_UKB | eczema;dermatitis | Height | height |
| Psoriasis_UKB | psoriasis | Hip_Circumference_EUR | hip circumference |
| Inflammatory_Bowel_Disease_UKB | inflammatory bowel disease | Waist_Circumference_EUR | waist circumference |
| Crohns_Disease_UKB | crohn's disease | Waist-to-Hip_Ratio_EUR | waist-to-hip |
| Ulcerative_Colitis_UKB | ulcerative colitis | HDL_Cholesterol | hdl cholesterol |
| Rheumatoid_Arthritis_UKB | rheumatoid arthritis | LDL_Cholesterol | ldl cholesterol |
| Gout_UKB | gout | Triglycerides | triglycerides |
| High_Cholesterol_UKB | total cholesterol | Neuroticism | neuroticism |
| Insomnia_UKB | insomnia | Heart_Rate | heart rate |
| Fluid_Intelligence_Score_UKB | intelligence | Crohns_Disease | crohn's disease |
| Birth_Weight_UKB | birth weight | Inflammatory_Bowel_Disease | inflammatory bowel disease |
| Neuroticism_UKB | neuroticism | Ulcerative_Colitis | ulcerative colitis |
| BMI_UKB | bmi;body mass index | Alzheimers_Disease | alzheimer |
| Body_Fat_Percentage_UKB | body fat | Epilepsy | epilepsy |
| Balding_Pattern_2_UKB | | Celiac_Disease | celiac disease |
| Balding_Pattern_3_UKB | | Multiple_Sclerosis | multiple sclerosis |
| Balding_Pattern_4_UKB | | Systemic_Lupus_Erythematosus | systemic lupus erythematosus |
| Mothers_Age_At_Death_UKB | aging | Stroke | stroke |
| Standing_Height_UKB | height | Chronotype | chronotype |
| Heart_Attack_UKB | | Sleep_Duration | sleep duration |
| | | Fasting_Glucose | fasting glucose; fasting plasma glucose |
| Pulmonary_Embolism_UKB | | Fasting_Insulin | fasting insulin |
| Asthma_UKB | asthma | CH2DB_NMR | |
| Hayfever_UKB | | HDL_Cholesterol_NMR | hdl cholesterol |
| Epilepsy_UKB | epilepsy | Triglycerides_NMR | triglycerides |
| Migraine_UKB | migraine | LDL_Cholesterol_NMR | ldl cholesterol |
| Hypertension_UKB | hypertension | Attention_Deficit_Hyperactivity_Disorder | attention deficit hyperactivity disorder |
| Adiponectin | adiponectin | Autism_Spectrum_Disorder | autism |
| Eosinophil_Count | eosinophil count | Schizophrenia | schizophrenia |
| Granulocyte_Count | | Rheumatoid_Arthritis | rheumatoid arthritis |
| High_Light_Scatter_Reticulocyte_Count | | Depressive_Symptoms | depression |
| Lymphocyte_Count | lymphocyte | Education_Years | education |
| Monocyte_Count | monocyte count;monocytes | Asthma_TAGC_EUR | asthma |
| Myeloid_White_Cell_Count | | Systolic_Blood_Pressure | systolic blood pressure |
| Neutrophil_Count | neutrophil count;neutrophils | Diastolic_Blood_Pressure | diastolic blood pressure |
| Platelet_Count | platelet counts | ER-negative_Breast_Cancer | breast cancer |
| Red_Blood_Cell_Count | red blood cell count | ER-positive_Breast_Cancer | breast cancer |
| Reticulocyte_Count | | Breast_Cancer | breast cancer |
| Sum_Basophil_Neutrophil_Ct | | Smoker | smoking behavior |

**Table A.4: Keywords of GWAS traits used for mapping with the GWAS catalog.** Keywords of all 114 GWAS traits used for OMIM-based curation and analyses are listed.

### A.5.13.2 Rare variant association-based curation of causal genes

In addition to the OMIM-based curation, we collected a set of genes in which rare protein-coding variants were reported to be significantly associated with our list of complex traits. Given the power of existing rare variant association studies, we focused on height and lipid

traits (low-density lipid cholesterol, high-density lipid cholesterol, triglycerides, and total cholesterol levels) [96, 85, 89].

We collected significant coding/splicing variants reported previously [96] and kept variants with effect allele frequency $< 0.01$ (Table S6 in [96]: ExomeChip variants with $P_{\text{discovery}} < 2 \times 10^{-7}$ in the European-ancestry meta-analysis (N=381,625)). Similarly, we collected significant variants reported by [85] (table S12 therein: Association Results for 444 independently associated variants with lipid traits) and filtered out variants with minor allele frequency $< 0.01$. For the whole-exome sequencing study conducted in Finnish isolates [89], we extracted significant genes identified by a gene-based test using protein truncating variants (Table S9 in [89]: Gene-based associations from aggregate testing with EMMAX SKAT-O with $P < 3.88 \times 10^{-6}$) and significant variants (Table S7 in [89]: A review of all variants that pass unconditional threshold of $P < 5 \times 10^{-7}$ for at least one trait) with gnomAD MAF $< 0.01$. The full list of trait-gene pairs constructed from the process is available in table A.10.

#### A.5.13.3 Setting up the classification problem to quantify performance for identifying causal genes

We partitioned the genome into approximately independent LD blocks [12] and for each GWAS trait, we kept only genes located in LD blocks where there were at least one silver standard gene and a GWAS significant hit for the trait as illustrated in Figure A.22. Then, we labelled the silver standard genes as 1 and all the others were labelled as 0, as represented schematically in Figure A.23. We calculated the ROC and precision recall curves for classifying the silver standard gene correctly. Note that we used a universal cutoff across all GWAS loci, hence highly correlated genes would be classified as causal or non-causal as a cluster.

In more detail, for each of tested gene-trait pairs, we obtained the gene-level statistics for the corresponding trait from the application of various methods, *i.e. enloc*, *coloc*, SMR,

and PrediXcan-*mashr*. Since we had results across tissues, we selected the 'best' scores (highest regional colocalization probability (rcp) in *enloc*; highest posterior probability under hypothesis 4 in *coloc*; smallest p-value in SMR and PrediXcan-*mashr*) to build the table A.23. For splicing (with statistics reported at the intron excision event level), we obtained gene-level statistics by taking the 'best' score among all splicing events of the gene, across all tissues.

The full list of silver standard genes can be found in tables A.13 and A.14. The number of GWAS loci and silver standard genes that remained after the above filtering steps can be found in table A.5. The number of genes tested per LD block is shown in Figure A.21.

| silver standard | trait | nloci | ngene | silver standard | trait | nloci | ngene |
|---|---|---|---|---|---|---|---|
| rare variant | Standing_Height_UKB | 29 | 35 | OMIM | Monocyte_Count | 1 | 1 |
| rare variant | LDL_Cholesterol | 7 | 10 | OMIM | Neutrophil_Count | 14 | 17 |
| rare variant | High_Cholesterol_UKBS | 6 | 8 | OMIM | White_Blood_Cell_Count | 16 | 17 |
| rare variant | HDL_Cholesterol | 12 | 18 | OMIM | Coronary_Artery_Disease | 12 | 13 |
| rare variant | Triglycerides | 6 | 9 | OMIM | Type_2_Diabetes | 11 | 12 |
| OMIM | Deep_Venous_Thrombosis | 2 | 2 | OMIM | Waist_Circumference_EUR | 6 | 6 |
| OMIM | Asthma_UKBS | 10 | 12 | OMIM | LDL_Cholesterol | 7 | 9 |
| OMIM | Type_1_Diabetes_UKBS | 1 | 2 | OMIM | Triglycerides | 11 | 11 |
| OMIM | Hypothyroidism_UKBS | 14 | 14 | OMIM | Inflammatory_Bowel_Disease | 7 | 8 |
| OMIM | Eczema_UKBS | 4 | 5 | OMIM | Ulcerative_Colitis | 4 | 4 |
| OMIM | Psoriasis_UKBS | 2 | 2 | OMIM | Alzheimers_Disease | 2 | 2 |
| OMIM | Gout_UKBS | 1 | 1 | OMIM | Systemic_Lupus_Erythematosus | 3 | 5 |
| OMIM | High_Cholesterol_UKBS | 6 | 8 | OMIM | Schizophrenia | 1 | 1 |
| OMIM | BMI_UKB | 35 | 35 | OMIM | Rheumatoid_Arthritis | 3 | 3 |
| OMIM | Hypertension_UKBS | 19 | 24 | OMIM | Systolic_Blood_Pressure | 2 | 2 |
| OMIM | Eosinophil_Count | 7 | 7 | OMIM | Diastolic_Blood_Pressure | 3 | 3 |
| OMIM | Lymphocyte_Count | 2 | 2 | | | | |

**Table A.5: Count of GWAS loci with predicted causal effects overlapping likely functional genes.** The number of GWAS loci and the number of silver standard genes included for analysis after taking the intersection between GWAS loci and silver standard genes are shown.

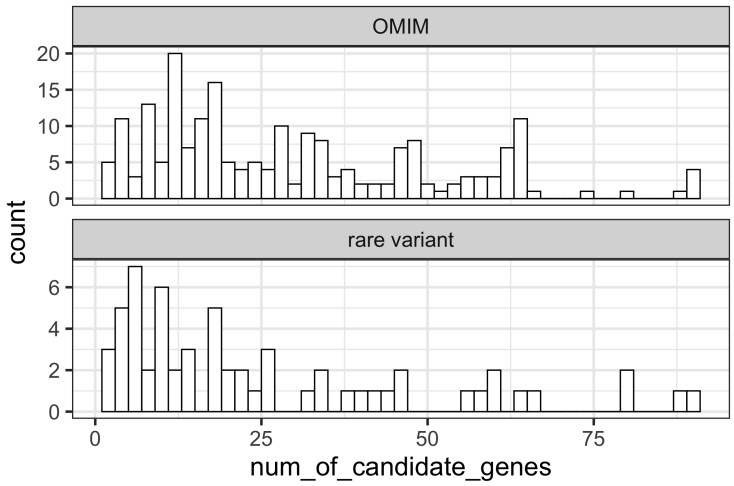



**Figure A.21: Distribution of the number of tested genes per GWAS locus overlapping OMIM- and rare variant-based silver standard.** The distributions of the number of candidate genes per GWAS locus are shown for OMIM-based curation (top) and rare variant association-based curation (bottom).

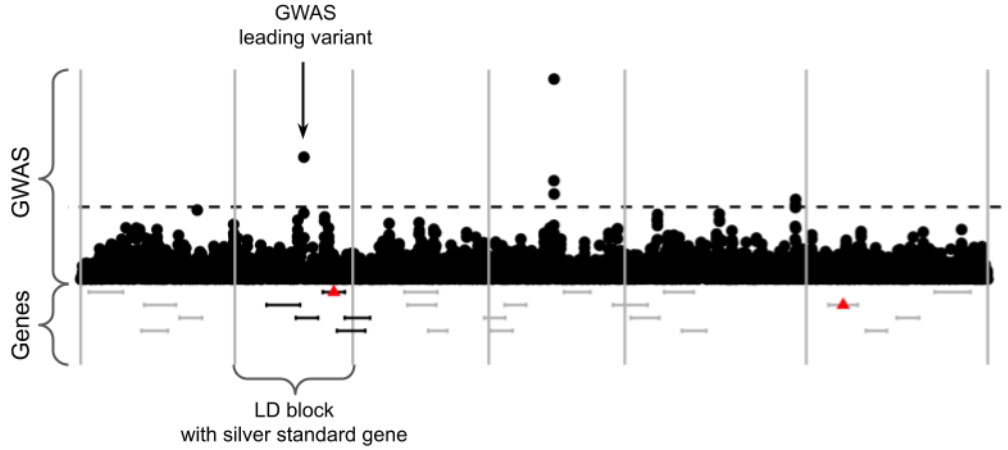



**Figure A.22: Selection of genes to assess ability to identify silver standard genes.** The GWAS summary statistics were binned into independent LD blocks (boundaries of LD block are shown as gray vertical lines). Only genes within LD blocks that contain both a silver standard gene (red triangle) and a GWAS significant variant (points above $-\log_{10}(p) > -\log_{10}(5 \cdot 10^{-8})$) were used in the calculation of performance (ROC and PR curves).

| trait | gene | LD block | lead SNP | silver std status | enloc rcp | PrediXcan p-value | proximity |
|---|---|---|---|---|---|---|---|
| trait1 | gene1 | LD1 | SNP1 | 0 | 0.01 | 1E-04 | 0 |
| trait1 | gene2 | LD1 | SNP1 | 1 | 0.30 | 1E-08 | 1 |
| trait1 | gene3 | LD1 | SNP1 | 0 | 0.02 | 0.32 | 2 |
| trait1 | gene4 | LD1 | SNP1 | 0 | 0.10 | 0.01 | 3 |
| trait1 | gene5 | LD2 | SNP2 | 0 | 0.00 | 0.38 | 0 |
| trait1 | gene6 | LD2 | SNP2 | 0 | 0.00 | 0.26 | 1 |

**Figure A.23: Schematic representation of data used for classification.**

### A.5.13.4 AUC of the ROC curves

For expression, the areas under the curve (AUC) of were, in increasing performance, 0.553, 0.591, 0.669, and 0.672 for *coloc*, SMR, *enloc*, and PrediXcan using the OMIM silver standard A.4C. AUC were higher when using the rare variant silver standard with SMR at the bottom of the ranking followed by *coloc*, PrediXcan, and enloc at the top A.6. For splicing *enloc* had higher 0.650 vs. 0.632 for PrediXcan using OMIM silver standard and 0.714 and 0.686 using the rare variant silver standard.

| Regulation | Dataset | Method | ROC AUC |
|---|---|---|---|
| expression | OMIM | *coloc* | 0.553 |
| expression | OMIM | *enloc* | 0.669 |
| expression | OMIM | PrediXcan | 0.672 |
| expression | OMIM | SMR | 0.591 |
| expression | Rare variant | *coloc* | 0.661 |
| expression | Rare variant | *enloc* | 0.755 |
| expression | Rare variant | PrediXcan | 0.743 |
| expression | Rare variant | SMR | 0.629 |
| splicing | OMIM | *enloc* | 0.650 |
| splicing | OMIM | PrediXcan | 0.632 |
| splicing | Rare variant | *enloc* | 0.714 |
| splicing | Rare variant | PrediXcan | 0.686 |

**Table A.6: Enrichment and AUC of *coloc*, *enloc*, SMR, and PrediXcan.**

### A.5.13.5 Precision-recall curves of PrediXcan and *enloc* on silver standard gene sets

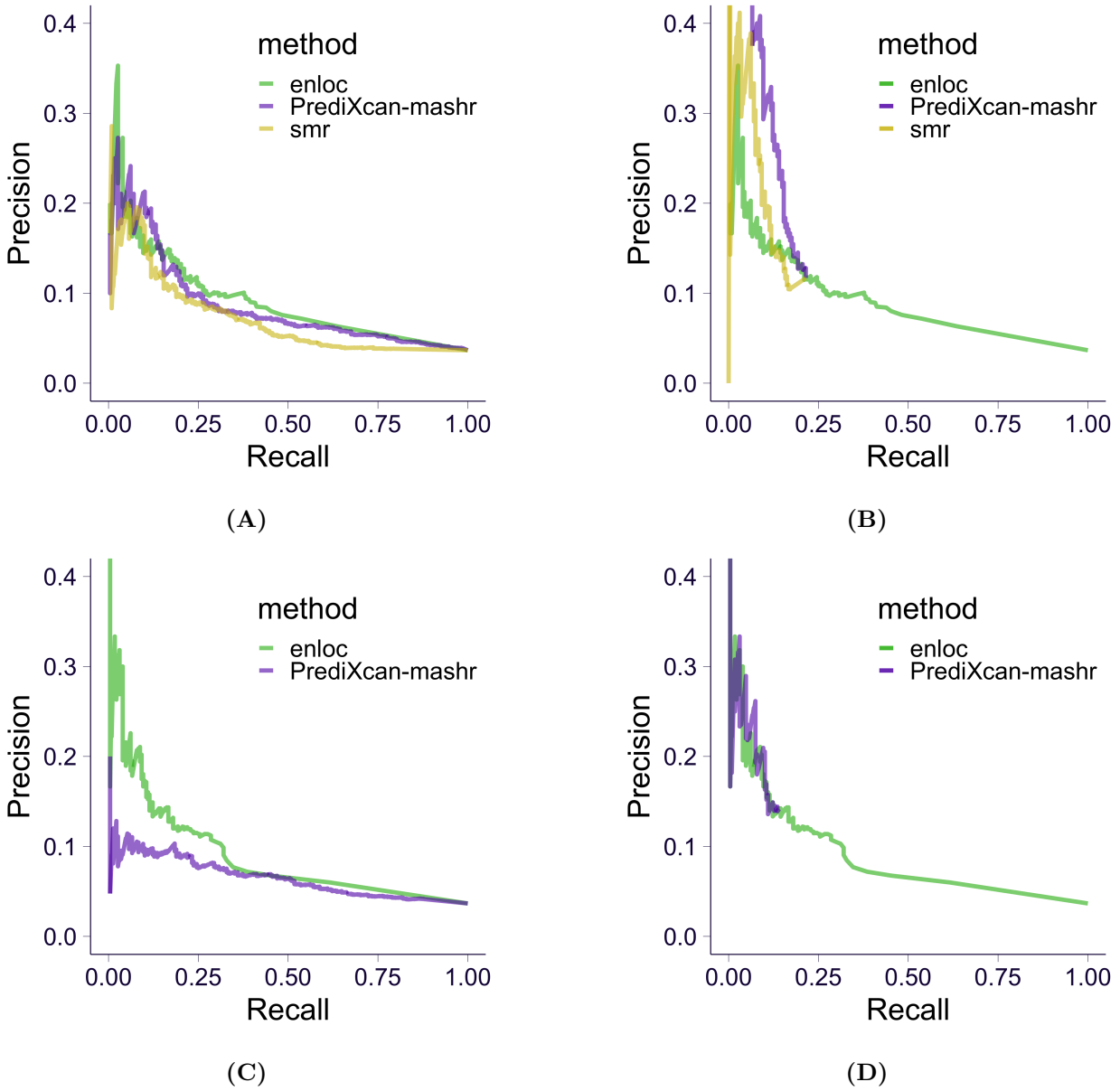



**Figure A.24: Precision-recall curves of colocalization/association based methods on OMIM silver standard.** The results on expression data are shown in top row and the ones on splicing data are shown in bottom row. **(A,C)** Precision-recall curve of colocalization/association based methods. **(B,D)** Precision-recall curve of association based methods when pre-filtering with *enloc* rcp $> 0.1$.

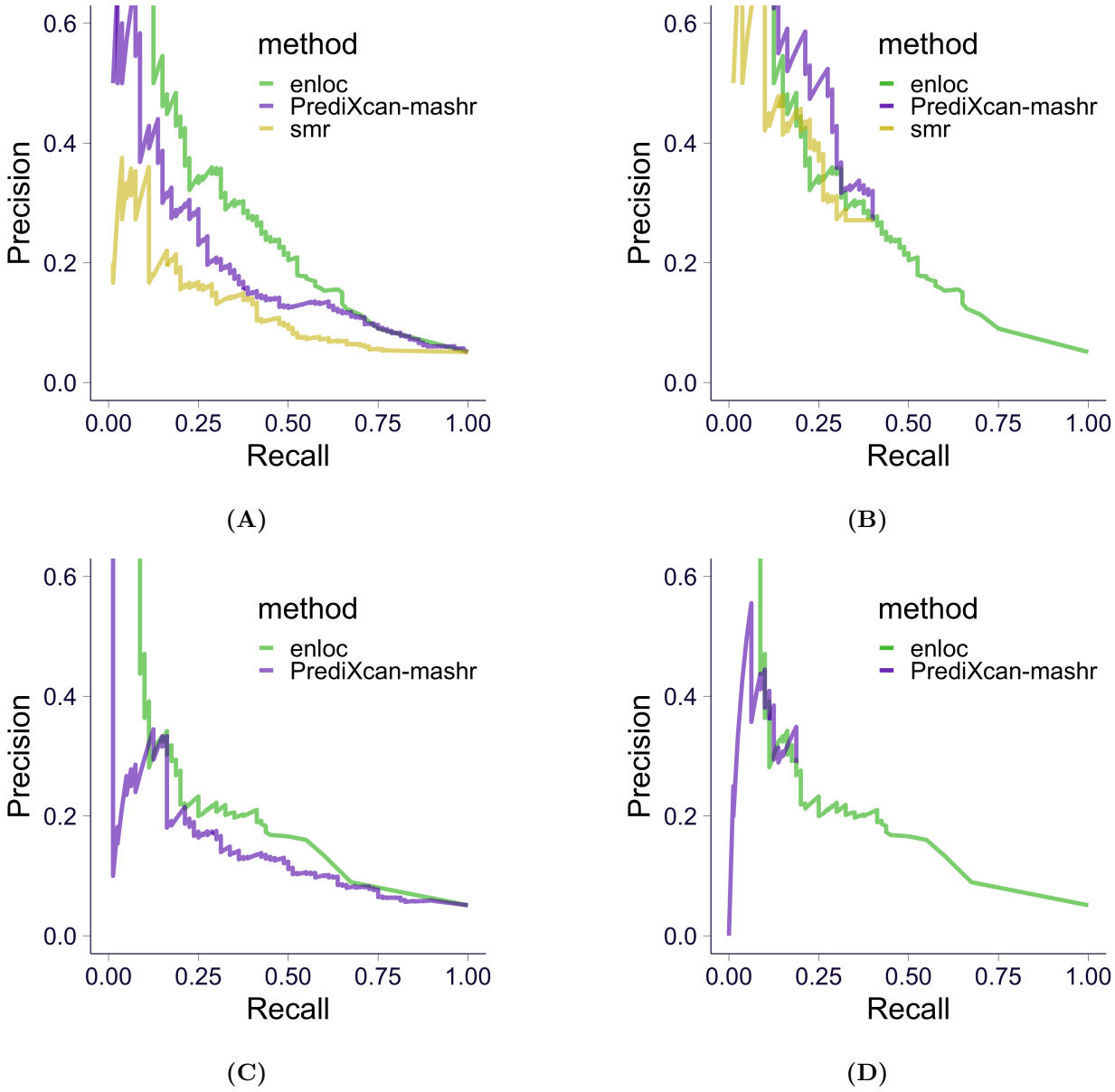
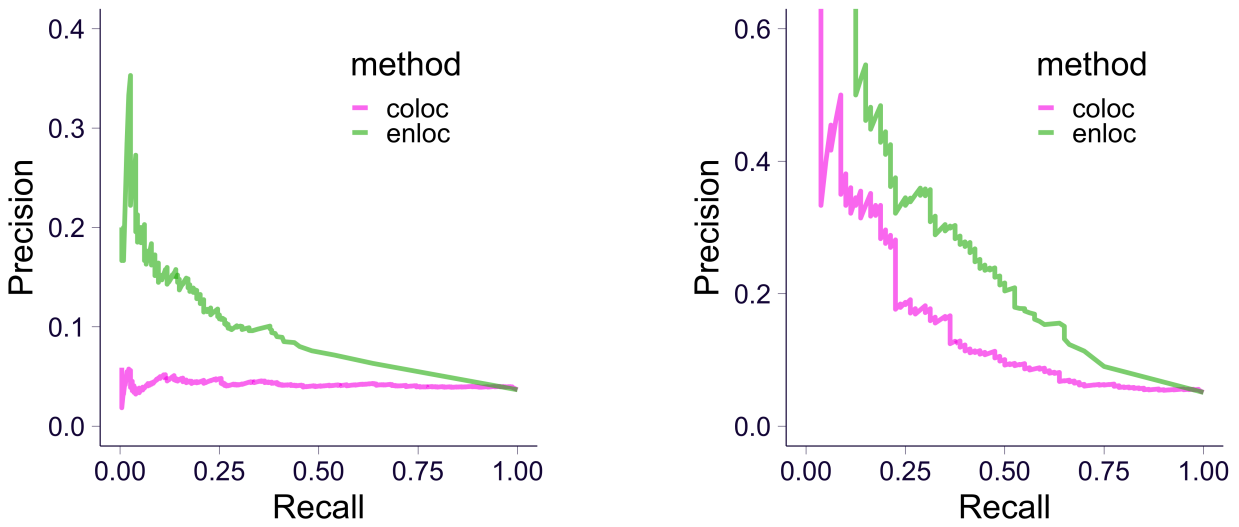



**Figure A.25: Precision-recall curves of colocalization/association based methods on rare variant-based silver standard.** The results on expression data are shown in top row and the ones on splicing data are shown in bottom row. **(A,C)** Precision-recall curve of colocalization/association based methods. **(B,D)** Precision-recall curve of association based methods when pre-filtering with *enloc* rcp $> 0.1$.

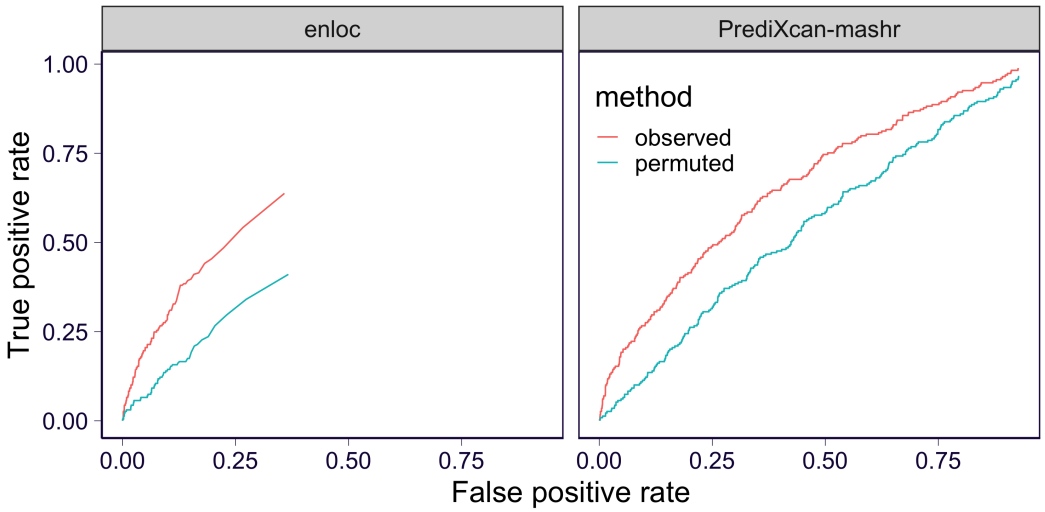



**(A)** **(B)**

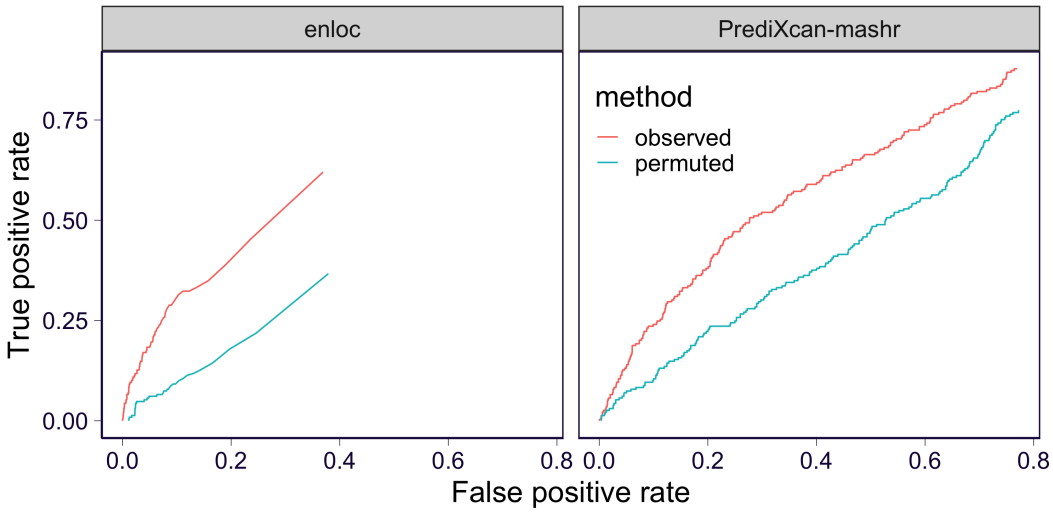
**Figure A.26: Precision-recall curves of *enloc* vs *coloc*.** Precision recall curve of *enloc* (blue) and *coloc* (green) with expression using OMIM silver standard (in **(A)**) and rare variant-based silver standard (in **(B)**).

### A.5.13.6 ROC and PR curves under permuted data

To examine if the shapes of PR and ROC curves were driven by the bias buried in the data, we plotted the PR and ROC curves under the permuted data. Specifically, for each GWAS locus, we permuted the genes that overlapped with the locus while keeping the scores unchanged. We compared the PR and ROC curves for observed and permuted data for OMIM silver standard as shown in Figure A.27 and Figure A.28.

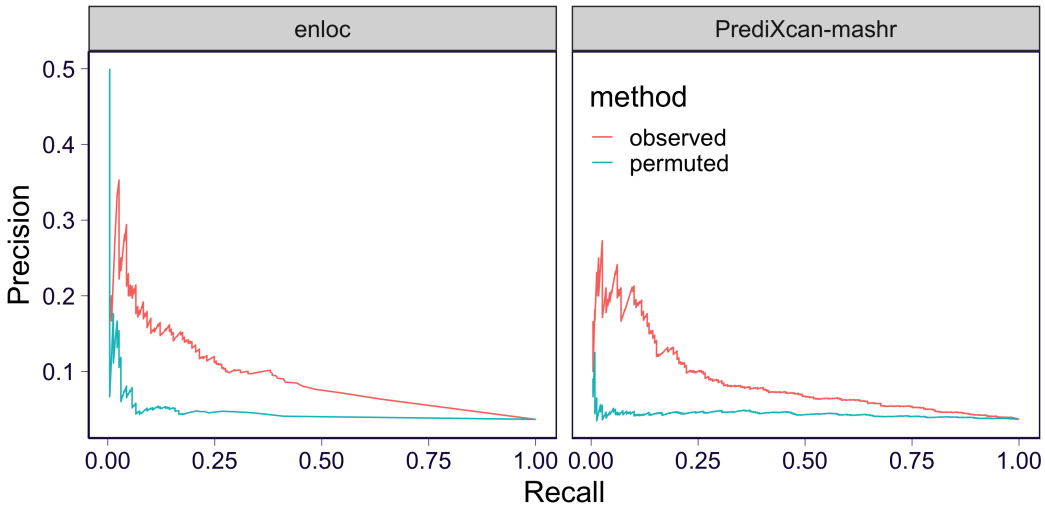



**(A)**

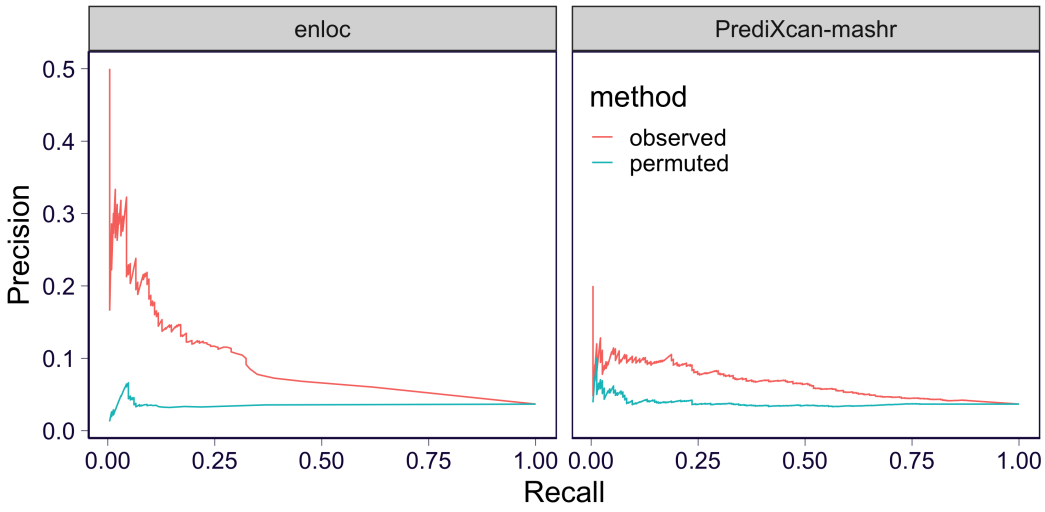



**(B)**

**Figure A.27: The ROC curves under permuted data based on OMIM silver standard.** The ROC curves under observed and permuted data are shown in **(A)** (for expression based analysis) and **(B)** (for splicing based analysis).




**(A)**




**(B)**

**Figure A.28: The PR curves under permuted data based on OMIM silver standard.** The PR curves under observed and permuted data are shown in **(A)** (for expression based analysis) and **(B)** (for splicing based analysis).

### A.5.13.7 Assessing the contribution of proximity, colocalization, and association significance

To investigate the usefulness of the colocalization and association statistics reported by *enloc* and PrediXcan respectively, we performed logistic regression, as described in Eq. A.28, to fit log odds of being a 'causal' gene against the ranking of: 1) proximity to GWAS lead variant (from close to distal), 2) rcp from *enloc* (from high to low), and 3) gene-level association

p-value from PrediXcan-*mashr* or SMR (from significant to non-significant).

$$\text{logit}(\Pr(\text{causal}_i)) = \beta_0 + \beta_1 \cdot \text{rank}(\text{proximity}_i) + \beta_2 \cdot \text{rank}(\text{rcp}_i) + \beta_3 \cdot \text{rank}(\text{P-value}_i), \tag{A.28}$$

in which non-zero $\beta_k$ meant that the $k$th variable contributed independently on predicting whether a gene was causal. Moreover, negative $\beta_k$ indicated that the direction of contribution of the variable was as expected.

We note that here the analysis is performed by LD blocks rather than genome-wide as was done for calculating the ROC and precision recall curves. More specifically, the ranking within each LD block is used rather than genome-wide.

| regulation | silver_standard | variable | coefficient | coefficient_se | pvalue |
|---|---|---|---|---|---|
| expression | OMIM | rank_proximity | -0.018 | 0.0081 | 0.03 |
| expression | OMIM | predixcan_mashr_eur | -0.038 | 0.008 | $2.2 \times 10^{-6}$ |
| expression | OMIM | enloc | -0.02 | 0.0093 | 0.031 |
| splicing | OMIM | rank_proximity | -0.026 | 0.0073 | 0.00031 |
| splicing | OMIM | predixcan_mashr_eur | -0.037 | 0.008 | $3.5 \times 10^{-6}$ |
| splicing | OMIM | enloc | -0.012 | 0.0086 | 0.17 |
| expression | rare variant | rank_proximity | -0.013 | 0.018 | 0.46 |
| expression | rare variant | predixcan_mashr_eur | -0.043 | 0.016 | 0.0084 |
| expression | rare variant | enloc | -0.043 | 0.02 | 0.032 |
| splicing | rare variant | rank_proximity | -0.048 | 0.015 | 0.0015 |
| splicing | rare variant | predixcan_mashr_eur | -0.018 | 0.013 | 0.15 |
| splicing | rare variant | enloc | -0.02 | 0.015 | 0.2 |

**Table A.7: Predictive value of different per-locus prioritization methods.** Results on regression-based test (logistic regression) in per-locus analysis are shown. The estimated log odds ratio of the rank of proximity (distance between GWAS leading variant and gene body), PrediXcan significance, and *enloc* rcp are shown in rows **rank_proximity**, **predixcan_mashr_eur**, and **enloc**.

### *A.5.14 Causal tissue analysis*

To identify tissues of relevance for the etiology of complex traits, we investigated the patterns of tissue specificity and tissue sharing of PrediXcan association results across 49 tissues. For each trait-gene pair, the PrediXcan z-score can be represented as a $49 \times 1$ vector with each entry being the gene-level z-score in the corresponding tissue (if the prediction model of the gene is not available in that tissue, we filled in zero). To explore the tissue-specificity of the PrediXcan z-score vector, we proceeded by assigning the z-score vector to a tissue-pattern

category and tested whether certain tissue-pattern categories were over-represented among colocalized PrediXcan genes as compared to non-colocalized genes. We used the FLASH factors identified from matrix factorization applied to the cis-eQTL effect size matrix, as described in Section A.5.9 (as PrediXcan and cis-eQTL shared similar tissue-sharing pattern, data not shown). To obtain a set of detailed and biologically interpretable tissue-pattern categories from the 31 FLASH factors, we manually merged them into 18 categories as shown in Figure A.29. For each trait, we projected the z-score vector of each gene to one of the 31 FLASH factors (as described in Section A.5.9) so that the gene was assigned to the corresponding tissue-pattern category. We defined a 'positive' set of genes as the ones with PrediXcan p-value that meets Bonferroni significance at $\alpha = 0.05$ in at least one tissue and *enloc* rcp $> 0.01$ in at least one tissue, which could be thought as a set of candidate genes affecting the trait through expression level. We chose a rather low threshold used for the rcp due to the stringent conservative nature of colocalization probabilities. We also constructed a 'negative' set of genes with *enloc* rcp $= 0$, which could be thought as a set of genes whose expressions were unlikely to affect the trait. We proceeded to test whether certain tissue-pattern categories were enriched in 'positive' set as compared to 'negative' set. Since the main focus of this analysis was tissue-specific patterns, we excluded *Factor1* (the cross-tissue factor) and *Factor25* (likely to be a tissue-shared factor capturing tissues with large sample size). Additionally, we excluded *Factor7* (testis), as it was unlikely to be the mediating tissue but might introduce false positives. We tested the enrichment of each tissue-pattern category by Fisher's exact test ('positive'/'negative' sets and in/not in tissue-patter category). Among 87 traits, 82 traits had *enloc* signal and the enrichment of these was calculated accordingly.

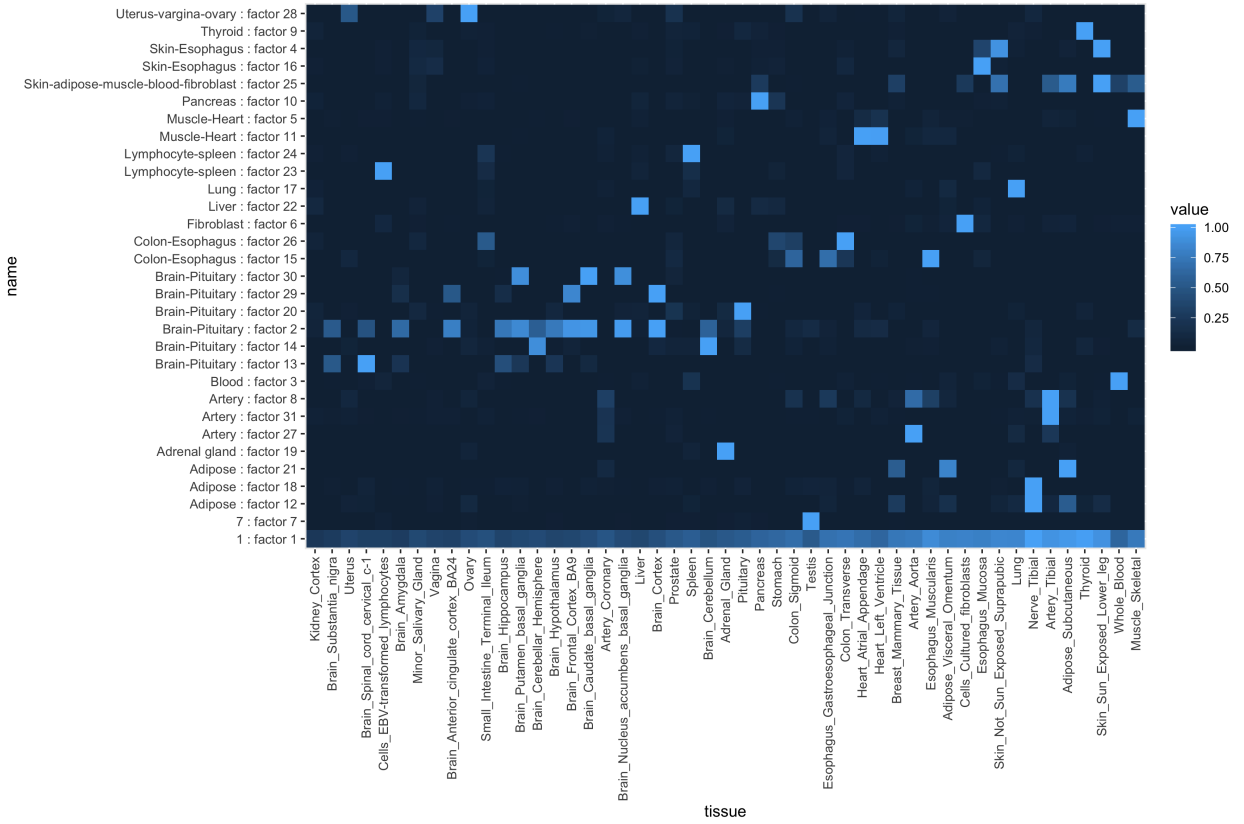



**Figure A.29: Patterns of tissue sharing identified via factor analysis using flashr.** Tissue-pattern categories generated from FLASH applied to the cis-eQTLs are shown. Factor 1 represents cross tissue category covering all tissues, with higher weight for larger sample size tissues. These tissue categories (on y-axis) were used in the analysis of causal tissue identification. Tissues are ordered by sample size.

### *A.5.15 Supplementary tables in spreadsheet*

Tables listed in this section can be found at `https://docs.google.com/spreadsheets/d/1DFf_UMfrGR3lrYdctdr1djs4Ef9Omh4onPVPlkE1FK8`.

**Table A.9: Presumed causal genes included in the OMIM database.** Columns are: **trait**: Tag used for the trait, **pheno_mim**: MIM ID of the phenotype mapped to GWAS trait, **mim**: MIM ID of the corresponding gene, **entry_type**: Entry type in the OMIM database, **entrez_gene_id**: Gene ID based on Entrez database, **gene_name**: Official gene symbol, **ensembl_gene_id**: Gene ID based on Ensembl database, **gene_type**: Gene type based on Gencode, **gene**: Trimmed Gene ID based on Ensembl database.

**Table A.8: GWAS Metadata.** contains relevant information concerning each GWAS study used. Analyses used the 87 traits with deflation=0 unless explictly said otherwise. Columns are: **Tag**: Internal name to identify the study, **Deflation**: Deflation status after imputation (0 for no deflation, 1 for moderate deflation, 2 for extreme deflation), **PUBMED_Paper_Link**: PUBMED entry, **Pheno_File**: name of downloaded file, **Source_File**: actual name of GWAS summary statistics (i.e. downloaded files might contain several traits), **Portal**: URL to GWAS study portal, **Consortium**: Name of Consortium if any, **Link**: download link for the file, **Notes**: any special comment on the GWAS trait, **Header**: GWAS summary statistics header in case the file is malformed, **EFO**: Experimental Factor Ontology [94] entry if applicable, **HPO**: Human Phenotype Ontology [69] entry if applicable, **Description**: optional description of the study, **Trait**: trait name, **Sample_Size**: number of individuals included in the study, **Population**: types of populations present (EUR for European, AFR for African, EAS for East Asian, etc), **Date**: Date the file was downloaded, **Declared_Effect_Allele**: column specifying effect allele, **Genome_Reference**: Human Genome release used as reference (i.e. hg19, hg38), **Binary**: wether the trait is dichotomous, **Cases**: number of cases if binary trait, **abbreviation**: short string for figure and table display, **new_abbreviation**: additional abbreviation, **new_Trait**: additional trait name, **Category**: type of trait, **Color**: Hexadecimal color code for display

**Table A.10: Genes suggested as causal by rare variant association studies.** Columns are: **gene**: Trimmed gene ID based on Ensembl database, **nobs**: Number of times gene has been observed in the trait, **trait**: Tag for the trait name.

**Table A.11: PrediXcan and enloc results for predicted causal genes selected based on OMIM.** Columns are: **lead_var**: the most significant variant within the LD block, **trait**: trait name, **gene**: Ensembl ID for the gene, **is_omim**: Is included in the OMIM database. TRUE if included, FALSE if not, **proximity**: 0 if variant is in the gene, otherwise BPS from the gene boundary, **rank_proximity**: ranking by proximity within LD block (rank starts from 0 and the closer the lower rank), **percentage_proximity**: rank_proximity / number of genes in the locus, **predixcan_mashr_eur_score**: -log10 p-value (most significant across tissues is used) of PrediXcan-MASH trained on European data, **enloc_score**: rcp (max across tissues), **predixcan_mashr_eur_rank**: PrediXcan significance ranking within LD block (rank starts from 0 and the higher significance the lower rank), **enloc_rank**: enloc rcp ranking within LD block (rank starts from 0 and the higher rcp the lower rank), **predixcan_mashr_eur_percentage**: predixcan_mashr_eur_rank / number of genes in the locus, **enloc_percentage**: enloc_rank / number of genes in the locus, **gene_name**: Official gene symbol, **gene_type**: Gencode annotsted gene type, **chromosome**: Chromosome for the gene, **start**: Gencode annotated gene start position. All isoforms are combined, **end:** Gencode annotated gene end position. All isoforms are combined, **strand**: Gencode annotated gene strand.

**Table A.12: PrediXcan and enloc results for presumed causal genes in the rare variant based silver standard.** Columns are: **lead_var**: the most significant variant within the LD block, **trait**: trait name, **gene**: Ensembl ID for the gene, **is_ewas**: Is included in the EWAS . TRUE if included, FALSE if not, **proximity**: 0 if variant is in the gene, otherwise BPS from the gene boundary, **rank_proximity**: ranking by proximity within LD block (rank starts from 0 and the closer the lower rank), **percentage_proximity**: rank_proximity / number of genes in the locus, **predixcan_mashr_score**: -log10 p-value (most significant across tissues is used) of PrediXcan-MASH trained on European data, **enloc_score**: rcp (max across tissues), **predixcan_mashr_rank**: PrediXcan significance ranking within LD block (rank starts from 0 and the higher significance the lower rank), **enloc_rank**: enloc rcp ranking within LD block (rank starts from 0 and the higher rcp the lower rank), **predixcan_mashr_percentage**: predixcan_mashr_eur_rank / number of genes in the locus, **enloc_percentage**: enloc_rank / number of genes in the locus, **gene_name**: Official gene symbol, **gene_type**: Gencode annotsted gene type, **chromosome**: Chromosome for the gene, **start**: Gencode annotated gene start position. All isoforms are combined, **end**: Gencode annotated gene end position. All isoforms are combined, **strand**: Gencode annotated gene strand.

**Table A.13: OMIM genes included in the analysis.** Columns are: **gene**, **trait**.

**Table A.14: Rare variant silver standard genes included in the analysis.** Columns are: **gene**, **trait**.

**Table A.15: BioVU.** Columns are: **gene**, **tissue**, **trait_map**: mapped trait, **pheno**: trait, **gene_name**, **p_discovery**, **rcp_discovery**, **beta_biovu**, **p_biovu**, **z_biovu**.

# REFERENCES


[1] 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*, 526(7571):68–74, 2015.

[2] Tiffany Amariuta, Kazuyoshi Ishigaki, Hiroki Sugishita, Tazro Ohta, Masaru Koido, Kushal K Dey, Koichi Matsuda, Yoshinori Murakami, Alkes L Price, Eiryo Kawakami, et al. Improving the trans-ancestry portability of polygenic risk scores by prioritizing variants in predicted cell-type-specific regulatory elements. *Nature genetics*, 52(12):1346–1354, 2020.

[3] Verneri Anttila, Bendik S Winsvold, Padhraig Gormley, Tobias Kurth, Francesco Bettella, George McMahon, Mikko Kallela, Rainer Malik, Boukje De Vries, Gisela Terwindt, et al. Genome-wide meta-analysis identifies new susceptibility loci for migraine. *Nature genetics*, 45(8):912–917, 2013.

[4] Hugues Aschard, Bjarni J Vilhjálmsson, Amit D Joshi, Alkes L Price, and Peter Kraft. Adjusting for heritable covariates can bias effect estimates in genome-wide association studies. *The American Journal of Human Genetics*, 96(2):329–339, 2015.

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

[6] Alvaro N Barbeira, Rodrigo Bonazzola, Eric R Gamazon, Yanyu Liang, YoSon Park, Sarah Kim-Hellmuth, Gao Wang, Zhuoxun Jiang, Dan Zhou, Farhad Hormozdiari, et al. Exploiting the gtex resources to decipher the mechanisms at gwas loci. *Genome biology*, 22(1):1–24, 2021.

[7] Alvaro N Barbeira, Scott P Dickinson, Rodrigo Bonazzola, Jiamao Zheng, Heather E Wheeler, Jason M Torres, Eric S Torstenson, Kaanan P Shah, Tzintzuni Garcia, Todd L Edwards, et al. Exploring the phenotypic consequences of tissue specific gene expression variation inferred from gwas summary statistics. *Nature communications*, 9(1):1–20, 2018.

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

[9] Alvaro N Barbeira, Milton Pividori, Jiamao Zheng, Heather E Wheeler, Dan L Nicolae, and Hae Kyung Im. Integrating predicted transcriptome from multiple tissues improves association detection. *PLoS genetics*, 15(1):e1007889, 2019.

[10] Lisa Bastarache, Jacob J Hughey, Scott Hebbring, Joy Marlo, Wanke Zhao, Wanting T Ho, Sara L Van Driest, Tracy L McGregor, Jonathan D Mosley, Quinn S Wells, et al. Phenotype risk scores identify patients with unrecognized mendelian disease patterns. *Science*, 359(6381):1233–1239, 2018.

[11] Christian Benner, Aki S Havulinna, Marjo-Riitta Jarvelin, Veikko Salomaa, Samuli Ripatti, and Matti Pirinen. Prospects of Fine-Mapping Trait-Associated Genomic Regions by Using Summary Statistics from Genome-wide Association Studies. *The American Journal of Human Genetics*, 101(4):539–551, October 2017.

[12] Tomaz Berisa and Joseph K. Pickrell. Approximately independent linkage disequilibrium blocks in human populations. *Bioinformatics*, 32(2):283–285, January 2016.

[13] David R Blair, Christopher S Lyttle, Jonathan M Mortensen, Charles F Bearden, Anders Boeck Jensen, Hossein Khiabanian, Rachel Melamed, Raul Rabadan, Elmer V Bernstam, Søren Brunak, Lars Juhl Jensen, Dan Nicolae, Nigam H Shah, Robert L Grossman, Nancy J Cox, Kevin P White, and Andrey Rzhetsky. A Nondegenerate Code of Deleterious Variants in Mendelian Loci Contributes to Complex Disease Risk. *Cell*, 155(1):70–80, September 2013.

[14] Jack Bowden, George Davey Smith, and Stephen Burgess. Mendelian randomization with invalid instruments: effect estimation and bias detection through egger regression. *International journal of epidemiology*, 44(2):512–525, 2015.

[15] Jack Bowden, George Davey Smith, Philip C Haycock, and Stephen Burgess. Consistent estimation in mendelian randomization with some invalid instruments using a weighted median estimator. *Genetic epidemiology*, 40(4):304–314, 2016.

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

[17] Brendan K Bulik-Sullivan, Po-Ru Loh, Hilary K Finucane, Stephan Ripke, Jian Yang, Nick Patterson, Mark J Daly, Alkes L Price, Benjamin M Neale, Schizophrenia Working Group of the Psychiatric Genomics Consortium, et al. Ld score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature genetics*, 47(3):291, 2015.

[18] Annalisa Buniello, Jacqueline A.L. Macarthur, Maria Cerezo, Laura W. Harris, James Hayhurst, Cinzia Malangone, Aoife McMahon, Joannella Morales, Edward Mountjoy, Elliot Sollis, Daniel Suveges, Olga Vrousgou, Patricia L. Whetzel, Ridwan Amode, Jose A. Guillen, Harpreet S. Riat, Stephen J. Trevanion, Peggy Hall, Heather Junkins, Paul Flicek, Tony Burdett, Lucia A. Hindorff, Fiona Cunningham, and Helen Parkinson. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research*, 2019.

[19] Stephen Burgess, Adam Butterworth, and Simon G Thompson. Mendelian randomization analysis with multiple genetic variants using summarized data. *Genetic epidemiology*, 37(7):658–665, 2013.

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

[21] Minal Caliskan, Elisabetta Manduchi, H. Shanker Rao, Julian A. Segert, Marcia Holsbach Beltrame, Marco Trizzino, YoSon Park, Samuel W. Baker, Alessandra Chesi, Matthew E. Johnson, Kenyaita M. Hodge, Michelle E. Leonard, Baoli Loza, Dong Xin, Andrea M. Berrido, Nicholas J. Hand, Robert C. Bauer, Andrew D. Wells, Kim M. Olthoff, Abraham Shaked, Daniel J. Rader, Struan F. A. Grant, and Christopher D. Brown. Genetic and Epigenetic Fine Mapping of Complex Trait Associated Loci in the Human Liver. *American journal of human genetics*, 105(1):89–107, July 2019.

[22] Latarsha J Carithers, 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, Helen M Moore, and on behalf of the GTEx Consortium. A Novel Approach to High-Quality Postmortem Tissue Procurement: The GTEx Project. *Biopreservation and Biobanking*, 13(5):311–319, October 2015.

[23] Taylor B Cavazos and John S Witte. Inclusion of variants discovered from diverse populations improves polygenic risk score transferability. *Human Genetics and Genomics Advances*, 2(1):100017, 2021.

[24] Ananyo Choudhury, Shaun Aron, Laura R Botigué, Dhriti Sengupta, Gerrit Botha, Taoufik Bensellak, Gordon Wells, Judit Kumuthini, Daniel Shriner, Yasmina J Fakim, et al. High-depth african genomes inform human migration and health. *Nature*, 586(7831):741–748, 2020.

[25] John H. Contois, Joseph P. McConnell, Amar A. Sethi, Gyorgy Csako, Sridevi Devaraj, Daniel M. Hoefner, and G. Russell Warnick. Apolipoprotein B and cardiovascular disease risk: position statement from the AACC Lipoproteins and Vascular Diseases Division Working Group on Best Practices. *Clinical chemistry*, 55(3):407–419, March 2009.

[26] Francis Crick. Central dogma of molecular biology. *Nature*, 227(5258):561–563, 1970.

[27] David Curtis. Polygenic risk score for schizophrenia is more strongly associated with ancestry than with schizophrenia. *Psychiatric genetics*, 28(5):85–89, 2018.

[28] Hans D Daetwyler, Beatriz Villanueva, and John A Woolliams. Accuracy of predicting the genetic risk of disease using a genome-wide approach. *PloS one*, 3(10):e3395, 2008.

[29] Andy Dahl, Vincent Guillemot, Joel Mefford, Hugues Aschard, and Noah Zaitlen. Adjusting for principal components of molecular phenotypes induces replicating false positives. *Genetics*, 211(4):1179–1189, 2019.

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

[31] Olivier Delaneau, Jonathan Marchini, and Jean-François Zagury. A linear complexity phasing method for thousands of genomes. *Nature methods*, 9(2):179–181, 2012.

[32] David S DeLuca, Joshua Z Levin, Andrey Sivachenko, Timothy Fennell, Marc-Danie Nazaire, Chris Williams, Michael Reich, Wendy Winckler, and Gad Getz. RNA-SeQC: RNA-seq metrics for quality control and process optimization. *Bioinformatics*, 28(11):1530–1532, June 2012.

[33] Joshua C Denny, Lisa Bastarache, Marylyn D Ritchie, Robert J Carroll, Raquel Zink, Jonathan D Mosley, Julie R Field, Jill M Pulley, Andrea H Ramirez, Erica Bowton, et al. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nature biotechnology*, 31(12):1102–1111, 2013.

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

[35] Georg B Ehret, Patricia B Munroe, Kenneth M Rice, Murielle Bochud, Andrew D Johnson, Daniel I Chasman, Albert V Smith, Martin D Tobin, Germaine C Verwoert, Shih-Jen Hwang, et al. Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. *Nature*, 478(7367):103–109, 2011.

[36] Lloyd T Elliott, Kevin Sharp, Fidel Alfaro-Almagro, Sinan Shi, Karla L Miller, Gwenaëlle Douaud, Jonathan Marchini, and Stephen M Smith. Genome-wide association studies of brain imaging phenotypes in uk biobank. *Nature*, 562(7726):210–216, 2018.

[37] Benjamin L Elsworth, Matthew S Lyon, Tessa Alexander, Yi Liu, Peter Matthews, Jon Hallett, Phil Bates, Tom Palmer, Valeriia Haberland, George Davey Smith, et al. The mrc ieu opengwas data infrastructure. *bioRxiv*, 2020.

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

[39] Ronald A Fisher. Xv.—the correlation between relatives on the supposition of mendelian inheritance. *Earth and Environmental Science Transactions of the Royal Society of Edinburgh*, 52(2):399–433, 1919.

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

[41] Christian Fuchsberger, Jason Flannick, Tanya M Teslovich, Anubha Mahajan, Vineeta Agarwala, Kyle J Gaulton, Clement Ma, Pierre Fontanillas, Loukas Moutsianas, Davis J McCarthy, et al. The genetic architecture of type 2 diabetes. *Nature*, 536(7614):41–47, 2016.

[42] Eric R Gamazon, Ayellet V Segrè, Martijn van de Bunt, Xiaoquan Wen, Hualin S Xi, Farhad Hormozdiari, Halit Ongen, Anuar Konkashbaev, Eske M Derks, François Aguet, Jie Quan, GTEx Consortium, Dan L Nicolae, Eleazar Eskin, Manolis Kellis, Gad Getz, Mark I McCarthy, Emmanouil T Dermitzakis, Nancy J Cox, and Kristin G Ardlie. Using an atlas of gene regulation across 44 human tissues to inform complex disease- and trait-associated variation. *Nature genetics*, 50(7):956–967, July 2018.

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

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

[45] John Michael Gaziano, John Concato, Mary Brophy, Louis Fiore, Saiju Pyarajan, James Breeling, Stacey Whitbourne, Jennifer Deen, Colleen Shannon, Donald Humphries, et al. Million veteran program: A mega-biobank to study genetic influences on health and disease. *Journal of clinical epidemiology*, 70:214–223, 2016.

[46] Tian Ge, Chia-Yen Chen, Yang Ni, Yen-Chen Anne Feng, and Jordan W Smoller. Polygenic prediction via bayesian regression and continuous shrinkage priors. *Nature communications*, 10(1):1–10, 2019.

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

[48] Richard A Gibbs, John W Belmont, Paul Hardenbol, Thomas D Willis, FL Yu, HM Yang, Lan-Yang Ch'ang, Wei Huang, Bin Liu, Yan Shen, et al. The international hapmap project. *Nature*, 426(6968):789–796, 2003.

[49] Jing Gong, Shufang Mei, Chunjie Liu, Yu Xiang, Youqiong Ye, Zhao Zhang, Jing Feng, Renyan Liu, Lixia Diao, An-Yuan Guo, Xiaoping Miao, and Leng Han. PancanQTL: systematic identification of cis-eQTLs and trans-eQTLs in 33 cancer types. *Nucleic acids research*, 46(D1):D971–D976, January 2018.

[50] Hui Guo, Mary D. Fortune, Oliver S. Burren, Ellen Schofield, John A. Todd, and Chris Wallace. Integration of disease association and eQTL data using a Bayesian colocalisation approach highlights six candidate causal genes in immune-mediated diseases. *Human Molecular Genetics*, 24(12):3305–3313, 03 2015.

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

[52] Ada Hamosh, Alan F. Scott, Joanna S. Amberger, Carol A. Bocchini, and Victor A. McKusick. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research*, 2005.

[53] Gibran Hemani, Jie Zheng, Benjamin Elsworth, Kaitlin H Wade, Valeriia Haberland, Denis Baird, Charles Laurin, Stephen Burgess, Jack Bowden, Ryan Langdon, et al. The mr-base platform supports systematic causal inference across the human phenome. *Elife*, 7:e34408, 2018.

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

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

[56] Margaux LA Hujoel, Steven Gazal, Po-Ru Loh, Nick Patterson, and Alkes L Price. Liability threshold modeling of case–control status and family history of disease increases association power. Technical report, Nature Publishing Group, 2020.

[57] Abhay Hukku, Milton Pividori, Francesca Luca, Roger Pique-Regi, Hae Kyung Im, and Xiaoquan Wen. Probabilistic colocalization of genetic variants from complex and molecular traits: Promise and limitations. *bioRxiv*, 2020.

[58] International HapMap 3 Consortium and others. Integrating common and rare genetic variation in diverse human populations. *Nature*, 467(7311):52, 2010.

[59] Ahmet Turan Isik. Late onset alzheimer's disease in older people. *Clinical interventions in aging*, 5:307, 2010.

[60] Philip R Jansen, Mats Nagel, Kyoko Watanabe, Yongbin Wei, Jeanne E Savage, Christiaan A de Leeuw, Martijn P van den Heuvel, Sophie van der Sluis, and Danielle Posthuma. Genome-wide meta-analysis of brain volume identifies genomic loci and genes shared with intelligence. *Nature communications*, 11(1):1–12, 2020.

[61] Roby Joehanes, Xiaoling Zhang, Tianxiao Huan, Chen Yao, Sai-xia Ying, Quang Tri Nguyen, Cumhur Yusuf Demirkale, Michael L Feolo, Nataliya R Sharopova, Anne Sturcke, et al. Integrated genome-wide analysis of expression quantitative trait loci aids interpretation of genomic association studies. *Genome biology*, 18(1):16, 2017.

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

[63] Konrad J Karczewski, Laurent C Francioli, Grace Tiao, Beryl B Cummings, Jessica Alföldi, Qingbo Wang, Ryan L Collins, Kristen M Laricchia, Andrea Ganna, Daniel P Birnbaum, et al. Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. *BioRxiv*, page 531210, 2019.

[64] Alon Keinan and Andrew G. Clark. Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science*, 336(6082):740–743, 2012.

[65] Sinead Kelly, Neda Jahanshad, A Zalesky, P Kochunov, Ingrid Agartz, C Alloza, OA Andreassen, C Arango, N Banaj, S Bouix, et al. Widespread white matter microstructural differences in schizophrenia across 4322 individuals: results from the enigma schizophrenia dti working group. *Molecular psychiatry*, 23(5):1261–1269, 2018.

[66] Amit V Khera, Mark Chaffin, Krishna G Aragam, Mary E Haas, Carolina Roselli, Seung Hoan Choi, Pradeep Natarajan, Eric S Lander, Steven A Lubitz, Patrick T Ellinor, et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nature genetics*, 50(9):1219, 2018.

[67] S. Kidambi and S. B. Patel. Cholesterol and non-cholesterol sterol transporters: ABCG5, ABCG8 and NPC1l1: a review. *Xenobiotica; the fate of foreign compounds in biological systems*, 38(7-8):1119–1139, July 2008.

[68] Katherine A Knutson, Yangqing Deng, and Wei Pan. Implicating causal brain imaging endophenotypes in alzheimer's disease using multivariable iwas and gwas summary data. *NeuroImage*, 223:117347, 2020.

[69] Sebastian Köhler, Sandra C Doelken, Christopher J Mungall, Sebastian Bauer, Helen V Firth, Isabelle Bailleul-Forestier, Graeme CM Black, Danielle L Brown, Michael Brudno, Jennifer Campbell, et al. The human phenotype ontology project: linking molecular biology and disease through phenotype data. *Nucleic acids research*, 42(D1):D966–D974, 2013.

[70] Natsuhiko Kumasaka, Andrew J Knights, and Daniel J Gaffney. Fine-mapping cellular qtls with rasqual and atac-seq. *Nature genetics*, 48(2):206, 2016.

[71] Brian W Kunkle, Benjamin Grenier-Boley, Rebecca Sims, Joshua C Bis, Vincent Damotte, Adam C Naj, Anne Boland, Maria Vronskaya, Sven J Van Der Lee, Alexandre Amlie-Wolf, et al. Genetic meta-analysis of diagnosed alzheimer's disease identifies new risk loci and implicates a$\beta$, tau, immunity and lipid processing. *Nature genetics*, 51(3):414–430, 2019.

[72] Eric S Lander, Lauren M Linton, Bruce Birren, Chad Nusbaum, Michael C Zody, Jennifer Baldwin, Keri Devon, Ken Dewar, Michael Doyle, William FitzHugh, et al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, 2001.

[73] Cue Hyunkyu Lee, Seungho Cook, Ji Sung Lee, and Buhm Han. Comparison of two meta-analysis methods: inverse-variance-weighted average and weighted sum of z-scores. *Genomics & informatics*, 14(4):173, 2016.

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

[75] Sang Mee Lee, Theodore G Karrison, Nancy J Cox, and Hae Kyung Im. Quantitative allelic test—a fast test for very large association studies. *Genetic epidemiology*, 37(8):831–839, 2013.

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

[77] Mitch Leslie. To help save the heart, is it time to retire cholesterol tests? *Science (New York, N.Y.)*, 358(6368):1237–1238, December 2017.

[78] Bo Li and Colin N. Dewey. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, 12(1):323, August 2011.

[79] Yang I. Li, David A. Knowles, Jack Humphrey, Alvaro N. Barbeira, Scott P. Dickinson, Hae Kyung Im, and Jonathan K. Pritchard. Annotation-free quantification of RNA splicing using LeafCutter. *Nature Genetics*, 50(1):151–158, January 2018.

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

[81] Yanyu Liang, François Aguet, Alvaro N Barbeira, Kristin Ardlie, and Hae Kyung Im. A scalable unified framework of total and allele-specific counts for cis-qtl, fine-mapping, and prediction. *Nature communications*, 12(1):1–11, 2021.

[82] Yanyu Liang, Owen Melia, Thomas Brettin, Andrew Brown, and Hae Kyung Im. Brainxcan identifies brain features associated with behavioral and psychiatric traits using large scale genetic and imaging data. *medRxiv*, 2021.

[83] Yanyu Liang, Milton Pividori, Ani Manichaikul, Abraham A Palmer, Nancy J Cox, Heather E Wheeler, and Hae Kyung Im. Polygenic transcriptome risk scores improve portability of polygenic risk scores across ancestries. *Biorxiv*, 2020.

[84] Thomas J Littlejohns, Jo Holliday, Lorna M Gibson, Steve Garratt, Niels Oesingmann, Fidel Alfaro-Almagro, Jimmy D Bell, Chris Boultwood, Rory Collins, Megan C Conroy, et al. The uk biobank imaging enhancement of 100,000 participants: rationale, data collection, management and future directions. *Nature Communications*, 11(1):1–12, 2020.

[85] Dajiang J Liu, Gina M Peloso, Haojie Yu, Adam S Butterworth, Xiao Wang, Anubha Mahajan, Danish Saleheen, Connor Emdin, Dewan Alam, Alexessander Couto Alves, et al. Exome-wide association study of plasma lipids in¿ 300,000 individuals. *Nature genetics*, 49(12):1758, 2017.

[86] Jimmy Z Liu, Yaniv Erlich, and Joseph K Pickrell. Case–control association mapping by proxy using family history of disease. *Nature genetics*, 49(3):325, 2017.

[87] Yaowu Liu, Sixing Chen, Zilin Li, Alanna C Morrison, Eric Boerwinkle, and Xihong Lin. Acat: A fast and powerful p value combination method for rare-variant analysis in sequencing studies. *The American Journal of Human Genetics*, 104(3):410–421, 2019.

[88] Luke R Lloyd-Jones, Jian Zeng, Julia Sidorenko, Loïc Yengo, Gerhard Moser, Kathryn E Kemper, Huanwei Wang, Zhili Zheng, Reedik Magi, Tonu Esko, et al. Improved polygenic prediction by bayesian multiple regression on summary statistics. *Nature communications*, 10(1):1–11, 2019.

[89] Adam E Locke, Karyn Meltz Steinberg, Charleston WK Chiang, Susan K Service, Aki S Havulinna, Laurel Stell, Matti Pirinen, Haley J Abel, Colby C Chiang, Robert S Fulton, et al. Exome sequencing of finnish isolates enhances rare-variant association power. *Nature*, page 1, 2019.

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

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

[92] James R Lupski, John W Belmont, Eric Boerwinkle, and Richard A Gibbs. Clan Genomics and the Complex Architecture of Human Disease. *Cell*, 147(1):32–43, September 2011.

[93] Timothy Shin Heng Mak, Robert Milan Porsch, Shing Wan Choi, Xueya Zhou, and Pak Chung Sham. Polygenic scores via penalized regression on summary statistics. *Genetic epidemiology*, 41(6):469–480, 2017.

[94] James Malone, Ele Holloway, Tomasz Adamusiak, Misha Kapushesky, Jie Zheng, Nikolay Kolesnikov, Anna Zhukova, Alvis Brazma, and Helen Parkinson. Modeling sample variables with an Experimental Factor Ontology. *Bioinformatics*, 26(8):1112–1118, 2010.

[95] Scott Marek, Brenden Tervo-Clemmens, Finnegan J Calabro, David F Montez, Benjamin P Kay, Alexander S Hatoum, Meghan Rose Donohue, William Foran, Ryland L Miller, Eric Feczko, et al. Towards reproducible brain-wide association studies. *BioRxiv*, 2020.

[96] Eirini Marouli, Mariaelisa Graff, Carolina Medina-Gomez, Ken Sin Lo, Andrew R Wood, Troels R Kjaer, Rebecca S Fine, Yingchang Lu, Claudia Schurmann, Heather M Highland, et al. Rare and low-frequency coding variants alter human adult height. *Nature*, 542(7640):186, 2017.

[97] Alicia R Martin, Masahiro Kanai, Yoichiro Kamatani, Yukinori Okada, Benjamin M Neale, and Mark J Daly. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nature genetics*, 51(4):584, 2019.

[98] Shane McCarthy, Sayantan Das, Warren Kretzschmar, Olivier Delaneau, Andrew R Wood, Alexander Teumer, Hyun Min Kang, Christian Fuchsberger, Petr Danecek, Kevin Sharp, et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nature genetics*, 48(10):1279–1283, 2016.

[99] Gregor Mendel. Experiments in plant hybridization (1865). *Verhandlungen des naturforschenden Vereins Brünn.) Available online: www. mendelweb. org/Mendel. html (accessed on 1 January 2013)*, 1996.

[100] Lauren S Mogil, Angela Andaleon, Alexa Badalamenti, Scott P Dickinson, Xiuqing Guo, Jerome I Rotter, W Craig Johnson, Hae Kyung Im, Yongmei Liu, and Heather E Wheeler. Genetic architecture of gene expression traits across diverse populations. *PLoS genetics*, 14(8):e1007586, 2018.

[101] Pejman Mohammadi, Stephane E Castel, Andrew A Brown, and Tuuli Lappalainen. Quantifying the regulatory effect size of cis-acting genetic variation using allelic fold change. *Genome research*, 27(11):1872–1884, 2017.

[102] Pejman Mohammadi, Stephane E Castel, Beryl B Cummings, Jonah Einson, Christina Sousa, Paul Hoffman, Sandra Donkervoort, Zhuoxun Jiang, Payam Mohassel, A Reghan Foley, Heather E Wheeler, Hae Kyung Im, Carsten G Bonnemann, Daniel G MacArthur, and Tuuli Lappalainen. Genetic regulatory variation in populations informs transcriptome analysis in rare disease. *Science*, 366(6463):eaay0256–356, October 2019.

[103] Felix Mölder, Kim Philipp Jablonski, Brice Letcher, Michael B Hall, Christopher H Tomkins-Tinch, Vanessa Sochat, Jan Forster, Soohyun Lee, Sven O Twardziok, Alexander Kanitz, et al. Sustainable data analysis with snakemake. *F1000Research*, 10:33, 2021.

[104] Jill E Moore, Michael J Purcaro, Henry E Pratt, Charles B Epstein, Noam Shoresh, Jessika Adrian, Trupti Kawli, Carrie A Davis, Alexander Dobin, Rajinder Kaul, et al. Expanded encyclopaedias of dna elements in the human and mouse genomes. *Nature*, 583(7818):699–710, 2020.

[105] Kiran Musunuru, Alanna Strong, Maria Frank-Kamenetsky, Noemi E. Lee, Tim Ahfeldt, Katherine V. Sachs, Xiaoyu Li, Hui Li, Nicolas Kuperwasser, Vera M. Ruda, James P. Pirruccello, Brian Muchmore, Ludmila Prokunina-Olsson, Jennifer L. Hall, Eric E. Schadt, Carlos R. Morales, Sissel Lund-Katz, Michael C. Phillips, Jamie Wong, William Cantley, Timothy Racie, Kenechi G. Ejebe, Marju Orho-Melander, Olle Melander, Victor Koteliansky, Kevin Fitzgerald, Ronald M. Krauss, Chad A. Cowan, Sekar Kathiresan, and Daniel J. Rader. From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature*, 466(7307):714–719, August 2010.

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

[107] Luke J O'Connor, Armin P Schoech, Farhad Hormozdiari, Steven Gazal, Nick Patterson, and Alkes L Price. Extreme polygenicity of complex traits is explained by negative selection. *The American Journal of Human Genetics*, 105(3):456–476, 2019.

[108] All of Us Research Program Investigators. The "all of us" research program. *New England Journal of Medicine*, 381(7):668–676, 2019.

[109] Halit Ongen, Andrew A Brown, Olivier Delaneau, Nikolaos I Panousis, Alexandra C Nica, Emmanouil T Dermitzakis, GTEx Consortium, et al. Estimating the causal tissues for complex traits and diseases. *Nature genetics*, 49(12):1676, 2017.

[110] Halit Ongen, Alfonso Buil, Andrew Anand Brown, Emmanouil T Dermitzakis, and Olivier Delaneau. Fast and efficient qtl mapper for thousands of molecular phenotypes. *Bioinformatics*, 32(10):1479–1485, 2015.

[111] Oliver Pain, Kylie P Glanville, Saskia Hagenaars, Saskia Selzam, Anna Fürtjes, Jonathan RI Coleman, Kaili Rimfeld, Gerome Breen, Lasse Folkersen, and Cathryn M Lewis. Imputed gene expression risk scores: a functionally informed component of polygenic risk. *Human Molecular Genetics*, 30(8):727–738, 2021.

[112] Bogdan Pasaniuc, Noah Zaitlen, Huwenbo Shi, Gaurav Bhatia, Alexander Gusev, Joseph Pickrell, Joel Hirschhorn, David P. Strachan, Nick Patterson, and Alkes L. Price. Fast and accurate imputation of summary statistics enhances evidence of functional enrichment. *Bioinformatics*, 30(20):2906–2914, 2014.

[113] Evanthia E. Pashos, YoSon Park, Xiao Wang, Avanthi Raghavan, Wenli Yang, Deepti Abbey, Derek T. Peters, Juan Arbelaez, Mayda Hernandez, Nicolas Kuperwasser, Wenjun Li, Zhaorui Lian, Ying Liu, Wenjian Lv, Stacey L. Lytle-Gabbin, Dawn H. Marchadier, Peter Rogov, Jianting Shi, Katherine J. Slovik, Ioannis M. Stylianou, Li Wang, Ruilan Yan, Xiaolan Zhang, Sekar Kathiresan, Stephen A. Duncan, Tarjei S. Mikkelsen, Edward E. Morrisey, Daniel J. Rader, Christopher D. Brown, and Kiran Musunuru. Large, Diverse Population Cohorts of hiPSCs and Derived Hepatocyte-like Cells Reveal Functional Genetic Variation at Blood Lipid-Associated Loci. *Cell stem cell*, 20(4):558–570.e10, April 2017.

[114] Gina M. Peloso Peloso, Akihiro Nomura, Amit V. Khera, Mark Chaffin, Hong-Hee Won, Diego Ardissino, John Danesh, Heribert Schunkert, James G. Wilson, Nilesh Samani, Jeanette Erdmann, Ruth McPherson, Hugh Watkins, Danish Saleheen, Shane McCarthy, Tanya M. Teslovich, Joseph B. Leader, H. Lester Kirchner, Jaume Marrugat, Atsushi Nohara, Masa-aki Kawashiri, Hayato Tada, Frederick E. Dewey, Aris Carey, David J. Baras, and Sekar Kathiresan. Rare protein-truncating variants in apob, lower low-density lipoprotein cholesterol, and protection against coronary heart disease. *Circulation: Genomic and Precision Medicine*, 2019.

[115] Milton Pividori and Hae Kyung Im. ukbrest: efficient and streamlined data access for reproducible research in large biobanks. *Bioinformatics*, 35(11):1971–1973, 2019.

[116] Milton Pividori, Padma S Rajagopal, Alvaro Barbeira, Yanyu Liang, Owen Melia, Lisa Bastarache, YoSon Park, GTEx Consortium, Xiaoquan Wen, and Hae K Im. PhenomeXcan: Mapping the genome to the phenome through the transcriptome. *Science advances*, 6(37):eaba2083, September 2020.

[117] Robert M Plenge, Edward M Scolnick, and David Altshuler. Validating therapeutic targets through human genetics. *Nature Publishing Group*, 12(8):581–594, July 2013.

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

[119] Junyang Qian, Yosuke Tanigawa, Wenfei Du, Matthew Aguirre, Chris Chang, Robert Tibshirani, Manuel A Rivas, and Trevor Hastie. A fast and scalable framework for large-scale and ultrahigh-dimensional sparse regression with application to the uk biobank. *PLoS genetics*, 16(10):e1009141, 2020.

[120] Stephan Ripke, James TR Walters, Michael C O'Donovan, Schizophrenia Working Group of the Psychiatric Genomics Consortium, et al. Mapping genomic loci prioritises genes and implicates synaptic biology in schizophrenia. *MedRxiv*, 2020.

[121] Mark D. Robinson and Alicia Oshlack. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, 11(3):R25, March 2010.

[122] D. M. Roden, J. M. Pulley, M. A. Basford, G. R. Bernard, E. W. Clayton, J. R. Balser, and D. R. Masys. Development of a large-scale de-identified DNA biobank to enable personalized medicine. *Clinical Pharmacology and Therapeutics*, 2008.

[123] Aabida Saferali, Jeong H. Yun, Margaret M. Parker, Phuwanat Sakornsakolpat, Robert P. Chase, Andrew Lamb, Brian D. Hobbs, Marike H. Boezen, Xiangpeng Dai, Kim de Jong, Terri H. Beaty, Wenyi Wei, Xiaobo Zhou, Edwin K. Silverman, Michael H. Cho, Peter J. Castaldi, Craig P. Hersh, COPDGene Investigators, and the International COPD Genetics Consortium Investigators. Analysis of genetically driven alternative splicing identifies fbxo38 as a novel copd susceptibility gene. *PLOS Genetics*, 15(7):1–19, 07 2019.

[124] Andrey A Shabalin. Matrix eqtl: ultra fast eqtl analysis via large matrix operations. *Bioinformatics*, 28(10):1353–1358, 2012.

[125] Xueyi Shen, David M Howard, Mark J Adams, W David Hill, Toni-Kim Clarke, Ian J Deary, Heather C Whalley, and Andrew M McIntosh. A phenome-wide association and mendelian randomisation study of polygenic risk for depression in uk biobank. *Nature communications*, 11(1):1–16, 2020.

[126] Alana M Shepherd, Kristin R Laurens, Sandra L Matheson, Vaughan J Carr, and Melissa J Green. Systematic meta-review and quality assessment of the structural brain alterations in schizophrenia. *Neuroscience & Biobehavioral Reviews*, 36(4):1342–1356, 2012.

[127] Huwenbo Shi, Kathryn S Burch, Ruth Johnson, Malika K Freund, Gleb Kichaev, Nicholas Mancuso, Astrid M Manuel, Natalie Dong, and Bogdan Pasaniuc. Localizing components of shared transethnic genetic architecture of complex traits from gwas summary data. *The American Journal of Human Genetics*, 2020.

[128] Stephen Smith, Fidel Alfaro-Almagro, and Karla Miller. Uk biobank brain imaging documentation. `https://biobank.ctsu.ox.ac.uk/crystal/crystal/docs/brain_mri.pdf`, 2020.

[129] Stephen M Smith, Gwenaëlle Douaud, Winfield Chen, Taylor Hanayik, Fidel Alfaro-Almagro, Kevin Sharp, and Lloyd T Elliott. An expanded set of genome-wide association studies of brain imaging phenotypes in uk biobank. *Nature neuroscience*, pages 1–9, 2021.

[130] Oliver Stegle, Leopold Parts, Richard Durbin, and John Winn. A bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eqtl studies. *PLoS computational biology*, 6(5):e1000770, 2010.

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

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

[133] Wei Sun. A statistical framework for eqtl mapping using rna-seq data. *Biometrics*, 68(1):1–11, 2012.

[134] Atsushi Takata, Naomichi Matsumoto, and Tadafumi Kato. Genome-wide identification of splicing QTLs in the human brain and their enrichment among schizophrenia-associated loci. *Nature Communications*, 8(1):14519, February 2017.

[135] Daniel Taliun, Daniel N Harris, Michael D Kessler, Jedidiah Carlson, Zachary A Szpiech, Raul Torres, Sarah A Gagliano Taliun, André Corvelo, Stephanie M Gogarten, Hyun Min Kang, et al. Sequencing of 53,831 diverse genomes from the nhlbi topmed program. *BioRxiv*, page 563866, 2019.

[136] Amaro Taylor-Weiner, François Aguet, Nicholas J Haradhvala, Sager Gosai, Shankara Anand, Jaegil Kim, Kristin Ardlie, Eliezer M Van Allen, and Gad Getz. Scaling computational genomics to millions of individuals with gpus. *Genome biology*, 20(1):1–5, 2019.

[137] the CARDIoGRAMplusC4D Consortium. A comprehensive 1000 genomes-based genome-wide association meta-analysis of coronary artery disease. *Nature genetics*, 47(10):1121, 2015.

[138] The GTEx Consortium. Genetic effects on gene expression across human tissues. *Nature*, 550(7675):204–213, 2017.

[139] The GTEx Consortium. The gtex consortium atlas of genetic regulatory effects across human tissues. *Science*, 369(6509):1318–1330, 2020.

[140] Sarah M Urbut, Gao Wang, Peter Carbonetto, and Matthew Stephens. Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. Technical report, Nature Publishing Group, 2018.

[141] Bryce Van De Geijn, Graham McVicker, Yoav Gilad, and Jonathan K Pritchard. Wasp: allele-specific software for robust molecular quantitative trait locus discovery. *Nature methods*, 12(11):1061, 2015.

[142] Theo GM van Erp, Derrek P Hibar, Jerod M Rasmussen, David C Glahn, Godfrey D Pearlson, Ole A Andreassen, Ingrid Agartz, Lars T Westlye, Unn K Haukvik, Anders M Dale, et al. Subcortical brain volume abnormalities in 2028 individuals with schizophrenia and 2540 healthy controls via the enigma consortium. *Molecular psychiatry*, 21(4):547–553, 2016.

[143] J Craig Venter, Mark D Adams, Eugene W Myers, Peter W Li, Richard J Mural, Granger G Sutton, Hamilton O Smith, Mark Yandell, Cheryl A Evans, Robert A Holt, et al. The sequence of the human genome. *science*, 291(5507):1304–1351, 2001.

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

[145] Urmo Võsa, Annique Claringbould, Harm-Jan Westra, Marc Jan Bonder, Patrick Deelen, Biao Zeng, Holger Kirsten, Ashis Saha, Roman Kreuzhuber, Silva Kasela, et al. Unraveling the polygenic architecture of complex traits using blood eqtl meta-analysis. *bioRxiv*, page 447367, 2018.

[146] Michael Wainberg, Nasa Sinnott-Armstrong, Nicholas Mancuso, Alvaro N. Barbeira, David A. Knowles, David Golan, Raili Ermel, Arno Ruusalepp, Thomas Quertermous, Ke Hao, Johan L. M. Bjorkegren, Hae Kyung Im, Bogdan Pasaniuc, Manuel A. Rivas, and Anshul Kundaje. Opportunities and challenges for transcriptome-wide association studies. *Nature genetics*, 51(4):592–599, April 2019.

[147] G. Walldius and I. Jungner. Apolipoprotein B and apolipoprotein A-I: risk indicators of coronary heart disease and targets for lipid-modifying therapy. *Journal of internal medicine*, 255(2):188–205, February 2004.

[148] Austin T Wang, Anamay Shetty, Edward O'Connor, Connor Bell, Mark M Pomerantz, Matthew L Freedman, and Alexander Gusev. Allele-specific qtl fine mapping with plasma. *The American Journal of Human Genetics*, 106(2):170–187, 2020.

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

[150] Wei Wang and Matthew Stephens. Empirical bayes matrix factorization. *arXiv preprint arXiv:1802.06931*, 2018.

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

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

[153] Xiaoquan Wen, Roger Pique-Regi, and Francesca Luca. Integrating molecular qtl data into genome-wide genetic association analysis: Probabilistic assessment of enrichment and colocalization. *PLoS genetics*, 13(3):e1006646, 2017.

[154] Heather E Wheeler, Keston Aquino-Michaels, Eric R Gamazon, Vassily V Trubetskoy, M Eileen Dolan, R Stephanie Huang, Nancy J Cox, and Hae Kyung Im. Poly-omic prediction of complex traits: Omickriging. *Genetic epidemiology*, 38(5):402–415, 2014.

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

[156] Quin F Wills, Kenneth J Livak, Alex J Tipping, Tariq Enver, Andrew J Goldson, Darren W Sexton, and Chris Holmes. Single-cell gene expression analysis reveals genetic associations masked in whole-tissue experiments. *Nature biotechnology*, 31(8):748–752, 2013.

[157] Kenneth R. Wilund, Liqing Yu, Fang Xu, Helen H. Hobbs, and Jonathan C. Cohen. High-level expression of ABCG5 and ABCG8 attenuates diet-induced hypercholesterolemia and atherosclerosis in Ldlr-/- mice. *Journal of lipid research*, 45(8):1429–1436, August 2004.

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

[159] Lang Wu, Wei Shi, Jirong Long, Xingyi Guo, Kyriaki Michailidou, Jonathan Beesley, Manjeet K Bolla, Xiao-Ou Shu, Yingchang Lu, Qiuyin Cai, et al. A transcriptome-wide association study of 229,000 women identifies new candidate susceptibility genes for breast cancer. *Nature genetics*, 50(7):968–978, 2018.

[160] Jian Yang, Beben Benyamin, Brian P McEvoy, Scott Gordon, Anjali K Henders, Dale R Nyholt, Pamela A Madden, Andrew C Heath, Nicholas G Martin, Grant W Montgomery, et al. Common snps explain a large proportion of the heritability for human height. *Nature genetics*, 42(7):565, 2010.

[161] Jian Yang, S Hong Lee, Michael E Goddard, and Peter M Visscher. Gcta: a tool for genome-wide complex trait analysis. *The American Journal of Human Genetics*, 88(1):76–82, 2011.

[162] Douglas W Yao, Luke J O'Connor, Alkes L Price, and Alexander Gusev. Quantifying genetic effects on disease mediated by assayed gene expression levels. *Nature Genetics*, pages 1–8, 2020.

[163] Liqing Yu, Robert E. Hammer, Jia Li-Hawkins, Klaus von Bergmann, Dieter Lutjohann, Jonathan C. Cohen, and Helen H. Hobbs. Disruption of Abcg5 and Abcg8 in mice reveals their crucial role in biliary cholesterol secretion. *Proceedings of the National Academy of Sciences*, 99(25):16237–16242, 2002.

[164] Hui Zhang, Torben Schneider, Claudia A Wheeler-Kingshott, and Daniel C Alexander. Noddi: practical in vivo neurite orientation dispersion and density imaging of the human brain. *Neuroimage*, 61(4):1000–1016, 2012.

[165] Xiaoling Zhang, Hinco J Gierman, Daniel Levy, Andrew Plump, Radu Dobrin, Harald HH Goring, Joanne E Curran, Matthew P Johnson, John Blangero, Stuart K Kim, et al. Synthesis of 53 tissue and cell line expression qtl datasets reveals master eqtls. *BMC genomics*, 15(1):532, 2014.

[166] Bingxin Zhao, Tengfei Li, Yue Yang, Xifeng Wang, Tianyou Luo, Yue Shan, Ziliang Zhu, Di Xiong, Mads E Hauberg, Jaroslav Bendl, et al. Common genetic variation influencing human white matter microstructure. *Science*, 372(6548), 2021.

[167] Bingxin Zhao, Tianyou Luo, Tengfei Li, Yun Li, Jingwen Zhang, Yue Shan, Xifeng Wang, Liuqing Yang, Fan Zhou, Ziliang Zhu, et al. Genome-wide association analysis of 19,629 individuals identifies variants influencing regional brain volumes and refines their genetic co-architecture with cognitive and mental health traits. *Nature genetics*, 51(11):1637–1644, 2019.

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

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

[170] Jennifer Zou, Farhad Hormozdiari, Brandon Jew, Stephane E Castel, Tuuli Lappalainen, Jason Ernst, Jae Hoon Sul, and Eleazar Eskin. Leveraging allelic imbalance to refine fine-mapping for eqtl studies. *PLoS Genetics*, 15(12), 2019.